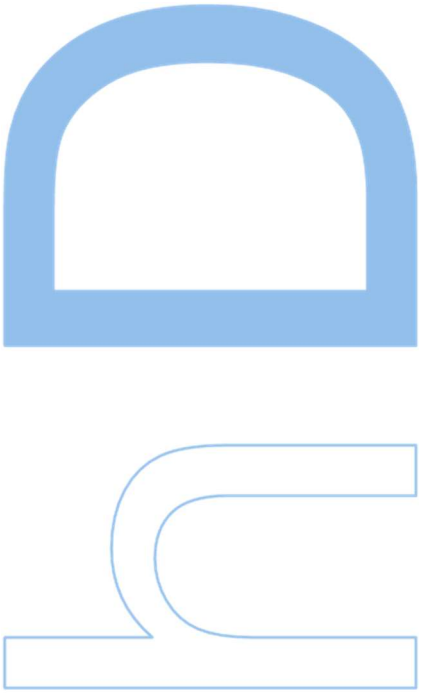
Causal Reasoning in Data



Ana Rita Dias Nogueira
Doctoral Program in Computer Science
Department of Computer Science
2023

Supervisor
João Gama, Full Professor, Faculdade de Economia da Universidade do Porto

Co-supervisor
Carlos Ferreira, Coordinating Professor, Instituto Superior de Engenharia do Porto

To my family

Sworn Statement

I, Ana Rita Dias Nogueira, enrolled in the Doctor's Degree of Computer Science at the Faculty of Sciences of the University of Porto hereby declare, in accordance with the provisions of paragraph a) of Article 14 of the Code of Ethical Conduct of the University of Porto, that the content of this thesis reflects perspectives, research work and my own interpretations at the time of its submission.

By submitting this thesis, I also declare that it contains the results of my own research work and contributions that have not been previously submitted to this or any other institution.

I further declare that all references to other authors fully comply with the rules of attribution and are referenced in the text by citation and identified in the bibliographic references section. This thesis does not include any content whose reproduction is protected by copyright laws.

I am aware that the practice of plagiarism and self-plagiarism constitute a form of academic offense.

Ana Rita Dias Nogueira

6th March 2023

Acknowledgements

The work presented in this thesis would not have been possible without the support of many people and institutions. Here I thank them.

First, I would like to thank professor João Gama and Professor Carlos Ferreira for proposing this incredible challenge and for believing in me and in my capabilities. But, most of all, I thank them for making me stay objective and focused on the target and for sharing their knowledge and wisdom with me.

I also would like to thank INESCTEC, specially LIAAD, for being my host organization during my PhD. But, of course, a host organization is nothing without its people, and because of that, I would like to thank all my colleagues for their unconditional support and shared knowledge.

I sincerely thank my family for supporting me in this journey and lifting me whenever I wanted to give up. Specially to my grandmother, who is no longer with us, I would like to thank her for always believing in me. I finally finished it grandma!

Finally, I thank the projects that contributed to making this thesis possible. The work presented in this thesis was partially supported by national funds through the FCT, with the PhD grant SFRH/BD/146197/2019.

NanoSTIMA Project (NORTE-01-0145-FEDER-000016) which was financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF), also contributed with financial support.

The COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, National Funds through the FCT as part of project UID/EEA/50014/2013, the project FailStopper (DSAIPA/DS/0086/2018) and Humane AI (grant # 820437) also contributed to the publication of several papers within this thesis' scope.

Thank you all for the support!

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores

PhD. Computer Science

Causal Reasoning in Data

by Ana NOGUEIRA

Determining the cause for a particular event has been a case study for several researchers over the years. Finding out why an event happens (its cause) signifies that, for example, if we remove the cause from a system, we can stop the effect. If it is replicated, we can then create a subsequent effect. The application areas for causal discovery methodologies are immense, from its use in climate research to business and bio-medical, among many other areas. For example, in the medical field, this type of causal analysis is quite relevant in diagnosing certain diseases. If a patient has a specific set of symptoms, and a given disease A is known for having the same symptoms the patient is experiencing, then it is possible to infer this patient has, in fact, the disease A.

This thesis's primary goal is to study how to extract causal relationships from data. Several solutions attempt to answer this problem; however, many of these solutions are primarily based on cross-sectional data, meaning that these algorithms are not prepared to evolve. To achieve this goal, we analyzed the state of art methods and proposed four different approaches.

Firstly, we analyzed the potential usage of association rules to infer causal relationships from observational data. In this topic we proposed CRPA-UC, an association rules global causal discovery methodology. Compared to other methods, the results suggested that causal association rules uncover potential causal relationships in data more accurately.

As causal discovery methodologies are highly interpretable but perform poorly in prediction problems, we analyzed the potential application of causal discovery in decision trees to create a semi-causal approach that represents causal relationships but maintains

a high predictive power. The results showed a resemblance with the traditional method's accuracy while creating significantly smaller trees.

Causal discovery methodologies can also be applied to other machine learning tasks, such as feature engineering. Regarding this topic, we investigated the potential application of causal discovery methods to generate new features that entail the supposed causal information about the relations between a target variable and the remaining ones. The results obtained show that, in the presented problems, the usage of these new features positively impacts the classification algorithm's performance.

Finally, we studied the potential conversion of cross-sectional causal methodologies to be used in time-series data. We analyzed the potential application of such a method in medical data (data comprised of static and time-series data, measured in irregular time intervals). We designed and proposed a method to deal with this type of data. The results showed that the method had a significant gain in terms of performance when compared to other methods.

As causal discovery is a broad topic, so are the methods that apply it. While cross-sectional Causal Bayesian Networks continue to be the norm for applying causal discovery to problems, given their properties, they are not the only methods available. In this thesis, we explored methodologies and techniques different from these traditional methods to infer causal relationships from data and demonstrated that, in certain situations, they have better results than the Bayesian Networks.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores

Doutoramento em Ciência de Computadores

Raciocínio Causal em Dados

por Ana NOGUEIRA

Determinar a causa de um determinado evento tem sido caso de estudo para diversos investigadores ao longo dos anos. Descobrir por que um evento acontece (a sua causa) significa que, por exemplo, se removermos a causa de um sistema, podemos interromper o efeito. Se a causa for replicada, podemos criar o subsequente efeito. As áreas de aplicação da descoberta causal são imensas, desde o seu uso em investigação da área do clima, até *business* e biomédica, entre muitas outras. Por exemplo, na área médica, este tipo de análise causal é relevante no diagnóstico de determinadas doenças. Se um paciente tem um conjunto específico de sintomas, e uma determinada doença A é conhecida por ter os mesmos sintomas que o paciente está a sofrer, então é possível inferir que esse paciente tem, de facto, a doença A.

O objetivo principal desta tese é estudar como extrair relações causais de dados. Várias soluções tentam responder a este problema; no entanto, muitas dessas soluções são baseadas principalmente em dados *cross-sectional*, o que significa que esses algoritmos não estão preparados para evoluir ao longo do tempo. Para atingir esse objetivo, analisamos o estado da arte e propusemos quatro abordagens diferentes.

Em primeiro lugar, analisamos o potencial uso de regras de associação para inferir relações causais a partir de dados observacionais. Neste tópico propusemos CRPA-UC, uma metodologia de regras de associação que aplica descoberta causal global. Em comparação com outros métodos, os resultados sugerem que as regras de associação causal descobrem possíveis relações causais nos dados com mais precisão.

Como as metodologias de descoberta causal são altamente interpretáveis, mas apresentam baixo desempenho em problemas de previsão, analisamos a potencial aplicação de descoberta causal em árvores de decisão para criar uma abordagem semi-causal que representa relações causais, mas mantém um alto poder preditivo. Os resultados demostraram uma semelhança com a precisão do método tradicional, mesmo tempo que cria árvores significativamente menores.

As metodologias de descoberta causal também podem ser aplicadas a outras tarefas de *machine learing*, como *feature engineering*. Em relação a este tópico, investigamos a potencial aplicação de métodos de descoberta causal para gerar novas variáveis que representam as supostas informações causais sobre as relações entre uma variável alvo e as demais. Os resultados obtidos demostram que, nos problemas apresentados, o uso desses novos recursos tem um impacto positivo no desempenho do algoritmo de classificação.

Finalmente, estudamos a conversão potencial de metodologias causais utilizadas em dados *cross-sectional* para poderem ser usados em dados séries temporais. Analisamos a potencial aplicação de tal método em dados médicos (dados compostos por dados estáticos e de séries temporais, medidos em intervalos de tempo irregulares). Projetamos e propusemos um método para lidar com este tipo de dados. Os resultados mostraram que o método teve um ganho significativo em termos de desempenho quando comparado a outros métodos.

Como a descoberta causal é um tópico amplo, os métodos que a aplicam também o são. Embora as Redes Bayesianas Causais para dados *cross-sectional* continuem a ser a norma para aplicar a descoberta causal, dadas suas propriedades, elas não são os únicos métodos disponíveis. Nesta tese, exploramos metodologias e técnicas diferentes desses métodos tradicionais para inferir relações causais a partir de dados e demonstramos que, em determinadas situações, elas apresentam resultados melhores que as Redes Bayesianas.

Contents

List of Figures

List of Tables

List of Algorithms

List of Abbrevations

G^2 G-square

χ^2 Chi-square

Adaptive Anytime FCI Adaptive Anytime FCI Fast Causal Inference

AGES aggregative greedy equivalence search

Anytime FCI Anytime Fast Causal Inference

ARM Association Rule Mining

BES Backward Equivalence Search

BIC Bayesian Information Criterion

BN Bayesian Networks

CDT Causal Decision Tree

CDT-PS Causal Decision Tree with Perfect Stratification

CDT-SPS Causal Decision Tree with Stratification on Propensity Scores

CIM Causal Inference over Mixtures

CMH Cochran-Mantel-Haenszel

CNN Causal Neural Networks

CPDAG Completed Partially Directed Acyclic Graph

CR-CS Causal Rule Discovery with Cohort Studies

CR-PA Causal Rule Discovery with Partial Association

CRPA-UC Causal Rule Discovery with Partial Association and Uncertainty Coefficient

DAG Directed Acyclic Graphs

DBN Dynamic Bayesian Network

FCI Fast Causal Inference

FES Forward Equivalence Search

FGS or FGES Fast Greedy Equivalence Search

FI Input layer input Features

FO Output layer input Features

GCMH Generalised Cochran-Mantel-Haenszel

GDS Greedy DAG Search

GES Greedy Equivalence Search

GFCI Greedy Fast Causal Inference

GIES Greedy Interventional Equivalence Search

ICU Intensive Care Unit

IG Information Gain

IGR Information Gain Ratio

ITBN Irregular Time Bayesian Network

ItsPC Irregular time-series PC

JCI Fast Causal Inference with Joint Causal Inference

JCI Joint Causal Inference

LHS Left-hand Side

MMPC Min-Max Parents and Children

PC Peter and Clark

RCTs Randomised Controlled Trials

RFCI Really Fast Causal Inference

RFCI-BSC Really Fast Causal Inference with Bayesian Score Constraints

RHS Right-hand Side

SC Tree Semi-Causal decision Tree

SC-IG Semi-causal Information Gain

SC-IGR Semi-causal Information Gain Ratio

SHD Structural Hamming Distance

SID Structural Intervention Distance

TBN Temporal Bayesian Network

TI Target outputs from the Input layer

TnBN Temporal node Bayesian Network

tsFCI time-series Fast Causal Inference

UC Uncertainty Coefficient

Notations

$\perp\!\!\!\perp$	Conditional independence
$\mathrm{Pa}(X)$	Parents of X
$\mathrm{De}(X)$	Descendants of X
$\emptyset$	Empty set
ct	critical value
α	alpha
β	Correlational weight
θ	Causal weight

Chapter 1

Introduction

Determining the cause for a particular event has been a case study for several researchers over the years. Finding out why an event happens (its cause) signifies that, for example, removing the cause from a system can stop the effect from happening. If it is replicated, we can then create a subsequent effect.

Humanity has always shown an interest in understanding how an event can cause another. This curiosity led some well-known minds of the past, such as Descartes or Aristotle, to make essential contributions to this matter [1, 2].

Recently, two authors stood out: Clive Granger and Judea Pearl. These authors distinguish themselves from many others for transposing the definition of causality, which was restricted to philosophy, to the computational domain, thereby finding a way to quantify causality through data [3, 4].

Nevertheless, what is the definition of causality? Causality is a connection between two different events (cause and effect), which are temporally distant [5]. This temporal notion of past and future is often one of the critical points in discovering the causes of a given event.

Causality is much more than the study of mere correlation, being instead of the study of the actions that take place between events, and from there, we can extract relevant information about these events, more specifically if they have a cause-effect relationship and hence use this information to prevent or even cause certain events. Besides studying

the relations between events, causality is to be applied to, for example, variable selection and classification, among others.

This type of study can also be applied to the most varied areas, such as climate research, business, and bio-medical.

1.1 Motivation

The search for an explanation for certain events has been the human object of study since the beginning. Finding an event's causes makes the world a more understandable place. Causality predicts the future based on the past and has the potential to alter or halt a particular outcome. This temporal notion of past and future can often be a critical point in discovering the causes of a given event and can be viewed as prior knowledge. Despite that, since time can be viewed as background knowledge, there are some instances where it is possible to surpass this and not use time (this can happen because time is unavailable).

This perception of temporality in the cause-effect relationship is evident in the medical field: for example if a patient takes a medication. Then, after a short time, it begins to show specific symptoms, it is possible to affirm with some degree of certainty that the cause is the medication taken.

With that said, the motivation to study this topic is related to the fact that studying causality can be relevant to several problems (more specifically, the study of the causal relationships inherent to the problem), for example, medical problems, since it may be possible to find out what impact the changes can have on the system to study.

In addition, and with the technological advances witnessed in recent years, causality's study may be based on the study of cross-sectional data and time-series data that is continuously obtained. This data collection and modelling variability present a challenge, as it is necessary to adapt existing models to deal with the continuous arrival of data to keep the models updated.

1.2 Objectives

This thesis's primary goal is to study how causality can be extracted from data. Several solutions attempt to answer this problem; however, many are primarily based on cross-sectional models, meaning that these algorithms are not prepared to evolve.

To achieve this goal, we will investigate some techniques that can be applied to this problem:

1. Techniques designed to deal with causality in cross-sectional data;
2. Techniques that, although not prepared to deal with causality, can be adapted to explore causal relationships;
3. Techniques that can be applied to different machine learning tasks, such as feature engineering;
4. Techniques designed to deal with causality in time-series data and/or can be adapted to deal with such data.

As study data, it can be divided into three different sets:

- Data sets associated with ground-truth models;
- Public data sets that have already been used in causally related tasks, hence have proven causal relationships;
- Real-world data sets, associated with key problems, suitable for causal studies (for example, medical data).

Finally, as application problem to analyse the application of causal discovery, we propose medical data since it has the potential of having causal relationships [6] (and in some cases, the existence of such causal relations is even proven), besides this, in some areas of medicine, it is common to register data sequentially. Therefore it is possible to analyse how the causal relations change over time. However, despite being an ideal case study, the use of this type of data represents a challenge for the following reasons:

1. Clinical data is composed of thousands of constantly changing variables, and not all of these variables are relevant to the event in question;

2. At any given time, the data available may not be the best to characterise the event and/or might be missing;

3. The environment changes over time;

4. Clinical events must be detected as soon as possible to prevent further damage.

Other possible study cases can be analysed, such as climatology and palaeontology, which, given their diversity, alone represent a challenge.

1.3 Research Questions

In this thesis, we try to answer the following research questions:

RQ.1 Is it possible to extract causal relationships from data? How?

In this question, we want to investigate existing approaches to studying causality in data, such as the approaches derived from Judea Pearl's idea (PC, FCI, among others [7]). Furthermore, we intend to investigate other less common approaches, such as Causal Decision Trees or Causal Rules Discovery, which are adaptations of known algorithms to the causal domain.

More specifically, we hypothesise that these algorithms may have better results than more traditional methodologies since, besides causal mechanisms, they have other inherent means that may help, for example, summarising the model, making it simpler to read and interpret [8].

RQ.2 Is it possible to obtain more interpretable models by using causal discovery?

We want to study how causal discovery methodologies can improve the models' interpretability in this question. This is a recent and trendy topic, as users are increasingly interested in understanding why the methods make confident decisions. Moreover, we hypothesise that the usage of causal discovery methodologies in correlational methods may increase their interpretability.

RQ.3 In what other situations can we apply causality beyond causal discovery?

In this question, we want to study areas of machine learning where it is possible to apply causal discovery algorithms (for example, feature selection, feature engineering, and classification).

More specifically, we hypothesise that by creating models in which the variables are strongly related, it might be possible to create features that retain information on how the variables relate to each other.

RQ.4 Can we create causal models from sequential data?

In this question, we intend to analyse how we can apply causal discovery in time-series data, especially how the links between different variables will appear, disappear, and change direction since the paradigm changes in this type of data. This issue is fascinating and pertinent nowadays since we continuously generate data from a wide range of equipment. Creating a single causal model is not feasible in these cases since the paradigm can change over time. As a result, we may lose relevant information that can prevent, for example, the bad administration of a particular medication.

More specifically, in this research question, we are interested in studying how a system will change over time, especially how the links between different variables will appear, disappear and change direction. We hypothesise that, with the application of causal discovery in time-series data, it will be possible to create more accurate models that reflect the system's current state.

RQ.5 Are causal relationships helpful, and can they bring significant gains?

Finally, in this research question, we want to understand if studying the causal relationships can help improve several machine learning tasks.

Moreover, we want to comprehend if variations in causal methodologies can help improve prediction and interpretability when compared to both causal and non-causal approaches.

1.4 Research Contributions

This thesis can be divided into four phases, as described next.

1.4.1 Causal rules with partial association an uncertainty coefficient

To answer the research questions **RQ.1** and **RQ.5**, we studied the potential application of association rule mining to uncover potential causal relationships in cross-sectional data [9] (Chapter 3).

For this, we thoroughly analysed the current solutions, where a void was found regarding global discovery using causal association rules in discrete observational data, as the available methods can only be used in local discovery for binary data. This led to the proposal of CRPA-UC. Compared to other methods, the results suggested that causal association rules more accurately uncover potential causal relationships in data.

Contributions

- Provide a causal association rule mining technique that can generate causal rules for all the data sets' variables;
- Can be applied in binary and non-binary discrete data.

1.4.2 Semi-causal decision trees

To answer research questions **RQ.2** and **RQ.5** regarding the usage of causal discovery in methods not suited to infer such relationships, we analysed the potential application in decision trees [10] (Chapter 4). Moreover, in this work, we also wanted to study the usage of causal discovery to boost the interpretability of correlation-based methodologies.

We designed the semi-causal decision trees with this information, a tree-based method that uses a customised statistical test that merges correlation and causality. The results showed an unmistakable resemblance with the traditional method in accuracy while creating significantly smaller trees.

Contributions

- Empirical performance analysis of the different decision tree types (correlation-based, causality-based and semi-causal-based).

1.4.3 Empirical study of the application of causal discovery in feature generation

To answer research questions **RQ.3** and **RQ.5**, we studied the application methodologies to generate new features that entail the supposed causal information about the relations between a target variable and the remaining ones [11] (Chapter 5). The results show that, in the presented problems, the usage of these new features positively impacts Random Forest's performance.

Contributions

- Study of the impact of causal features on Random Forest's performance. This analysis is made using experiences;
- The proposal of a framework to be used for the features' generation.

1.4.4 Irregular Time-series PC

To answer research questions **RQ.4** and **RQ.5**, we studied the potential PC's transformation, from a cross-sectional method to a time-series method (Chapter 6). Irregular time-series PC (ItsPC) models time by incorporating it into the variables' values (instead of creating new variables representing the stages in a particular timestamp). In addition to reducing the number of nodes generated, this method can model time series in which each variable is measured in a different time interval.

We analysed its potential implementation in ICU patients' survival as an application. This type of data comprises static and time-series data, measured in irregular intervals. The results showed that the method had a significant gain in terms of performance compared to other methods.

Contributions

- Provide a causal method that can deal with irregular time-series data;
- Empirical analysis of ItsPC's usage in a real-world problem.

1.5 Contributions to the Community

We created four resources that researchers and practitioners can use to learn about causal discovery and inference's most used mechanisms, with practical examples (these resources are part of [12]):

- Practical guide: https://github.com/AnaRitaNogueira/Methods-and-Tools-for-Causal-Discovery-and-Causal-Inference;
- Data sets used in causal related tasks: https://github.com/AnaRitaNogueira/Causality-Repository-data-sets-;
- List of the more used software: https://github.com/AnaRitaNogueira/Causality-Repository-software;
- List of current causal-related surveys: https://github.com/AnaRitaNogueira/-Causality-Repository-research-papers.

1.6 Publications

As part of the progress of this work, the following publications were submitted:

1.6.1 Journals

- Nogueira, A.R., Gama, J., & Ferreira, C.A. (2021). Causal discovery in machine learning: Theories and applications. Journal of Dynamics & Games. https://www.aimsciences.org/article/doi/10.3934/jdg.2021008;
- Nogueira, A.R., Ferreira, C.A. & Gama, J. Semi-causal decision trees. Prog Artif Intell (2021). https://doi.org/10.1007/s13748-021-00262-2;
- Nogueira, A. R., Pugnana, A., Ruggieri, S., Pedreschi, D., & Gama, J. (2022). Methods and tools for causal discovery and causal inference. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1449;
- Nogueira, A.R., Ferreira, C.A. & Gama, J. Generalized Partial Association Approach to Constraint-based Causal Discovery. In preparation.

1.6.2 Conferences

- Nogueira, A.R., Gama, J., & Ferreira, C.A. (2020, April). Improving Prediction with Causal Probabilistic Variables. In International Symposium on Intelligent Data Analysis (pp. 379-390). Springer, Cham;
- Costa, P., Nogueira, A.R., & Gama, J. (2021, September). Modelling Voting Behavior During a General Election Campaign Using Dynamic Bayesian Networks. In EPIA Conference on Artificial Intelligence (pp. 524-536). Springer, Cham.;
- Nogueira, A.R., Ferreira, C., Gama, J., & Pinto, A. (2021, September). Generalised Partial Association in Causal Rules Discovery. In EPIA Conference on Artificial Intelligence (pp. 485-497). Springer, Cham.;
- Nogueira, A. R., Ferreira, C. & Gama, J. Causal Temporal Nodes for in ICU Outcome Prediction. Accepted at the EPIA 2022 conference.
- Teixeira, S., Nogueira, A. R. & Gama, J. Fairness analysis in causal models: An application to public procurement. In preparation.

1.6.3 Proceedings

- Bifet, A., Berlingerio, M., Gama, J., Read, J., Nogueira, AR. (2020). Proceedings of the 8th International Workshop on Big Data, IoT Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications co-located with the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD, 2019), Anchorage, Alaska, August 4 -8, 2019.

1.6.4 Communications

- A Full Causal Parallel Approach to Markov Blanket Variable Selection presented at the INFORUM 18 conference;
- Predictive models using causal networks presented at the IDA 21 conference;
- Causal Reasoning in Data presented at the DSAA 2021 conference.

1.7 Document Structure

The remaining document is divided into six different sections:

- Chapter 2 provides background explanations and key notions regarding causality;
- Chapter 3 presents *Causal Rule Discovery with Partial Association and Uncertainty Coefficient* (CRPA-UC), a causal association rules methodology;
- Chapter 4 presents the *Semi-causal Trees* (SC Tree), a decision tree method that merges causality and correlation;
- Chapter 5 presents a framework to generate causal features from data;
- Chapter 6 presents the *Irregular Time-series PC* (ItsPC), an irregular time-series causal method. Besides this, we also present its implementation to the problem of ICU patients' survival analysis;
- Finally, Chapter 7 presents the conclusions and future work.

Chapter 2

State of the Art

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, humanity has always been interested in understanding and explaining its 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 last 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 investigating the causal relationships is in their observation, meaning that, to discover potential causal relationships, it is necessary to observe them first. Ideally, these observations are performed in a controlled environment and through exhaustive testing so that we can isolate the desired behaviours (these types of experiments are called Randomised Controlled Trials (RCTs)). Unfortunately, this is not always possible, either because it is impossible to follow a particular action during the necessary time to happen or because it is not ethical or even prohibited. We must deal with the available information and draw conclusions from it in these cases. In such cases, several authors advocate using observational data over RCTs data [13] 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 (BN) [14]. 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 following sections).

The application of causality in the machine learning domain is not as trivial as it may seem, since it is necessary to distinguish between cause [15] and correlation [16]. This distinction is so important that there is even a very famous sentence in statistics that is assumed to be an absolute truth. This sentence is: **"correlation is not causation"**. Correlation is not the same as causation [17] because, although there might be a causal relationship when there is a strong correlation between events, two events occur 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 and links 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. There are cases where there is a clear causal relationship between two events, but there is no clear evidence of a correlation. This is the case of the Simpson Paradox [18].

This paradox is a statistical 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). This means that if two variables A and B are associated in a given data set, it does not mean we can extrapolate the same relationship in any of its subsets.

There are two ways to deal with this paradox: proving the causal relationship is wrong or denying the premise that the standard probability calculus governs this relationship.

The application areas 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 diagnosing certain diseases. For example, if a patient has a set of symptoms, we can prove that this specific combination of symptoms is caused by disease B and only by this disease can we infer that the patient has disease B.

2.1 The evolution of causality

Causality is a concept that dates back to Ancient Greece, with one of the first known definitions being attributed to Plato: *"everything that becomes or changes must be so owing to some cause"* [19]. Later, this definition became the foundation of many other philosophers' ideas.

Despite Plato being the one who first defined causality, a more in-depth study was performed by Aristotle, who interpreted causality as a four-shaped concept: material cause, formal cause, efficient cause, and final cause [2, 20].

In the middle ages, new ideas about causality rose. Aristotle's interpretation of the matter was reformulated to accommodate only two of the four original forms (efficient and final cause) [21].

In the 18th century, philosophers proposed a more empirical view of causality. However, at this time, philosophers only partially agreed on the concept, rejecting other ideas: while Hume [22] 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 [23], and Newton [24] defended that causality does not involve a necessary connection.

In more recent years, causality went trough a shift, going from a purely philosophical abstract concept to a more precise and quantified one, combining statistics, machine learning, data mining and several other quantitative disciplines to search for potential cause-effect relationships in observational data [25]. In these fields, it is seen as an influence for events production, and where a cause is responsible for creating an effect, being the second a consequence of the first's occurrence. For instance, if we consider two events A and B, where B is a consequence of A, A is required for B to exist, but the opposite is invalid. Therefore this subject's study implies understanding how different events interact.

Causality can be further divided into causal discovery and causal inference, with the first being in charge of analyzing and generating models that represent the data's relationships and the latter studying the potential effects occurring when there are changes in the system [26].

These causal models can be defined as "mathematical models representing causal relationships within an individual system or population" [27], with causal relationships represented in them entailing:

1. The variables' probabilistic (in)dependences;
2. The intervention's effect;
3. Or hypothetical interventions, such as counterfactual claims.

These models can be seen from several different perspectives. For example, Rubin [28] proposed a causal inference model for randomized and non-randomized studies. In a randomized study, the treatment's causal effect in a study object (unit) is the difference between the variables' post-exposure if the treatment ($Y_t(u)$) is applied and the response variable if the control ($Y_c(u)$) is employed (see math formula (2.1)).

$$Y_t(u) - Y_c(u) \tag{2.1}$$

In non-randomized studies, post-exposure responses cannot be measured. To deal with this, Rubin defined the treatment's causal effect as the measured control's and treatment's expected (E) causal effect T in the set of all units:

$$E(Y_t - Y_c) = T \tag{2.2}$$

Simultaneously, Clive Granger proposed the granger causality test, a statistical test for time-series data that uses past events to infer present and future events.

In 1988, Pearl proposed the Bayesian Networks [14]. Conventionally, these networks represent probability distributions. Nevertheless, in some cases, the depicted relationships can be perceived as causal [29]. Being a graphical representation of conditional probabilistic dependencies, these graphs have a particularity of not having cycles (it is impossible to start on a node and return to this same node in a sequence) and are called DAG.

TABLE 2.1: Excerpt from the *Abalone* cross-sectional dataset

Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
M	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15
M	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9
M	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10
I	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7
I	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16

2.2 Cross-sectional Methodologies

Cross-sectional causal relationship search has been one of the most researched topics in causal discovery. Cross-sectional data is described as the collection of observations of several subjects simultaneously, thus disregarding time as a variable. This data type can be continuous, discrete, binary, or text. An excerpt is shown in Table 2.1. The *Abalone* data set is a collection of physical features used to characterize and differentiate each specimen. In this data set, each entry represents a different animal.

Definition 2.1 (Cross-sectional data). Observation of subjects at one point or period of time, or for which the analysis has no regard to differences in time among the observations [30].

Despite being the most commonly used and the most developed algorithms, cross-sectional data has a significant downside. Because it represents a snapshot of a moment in time, causal precedence does not apply (A causes B if A happens before B). An extra step is needed to infer the relationship's direction. Various methods exist, covering all types of variables (binary, discrete, continuous, and mixed).

The following sections present several methods to deal with such data and the most common evaluation metrics available.

2.2.1 Causal Bayesian Networks

A particular case of the Bayesian Networks is the Causal Bayesian Networks [31]. In these Bayesian Networks, the nodes represent the studied variables and the edges the causal relationships between them. 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 (an example of a graph can be seen in Figure 2.1).

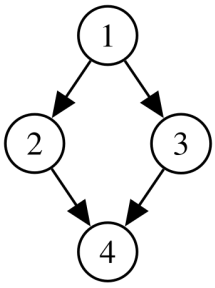


FIGURE 2.1: Example of a graph

Nomenclatures

Before presenting the most commonly used causal Bayesian Networks, several concepts need to be comprehended. Therefore, we introduce some of the main causal models, starting with the necessary graph's background. They are a powerful tool to represent the relationships across variables in a system visually.

A **graph** $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ is defined by a set of **nodes** $\mathbf{V}$ and a set of **edges** $\mathcal{E} \subseteq \{(\{U, V\}, M) \,|U, V \in \mathbf{V}, U \neq V, M \in \mathbf{M}\}$, where $\mathbf{M}$ is a set of **labels**. In particular, an edge can be directed, undirected, or bi-directed, graphically represented as $U \rightarrow V$ or $V \rightarrow U$, $U - V$, and $U \leftrightarrow V$, with U and V being **adjacent**. The $\mathcal{E}$ or edge relationship is a partial function, meaning that no more than one label can be assigned to the adjacent nodes. A graph $\mathcal{G}$ is considered as **directed** if all the edges contained in it are directed. On the other hand, if each edge can be either directed or undirected, it is considered as a **pattern**.

A node $U \in \mathbf{V}$ is a **parent** of another node $V \in \mathbf{N}$ (considered to be U's child) if $U \rightarrow N \in \mathcal{E}$. In this context, we formulate $\mathrm{Pa}(N)$ as the set of parents of N, and $\mathrm{Ch}(U)$ as the set of children of U.

A **(acyclic) path** in $\mathcal{G}$ is a sequence of vertices $N_1, ..., N_n$ such that an edge $(\{V_j, V_{j+1}\}, M_j)$ between two vertices is in $\mathcal{E}$, for $j = 1, ..., n-1$. If all the edges are directed as $V_j \rightarrow V_{j+1}$, the path is a **directed path**. In these cases, the node V_1 is an ancestor of V_n, while V_n is a descendant of V_1. The set of all the ancestors of N is denoted as $\mathrm{An}(V)$ while the set of descendants is drafted as $\mathrm{De}(V)$. At this moment it is important to understand that $V \in \mathrm{An}(V)$ and $V \in \mathrm{De}(V)$. A direct graph is called a DAG if there is no directed cycle, i.e., no pair of vertices $V \neq U$ with a directed path from V to U and from U to V.

DAGs were adopted by Judea Pearl [14] as a graphical representation for the constrained joint probability distribution of a set of random variables.

If we consider i random variables $\mathbf{X} = (R_1, ..., R_p)$ with a joint distribution $P(\mathbf{X})$, being $P(X_i|\mathbf{S})$ the marginal distribution of X_i conditional to $\mathbf{S} \subseteq \mathbf{X}$.

Definition 2.2. Given a DAG $\mathcal{G} = (\mathbf{X}, \mathcal{E})$, the random variables $\mathbf{X}$ are a Bayesian network concerning $\mathcal{G}$ if:

$$P(\mathbf{X}) = \prod_{X \in \mathbf{X}} P(X|\text{Pa}(X)) \tag{2.3}$$

Bayesian networks are graphical representations of probabilistic relationships among variables, the nodes a representation of these variables and the edges the conditional dependencies between these variables. It is worth noting that such a representation is advantageous, as it allows the model to represent how the variables interact. For instance, let us consider an example of a survival problem where T is a binary variable of treatment (received the treatment/did not receive the treatment) and Y the outcome (recovered/died), such that $T \rightarrow Y$ if we acknowledge the existence of a third confounding variable U (the patient has a certain disease or not), such that $U \rightarrow T$ and $U \rightarrow Y$. This information can be represented through the DAG shown in Figure 2.2.

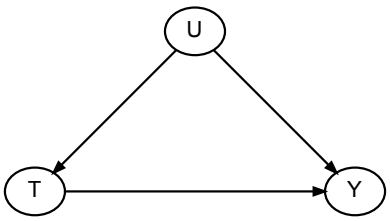


FIGURE 2.2: DAG representation: U is a confounder of T and Y

The factorisation formula (2.3) is equivalent for DAGs [32] to the Markov condition, that states that a variable is conditionally independent ($\perp\!\!\!\perp$) of its nondescendents, given its parents.

Definition 2.3 (Markov Condition). Given a DAG $\mathcal{G} = (\mathbf{X}, \mathcal{E})$, the random variables $\mathbf{X}$ satisfy the Markov Condition if for every $X \in \mathbf{X}$, $X \perp\!\!\!\perp \mathbf{X} \backslash (\text{De}(X) \cup \text{Pa}(X)) | \text{Pa}(X)$ [7].

The Markov Condition is insufficient to remove all conditional (in)dependencies in a Bayesian network. As a consequence, the d-separation is needed. Let us first introduce the notion of a blocking set.

Definition 2.4 (Blocking set). A path $V_1, \ldots, V_n$ in a DAG $\mathcal{G}$ is blocked by a set of nodes $\mathbf{Z}$ (not containing neither V_1 nor V_n) if there exists a node V_k in the path such that one of the following conditions hold [4]:

(i) V_k is a non-collider, i.e. $V_{k-1} \to V_k \to V_{k+1}$ or $V_{k-1} \leftarrow V_k \leftarrow V_{k+1}$ or $V_{k-1} \leftarrow V_k \to V_{k+1}$, and $V_k \in \mathbf{Z}$;

(ii) V_k is a collider, i.e., $V_{k-1} \to V_k \leftarrow V_{k+1}$, and $\mathrm{De}(V_k) \cap \mathbf{Z} = \emptyset$, i.e., neither V_k nor any of its descendants is in $\mathbf{Z}$.

Definition 2.5 (d-separation). In a DAG $\mathcal{G}$, we say that two sets of nodes $\mathbf{L}$ and $\mathbf{M}$ are *d*-separated by a third set of nodes $\mathbf{Z}$, where $\mathbf{L}$, $\mathbf{M}$ and $\mathbf{Z}$ are pairwise disjoint, if $\mathbf{Z}$ is blocking all the paths between nodes in $\mathbf{L}$ and $\mathbf{M}$. This is denoted as [7]: $\mathbf{L} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{M}|\mathbf{Z}$.

For example, if a variable Y d-separates X and Z, the respective generated graphical representation will be similar to Figure 2.3.

FIGURE 2.3: Sample DAG: Y d-separates X and Z

The Markov condition can be further extended to the whole DAG by factorizing once again the (2.3) formula, transforming it into the Global Markov Condition [32].

Definition 2.6 (Global Markov Condition). Given a DAG $\mathcal{G} = (\mathbf{X}, \mathcal{E})$, the random variables $\mathbf{X}$ satisfy the Global Markov Condition if for every pairwise disjoint $\mathbf{L}, \mathbf{M}, \mathbf{Z} \subseteq \mathbf{X}$, if $\mathbf{L} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{M}|\mathbf{Z}$ then $\mathbf{L} \perp\!\!\!\perp \mathbf{M}|\mathbf{Z}$ [33].

The Faithfulness assumption reverses the direction shown above so that the graph's conditionally independent variables are d-separated.

Definition 2.7 (Faithfulness). Given a DAG $\mathcal{G} = (\mathbf{X}, \mathcal{E})$, the random variables $\mathbf{X}$ satisfy the Faithfulness assumption if for every pairwise disjoint $\mathbf{L}, \mathbf{M}, \mathbf{Z} \subseteq \mathbf{X}$, if $\mathbf{L} \perp\!\!\!\perp \mathbf{M}|\mathbf{Z}$ then $\mathbf{L} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{M}|\mathbf{Z}$ [7].

Finally, **Causal Sufficiency** assumption describes that all the common causes of a pair of variables must be measured. Despite this assumption being commonly applied by most causal models, it cannot always be satisfied due to hidden variables in the system. To deal with this, some methods incorporate these latent variables into the system.

Definition 2.8 (Causal Sufficiency). For a pair of observed variables X and Y, all their common causes must also be observed in the data (and modelled in a graph $\mathcal{G}$) [7].

Although Bayesian networks model the edges as conditional probabilities (*condition on observing*), these probabilities do not necessarily represent causal effects (known as *interventions*), intuitively, a variable S has a causal effect on T if through manipulation S can change the T's distribution. Because of this, Causal Bayesian network are designed to account for *do*-interventions: $P(\mathbf{X}|do(\mathbf{W} = \mathbf{w})) = \prod_{X \in \mathbf{X} \setminus \mathbf{W}} P(X|\text{Pa}(X))\mathbb{1}_{\mathbf{W}=\mathbf{w}}$. The **do-operator**, represented as $do(W = w)$ and proposed by Judea Pearl [34], represents the symbolic operation of changing the definition W to the constant value w defined a priori (*atomic intervention*). It is important to note that intervention distributions $P(X|do(W = w)$ are not always equivalent to conditional distributions $P(X|W = w)$ counterparts.

Common Implementations

As was previously mentioned, causal Bayesian Networks must fulfil certain assumptions. These assumptions are Causal sufficiency, Causal Markov condition and Faithfulness.

Typically, a causal discovery algorithm is composed of three stages [35]: creating a skeleton that connects the variables with undirected edges, searching for v-structures (X $\rightarrow$ Y $\leftarrow$ Z) and orientation of all the possible edges.

In the first stage (skeleton's creation), two different approaches can be implemented depending on the situation. The first approach (global approach) builds an undirected graph with all the variables by applying independence tests. Typically the algorithm starts with a fully connected undirected graph and, in each iteration, some edges are removed if both variables are deemed independent from each other by an independence test (first tests on all variables with one conditional variable, then two conditional variables, and so on). For the second approach, the local approach, the algorithm searches these skeletons locally for one or more variables. Usually, the selected nodes are adjacent nodes or Markov Blanket* of the studied variable(s). Next, the algorithm aggregates all local skeletons into a global skeleton [36]. The first approach is usually used in relatively small data sets (number of variables), and the second one is when we have data sets with a large number of variables.

*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 the node's father constitutes, children and parents of the children (the so-called spouse nodes).

In the second stage (search for v-structures), its objective is to find connections that can be transformed into a V-like structure. To generate a v-structure, 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 must exist. If this proposition holds, the edge can be direct as $X \rightarrow Z \leftarrow Y$. This process of finding v-structures is performed by applying the d-separation assumption.

Finally, in the third step, the remaining undirected edges are oriented. This can be performed in three different forms [35]:

1. Using a set of established rules that instruct the algorithm on how to orient the edges employing specific patterns;

2. Using experimental data to orient the edges by manipulating the variables and obtaining the statistical association;

3. Using a mixture of both the previous approaches (orient the edges with the first method and then use the second method to orient the remaining undirected edges).

Depending on the causal algorithm's construction, it can be classified as either constraint-based or score-based. This classification is usually applied to Bayesian-like methods, but it can be extrapolated to other methods, provided they have a similar structure.

Constraint-based algorithms

Constraint-based algorithms employ independence tests to identify a set of edge constraints for the graph using observational data, e.g., using the G-square ($\boldsymbol{G^2}$) test [7]. Further rules then determine the direction of the found relationships. In exceptional cases, the rule phase is skipped to create undirected graphs. These graphs are usually local, meaning they only convey a particular node's (undirected) relationships.

Perhaps the most known constraint-based causal discovery algorithm is **Peter and Clark (PC)** (named after its authors, Peter and Clark) [7]. It relies upon the faithfulness assumption to create the models, meaning that all independencies must obey the d-separation criterion. Like most constraint-based methods, this methodology consists of two phases: searching for (in)dependencies (also called *skeleton*[*] *phase*) and orienting dependencies.

[*]A skeleton is a graph with only undirected edges.

Algorithm 2.1: PC algorithm

Input: complete undirected graph G'
Output: Completed partially directed acyclic graph

```
1  i = 0;
2  repeat
3      for each X ∈ X do
4          for each Y ∈ ADJ_X do
5              Test whether ∃S ⊆ ADJ_X − {Y} with |S| = i and (X ⫫ Y|S);
6              if this set exists then
7                  Make S_XY = S;
8                  Remove X – Y from G';
9      i = i + 1;
10 until |ADJ_X| ≤ i, ∀X;
11 for all pair of non-adjacent variables X and Y with a common neighbour Z do
12     if Z ∉ S_XY then
13         replace the links X – Z – Y by X → Z ← Y;
14 Direct the remaining undirected edges with the following rules:
15 • If X − Y and direct path between X and Y then X → Y;
16 • If X − Y and Y → Z then X → Y;
```

The algorithm (the pseudo-code can be seen in Algorithm 2.1) starts with a fully connected undirected graph in the first phase. For each pair of adjacent variables A and B, it tests if the conditional independence $A \perp\!\!\!\perp B|C$ for a set C of variables all adjacent to A (or, equivalently, all adjacent to B). Tests start with $C = \emptyset$ (unconditional independence) and iterate over sets of increasing size. If conditional independence holds, the undirected edge between A and B is removed.

The orientation phase applies several rules to direct edges [7]:

1. Consider variables A, B, C such that $A - B - C$, namely A and B, B and C are adjacent, but A and C are not adjacent, i.e, it holds in the skeleton phase that $A \perp\!\!\!\perp C|D$ for some D. If $B \notin D$, we orient the edges as $A \rightarrow B \leftarrow C$. The triple A, B, C is called a v-structure;

2. If there is a directed edge $A \rightarrow B$, and B and C are adjacent ($B - C$), but A and C are not adjacent, then $B - C$ is oriented as $B \rightarrow C$;

3. If there is a direct path between A and B and an undirected edge between A and B, orient $A - B$ as $A \rightarrow B$.

PC-stable [37] tackles a known problem inherent to PC known as order dependence. PC output depends on the order the variables are analyzed in the skeleton phase. This means that, if we have a $order_1(V) = \{A, B, C, D, E\}$ and $order_2(V) = \{A, D, B, E, C\}$, the resulting skeletons will not be the same. PC-stable tackles this by saving discarded nodes in a separate list instead of removing them immediately at each iteration. The saved nodes are only removed permanently in the next iteration. This way, removing edges is no longer affected by the order of the independence tests at an iteration.

Another variant is the **conservative PC**. After creating the skeleton, this algorithm tests every potential v-structure $X - Y - Z$ by checking if $X \perp\!\!\!\perp Z|N$ where N includes all the neighbours of X and Z. If Y is not in all the separating sets or there are no variables in the set, $X - Y - Z$ is marked as *ambiguous*, and it is not directed. On the other hand, if Y is not in any separating set, the method continues as PC.

Although **PC** (and its variants) is a powerful tool to uncover causal relationships, it does not scale to high dimensional data. For example, in the **PC-select** (sometimes called PC-simple) method [38], the second phase is removed, and the conditional independence test is only applied to a target variable. Furthermore, the output is an undirected graph because the method does not include an orientation phase.

Another strategy to tackle high dimensional data is to search for causal relations only locally to a target variable. The **Min-Max Parents and Children (MMPC)** [39] adopts this approach using a Min-Max heuristic as a conditional independence test.

Although PC is considered a benchmark algorithm for this type of data, it assumes causal sufficiency (Definition 2.8), meaning that it does not allow for open systems (systems with latent variables). For cases where the causal assumption cannot be fulfilled, Fast Causal Inference (FCI) can be used [40]. This method applies the same phases of PC: the skeleton and orientation phases. First, FCI applies a conditional independence test to find all the potential causal relationships in the skeleton phase. It is in the second phase that FCI differs the most from PC: instead of assuming that a relationship must have a direction [41], the method tests possible d-separations $X \perp\!\!\!\perp Y|Z$ in the skeleton. If there is at least a variable in Z that d-separates the edge, then it is removed. After this, FCI

applies several rules to direct the edges [42]. FCI also differs from PC in the way it represents relationships. Instead of two types of relationships ($\rightarrow$ and $-$), FCI's current implementations have four:

- $\boldsymbol{X \rightarrow Y}$ that represents X *causes* Y;
- $\boldsymbol{X \leftrightarrow Y}$ that represents that there are unmeasured confounders from both variables;
- $\boldsymbol{X \circ\!\!\rightarrow Y}$ that represents either X *causes* Y or there are unmeasured confounders from both variables;
- $\boldsymbol{X \circ\!\!-\!\!\circ Y}$ can represent: (1) X *causes* Y, (2) Y *causes* X, (3) there is unmeasured confounders from both variables, (4) X *causes* Y and there are unmeasured confounders from both variables or (5) Y *causes* X and there are unmeasured confounders from both variables.

The **Anytime Fast Causal Inference (Anytime FCI)** is a slight modification of FCI that restricts the maximum number of variables in the separation set used to perform the conditional independence tests to a user-defined threshold.

The **Adaptive Anytime FCI Fast Causal Inference (Adaptive Anytime FCI)** [43] is similar to Anytime FCI in the sense that it restrains the number of variables in the separation set. The critical difference is that, instead of the user defining this maximum, it is calculated by the algorithm, using $K = max_i(|adj(C_1, X_i)| - 1)$, where C_1 represents the initial skeleton, X_i a vertice of C_1 and adj represents the list of adjacencies from X_i in C_1.

FCI and its variants can benefit from data preparation according to the **Joint Causal Inference (JCI)** [44] approach. This method extracts the context from several datasets, thus creating a pooled dataset where a traditional causal discovery method can be applied. This allows the generated model to encapsulate information about the variables and the system from where these variables were measured. It is essential to understand that JCI is not a causal discovery method but a tool to prepare the data for it. The authors advocate its use with any causal discovery method but suggest using FCI specifically (hence FCI-JCI).

The **Really Fast Causal Inference (RFCI)** [43] is another FCI-like method that performs an additional test to the conditional independences before the v-structures phase: in this extra phase, the algorithm checks every unshielded triplet $X - Y - Z$ and examines $X \perp$

$\perp Y|Z$ and $Y \perp\!\!\!\perp Z|X$. If this holds and Y is not in the separating set of X and Z, then this triplet is directed as $X \rightarrow Y \leftarrow Z$.

The **Really Fast Causal Inference with Bayesian Score Constraints (RFCI-BSC)** [45] is a modification of RFCI, in which the Bayesian Scores Constraints (BSC) is used as a conditional independence test.

Score-based algorithms

Score-based algorithms assign a relevance score to candidate graphs through some adjustment measures, such as the Bayesian Information Criterion (BIC). However, these algorithms are computationally expensive since they have to enumerate (and score) every possible graph among the given variables. In addition, greedy heuristics are applied to restrict the number of candidates.

The **Greedy Equivalence Search (GES)** [46] is a score Bayesian-based method. It scales to high-dimensional data since it does not consider all existing patterns. This algorithm first 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. Besides this, it also directs every edge of neighbour T of Y and not adjacent to X as $T \rightarrow Y$. Secondly, the method removes the best link in each iteration using the following criteria: it deletes every edge $X - Y$ or $X \rightarrow Y$ if there is a subset of neighbours of Y, Z that is adjacent to X. Besides, the algorithm transforms all edges $Z - Y$ as $Z \rightarrow Y$ and all edges $X - Z$ as $X \rightarrow Z$.

The **Greedy Interventional Equivalence Search (GIES)** [47] is an improvement of GES. Besides adding and removing edges, this method has a third phase. The algorithm elongates the DAG sequence in this phase by continuously modifying the original graph without altering the graph's skeleton. This new graph has the same number of edges and can be transformed into the original one by only changing one arrow.

The **Fast Greedy Equivalence Search (FGS or FGES)** [48] is another modification of GES that uses parallelisation to optimise the algorithm's runtime.

The **Greedy Fast Causal Inference (GFCI)** [49] is a combination between the FGES and FCI. In this new method, both algorithms' skeleton and orientation phases are used: first, the skeleton phase of FGES is applied to the data, and then FCI is used to perfect the skeleton. The same happens in the orientation phase: initially, the algorithm accesses all

the directed edges using FGES. This information is given to FCI, so it can use to correct the edges' direction further.

2.2.2 Causal Neural Networks

The Causal Neural Networks (CNNs) [50] are an algorithm that adapts a neural network to perform causal discovery. This is done by altering the *feedforward* phase to be more like a Bayesian Network, hence representing causal relationships. Furthermore, the CNNs are structured to represent the input variables as output and vice-versa. Accordingly, it can represent causes in the input layer and effects in the output layer. In Figure 2.4 an example of a CNN is presented.

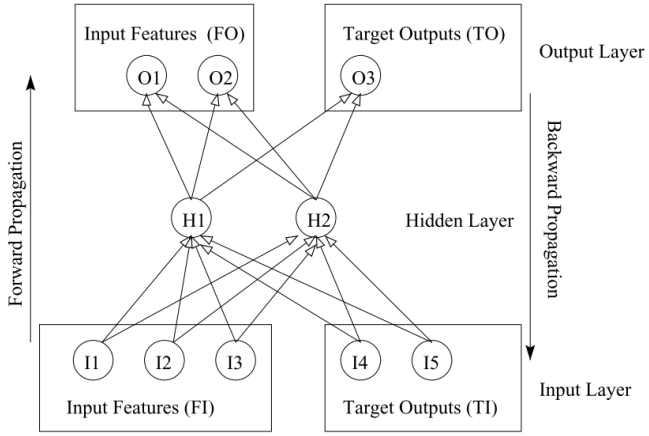


FIGURE 2.4: Example of a CNN with one hidden layer (source [50])

The algorithm splits the input and output features (Input layer input Features (FI) and Output layer input Features (FO) in Figure 2.4) between input and output layers with 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 Rumelhart *et al.* [51] *backpropagation*. The input features are inserted in the FI and the FO's output layers in this sub-algorithm. The unknown features of the input features are inserted in the Target outputs from the Input layer (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.

2.2.3 Association Rule Mining

Association rule mining is inserted in the field of *data mining/machine learning* and the algorithms that fall into this category create $if \Rightarrow then$ rules. Related to association rule mining, there are a set of measures that measure the effectiveness of a rule: **support**, which measures the historical data support (how much the data supports the proposed rule) the rule has, **confidence**, which measures how confident the algorithm is in the rule, and **lift**, which is the ratio between confidence and support.

One of the best-known association rules algorithms is the *apriori*, proposed by Agrawal and Srikant [52] since several authors have applied it in their research.

Within the association rule mining, we have a special case: causal association rule mining. In this category of association rules algorithms, the interest is not in searching for rules $\{attribute = value\} \Rightarrow \{attribute = value\}$, but instead $\{attibute\} \Rightarrow \{attribute\}$. The change from the original association rule's definition is because, in this type of algorithm, the objective is to find *"hypothesized causal relationships around a given target"* [53] and not between attributes' values. This is evident given the definition of causal rule: *"Association rule $x \rightarrow z$ is a causal rule if there exists a significant partial association between variables X and Z"* [53]. This means that, unlike the traditional association rules, these algorithms define the rules by the partial association between the Right-hand Side (RHS) and the Left-hand Side (LHS).

One example of this association rules approach is the work of Jin et al. [54]: Causal Rule Discovery with Partial Association (CR-PA). This algorithm searches for potential causal rules for a target variable through independence tests' application: Chi-square ($\boldsymbol{\chi^2}$) and the Cochran-Mantel-Haenszel (CMH) [55] (see Appendix A). The χ^2 test is applied so that it is possible to determine if two variables are related to each other. If they are not, applying the second independence test is unnecessary. The CMH is applied to the variables selected in this phase. This test is applied to contingency tables of type $K \times 2 \times 2$.

The authors of CR-PA have also proposed a similar algorithm: Causal Rule Discovery with Cohort Studies (CR-CS) [56]. This approach exchanges the independence tests for retrospective cohort studies (odds ratio) to find causal rules. To create these cohort studies, the algorithm selects two types of samples (exposure samples and control samples) and tries to match them so that the distribution of the control variables of the two groups

is as similar as possible. The association of two variables is defined by a support threshold and a minimum odds ratio.

2.2.4 Causal Decision Trees

Although originally, causal discovery in machine learning was applied in Bayesian Networks [57], in more recent years, several authors applied the same processes to other algorithms with good results. This is the case of the Causal Decision Tree (CDT) proposed by Li *et al.* [8]. This algorithm uses the traditional Decision Trees and alters them so that the created tree represents the causal relationships between each variable and the outcome variable. The authors proposed two different variants: CDT-PS and CDT-SPS.

The first approach (CDT-PS) uses $K \times 2 \times 2$ contingency tables (called in this approach perfect stratification) to apply the CMH with one degree of freedom for each variable in each split, to understand whether the variables are causally related with the target node. Then, it orders them by the test's value (in this case, the authors use the test's critical value instead of its p-value) and select the variable with the highest partial association. If the CMH test value is higher than a stipulated significance level, this variable will be a new node in the tree. If the value is lower, the algorithm stops the splitting in that sub-tree.

In the second approach (CDT-SPS), the authors substitute the $2 \times 2 \times J$ tables by propensity scores. In this case, the propensity scores are calculated through logistic regression to the set of attributes correlated with the target to measure each variable's importance. After this, the CMH test is applied to the previous step result, being the chosen variable the one with the highest value.

2.2.5 Evaluation Metrics

Several metrics are used to evaluate causal discovery methodologies. These metrics are usually called *pattern metrics* as they search for common patterns between the ground-truth model that explains the data (or from which the data was generated) and the model generated by the method. Since the ground truth model is generally represented in network form (DAGs, for example), these metrics are also related to network metrics. Despite this restriction, some models generated by non-Bayesian methods can be transformed into networks as long as the generated model is a rule-like model (such as association rule models) and given that all the generated relationships are simple (for example,

rules such as $\{A, B\} \rightarrow \{C\}$ are not allowed). Table 2.2 reports a collection of pattern metrics [58, 59].

TABLE 2.2: Pattern metrics used for cross-sectional causal discovery methods

Metric	Description
Missing edges	Number of edges that are present in the original model but not in the generated one
Extra edges	Number of edges that are present in the generated model but not in the original one
Incorrect Adjacencies (undirected edges)	Number of undirected edges that are present in the generated model but not in the original one
Correct directed edges	Number directed edges present in the generated model that were correctly directed
Incorrect directed edges	Number directed edges present in the generated model that were incorrectly directed
Structural Hamming Distance (SHD)	Sum of missing edges, extras edges and incorrectly directed edges
Structural Intervention Distance (SID)	For each pair X and Y checks whether the parents of X in the generated model are a valid adjustment set [4] in the true model. If it is, it is counted as a correct procedure. If it is not, it is counted as a mistake
Adjacency Precision	$Adj\ Precision = \frac{correctly\ predicted\ adjacencies^{a}}{predicted\ adjacencies^{b}}$
Adjacency Recall	$Adj\ Recall = \frac{correctly\ predicted\ adjacencies}{true\ adjacencies^{c}}$
Arrowhead Precision	$Arrhd\ Precision = \frac{correctly\ predicted\ arrowheads^{d}}{predicted\ arrowheads^{e}}$
Arrowhead Recall	$Arrhd\ Recall = \frac{correctly\ predicted\ arrowheads}{true\ arrowheads^{f}}$

[a] number of undirected edges present in both the generated model and the original one.
[b] all the edges found in the predicted model.
[c] all the edges found in the original model.
[d] number of directed edges present in both the generated model and the original one.
[e] all the directed edges in the predicted model.
[f] all the directed edges found in the original model.

Whenever a ground-truth model is unavailable, causal discovery methods can be evaluated regarding their performance in classification or regression tasks. In these cases, the traditional classification performance metrics are adopted [60].

2.2.6 Software Tools

The three most known tools/libraries for causal discovery in cross-sectional data are: *pcalg*, *bnlearn* and *Tetrad*.

Beginning with `pcalg` [61], this package has implemented several causal methods, such as PC (original, conservative and stable versions), GES, GIES, Greedy DAG Search (GDS), aggregative greedy equivalence search (AGES), FCI (original, Anytime FCI, Adaptive Anytime FCI and FCI-JCI, FCI+ and RFCI). Depending on the type of data used, this package offers default conditional independence tests for binary (G^2 test), discrete (G^2 test) and continuous (Fisher's z-transformation) data. Moreover, it is possible to adapt other conditional dependence tests in this framework. For score-based methods (such as GES), *pcalg* includes the ℓ_0-penalised Gaussian maximum likelihood estimator for both discrete and continuous data.

`bnlearn` is a widely known and used R package [62]. This package provides an implementation for PC stable and MMPC, and it is possible to accommodate discrete, continuous and mixed data by changing the conditional independence test. `bnlearn` implements several conditional independence tests. For discrete data, `bnlearn` has the following tests available: mutual information (information-theoretic distance measure), shrinkage estimator for the mutual information [63] and Pearson's χ^2 (classical version for contingency tables). For continuous data, Pearson's linear correlation, Fisher's Z (transformation of the linear correlation with asymptotic normal distribution), mutual information (information-theoretic distance measure) and shrinkage estimator for the mutual information [64] are available. Finally, mutual information (information-theoretic distance measure)is available for mixed data.

Finally, `Tetrad` [65] is one of the most complete graphical tools for cross-sectional causal discovery. This tool implements the following methods: FCI, RFCI-BSC, FGES, GFCI, PC, and RFCI. `Tetrad` 's methods can be applied in continuous, discrete, and mixed data by choosing the correspondent independence tests/score methods. For constraint-based algorithms, *Tetrad* also implements the following conditional independence tests. For discrete data, the conditional Gaussian test, χ^2 test, degenerate Gaussian likelihood ratio test, G^2 test, and probabilistic test are available. For continuous data, `Tetrad` presents the following tests: conditional correlation independence, conditional gaussian test, degenerate Gaussian likelihood ratio test, fisher Z test and kernel conditional independence.

TABLE 2.3: Overview of software and methods for causal discovery in observational data

Software		Data					Type of Algorithm	
		Categorical Data	Continuous Data	Mixed Data	Causal Sufficiency	Contraint-based	Score-based	Non-Bayesian
bnlearn	MMPC	✓	✓	✓	✓	✓		
	PC	✓	✓	✓		✓	✓	
pcalg	AGES	✓	✓	✓		✓	✓	
	FCI	✓	✓	✓		✓		
	FCI-JCI	✓	✓	✓		✓		
	Anytime FCI	✓	✓	✓		✓		
	Adaptative Anytime FCI	✓	✓	✓		✓		
	FCI+	✓	✓	✓		✓		
	GDS	✓	✓	✓	✓		✓	
	GES	✓	✓	✓	✓		✓	
	GIES	✓	✓	✓	✓		✓	
	LINGAM	✓	✓	✓				✓
	PC	✓	✓	✓	✓	✓		
	CPC	✓	✓	✓	✓	✓		
	PC Select (PC simple)	✓	✓	✓	✓	✓		
	RFCI	✓	✓	✓		✓		
Tetrad	PC and PCStable	✓	✓	✓	✓	✓		
	CPC and CPCStable	✓	✓	✓	✓	✓		
	PcMax	✓	✓	✓	✓	✓		
	FGES/FGES-MB	✓	✓	✓	✓		✓	
	IMaGES	✓	✓	✓			✓	
	FCI	✓	✓	✓		✓		
	RFCI/RFCI-BSC	✓	✓	✓		✓		
	GFCI	✓	✓	✓			✓	
	MBFS	✓	✓	✓	✓	✓		
	GLASSO	✓	✓	✓				✓
	FOFC	✓	✓	✓	✓	✓		
	FTFC	✓	✓	✓				
	LiNGAM	✓	✓	✓				✓

Finally, the following tests are available for mixed data: conditional gaussian test and degenerate Gaussian likelihood ratio test. For score-based causal algorithms, *Tetrad* also offers several scoring methods. For discrete data, `Tetrad` offers the following tests: BDeu score, BIC score, conditional gaussian BIC score and degenerate gaussian BIC score. For continuous data, `Tetrad` has CCI-score, extended BIC (EBIC) score, conditional gaussian BIC score and degenerate gaussian BIC score. Finally, conditional gaussian BIC score and degenerate gaussian BIC score are available for mixed data.

A summary overview of these frameworks can be found in Table 2.3.

2.2.7 Applications

As cross-sectional causal discovery is one of the most developed sub-areas, many authors chose this method to apply to their problems.

This is the case of Miley et al. [66], who used GFCI to identify treatments to early schizophrenia patients. This study proved that the generated model found relationships supported by an early data analysis.

A different application can be seen in the work of Shen et al. [67] where the authors analyze the usage of FCI, FGS or FGES and Structural equation modeling (SEM) to access their ability to discover the underlying structure in data collected by the Alzheimer's Disease Neuroimaging Initiative (ADNI). In their study, the authors found that both FCI and FGS or FGES outperformed SEM, which is not a causal discovery methodology.

Finally, another related work belongs to Afrianto et al. [68]. In this work, the authors analyzed the usage of PC and GES in three different clinical data sets: Heart Disease, Diabetes, and Hepatitis*.

2.3 Time-series data Methodologies

Time-series data can be seen as a sequence of observations about a single subject multiple times.

Definition 2.9 (Time-series data). Observations about a single subject at multiple points or periods of time, indexes in time order. We write X_t for the observation of random variable X at time t [69].

This type of data is characterized by the fact that they are collected in adjacent time periods, and there may be a correlation between distinct observations. Data collected continuously usually does not fall under the assumptions of conventional statistical methods, thus requiring different methods and tools. These types of data are univariate (only one variable is measured) or multivariate (multiple variables are measured), and the variables can be continuous, discrete, binary, or text, among other types, as seen in Table 2.4.

In recent years, the search for causal relationships among variables in time-series data has seen an exponential increase in interest, with sequential data collection becoming a common practice. Causal discovery from this type of data can overcome the problems found in cross-sectional data. Furthermore, since there is a time component, we can assume causal precedence: events in the present cannot cause events in the past. Thus, when

*these three data sets are available on kaggle.com

TABLE 2.4: Excerpt from the *Air Quality* time-series dataset

Date	Time	CO (GT)	PT08.S1 (CO)	NMHC (GT)	C6H6 (GT)	PT08.S2 (NMHC)	NOx (GT)	PT08.S3 (NOx)	NO2 (GT)	PT08.S4 (NO2)	PT08.S5 (O3)	T	RH	AH
10/03/2004	18.00.00	2.6	1360	150	11.9	1046	166	1056	113	1692	1268	13.6	48.9	0.7578
10/03/2004	19.00.00	2	1292	112	9.4	955	103	1174	92	1559	972	13.3	47.7	0.7255
10/03/2004	20.00.00	2.2	1402	88	9	939	131	1140	114	1555	1074	11.9	54	0.7502
10/03/2004	21.00.00	2.2	1376	80	9.2	948	172	1092	122	1584	1203	11	60	0.7867
10/03/2004	22.00.00	1.6	1272	51	6.5	836	131	1205	116	1490	1110	11.2	59.6	0.7888
10/03/2004	23.00.00	1.2	1197	38	4.7	750	89	1337	96	1393	949	11.2	59.2	0.7848
11/03/2004	00.00.00	1.2	1185	31	3.6	690	62	1462	77	1333	733	11.3	56.8	0.7603

faced with an identified (undirected) dependence, it is safe to assume the relationship's direction as $past \rightarrow future$.

Several methods are specifically designed to solve the task of finding causal relationships in sequential observational data. One of the most known frameworks is the Granger causality, proposed by Granger [3]. Intuitively, X Granger-causes Y if predicting Y based on its past observations and the past observations of X perform better than predicting Y based on its past only. Mathematically, this relationship can be formalized by testing that in the auto-regression:

$$Y_t = \sum_{j=1}^{m} a_j Y_{t-j} + \sum_{j=1}^{m} b_j X_{t-j} + \varepsilon_t \tag{2.4}$$

the coefficients b_j's are statistically significant.

In this equation, m represents the model order or the maximum number of lags to be used, a_j's and b_j's are the contributions of the delayed observation of Y and X respectively.

More recent approaches include **time-series Fast Causal Inference (tsFCI)** [70], which is an adaptation of FCI for time-series data. This method uses sliding windows to transform the original time series into different subsets of consecutive timestamps, disregarding the time component in each subset and treating them as cross-sectional. The method creates a model for each subset of data using the models from previous timestamps as prior knowledge. Besides this, if a relationship disappears from the model m_t, this relation will be disregarded in the latter timestamps.

The **PCMCI** [71] is a causal graphical method designed to deal with linear and non-linear time series. This algorithm is divided into two phases corresponding to a different conditional independence test: the PC_1 and MCI phases. First, in the PC_1 phase, the algorithm applies the conditional independence strategy implemented by PC (skeleton phase) to uncover potential dependencies between each variable in a specific timestamp and all the

other variables in all the previous timestamps, e.g., $X_t \perp\!\!\!\perp Y_{t-1}|Z, X_t \perp\!\!\!\perp Y_{t-2}|Z$, among others, where t is the specific timestamp. Next, the method applies the MCI (momentary conditional independence) test [71] further to determine causal relationships between variables in different timestamps while taking into account auto-correlation and incorrect edge detections.

PCMCI+ [72] is an extension of PCMCI, which admits the existence of contemporaneous links (a causal relationship between variables in the same timestamp). Because of this, PCMCI+ divides the skeleton search by type of relationships, namely, lagged and contemporaneous relationships are found separately.

LPCMCI [73] is yet another PCMCI extension specifically designed to deal with latent variables. This method uses an FCI-like approach to represent the latent variables that are present in the relationships.

Time-series data is a particular case of longitudinal data (Definition 2.10) [74, Chapter 1].

Definition 2.10 (Longitudinal data). Observations about several subjects at multiple points or periods of time, indexes in time order, and subject [75].

This type of data is characterized by collecting information about the same individual at different points in time. This means that, for each subject in a dataset, a set of time-series variables characterizes him. The variables in longitudinal data can be continuous, discrete, binary, and text, among other types, as seen in Table 2.5.

TABLE 2.5: Excerpt from the *National Footprint Accounts 2018* longitudinal dataset

country	ISO alpha-3 code	UN region	UN subregion	year	record	crop land	grazing land	forest land	fishing ground	built up land	carbon	total	Percapita GDP (2010 USD)	population
Armenia	ARM	Asia	Western Asia	1992	BiocapPerCap	0.16	0.14	0.08	0.01	0.03	0	0.43	949.03	3449000
Armenia	ARM	Asia	Western Asia	1992	BiocapTotGHA	555812.97	465763.33	289190.66	47320.22	116139.60	0	1474226.80	949.03	3449000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Armenia	ARM	Asia	Western Asia	2014	EFProdPerCap	0.35	0.17	0.20	0.0006	0.062	0.62	1.40	3827.34	3006000
Armenia	ARM	Asia	Western Asia	2014	EFProdTotGHA	1062873.66	516394.76	595089.72	1692.15	185046.34	1856992.85	4218089.49	3827.34	3006000
Afghanistan	AFG	Asia	Southern Asia	1961	BiocapPerCap	0.54	0.68	0.07	0	0.03	0	1.32		9165000
Afghanistan	AFG	Asia	Southern Asia	1961	BiocapTotGHA	4990784.71	6212850.07	654431.08	0	272261.57	0	12130327.43		9165000
...	...	...	...	...	...	...	...	...	...	...	...	...		
Afghanistan	AFG	Asia	Southern Asia	2014	EFProdPerCap	0.25	0.18	0.06	4.86×10^{-5}	0.05	0.11	0.65	610.24	31628000
Afghanistan	AFG	Asia	Southern Asia	2014	EFProdTotGHA	7960359.55	5704672.32	1920868.33	1536.006	1458818.88	3372775.04	20419030.13	610.24	31628000

2.3.1 Evaluation Metrics

The pattern metrics presented in Section 2.2.5 can also be applied to time-series methods if there is a ground-truth model that represents the causal relationships present in the data.

Table 2.6 shows a set of performance metrics specific of time-series data [76], to be used when this information is not available.

The *accuracy* is a metric used to evaluate classification models and can be defined as the fraction of correct predictions made by the model. Usually, this measure takes values between 0 and 1.

The *mean* and *median errors* are metrics that encapsulate the fraction of times the model got some response wrong. This error can be calculated in several ways, the simplest one $1 - accuracy$. These metrics can be valued between 0 and 1.

The *euclidean distance* [77] is another symmetric metric that calculates the distance between two time series $\overrightarrow{x}$ and $\overrightarrow{y}$ (the predicted and the ground-truth). This metric is usually used for regression problems. This metric is valued between 0 and a maximum possible discrepancy, which needs to be calculated [78].

The *longest common subsequence* [79] is an asymmetric metric that measures the number of correct predictions in sequence and reports the highest number. This metric is usually used in regression problems since it uses the euclidean distance to calculate the difference between the predictions and ground truth. This is performed by reducing the difference to 0 or 1 depending on the distance. They are considered equal if the Euclidean distance between two values is smaller than a defined threshold. Hence the distance is 0. On the other hand, if the difference is higher than the threshold, then the distance is 1.

The *Edit Distance with real penalty* [80] is another distance metric that reports the number of edits that are needed to transform the series of predictions into the ground truth. This metric can be valued between 0 and ∞.

Finally, the *Dynamic Time Warping* [81] is a distance metric that calculates the difference between two-time series, taking into account the potential differences in measurement in the timestamps (for example, different frequencies). This is done by comparing each timestamp t from one time series with $t + 1$, $t + 2$ and so on from the second time series.

Concerning metrics for evaluating causal methods for longitudinal data, there are two options. First, the evaluation metrics presented in Section 2.2.5 can be applied if there is a ground-truth structure to compare. Second, since time-series data is a particular type of longitudinal data, the evaluation metrics presented in this section can also be applied.

TABLE 2.6: Pattern metrics used in causal discovery from time-series data.

Metric	Description
Accuracy	$\frac{TruePositive+TrueNegative}{TruePositive+FalsePositive+FalseNegative+TrueNegative}$
Mean/Median Error	Measures the differences between the predicted and ground truth. In this category, we can have all the variances of mean and median measures (root, squared, *etc.*)
Longest Common SubSequence	Measures the size of the longest sequence of events in a time-series model
Edit Distance with Real Penalty	Measure the number of changes to transform one series into another, with a user-defined penalty
Euclidean Distance	Measures the distance between each step of the series $d_E(\overrightarrow{x}, \overrightarrow{y}) = \sqrt{(\overrightarrow{x} - \overrightarrow{y})(\overrightarrow{x} - \overrightarrow{y})'}$
Dynamic Time Warping	Measures the distance between two sequences. Being a sequence of a set of time points, the distance between each point is measured using the euclidean distance

2.3.2 Software Tools

Several libraries are offered in different programming languages to solve the task of finding causal relationships in time-series data.

`lmtest` [82] is an R package known mainly by its implementation of the Granger causality, as well as the data set *ChickEgg*.

`NlinTS` [83] is another R package. Similarly to `lmtest`, this package implements a version of the Granger causality. Besides this, `NlinTS` implements a non-linear version of this test.

`Tetrad`, the tool presented in Section 2.2.6, has also implementations for several methods that deal with time-series data, including TsFCI, FASK and TsGFCI.

`Tigramite` [84] is a Python framework for causal discovery in time-series data. This tool implements three different causal discovery methods (PCMCI, PCMCI+ and LPCMCI) and the following conditional independence test (all these tests can be used together with

TABLE 2.7: Overview of software and methods for causal discovery in time series data

Software		Data					Type of Algorithm		
		Categorical Data	Continuous Data	Mixed Data	Time-series data	Causal Sufficiency	Contraint-based	Score-based	Non-Bayesian
Tetrad	TsFCI	✓	✓	✓	✓		✓		
	TsGFCI	✓	✓	✓	✓			✓	
	TsIMaGES	✓	✓	✓	✓			✓	
	MultiFASK				✓			✓	
Tigramite	PCMCI				✓		✓		
	PCMCI+				✓		✓		
	LPCMCI				✓		✓		
NlinTS			✓		✓				✓
lmtest			✓		✓				✓

the causal discovery methods): ParCorr [85], GPDC / GPDCtorch [86], CMIknn [87] and CMIsymb [87].

Unlike the previous data types, to the best of our knowledge, there is no tool currently available to deal specifically with longitudinal data. However, a few theoretical frameworks have been proposed for this data type. One such framework is the **Causal Inference over Mixtures (CIM)** [88]. This method infers the causal structure by creating a mixture of DAGs using the Global Markov Condition (Definition 2.6). Explicitly designed for longitudinal medical data, it allows for cycles. Besides this, it applies the skeleton phase of [37]. The orientation phase proposed by the authors is similar to FCI.

These libraries can be overviewed in Table 2.7.

2.3.3 Applications

Although time-series causal discovery is not a subject as developed as cross-sectional causal discovery, many authors still used the proposed methods in their research.

This is the case of [89], where the authors compared Granger causality and transfer entropy in oscillation diagnosis. This work found that, while transfer entropy seems more accurate than Granger causality, this method was easy to automate and interpret.

Another Granger causality application is the work from Troster et al. [90], where the authors analyze the potential causal relationship between renewable energy consumption, oil prices, and economic activity in the United States from 1989 to 2016. In this analysis,

the authors found relationships between energy consumption and economic growth as well as between oil prices and economic growth and oil prices and energy consumption.

A different work was proposed by Krich et al. [91], where the authors used PCMCI+ to infer the decoupling between photosynthesis and transpiration in trees at high temperatures was already identified through experimental analysis. The results pinpointed several critical issues in some ecosystems.

2.4 Remarks

As presented in Chapter 1, this thesis aims to study how researchers can apply causal discovery methodologies to cross-sectional and time-series observational data. To do such a task, we analyzed and divided these methods according to the data type they can be applied to.

Most proposed methods can be classified as (Causal) Bayesian Networks regarding cross-sectional causal discovery algorithms. These methods can be further divided into constraint-based and score-based, depending on the tests they implement to retrieve causal relationships from data. Despite being very versatile, these methods still rely on developing new statistical tests to improve further. Moreover, they cannot be applied to all data set types (for example, data sets with continuous + discrete/binary data).

Although Causal Bayesian Networks continue to be the norm in cross-sectional causal discovery, more recently, some authors have proposed altering correlation-based methods, such as Association Rules or Decision Trees, to be able to handle causal relationships. However, despite maintaining the key features from the original methods, the currently proposed approaches have two significant disadvantages: first, they naively assume the relationships as $variable \rightarrow target$, and second, they can only be used in binary data.

Regarding time-series causal discovery methods, this is a subject not as developed as the previous one, with the most known methodology being the Granger Causality. Even though some successful time-series methodologies have been developed, one disadvantage found in these methods is that most publicly available algorithms can only deal with time-series data, not longitudinal data.

Chapter 3

Generalized Partial Association in Causal Rules Discovery

Causal discovery aims to study the possible cause-and-effect relationships between variables in a data set [56]. These causal relationships can be found through several methods, with the most commonly applied algorithms based on Bayesian networks. Despite being the most widely used algorithms for searching for causal relationships in observational data, more and more causal discovery algorithms that do not fall into this category have appeared in recent years.

In this chapter we aim at answering **RQ.1** and **RQ.5** (Section 1.3): *Is it possible to extract causal relationships from data? How?* and *Are causal relationships helpful, and can they bring significant gains?*. We started by studying the potential usage of association rules to infer causal relationships from cross-sectional observational data to answer these questions.

Association Rule Mining (ARM) is a technique used to find correlations between variables in data [92]. Within the association rules algorithms, a subcategory known as causal association rules applies independence tests to determine if there is a causal relationship between two or more variables [53]. Approaches like these have the advantage of being able to create causal hypotheses when dealing with large amounts of data [54]. There are already a few approaches that combine association rule mining and causal discovery. However, they have some restrains/limitations as they can only be applied to a niche type of data (binary) and employ a naive approach to direct rules.

To deal with these limitations, we propose CRPA-UC. This approach applies the GCMH test (conditional independence teest designed to infer dependences in discrete data; see Appendix A), combined with the χ^2 so that it is possible to apply this method in any discrete data set. Finally, and since both independence tests are symmetrical, we propose using the UC (Appendix B) [93] that will act as an orientation method. We also provide an extensive evaluation of this approach using several public data sets, where the proposed approach outperforms the state-of-the-art method (in this case, PC [94]).

3.1 Problem

As stated previously, the current causal association rules CR-PA and CR-CS, presented in Section 2.2.3) have several limitations that restrain their usage to a few potential cases:

1. It is only possible to apply these algorithms to binary data sets;
2. Only to one variable (local structure discovery) [95];
3. They assume a naive approach as an orientation method since they assume that all the rules are $variable \Rightarrow target$, which is not always true.

These limitations mean that it is not possible to apply these algorithms to, for example, non-binary discrete data. This data can be binary (gender), or it can encode stages ($\{normal, risk, failure\}$).This fact implies the need to binarise the data to generate such causal association rules, leading to a data set size increase and, consequently, a run-time increase.

Another critical issue presented by previous methodologies is that it is impossible to infer causal direction [96]. With these methodologies, it is only possible to create and evaluate the undirected relations of the variables with a chosen target, implying the necessity of having a clear idea of what variable is the target to apply these methodologies. However, in some instances, the entire environment's study (and evaluation) is the objective and not a specific outcome. In such cases, these approaches cannot be applied.

Data

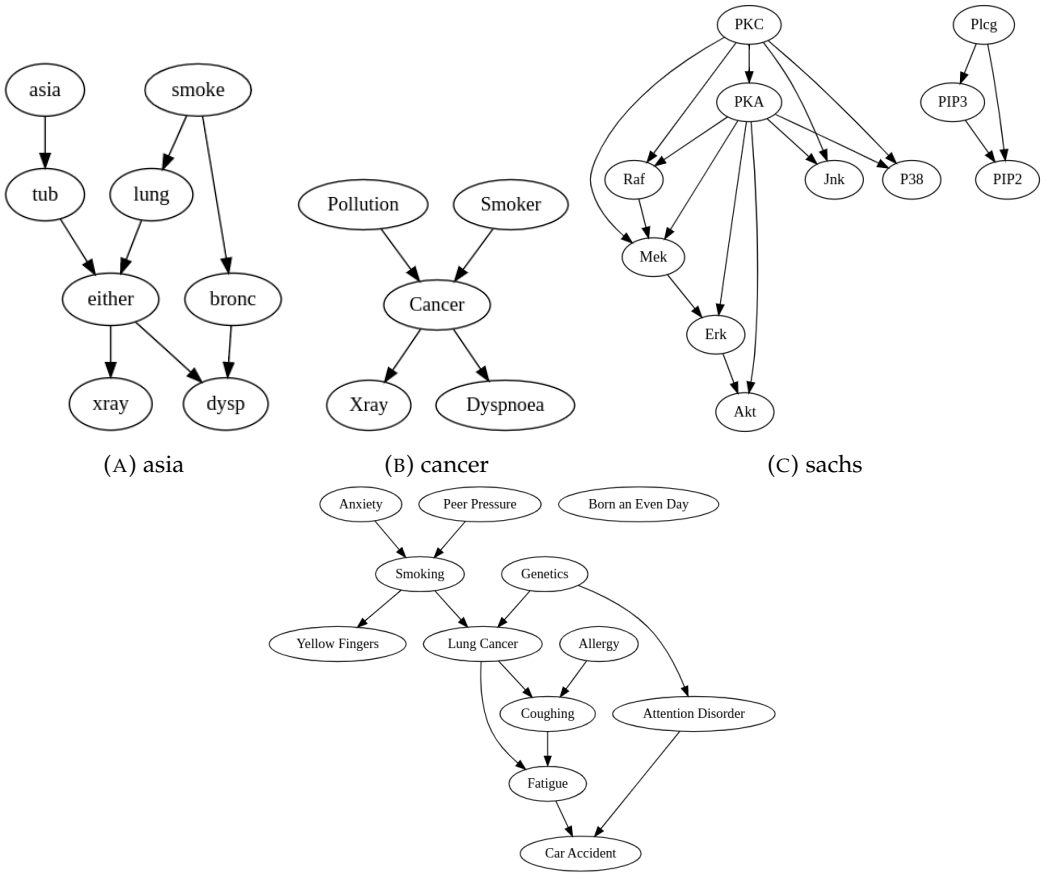
To analyze and solve this problem, throughout the following sections, we will use publicly available datasets and networks, as shown in Figure 3.1 and Table 3.1.

TABLE 3.1: Data set description

Data set		Number of examples	Number of attributes	Number of classes
1	asia[a]	10000	8	0 (44%) 1 (56%)
2	cancer[a]	10000	5	0 (1%) 1 (99%)
3	coronary [b]	1841	6	0 (86%) 1 (14%)
4	earthquake[a]	10000	5	0 (2%) 1 (98%)
5	GMB[c]	5000	5	0 (62%) 1 (38%)
6	lucas[d]	2000	12	0 (28%) 1 (72%)
7	monica[e]	6367	12	0 (55%) 1 (45%)
8	mux6[b]	128	7	0 (50%) 1 (50%)
9	PreSex[f]	1036	4	0 (77%) 1 (23%)
10	sachs[a]	10000	11	0 (60%) 1 (26%) 2 (13%)
11	survey[a]	10000	5	0(56%) 1(28%) 2(16%)
12	titanic[e]	1316	4	0 (62%) 1 (38%)
13	youth risk 2008[e]	500	5	0 (40%) 1 (59%)

[a] https://www.bnlearn.com/
[b] https://www.openml.org
[c] https://cran.r-project.org/web/packages/pcalg/index.html
[d] http://www.causality.inf.ethz.ch/data/LUCAS.html
[e] https://vincentarelbundock.github.io/Rdatasets/articles/data.html
[f] https://cran.r-project.org/web/packages/vcd/index.html

This data was gathered from several public databases and are publicly available for usage. The datasets *asia*, *cancer* and *sachs*, presented in Table 3.1, were randomly generated using the networks present in Figure 3.1. The *lucas* dataset is a public dataset that provides both the generated data and the correspondent network.



(A) asia (B) cancer (C) sachs

(D) lucas

FIGURE 3.1: Networks description

3.2 Causal Association Rules With Partial Association and Uncertainty Coefficient

In this section, we present CRPA-UC, a causal association rules algorithm. This approach can be used for both binary and non-binary discrete data. As conditional independence tests, it applies two tests: χ^2 and GCMH. Besides this, and since it is a global structure causal discovery algorithm, to direct the dependencies found by the conditional independence tests accordingly, the UC (see Appendix B) is applied. It is important to note that in this approach, testing a pair {A,B} and {B,A} will produce the same result. Hence each pair of variables is only tested once.

CRPA-UC (Algorithm 3.1) starts by searching, for each variable, for frequent itemsets (s_1, a) (with user-defined support) in which they are present and selects every variable a that meets this criterion. This pruning is performed because the objective is to generate

Algorithm 3.1: Causal Rules with Partial Association and Uncertainty Coefficient: CRPA-UC

Input: Let **D** be a data set with a set of variables $S = \{s_1, s_2, ..., s_n\}$. Let $\boldsymbol{\alpha}$ be the significance level for the conditional independence tests and **ct** the correspondent critical value. Let $\mathbf{m_{supp}}$ be the minimum support. Let $\mathbf{u_{coef}}$ be the minimum accepted coefficient.

Output: **R**, a set of cause rules

```
1 for each variable s1 in D do
2     Search for frequent itemsets in D containing s1, with support higher than m_supp,
        and save them in F
3     for each pair {s1, s2} in F, with distribution d = dist(s1, s2) do
4         if χ²(d) ≥ ct verifies then
5             if Generalised_CMH(d) ≥ ct verifies then
6                 Verify {s1, s2} direction using the uncertainty coefficient (UC)
7                 if the coefficient is higher than u_coef then
8                     Save rule in R
9 return R
```

rules representing the data's frequent behaviours. These relations can be discarded since rare items only generate more infrequent supersets.

Next, CRPA-UC applies the χ^2 test (line 4). It defines that two variables are associated if the value resulting from the test is greater than or equal to its critical value ct, with significance level α. If the two variables are not dependent, the second and third tests are ignored. In this case, χ^2 acts as a pre-processing method, in a way that GCMH is more computationally demanding and if the algorithm determines *apriori* that two variables are not related, there is no need to apply it.

In line 5, to variables selected by χ^2, the GCMH test is applied. This test checks if two variables remain dependent, given the other variables' influence.

After determining all the potential partial associations, the UC is applied to determine the associations' direction (lines 7 and 8). The direction is obtained by testing both options ($A \Rightarrow B$ and $B \Rightarrow A$), with the selected option being the one with the highest coefficient (for the sake of consistency, the chosen value must also be higher than a minimum user-defined coefficient).

An Illustrative Example

To explain in more detail how this approach works, we will use as an example a data set*

*The data set is available in https://tinyurl.com/gitbub

with three discrete variables (A, B and C), with 10 000 instance and values comprehended in $\{0, 1, 2\}$. This data set can be represented as $B \leftarrow A \rightarrow C$, meaning that A is a common cause of B and C. In this example, we will set the minimum support and α as 1 % (being the correspondent critical value 6.64 for one degree of freedom), and the minimum accepted coefficient as 0.60.

As we are looking for causal rules for all variables in the data set (in this case, A, B and C), this algorithm will have three iterations: one to search for A's rules, another to find B's rules and a third for C's rules. Since both GCMH test and χ^2 tests are symmetric, searching for the direction of the relation between A and B (when A is the target), and B and A (when B is the target) will have the same result (i.e. we will have $A \Rightarrow B$ or $B \Rightarrow A$ duplicated). To solve this results' duplication, the already tested variables are discarded. A is tested with B and C, B is only tested with C and C is not tested at all (as mentioned in the previous section).

We will start with variable A: in the first phase, the algorithm looks for the frequent itemsets in which the variable A is present. The method does not remove any variables since the minimum support is 100 ($number\ of\ instances \times support$). Furthermore, both B and C have higher support (4039 and 3653 respectively) and therefore are not removed.

In the second phase (line 3 in the algorithm), the χ^2 test is applied. In this case, the value obtained for B and C are 1104.83 and 2758.66 respectively. Since we set α as 1 % (with the correspondent critical being 6.64), this means that there is a (still undirected) dependence between B, C and A. Because of this, these two variables are selected for the next step: the GCMH test. The values obtained from this test are: 985.30 for A-B and 2690.41 for A-C. Since they are both higher than the critical value, this means that again B and C are associated with A.

TABLE 3.2: UC for variable A

Variable	$A \Rightarrow Variable$	$Variable \Rightarrow A$
B	0.70	0.60
C	0.78	0.70

Finally, the UC is applied to B and C. As we can see in Table 3.2, the rules $\boldsymbol{A \Rightarrow B}$ and $\boldsymbol{A \Rightarrow C}$ are selected since the coefficient in both of these rules is the highest and is higher than the minimum acceptable coefficient.

After the discovery of A's rules, the same process is repeated for B. First, the algorithm looks for frequent itemsets with variable B and scores the variables accordingly. In this case, and as stated before, the algorithm ignores variable A and only tests variable C. Being that this variable has the support of 3863 is not removed. After that, the first test is applied between variables B and C, and the correspondent value is 106.46, meaning that these two variables are dependent. This means that this variable is selected to be tested with the GCMH test, which returns the value 4.21. Since this value is inferior to the critical value (6.64) so B and C are independent.

Since there are no variables to test with C (since after testing, the variables are removed), the algorithm ends with the following rules: $\boldsymbol{A} \Rightarrow \boldsymbol{B}$ and $\boldsymbol{A} \Rightarrow \boldsymbol{C}$.

3.3 Experimental Setup

To evaluate the proposed approach and make a comparative study, we draft the following experimental setup: first, we access the algorithm's performance in terms of patterns generated. To do that we employ the pattern metrics presented in Chapter 2, Section 2.2.5. We have selected four public networks presented in Figure 3.1 in Section 3.1 to test this approach. As stated before, for *asia*, *cancer* and *sachs* datasets these networks were used to generate random data that represents the relationship in the networks (details about the number of instances can be seen in Table 3.1). The *lucas* dataset source website provided both the network and data.

Second, we demonstrate another side of the causal discovery: prediction. To do that, and since all the rules are simple and there are no cycles between them, we converted each rule generated by CRPA-UC into an edge of the equivalent network. This time, we compared the proposed approaches with PC using 10-fold cross-validation in the dataset group presented in Table 3.1 in Section 3.1.

Since PC produces a partially directed acyclic graph (PDAG), to be able to use the models for prediction, these models were extended to directed acyclic graphs (DAG)[97].

A sensitivity analysis was performed to choose the optimal parameters for the approaches presented in the following sections. This analysis consisted of obtaining the error (*1 - accuracy*) for the presented data sets (by dividing them into 70 % train, 30 % test). In the

case of PC, this test was repeated for significance levels 1 % and 5 %. In the case of CRPA-UC, the combination of significance level (1 % and 5 %), minimum support (1 % and 5 %) and uncertainty coefficient (0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70 and 0.80) were tested. These tests concluded that the algorithm's error in the three data sets did not change much when the parameters were changed. For this reason, for all the data sets, we selected a significance level and minimum support of 1 % and a minimum accepted coefficient of 0.60 (since this coefficient represents the strength of the relation, with this value we can find relationships that are moderately strong and avoid the weak ones). For easier comparison, in the tests presented in the following sections, only simple rules ($A \Rightarrow B$) will be considered for CRPA-UC.

3.4 Results

Next, we evaluate the proposed approach by studying first the rules generated by the proposed approach and accessing their quality. Secondly, we study the application of CRPA-UC in prediction problems.

These algorithms' performance was compared in terms of error rate (Table 3.4). This comparison was performed using the PC algorithm as a reference. The performance of CRPA-UC in each data set was compared to the reference using the Wilcoxon signed ranked-test. The sign $+/-$ indicates that the algorithm is significantly better/worse than the reference with a p-value of less than 5 %. Besides this, the algorithms are also compared in terms of the average and geometric mean of the errors, average ranks, average error ratio, win/losses, significant win/losses (number of times that the reference was better or worse than the algorithm, using signed ranked-test) and the Wilcoxon signed ranked-test. For the Wilcoxon signed ranked-test, we also consider a p-value of 5 %.

3.4.1 Pattern Metrics Evaluation

As stated before, first, we assess the validity of rules generated by CRPA-UC and the overall quality of the generated models. We compared this methodology with PC, used as a baseline, in four publicly available datasets. As for evaluation metrics, in these tests the metrics presented in Chapter 2, Section 2.2.5 were employed.

TABLE 3.3: Pattern Metrics for Asia (8 edges), Cancer (4 edges), Sachs (17 edges) and Lucas (12 edges) data set

Dataset	Sachs		Cancer		Asia		Lucas	
Algorithm	PC	CRPAUC	PC	CRPAUC	PC	CRPAUC	PC	CRPAUC
Missing Edges	7	2	0	0	3	4	0	0
Extra Edges	0	0	0	0	0	1	0	1
Correct Adjacencies	10	15	4	4	5	4	12	12
Incorrect Adjacencies	0	0	0	0	0	1	0	1
Correct Directed Edges	0	10	2	2	2	2	7	9
Incorrect Directed Edges	10	5	2	2	3	2	5	3
SHD	17	7	2	2	6	7	5	4
SID	60	54	10	10	30	33	41	38

If we analyze Table 3.3, we can see that, in general, CRPA-UC tends to have more edges that are correctly directed when compared with PC and fewer misdirected edges. However, despite this, our approach tends to have more extra relations than PC (except in the case of *sachs* data set).

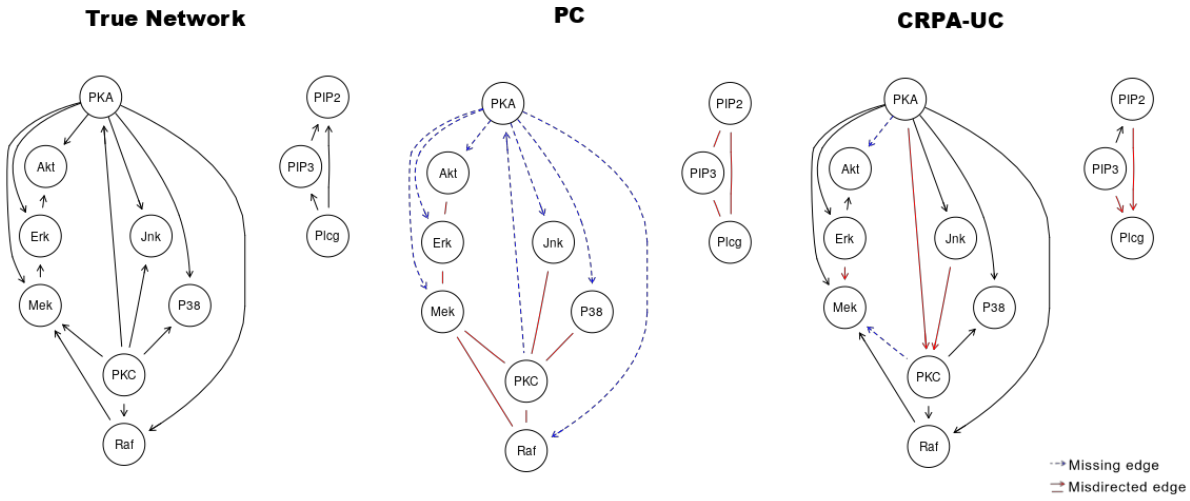


FIGURE 3.2: True networks and graphs generated by PC and CRPA-UC for data set *Sachs*

If we analyze Figure 3.2, which represents the comparison of the networks generated by PC and CRPA-UC with the true network for data set sachs, we can see that PC did not direct any edges and found a lower number of edges, when compared with CRPA-UC, that directed almost every edge it found correctly. This difference can be explained by the fact that these algorithms apply different independence tests (PC usually applies the G^2, which is similar to χ^2), which means that in theory, they can obtain different dependencies. The orientation method is also different: PC applies a set of orientation rules [94],

whereas CRPA-UC applies a coefficient that explains how a variable can predict another, and that is why it transmits more information about the relationship, such as its strength. Besides this, if we analyze the data itself, we can see that asia, cancer and lucas data sets are binary. At the same time, sachs has non-binary categorical data (three categories per variable). From this, we can conclude that both approaches work similarly in binary data, while CRPA-UC appears to find more correct relations than PC in non-binary data. This number of correct relationships might happen because it is impossible to presume any order in the binary data sets. In sachs's case, the change in category is intrinsically connected with the changes in other variables (for example, if one gene takes the value of high, another gene can go low) [98].

Analysing now the two measures also presented in Table 3.3 (SHD and SID), it is possible to see that CRPA-UC in general has a better performance than PC (only having worse performance in *asia* data set).

3.4.2 Prediction

As the predictive ability of an algorithm is essential to determine its ability to solve problems, we also analyzed the proposed method in several predictive problems, using the datasets presented in Table 3.1.

If we analyze Table 3.4, it is possible to see that, in general, the CRPA-UC has a better performance than PC since the value obtained in the Wilcoxon test is 0.008 73 or 0.873 % (less than the p-value of 5 %), which means that the difference between the performance is significant. This difference can also be seen in the values of the average and geometric ranks. More specifically, if we look at the average ranks, we can see that CRPA-UC has lower ranks (on average) than PC (1.077 against 1.769). The result obtained in these tests reinforces what was demonstrated in the previous section: the combination of *independence test-orientation method* has a beneficial impact on performance. This fact can be explained in two ways: first, by the difference in the way G^2 and GCMH calculate the dependencies. While G^2 is based on the log-likelihood-ratio, GCMH is a generalisation of the McNemar test [99]. This can explain the difference in the found relationships. Second, we use an orientation method that, besides orienting the relationships, can also find dependence between the variables (eliminating the weaker relations) to obtain more information about them.

TABLE 3.4: Error rates of PC and CRPA-UC in classification problems

Data set		PC	CRPA-UC
1	asia	15.18 ± 1.47	15.18 ± 1.47
2	cancer	1.05 ± 0.28	1.00 ± 0.25
3	coronary	14.13 ± 2.43	14.13 ± 2.43
4	earthquake	0.77 ± 0.33	0.67 ± 0.33
5	gmb	37.34 ± 4.92	+ 26.50 ± 5.07
6	lucas	20.02 ± 3.38	18.15 ± 4.02
7	monica	44.48 ± 3.35	+ 14.50 ± 0.67
8	mux6	61.86 ± 9.66	+ 45.26 ± 14.94
9	pre sex	24.64 ± 3.90	23.84 ± 3.32
10	sachs	39.61 ± 1.23	+ 34.49 ± 1.34
11	survey	43.95 ± 0.91	44.17 ± 1.12
12	titanic	24.16 ± 7.28	22.64 ± 3.45
13	youth risk 2009	40.80 ± 7.07	40.40 ± 6.31
Average Mean		25.385	20.602
Geometric Mean		17.594	14.709
Average Ranks		1.769	1.077
Average Error Ratio		1	0.842
Wicoxon test			0.00872
Win/Losses			10/1
Significant win/losses			4/0

3.5 Summary

Causality has become an increasingly studied topic in machine learning/data mining. Although Bayesian networks are among the favourite algorithms for applying causal discovery in observational data, more and more causal discovery algorithms have appeared that do not fall into this category in recent years.

One example is the causal rule discovery algorithms. There are already a few approaches that combine causality with association rules. However, these methods have some disadvantages: they can only be used for local structure discovery in binary data and apply a naive approach as orientation method.

This work proposes a global causal association discovery algorithm for binary and non-binary discrete data: CRPA-UC. In this method, we apply a combination of two independence tests, as well as the UC as a direct method. We compared this approach with PC in the experiments using public data sets. From these results, we can conclude that applying a more powerful independence test with an orientation method gives information about the variables' dependency and positively impacts the method's performance.

Chapter 4

Semi-Causal Decision Trees

Most classification algorithms use correlation analysis to make decisions with satisfactory results. This correlation can be understood as a statistical association between two random variables and can be very advantageous for classification algorithms since it uncovers the predictive relationship between variables.

Although correlation is an essential predictive clue, the information retrieved often does not make sense by real-world standards. However, despite this lack of human rationality, these relations can be evidence of a stronger type of relations: causal relationships.

Causality, more specifically, causal discovery, is the field that combines machine learning, data mining and statistics to search for potential cause-effect relationships in observational data [25]. The application of causal discovery in the various tasks of machine learning can be a challenge, both at the level of the causal process application itself and at the sampling process to generate the observed data [41]. Despite this, causal discovery has been the study focus of several researchers over the years, given its importance and the potential impact that the causal relationships' discovery between events can have in the problem-solving, namely at an interpretability level since this type of analysis can potentially uncover the underlying relationships between the variables, thus being possible to explain and sustain decisions more easily.

With that said, it is crucial to understand the difference between cause and correlation. As stated in Chapter 2, correlation is not the same as causation because, although there might exist 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, thus not providing enough information about the occurrence of events. There are several reasons why these correlations are similar to causality, such as data and links that go against established rules. Nevertheless, correlation is still important in finding the true relationship between events.

In real-world problems, the data is a mixture of causal and correlation relationships [100]. However, traditional causal and correlations models usually ignore the other types of relationships since it is possible to have causal relationships without correlation and vice-versa. This fact can lead to a loss in the interpretability in correlation-based models' interpretability, and prediction power, in causal based models (sometimes, correlation-based classification algorithms have better results than causal discovery algorithms, even in data with clear causal relationships [101]).

In this chapter, we aim to answer **RQ.2** and **RQ.5**: *Is it possible to obtain more interpretable models by using causal discovery?* and *Are causal relationships helpful, and can they bring significant gains?* To answer these questions, we start by studying the potential usage of decision trees to infer causal relationships from cross-sectional observational data.

Decision trees are a notorious type of algorithm, widely used for most problem-solving tasks. The generated models are easily understandable since the models' information is presented in a tree-like structure. Despite this, the trees' presented information is based on correlation metrics, meaning that found relationships may not make sense for the user (for example, the number of Nicolas Cage movies influences the number of deaths in swimming-pools[*]). To deal with this problem, Li et al. [8] proposed the Causal Decision Trees, introducing causal discovery to split the data, thus creating trees in where all the non-leaf nodes present a causal relationship with the outcome. However, this method ended up losing some predictive power [102].

For this reason, we propose a semi-causal hybrid approach that takes advantage of the correlation's predictive power and the causal discovery's interpretability potential to improve the performance of Decision Trees in discrete data. This is done by defining a new data splitting score method that merges the information gain/gain ratio approach usually

[*] https://www.nationalgeographic.com/science/phenomena/2015/09/11/nick-cage-movies-vs-drownings-and-more-strange-but-spurious-correlations/

used in Decision Trees with a simple causal discovery approach (GCMH and the Uncertainty Coefficient as a directional method) that detects and scores potential causal relationships. With this approach, we intend to create more concise trees than the traditional decision trees that are easily understandable for the average human being but are not so small that it is impossible to obtain accurate predictions. We also want to assure that the algorithm prioritises variables that are causes of the target, securing that the higher nodes in the tree are causally related to the outcome. Furthermore, the correlation component (gain ratio) is applied in every split to achieve and maintain predictive power throughout the tree's construction. This means that every choice made by the algorithm is based on strong causal relationships between strongly correlated variables. Finally, this splitting score has the advantage of guaranteeing that, if at any moment, the causal relationships are missing from that split due to data division, if there is a strong correlation between the variables, the algorithm will still split by the highest correlation, thus assuring that it is possible to find causal relationships in further splits.

Ultimately, the goal is to create a tree with strong causal and correlational relationships in the top tree nodes but also allow the tree to further split, even without causal relationships in a given division, assuring:

1. if possible, find more causal relationships in further splits;

2. the tree is not prematurely ended because there are no causal relationships.

4.1 Problem

As stated previously, the current causal decision trees methodologies CDT-PS and CDT-SPS, presented in Section 2.2.4 have several limitations that restrain their usage to few potential cases. These methods do not assure that the variables chosen are, in fact, the causes of the outcome and not the other way around. In practice, this means that when upon a potential causal undirected relationship between two variables A and T ($A - T$) found by the conditional independence test, the algorithm naively always chooses the relationship's direction as $A \rightarrow T$, which is not always true. Besides, this method can only be used in binary data sets.

Finally, and as mentioned before, these methods lose predictive power due to creating too concise trees. As it is known, the number of causal relationships typically found in a data

TABLE 4.1: Binary data set description

Data set		Number of examples	Number of attributes	Number of classes
1	asia	10000	8	0 (44%) 1 (56%)
2	corral[a]	160	7	0 (56%) 1 (44%)
3	earthquake	10000	5	0 (2%) 1 (98%)
4	GMB	5000	5	0 (62%) 1 (38%)
5	lucas	2000	12	0 (28%) 1 (72%)
6	mux6	128	7	0 (50%) 1 (50%)
7	PreSex	1036	4	0 (77%) 1 (23%)
8	threeOf9[a]	512	10	0 (54%) 1 (46%)
9	Titanic*	2201	7	0 (68%) 1 (32%)
10	xd6[a]	973	10	0 (67%) 1 (33%)

[a] https://www.openml.org

set is lower than the number of correlational relationships. In some instances, the small size of the trees is so extreme that there is only one non-leaf node and two leave nodes ($T_1 \leftarrow A \rightarrow T_2$). In practice, this might be because, despite several causal relationships present in the data, the target only has a causal relationship with one variable, or the split translates into a perfect separation of the target's values. This means that, for achieving accurate results, a balance in the tree size (not too small that the results are too random and not too big that is difficult to understand) is needed.

Data

To analyse and solve this problem, throughout the following sections, we will use publicly available datasets as shown in Table 4.1 and Table 4.2. This data was gathered from several public databases and are publicly available for usage.

4.2 Methodology

Decision Trees are a well-known algorithm, widely used for various problem-solving tasks, for imitating human reasoning, and thus being easily understandable. In this algorithm, the leaves represent the learned classes, and the branches represent the conjunctions of characteristics of that class. This algorithm is a classification approach that

*This data set is provided with CDT Weka jar file as an example (nugget.unisa.edu.au/jiuyong/)

TABLE 4.2: Non-binary data set description

	Data set	Number of examples	Number of attributes	Number of classes
1	medpar[b]	1495	9	0 (66%) 1 (34%)
2	monica	6367	12	0 (55%) 1 (45%)
3	respiratory[b]	555	5	0 (51%) 1 (49%)
4	sachs	10000	11	0 (60%) 1 (26%) 2 (13%)
5	titanic	1316	4	0 (62%) 1 (38%)

[b] https://vincentarelbundock.github.io/Rdatasets/articles/data.html

chooses the optimal solution at each stage, meaning it decides at each iteration which is the best solution for that iteration. This solution predicts new nodes for the tree until it cannot predict a better solution than the previous one.

To decide which is the optimal variable to split in the tree, decision trees use the Information Gain (IG) to rank the attributes according to the information gained from the attribute's value. It is obtained by the difference between the expected information or entropy *E(P)* of the target variable P before splitting and the weighted entropy after splitting the data by the attribute values of A (4.1) (D represents the data set, N represents the number of classes and v each of A's value) [103, 104].

$$\begin{aligned} E(Y) &= -\sum_{i=1}^{N} y_i \, log_2(y_i) \\ IG(Y, A) &= E(P) - \sum_{v \in A} \frac{|D_v|}{|D|} \times E(Y_v) \end{aligned} \tag{4.1}$$

Since IG can be biased towards variables with a larger number of distinct values, some decision trees' versions like J48 (on which our approach is based) use a modified version of this metric, called information gain ratio (4.3).

$$IV(S) = -\sum_{j=1}^{N} \frac{|S_j|}{S} \times log_2 \frac{|S_j|}{S} \tag{4.2}$$

$$IGR(S) = \frac{IG(S)}{IV(S)} \tag{4.3}$$

This ratio corrects the IG obtained for a variable using the intrinsic information of the split (also called intrinsic value) [105], which is the representation of the potential information generated if we split a data set in N partitions (4.2).

Although this metric can quantify how much information can be obtained by each variable, thus maximising the information gained in each tree's split, it does not give any information about the potential causal relationships between the variables that, in conjunction with this correlational information [106], might boost its interpretability.

This section presents a new approach that merges information about potential causal relationships between the variables with the information gained from them: SC Tree.

In this modified version of Decision Trees, in each split, besides calculating each variable's potential gain of information, it also tests if there is a potential causal relationship between them and the target variable. This is done by applying a causal discovery methodology that searches for potential causal relationships between the variables.

4.2.1 Local Causal Discovery Module

As explained previously, the proposed methodology takes advantage of the potential causal relationships found in the data to create the model. The discovery of these relationships is made in every split with the available data at that moment. This causal module searches for potential causal relations $variable \rightarrow class$ (being $class$ the target variable) and uses the measured dependence output as an input to calculate the best attribute to split.

At this moment, and before going into more detail about the module, some key definitions must be reminded, such as what is a (potential) causal relationship and how it is possible to measure them. By causal relationship, we comprehend the relation between different events, in which one is identified as the cause of the other. This means that, in theory, if the first event occurs, we can also expect the second to occur. On the contrary, if the first event does not occur, it is expectable that the second does not occur also.

Generally, a causal algorithm can be classified as a local or global discovery algorithm, depending on its purpose and mechanism applied to find relationships.

Although there is not a well-defined answer for this distinction between causal algorithms, we can define a global causal discovery algorithm (also known as global structure

discovery) as an algorithm that tries to search for all the existing potential causal relationships between several variables [107]. This type of algorithm is usually used to study the general causal interactions in a given system.

As for local causal discovery algorithms (also called local structure discovery), their objective is only to find causal relationships for a specific variable instead of all the variables [107]. This algorithm is mainly used in two specific cases, such as problems with high dimensional data or feature selection problems.

Besides only searching for causal relationships with a target variable, local causal discovery differs from global causal discovery in another aspect. Typically local causal discovery algorithms return only undirected causal relationships, *i.e.*, they find relationships but do not give any information about the direction of those relationships. In contrast, global causal discovery algorithms perform an extra step to find the direction of these relationships.

Arguably, in machine learning, the most common form found in the literature to measure if a relation is causal or not is through conditional independence tests [41]. State of the art global causal algorithms such as PC, FCI, among others [94] apply these tests to undercover which variables are independent of which, hence remaining the potential dependent ones.

As causal relations tend to be maintained in the presence of others' influence, these conditional independence tests verify whether the potential relations are maintained when one, two, or several variables influence their values (thus not sustaining the claim that they are causally related).

Several conditional independence tests can be applied, but the most commonly used for discrete data are χ^2 and G^2 [94]. Although both of these methods are Chi-squared based independence tests and are widely used for being independent of order, there are several limitations. For example, in χ^2 case, by merely testing if two variables are dependent on each other only using the information from these two variables, it is not possible to say for sure if they are causally dependent (since, in general, a causal relationship remains, even when other factors are influencing the relationship, *i.e.* if we have three variables A, B and C, we can only say that A and B are causally related if this relation is maintained when C also influences it) [54, 108]. Although G^2 solved this problem by inserting the influence

of other variables in the dependence calculation, it seems to be sensitive to sample size [109, 110], meaning that it does not detect well relationships in small data sets.

Some tests search for this type of association, called partial association (statistical measure to find conditional independence in controlled experiments [111]). One example is the GCMH.

As mentioned earlier, despite conditional independence tests being a vital component to finding potential causal relations, it is important to note that these tests do not direct the potential dependences, *i.e.*, if A is the potential cause of B or vice versa, they only hint that there is a relationship. As stated in Chapter 2, this orientation can be done by using a set of established rules, by using experimental data to orient the edges or by using a mixture of both the previous approaches.

Returning our focus to the proposed module, although it can be classified as a local causal discovery since it searches for causal relationships for a specific variable, it is a crucial difference closer to the global causal discovery algorithms. Instead of searching and returning the indirect causal relationships, this module directs all causal relationships found using an asymmetric dependence measure and returns only the causes of the target variable ($variable \rightarrow target$) and the respective coefficient (Algorithm 4.1)*. With this extra orientation step, we assure that the chosen variables are the ones that influence our target directly (they cause it) and not only (causally) related to the target, like, for example, the CDT that choose any variable that is related to it, without verifying if it is its cause or if its caused by it.

In this module, we propose the usage of the GCMH test instead of the traditional χ^2 or G^2 regularly used in literature because it mitigates the problems presented previously. Besides this fact, the GCMH test can also be used in both binary and non-binary discrete data. As an orientation method, we propose the UC, which measures how dependent two variables are, with the values of this coefficient between 0 and 1. Since this coefficient is asymmetric, it is possible to determine the direction of the dependence by comparing the obtained value for $A \rightarrow B$ and $B \rightarrow A$, choosing the direction of the most significant dependency.

*it is important to note that from now on causal sufficiency and faithfulness are assumed

Algorithm 4.1: CAUSALM: module for finding potential causal relationships and respective UCs

Input: Let **D** be a data set with a set of variables $S = \{s_1, s_2, ..., s_n\}$ and a target variable t. Let $\boldsymbol{\alpha}$ be the significance level for the conditional independence tests and **ct** the correspondent critical value. Let $\mathbf{u_{coef}}$ be the minimum accepted coefficient.

Output: **R**, list of all causal relationships and respective UCs

1 **for** ***each*** *pair of variables* $\{s, t\}$ *in D, with distribution* $d = dist(s, t)$ **do**
2 **if** $Generalised\_CMH(d) \geq ct$ *verifies* **then**
3 Verify $s \rightarrow t$ and $t \rightarrow s$ directions using the uncertainty coefficient (UC)
4 **if** *the coefficient of* $s \rightarrow t$ *is higher than* $t \rightarrow s$ *and* u_{coef} **then**
5 Save s and the respective coefficient in R
6 **return** R

To find all possible causal relationships with the target variable, the causal module (Algorithm 4.1) starts by applying the independence test to all variables (with a level of significance defined *a priori*) to determine what are the possible relationships between these and the target (line 2). To all the chosen variables (that is, they are partially associated with the target), the UC is applied. This is done, by testing both $variable \rightarrow target$ and $target \rightarrow variable$. Suppose $variable \rightarrow target$ is the one with the highest coefficient, and its coefficient is higher than a user-defined minimum coefficient (to assure that only strong dependencies are chosen). The variable and respective coefficient are saved (this value will be used later in the variables' importance calculation).

4.2.2 Revisiting SC Trees

Returning now our attention to the proposed algorithm, its operation is similar to the traditional decision trees in that it applies a divide and conquer methodology to build the tree, as we can see in Algorithm 4.2.

The main difference between the traditional method and the proposed method lies in the way the attributes' importance is calculated: instead of using only the value of IG as a measure of importance, in this algorithm, we propose the use of the sum of IG with the uncertainty coefficient (if there is a causal relationship), with defined weights (denominated in this work as semi-causal information gain or SC_{IG}) in the information gain ratio's calculation (denominated in this work as semi-causal information gain ratio or SC_{IGR}).

Algorithm 4.2: SCT: Semi-Causal Tree

Input: Let **D** be a data set with a set of variables $S = \{s_1, s_2, ..., s_n\}$ and a target variable t

Output: **Tree**

1 Tree = {}
2 **if** *D is pure OR other stop criteria is met* **then**
3 | **return** *Tree*
4 Map all the potential causal relationships in D with the t using CAUSALM
5 **for** *all attribute a in D* **do**
6 | Compute criteria of semi-causal gain ratio if we split on a (4.5)
7 a_{best} = Best attribute according to the above-computed criteria (an attribute that maximises the gain ratio)
8 Tree = Create a decision node that tests a_{best} in the root
9 D_v= Induced subsets from D based on a_{best}
10 **for** *all* D_v **do**
11 | $Tree_v$= SCT(D_v)
12 | Attach $Tree_v$ to the correspondent branch of the Tree

$$SC_{IG}(X) = \begin{cases} (\beta \times IG_X) + (\theta \times UC_X) & if\ GCMH_X < \alpha \quad UC_X \geq uc_{coef} \\ \\ IG_X & otherwise \end{cases} \tag{4.4}$$

$$SC_{IGR}(X) = \frac{SC_{IG}(X)}{IV(X)} \tag{4.5}$$

Therefore, and as we can see in Algorithm 4.2, in each split, the algorithm begins by searching all potential relationships with the target (line 5). This information is then used to calculate the semi-causal information gain ratio (4.5) of each variable, then choosing the one with the highest value.

It is important to note that the first statement of (4.4) is only used in the gain ratio calculation if and only if there is evidence of a (strong) causal relationship between the target and the current variable, that is given first and foremost by the GCMH test (Algorithm 4.1), that is responsible for accessing if there is a causal dependence between a variable and the target. Only after this causal dependence is established is the UC applied only to assure that the relationship is strong and that the direction is the desired one ($variable \rightarrow target$). If the equation's condition does not hold, only the IG (the second statement) is used in the gain ratio calculation.

This process of choosing the optimal variable and splitting by it is repeated until the stop criteria are met. For example, in SC Tree, the default criteria used to stop the creation of the tree is the following: there is no attribute with a positive IG ratio, or the minimum number of instances per leaf was met [112].

4.3 Experimental Setup

To evaluate the proposed approach and make a comparative study, the following configuration of experiments was designed:

- First, we investigate how alterations in the proposed IG formula (4.4) influence the outcome. We compared the different alterations using 10-fold cross validation, in several public data sets (Table 4.1 and Table 4.2);
- Second, we compare the proposed approach with other causal and non-causal decision tree based approaches (CDT, explained in detail in Section 2.2.4, and J48), using 10-fold cross validation, in several binary public data sets (Table 4.2);
- Finally, we compared the proposed approach with the current state of the art causal discovery method, PC [94], using 10-fold cross validation, in several public data sets (Table 4.1 and Table 4.2).

A sensitivity analysis was performed to choose the optimal parameters for the approach presented in the following sections. This analysis consisted of obtaining the error (*1 - accuracy*) for the presented data sets (by dividing them into 70 % train, 30 % test). In the case of of the proposed approach this test was repeated for significance levels (1 % and 5 %) and UC (0.50, 0.60, 0.70 and 0.80). These tests concluded that the algorithms' errors in the data sets did not change significantly when the parameters were changed. For this reason, for all the data sets, we select and present a significance level of 5 % and UC of 0.60. Furthermore, so that the tree's growth is not restricted, we set the minimum number of instances per leaf node as 2.

As for the baselines (CDT, PC and J48), for CDT* the default setting proposed by the authors (maximum height of the decision tree of 5 and performing pruning). For J48†

*we used the WEKA jar file provided by the authors to compare with our methodology
†we used the WEKA implementation

we use pruning confidence of 0.25 and the minimum number of instances per leaf of 2. Finally, for PC, we use a 5 % significance level.

4.4 Results

Our proposed algorithm's performance is compared with several baselines in terms of error rate, tree size, and the average number of causal relationships found in the following sections. In addition, the performance of SC Tree in each data set was compared to the baselines using the Wilcoxon signed ranked-test. The sign $+/-$ indicates that the algorithm is significantly better/worse than the reference with a p-value of less than 5 %. Besides this, the algorithms are also compared in terms of the average and geometric mean of the errors, average ranks, average error ratio, win/losses, significant win/losses (number of times that the reference was better or worse than the algorithm, using signed ranked-test) and the Wilcoxon signed ranked-test. For the Wilcoxon signed ranked-test, we also consider a p-value of 5 %.

4.4.1 SC Tree's possible configurations

In the previous section, we presented the optimal values for the causal module (optimal α and optimal minimum UC). In this section, we present a thorough investigation of how different configurations of weights in (4.4) influence the overall performance (by influencing the value obtained in (4.5)) of the proposed methodology.

To test its performance, we propose the following configurations for weights β and θ:

1. $\beta = 1$ and $\theta = 1$: this configuration will be our reference, since it gives equal weights to IG and UC;

2. $\beta = 2$ and $\theta = 1$: in this configuration we give more importance to the IG than the UC;

3. $\beta = 1$ and $\theta = 2$: in this configuration we give more importance to the UC than the IG;

4. $\beta = 0$ and $\theta = 1$: this configuration differs from the previous since it uses only causal information if available, *i.e.*, if in a split there is at least one causal relationship the method uses that information to do the split; if there is no causal relationship

available, then it uses the IG. This configuration is presented in (4.6), where $cr(X,T)$ represents the list of all causal relationships detected in that split.

$$SMC_{IG}(X) = \begin{cases} UC_X \quad if\ \exists X.cr(X,T) \quad GCMH_X < \alpha \quad UC_X \geq uc_{coef} \\ \\ IG_X \qquad otherwise \end{cases} \tag{4.6}$$

We compared these configurations in terms of error rate (Table 4.3). For this comparison, we used the metric presented in Section 4.3.

TABLE 4.3: Error rates of SC Tree in several configurations

	Data set	SC Tree $(IG + uc)$	SC Tree $((2 \times IG) + uc)$	SC Tree $(IG + (2 \times uc))$	SC Mixed Tree
1	asia	14.60 ± 1.42	14.60 ± 1.42	14.60 ± 1.42	14.60 ± 1.42
2	corral	2.50 ± 4.37	2.50 ± 4.37	2.50 ± 4.37	2.50 ± 4.37
3	earthquake	0.23 ± 0.11	0.23 ± 0.11	0.23 ± 0.11	0.23 ± 0.11
4	GMB	15.00 ± 1.77	15.00 ± 1.77	14.98 ± 1.73	14.98 ± 1.73
5	lucas	14.50 ± 1.35	14.50 ± 1.35	14.50 ± 1.35	14.50 ± 1.35
6	medpar	32.84 ± 4.30	32.84 ± 4.30	32.84 ± 4.30	32.84 ± 4.30
7	monica	14.42 ± 1.99	14.42 ± 1.99	14.42 ± 1.99	14.42 ± 1.99
8	mux6	19.55 ± 12.91	18.72 ± 13.94	21.22 ± 12.37	28.21 ± 8.68
9	PreSex	21.52 ± 3.77	21.52 ± 3.77	21.52 ± 3.77	21.52 ± 3.77
10	respiratory	39.62 ± 3.57	39.62 ± 3.57	39.62 ± 3.57	39.62 ± 3.57
11	sachs	22.20 ± 1.65	22.20 ± 1.65	22.20 ± 1.65	22.20 ± 1.65
12	threeOf9	2.55 ± 2.46	2.55 ± 2.46	2.55 ± 2.46	2.55 ± 2.46
13	titanic	20.82 ± 4.06	20.82 ± 4.06	20.82 ± 4.06	20.82 ± 4.06
14	Titanic	21.86 ± 3.85	21.86 ± 3.85	21.86 ± 3.85	21.86 ± 3.85
15	xd6	0.31 ± 0.50	0.31 ± 0.50	0.31 ± 0.5	0.31 ± 0.50
Average Mean		16.19	16.11	16.28	16.74
Geometric Mean		8.66	8.63	8.70	8.87
Average Ranks		1.13	1.20	1.20	1.13
Average Error Ratio		1	0.99	1	1.03
Wicoxon test			1	1	1
Win/Losses			1/0	1/1	1/1
Significant Win/Losses			0/0	0/0	0/0

As it is possible to spot in Table 4.3 the different configurations are relatively similar, with configuration 2 ($\beta = 2$ and $\theta = 1$) being the best (in terms of average mean error rate) and configuration four being the worst. Despite being more explicable, this difference shows that purely causal approaches or approaches that privilege causal relationships do

not perform as well on classification problems as algorithms that privilege correlation. Although the results are different, this difference is not significant.

For this reason, in the following sections, we will use configuration one since it gives equal importance to both measures, not privileging one over the other.

4.4.2 Comparing Decision Tree Approaches

To understand whether the mixture between correlation and causality would improve the performance while maintaining causal coherence, we compared the proposed approach with three decision tree-like approaches CDT-PS, CDT-SPS and J48.

As it was explained in Section 2.2.4, CDTs can only be used in binary data. Although the proposed methodology can be used in both binary and non-binary discrete data, for the sake of consistency, in this section we will only compare SC Tree with both CDT approaches using binary data sets. We compared the proposed approach in terms of error rate in several binary discrete data sets (Table 4.2). For this comparison we used the metrics presented in Section 4.4 and used SC Tree as reference.

TABLE 4.4: Error rates for SC Tree, CDT-PS, CDT-SPS and J48

	Dataset	SC Tree	CDT-PS	CDT-SPS	J48
1	asia	14.60 ± 1.42	- 15.58 ± 1.65	- 33.55 ± 15.25	14.59 ± 1.42
2	corral	2.50 ± 4.37	- 43.13 ± 18.03	- 40.00 ± 18.45	1.25 ± 3.95
3	earthquake	0.23 ± 0.11	- 1.44 ± 0.37	- 2.45 ± 1.86	0.24 ± 0.14
4	GMB	15.00 ± 1.77	- 19.91 ± 4.83	15.18 ± 1.99	15.06 ± 1.95
5	lucas	14.50 ± 1.35	15.30 ± 4.31	15.30 ± 4.31	13.86 ± 1.38
6	mux6	19.55 ± 12.91	- 52.31 ± 8.54	- 54.68 ± 10.09	10.38 ± 10.34
7	PreSex	21.52 ± 3.77	- 23.26 ± 4.58	- 23.26 ± 4.58	22.49 ± 4.37
8	threeOf9	2.55 ± 2.46	- 23.83 ± 4.01	- 24.42 ± 3.96	2.93 ± 2.10
9	Titanic	21.86 ± 3.85	- 26.61 ± 6.46	- 26.61 ± 6.46	20.94 ± 3.74
10	xd6	0.31 ± 0.50	- 15.31 ± 4.66	- 15.93 ± 2.82	0.21 ± 0.43
Average Mean		11.26	23.67	25.14	10.20
Geometric Mean		5.16	17.95	19.96	4.40
Average Ranks		1.60	3.20	3.50	1.40
Average Error Ratio		1	9.07	9.71	0.89
Wicoxon test			0.002	0.002	0.375
Win/Losses			10 0	0/10	6/4
Significant Win/Losses			9 0	0/8	0/0

As it is possible to see in Table 4.4, in general, the proposed approach has better performance than the causal decision tree algorithms, since the values obtained from the Wilcoxon test are 0.002 for both (CDT-PS) and (CDT-SPS). This means that there is a significant positive difference between the errors of the three methods (the p-value is lower than the stipulated p-value of 5 %). This difference can also be spotted in the average and geometric ranks' values, especially in the average error ratios. The proposed approach has a significantly lower rank than the other methods.

Compared to J48, the proposed approach is very similar to it in terms of performance since the Wilcoxon test value is 0.375, meaning that, although the splitting method was changed, these changes did not have a significant impact in SC Tree's performance.

As it is possible to observe in Figure 4.1, that presents a critical difference diagram (using the Nemenyi test with 5 % significance), comparing SC Tree, both CDT methods and J48, the algorithm that presents the best performance is J48, since it is the algorithm that appears first in the diagram (the closer to 1 the algorithm is, the better it is the performance). This algorithm is closely followed by SC Tree, that presents similar results to J48 (this information is in line with what was presented in Table 4.4).

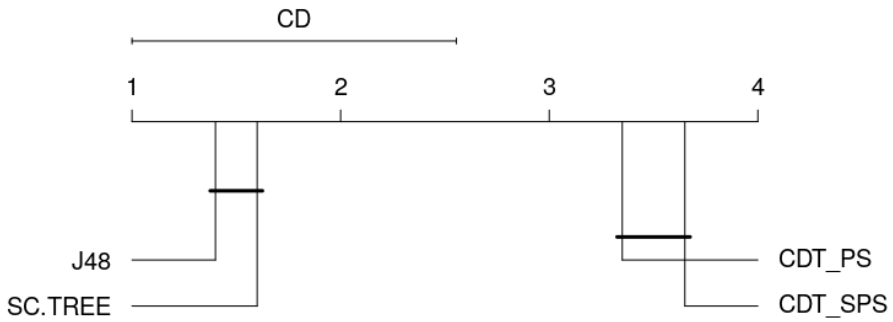


FIGURE 4.1: Critical difference diagram for SC Tree, CDT-PS, CDT-SPS and J48 (error rate)

As tree-based algorithms, some important metrics to investigate are the tree size, number of leaves, and depth (since it entails how much information is present in a tree and how interpretable a tree is for the average user) [113, 114]. This topic is closely related to Occam's razor, which states that one model should be preferred over another if the first has a simpler structure than the second, as it will generalise better [115].

As causal approaches, it is expected that CDT-PS and CDT-SPS generate smaller and more concise trees since the list from where they can retrieve relations is much smaller

(when compared to correlation relationships) [8]. As for the proposed approach, since it is a causal/correlation mixed approach that selects variables that are both strong causally related and correlated with the target, it is expected that the tree size lies between the CDT's and J48.

As it is possible to see in Table 4.5, that presents the average size of the trees generated for several binary data sets, using 10-fold cross-validation, in general, the pure causal approaches (CDT-PS and CDT-SPS) generate significantly smaller trees (with 0.002 in the Wilcoxon test for both approaches) than the SC Tree, while J48 generates significantly bigger trees than the proposed approach (0.002 in the Wilcoxon test). This same behaviour can be spotted in Table 4.6 and Table 4.7, which entail the average depth and the average number of leaves, respectively. In these tables, it is possible to see that CDT-PS and CDT-SPS generate shallow trees with a reduced number of leaves (this difference being significant when compared to SC Tree), while J48 creates deep trees with several leaves.

From these tables, we can assess that:

1. SC Tree creates bigger trees than the CDTs approaches, because it uses the IG as information whenever there is no causal information. Moreover, the usage of the semi-causal information ratio (4.5) leads to different splitting decisions, thus creating trees that evolve differently (this will be shown in Section 4.4.2);

2. SC Tree creates smaller trees than J48 because it restrains the tree's construction by adding further splitting constraints to the process.

Finally, it is essential to note that, although the CDTs approaches created relatively small and highly interpretable trees, this is done at the cost of giving almost no information about the system.

While there appears to be a difference between them, if we compare all four approaches using a critical difference diagram again, with significance level 5 % (Figure 4.2, Figure 4.3 and Figure 4.4). We can see that, while the difference in tree size is only significant when we compare J48 with CDT-PS and CDT-SPS, there is a significant difference between CDT-PS and SC Tree in both depth and number of leaves.

TABLE 4.5: Average tree size for SC Tree, CDT-PS, CDT-SPS and J48

	Dataset	SC Tree	CDT-PS	CDT-SPS	J48
1	asia	6.60 ± 0.80	+ 3.00 ± 0	+ 2.60 ± 1.84	7.00 ± 0
2	corral	12.40 ± 3.16	+ 1.00 ± 0	+ 1.40 ± 1.26	13.80 ± 2.53
3	earthquake	8.40 ± 3.13	+ 3.00 ± 0	+ 1.00 ± 0	- 12.00 ± 2.54
4	GMB	7.40 ± 1.26	+ 3.00 ± 0	-11.60 ± 3.29	- 12.20 ± 1.69
5	lucas	18.20 ± 4.02	+ 13.60 ± 4.43	+ 13.60 ± 4.43	- 37.60 ± 5.08
6	mux6	23.20 ± 6.89	+ 3.80 ± 2.86	+ 3.00 ± 2.82	- 39.20 ± 5.20
7	PreSex	5.00 ± 0	+ 1.00 ± 0	+ 1.00 ± 0	5.20 ± 0.63
8	threeOf9	48.60 ± 3.37	+ 22.20 ± 2.70	+ 21.20 ± 1.99	- 55.80 ± 4.73
9	Titanic	8.60 ± 0.84	+ 3.40 ± 2.80	+ 3.40 ± 2.80	9.20 ± 0.60
10	xd6	61.00 ± 0	+ 18.20 ± 1.93	+ 17.20 ± 2.74	- 71.67 ± 2.90
Average Mean		19.94	7.22	7.60	26.37
Geometric Mean		13.92	4.25	4.29	18.2
Average Ranks		2.9	1.5	1.3	4
Average Error Ratio		1	0.36	0.44	1.34
Wicoxon test			0.002	0.002	0.002
Win/Losses			10/0	9/1	0/10
Significant Win/Losses			10/0	9/1	0/6

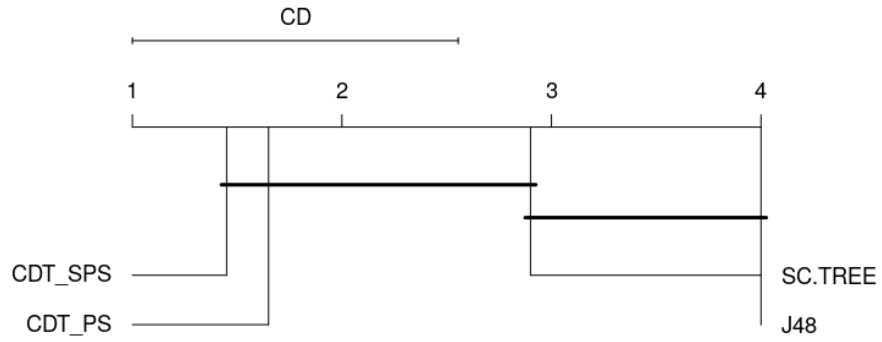


FIGURE 4.2: Critical difference diagram for SC Tree, CDT-PS, CDT-SPS and J48 (average tree size)

Finally, as a semi-causal approach, it is crucial to understand how SC Tree performs in finding (and using) causal relationships in its decisions. Because of that, we compared the approach with CDT-PS and CDT-SPS in terms of average causal relationships found. For both CDT approaches, since all the splits are done using causal information, the number of causal relationships was retrieved using the following formula: $tree\ size - number\ of\ leaves$.

If analyse Table 4.8, that represents the average number of causal relationships found by

TABLE 4.6: Average depth for SC Tree, CDT-PS, CDT-SPS and J48

	Dataset	SC Tree	CDT-PS	CDT-SPS	J48
1	asia	3.20 ± 1.03	+ 1.00 ± 0	0.70 ± 0.95	3.00 ± 0
2	corral	4.00 ± 0	+ 0 ± 0	+ 0.20 ± 0.63	4.00 ± 0
3	earthquake	3.60 ± 0.70	+ 1.00 ± 0	+ 0 ± 0	4.00 ± 0
4	GMB	3.40 ± 0.97	+ 1.00 ± 0	+ 3.00 ± 0	3.80 ± 0.42
5	lucas	6.0 ± 1.49	+ 3.70 ± 0.95	3.60 ± 1.26	- 9.30 ± 0.67
6	mux6	4.60 ± 0.52	+ 1.30 ± 1.49	+ 1.00 ± 1.41	- 5.70 ± 0.48
7	PreSex	2.00 ± 0	+ 0 ± 0	+ 0 ± 0	2.10 ± 0.32
8	threeOf9	7.00 ± 0	+ 4.00 ± 0	+ 4.00 ± 0	- 7.90 ± 0.32
9	Titanic	2.80 ± 0.42	1.10 ± 1.20	1.10 ± 1.20	3.00 ± 0
10	xd6	7.00 ± 0	+ 1.10 ± 1.20	1.10 ± 1.20	- 8.00 ± 0
Average Mean		4.36	1.17	1.76	5.08
Geometric Mean		4.05	0	0	4.56
Average Ranks		3.1	1.4	1.2	3.8
Wicoxon test			0.006	0.006	0.02
Win/Losses			10/0	10/0	1/8
Significant Win/Losses			10/0	6/0	0/4

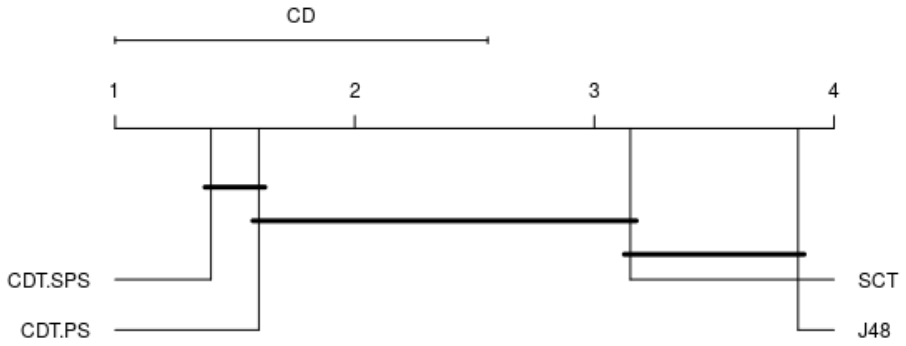


FIGURE 4.3: Critical difference diagram for SC Tree, CDT-PS, CDT-SPS and J48 (average depth)

the three approaches, we can see that, although the difference is not significant (0.1055 for CDT-PS and 1 for CDT-SPS in the Wilcoxon test) SC Tree finds (and uses) more causal relationships in its construction process than CDT-PS and CDT-SPS, finding more causal relationships than these two methods in 8 and 7 of 10 data sets (respectively). This is since the usage of the semi-causal gain ratio (4.5) enables the algorithm to find more causal relationships. What this means is that, if a particular split, there is no causal relationship available, the split is performed, using only the IG (4.4), enabling the algorithm to find more causal relationships in further splits.

TABLE 4.7: Average number of leaves for SC Tree, CDT-PS, CDT-SPS and J48

	Dataset	SC Tree	CDT-PS	CDT-SPS	J48
1	asia	4.20 ± 1.03	+2 ± 0	+ 1.70 ± 0.95	4 ± 0
2	corral	7.40 ± 1.26	+ 1.00 ± 0	+ 1.20 ± 0.63	7.40 ± 1.26
3	earthquake	7.50 ± 2.01	+ 2.00 ± 0	+ 1.00 ± 0	6.50 ± 1.27
4	GMB	5.40 ± 0.97	+ 2.00 ± 0	6.30 ± 0.48	6.60 ± 0.84
5	lucas	9.60 ± 2.01	7.30 ± 2.21	7.30 ± 2.21	- 19.30 ± 2.54
6	mux6	12.10 ± 3.45	+2.3 ± 1.49	+ 2.00 ± 1.41	- 21.10 ± 2.60
7	PreSex	3.00 ± 0	+1 ± 0	+ 1.00 ± 0	3.10 ± 0.32
8	threeOf9	24.90 ± 1.69	+ 11.60 ± 1.35	+ 11.10 ± 0.99	- 28.40 ± 2.37
9	Titanic	4.80 ± 0.42	+ 2.20 ± 1.40	+ 2.20 ± 1.40	5.10 ± 0.32
10	xd6	31.00 ± 0	+ 9.60 ± 0.97	+ 8.90 ± 1.91	- 36.10 ± 1.45
Average Mean		10.99	4.1	4.27	13.76
Geometric Mean		8.34	2.83	2.89	9.83
Average Ranks		3.1	1.5	1.3	3.7
Wicoxon test			0.002	0.004	0.06
Win/Losses			10/0	10/0	2/7
Significant Win/Losses			9/0	8/0	0/4

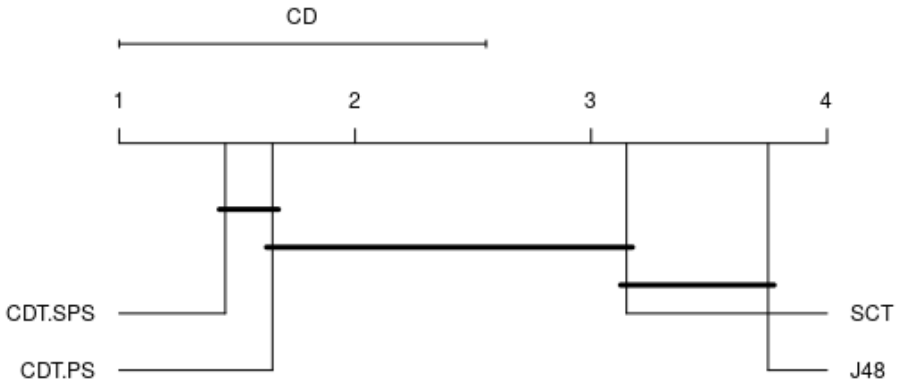


FIGURE 4.4: Critical difference diagram for SC Tree, CDT-PS, CDT-SPS and J48 (average number of leaves)

An Example

Despite this overall similarity with J48, SC Tree does not generate the same trees, as it is possible to see in Figure 4.6a and Figure 4.6c that represent the trees generated by J48 and SC Tree for data set GMB, represented by the network presented in Figure 4.5(the target variable is $V2$), a causal data set available in *pcalg*[*], a causal R package. In these figures, it is visible that both approaches choose different variables at different levels and that SC

[*] https://cran.r-project.org/web/packages/pcalg/index.html

TABLE 4.8: Average number of causal relationships found by SC Tree, CDT-PS and CDT-SPS

	Dataset	SC Tree	CDT-PS	CDT-SPS
1	asia	1.20 ± 0.42	1.00 ± 0	- 0.80 ± 0.92
2	corral	4.00 ± 0	- 0 ± 0	- 0.20 ± 0.63
3	earthquake	3.90 ± 0.32	- 1.00 ± 0	- 0 ± 0
4	GMB	3.30 ± 0.48	- 1.00 ± 0	+ 5.30 ± 0.82
5	lucas	7.90 ± 0.99	- 6.30 ± 2.21	- 6.30 ± 2.21
6	mux6	5.30 ± 0.48	1.40 ± 1.43	- 1.00 ± 1.41
7	PreSex	1.00 ± 0	0 ± 0	- 0 ± 0
8	threeOf9	8.00 ± 0	+ 10.60 ± 1.35	+ 10.10 ± 0.99
9	Titanic	2.00 ± 0	- 1.20 ± 1.40	- 1.20 ± 1.40
10	xd6	7.00 ± 0	+8.60 ± 0.97	+ 8.30 ± 2.6
Average Mean		4.36	3.11	3.32
Average Rank		1.5	2	2.2
Wilcoxon test			0.1055	1
Win/Losses			2/8	3/7
Significant Wins/Losses			2/5	3/7

Tree generates a slightly smaller tree. If we take the root node as an example, we can see why:

- In J48's case, the values for IG and IG ratio for the variables are: 0.17 and 0.17 for *V1*, 0.01 and 0.01 for *V3*, 0.005 and 0.01 for *V4*, and 0.28 and 0.29 for *V5*. In this case, the chosen root node is *V5*;

- In SC Tree's case, the IG values are the same (0.17, 0.01, 0.005 and 0.28). However, since SC Tree uses (4.4) and (4.5) to calculate the variables importance, first it must apply the causal module (Algorithm 4.1) to find which variables are causes of *V2*. *V3* and *V1* are excluded by this module (they are not causes fo *V2*) and because of that in the semi-causal IG ratio's calculation we only use their IG (and not $UC + IG$). *V3* was excluded because the module deemed that *V2* is not dependent from this variable (the value obtained from the GCMH was 2.19, with a p-value of 0.14, which is higher than the stipulated p-value of 5 %). *V1* was excluded because, although *V2* and *V1* are related, the direction of this relationship is $V2 \rightarrow V1$, meaning that *V2* is the cause of *V1* (the UC value obtained for $V1 \rightarrow V1$ was 0.84 and the value for

$V2 \rightarrow V1$ was 0.85). *V4* and *V5* are both causes of *V2*, with UC and semi-causal IG values of 0.95 and 0.96, and 0.86 and 1.16. After the semi-causal IG ratio calculation we obtain the following values: 2.13 for *V4* (intrinsic value of 0.45) and 1.22 for *V5* (intrinsic value of 0.95). This means that SC Tree selects *V4* as its root.

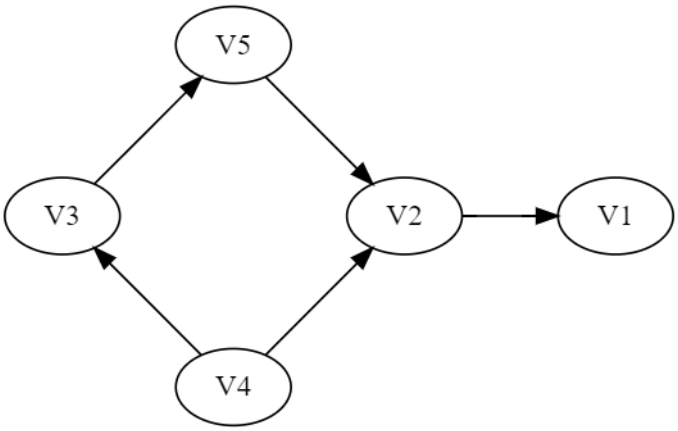


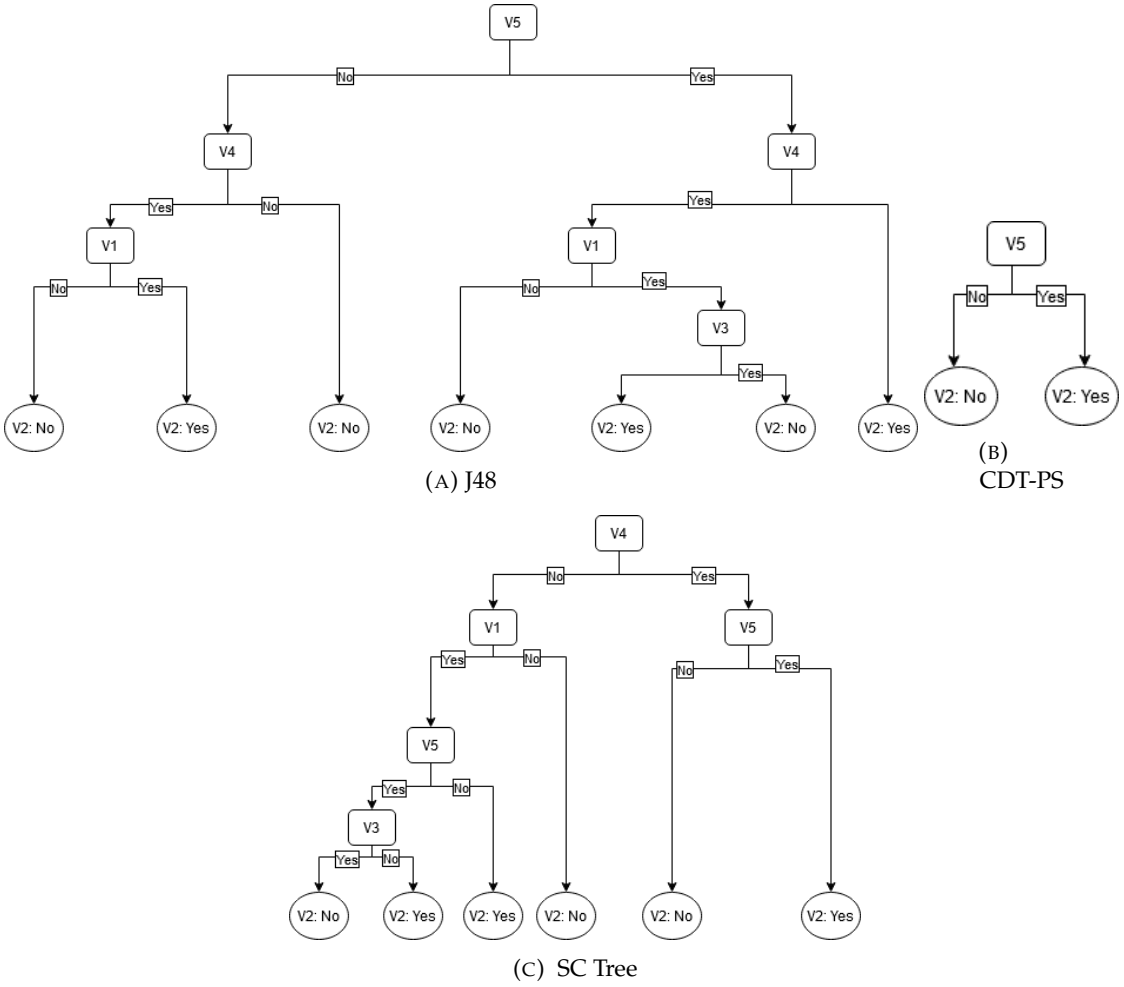
FIGURE 4.5: GMB's true network

While J48 and SC Tree are somewhat similar in terms of error rate but significantly different in terms of tree size, the CDT approaches (CDT-PS and CDT-SPS) are both significantly different from the proposed approach in terms of both error rate and tree size. This difference could be attributed to two different factors:

1. The usage of the UC helped to restrain the selected relationships by only choosing the variables that are direct causes of the class variable, and we are restraining the tree in that the algorithm will only use variables that transmit information to the target;

2. The usage of the UC, combined with the IG, gives extra information to the algorithms, thus boosting the classification.

If we take Figure 4.6c (presented previously) and Figure 4.6b (that represents the tree generated by CDT-PS) as example, it is possible to precisely what was stated previously: while CDT selects *V5* as root, since it is the variable with the highest value in the CMH test ($V5 = 1036/2.7 \times 10^{-277}$; $V4 = 7.425/0.006$ *and* $V1 = 457.4/1.769 \times 10^{-101}$; *V3* is excluded because its p-value is higher than 5 %). While this is also true in part for SC Tree, the UC deems *V4* as the strongest cause ($V4 = 0.95$; $V5 = 0.86$ and *V1* is ignored, since the relationship's direction is $V2 \rightarrow V1$).



(A) J48 (B) CDT-PS

(C) SC Tree

FIGURE 4.6: Trees generated by (a) J48, (b) CDT-PS and (c) SC Tree for data set GMB

The tree generated by CDT has another particularity: it is only composed of one root node and two leaf nodes. This happens because, although the algorithm finds relationships between *V4*, *V1* and *V2*, the resulting tree appears to have matching leaves on opposite sides of *V5* (only *'Yes'* on one side of the split and *'No'* in the other). This means that all the splits after *V5* are meaningless and shall be removed. Since SC Tree chooses another variable as root, it can split the tree further than CDT, thus obtaining better results.

From these results, we can make that SC Tree creates smaller and more interpretable trees (since the approach uses the causal relationships found in the splitting) without losing much accuracy.

4.4.3 SC Tree as a possible Causal Discovery Tool

To understand how the proposed approach would perform as an overall semi-causal approach, we compared it with the current state of the art causal algorithm: PC. We compared the proposed approach in terms of error rate in several discrete data sets (Table 4.1 and Table 4.2). For this comparison, we used the metrics presented in Section 4.4 and used PC as a reference.

TABLE 4.9: Error rates for PC and SC Tree

	Dataset	PC	SC Tree
1	asia	15.58 ± 1.68	+ 14.60 ± 1.42
2	corral	6.25 ± 13.18	2.50 ± 4.37
3	earthquake	0.84 ± 0.18	+ 0.23 ± 0.11
4	GMB	15.44 ± 1.61	+ 15.00 ± 1.77
5	lucas	17.70 ± 1.38	+ 14.50 ± 1.35
6	medpar	33.71 ± 3.95	32.84 ± 4.30
7	monica	43.74 ± 2.19	+ 14.42 ± 1.99
8	mux6	52.31 ± 6.83	+ 19.55 ± 12.91
9	PreSex	25.48 ± 5.59	+ 21.52 ± 3.77
10	respiratory	44.68 ± 7.03	+ 39.62 ± 3.57
11	sachs	39.61 ± 1.27	+ 22.20 ± 1.65
12	threeOf9	46.50 ± 5.15	+ 2.55 ± 2.46
13	titanic	22.65 ± 3.79	20.82 ± 4.02
14	Titanic	23.54 ± 4.38	21.86 ± 3.85
15	xd6	33.09 ± 3.71	+ 0.31 ± 0.50
Average Mean		28.07	16.17
Geometric Mean		20.91	8.66
Average Ranks		2	1
Average Error Ratio		1	0.62
Wicoxon test			1.00×10^{-4}
Win/Losses			15/0
Significant Win/Losses			11/0

In Table 4.9, in general, SC Tree as a significantly better performance than PC, with a value of 1×10^{-4} in the Wilcoxon test. In all the data sets, SC Tree manages to have better

results than PC, being these results significant in eleven out of fifteen data sets, having an improvement of 11 % in average mean over the reference.

These results can be attributed to several factors:

- *The conditional independence test:* the application of the GCMH (in conjunction with the UC, that functions as a double independence test), when compared to the G^2, used by PC has a positive impact in the discovery of the relationships;
- *The application of a mixed causal/correlation approach:* the IG's usage in SC Tree leads to an improvement in the overall accuracy since it uses correlation based information about the relationship between the target and the variables (besides causal information);
- *The overall process:* since decision trees and Bayesian networks implement different processes, it is natural that they find different relationships and, consequently, have different results.

As final a remark, we would like to point out that there was not much difference in runtime since the processes implemented are similar (*independence test* + *orientation phase*).

4.5 Summary

Usually, classification algorithms apply correlation in decision-making, typically obtaining satisfactory results. However, these algorithms' models often do not make sense by real-world standards, thus not being easily understandable to the typical user.

Causal discovery is the field that combines machine learning, data mining and statistics to study the potential cause-effect relationships in observational data.

In real-world problems, the data can be a mixture of causal and correlation relationships, so we hypothesise that we can benefit from the combination of both. For this reason, we proposed SC Tree, a decision tree approach that applies a semi-causal technique to select highly correlated features that are causally related to the target variable.

We compared the proposed approach with several causal and non-causal algorithms in classification problems with discrete data. SC Tree performed better than the causal algorithms, closely matching J48 results. From these results, we can conclude that applying

a mixture between causality and correlation positively impacts the interpretability of the trees without losing much predictive power.

Chapter 5

Improving Prediction with Causal Probabilistic Variables

In regular classification problems, a set of data classified with a finite set of classes is used as input so that a chosen classification algorithm can build a model that represents the learning set's behaviour. This classifier can have better or worse results, depending on the data and how the algorithm handles it.

Nevertheless, in many problems, applying only machine learning algorithms may not be the answer [116]. Instead, the use of feature engineering can be a way of improving these algorithms' performance.

Feature engineering is when new information is extracted from the available data to create new features. These new features are related to the original variables, but also with the target variable, being a better representation of the knowledge embedded in the data, hence helping the algorithms achieve more accurate results [116]. This type of solutions are usually problem-related, being that one solution might work in one particular problem but not in the other. However, one particular characteristic is common to many classification problems: causality.

In most observational data, there is the possibility causal relationships' existing between variables, especially in data related to medical problems (among others) [117, 118]. This fact should be considered when selecting or creating new features since it can give clues to which variables are the most important to the problem.

In this chapter, we aim to answer **RQ.3** and **RQ.5** (Section 1.3): *In what other situations can we apply causality beyond causal discovery?* and *Are causal relationships helpful, and can they bring significant gains?* To answer these questions, we started by studying the potential usage of causal discovery methodologies to create new features that represent the causal relationships between a target variable and the other ones.

By definition, causality, more specifically causal discovery, relates to the search for possible cause-effect relationships between variables [5]. The application of causal discovery in the various tasks of machine learning may be challenging, both at the causal process's level or the sampling process to generate the observed data [41]. Despite this, this subject has been the focus of several researchers over the years, given the importance and potential impact of discovering causal relationships between events in problem-solving. In the words of Judea Pearl: *"while probabilities encode our beliefs about a static world, causality tells us whether and how probabilities change when the world changes, be it by intervention or by an act of imagination"* [119]. Furthermore, by discovering causal relationships, it is possible to uncover correlations and relations that explain how and why the variables behave the way they do.

In this chapter, we propose a framework to create new features for discrete data sets (discrete features + discrete target) based on the causal relationships uncovered in the data. These attributes are created through the generation of a causal network, using a modified version of PC [94], and posterior probabilistic analysis of the relationships between a target variable and the variables considered relevant. Two different methods can choose the relevant variables: parents and children of the target and Markov blanket [120].

5.1 Problem

As stated before, feature engineering is an important part of machine learning, where raw data is manipulated and transformed to improve a model's prediction capability. New features can be created using several techniques such as feature splitting, aggregation, and one-hot encoding, among others, but tend to be more problem related than general.

Although most techniques are deemed as problem bound, one key element is common to many of these problems: causality. The usage of features that represent the supposed causal relationships between variables has several advantages:

TABLE 5.1: Data set description

Data set	Number of examples	Number of attributes	Number of classes
breast cancer[a]	286	10	0(70%) 1(30%)
cervical[a]	858	16	0(94%) 1(6%)
corral[b]	160	7	0(56%) 1(44%)
earthquake	10000	5	0(2%) 1(98%)
head injury[c]	3121	11	0(92%) 1(8%)
lucas	2000	12	0(28%) 1(72%)
medpar	1495	9	0(66%) 1(34%)
mifem[c]	1275	10	0(25%) 1(75%)
qualitative bankruptcy[a]	250	7	0(43%) 1(57%)
respiratory	555	5	0(51%) 1(49%)
survey	10000	6	0(56%) 1(28%) 2(16%)
titanic	1316	4	0(62%) 1(38%)
xd6	973	10	0(67%) 1(33%)

[a] https://archive.ics.uci.edu/ml/index.php
[b] https://www.openml.org
[c] https://vincentarelbundock.github.io/Rdatasets/articles/data.html
[d] http://www.causality.inf.ethz.ch/data/LUCAS.html
[e] https://vincentarelbundock.github.io/Rdatasets/articles/data.html
[f] https://cran.r-project.org/web/packages/vcd/index.html

1. Causal features can help to understand the data better, as they represent how the system behaves naturally;

2. Causal features can help the models infer stronger relationships between variables, as they represent those relationships;

3. Finally, causal features can help create more concise models since a single causal variable represents one or more relationships.

Data

To analyse and solve this problem, throughout the following sections, we will use publicly available datasets as shown in Table 5.1. This data was gathered from several public databases and are publicly available for usage.

5.2 Framework

In many machine learning problems, the application of only classification algorithms might not be the answer to obtaining satisfactory results [116]. Instead, the application of feature engineering can be a way of improving such results. There are already several methods to improve the overall algorithm's performance through the attributes' creation or modification, but, to the best of our knowledge, none of them explores the potential causal relationships between the target variable and the other variables.

The addition of these new inferred causal attributes may help improve the classification algorithms's performance since they encode the relationship between the target and the other variables, thus feeding more information about the data set and its behaviour to the model. Moreover, these features may also aid in the generated model's interpretability since they encode the underlying relationships between the variables, thus being possible to explain more easily the decisions made by them.

This section presents a new framework to create new features using causal probabilities retrieved from a model representing causal associations between variables. This framework can be divided into four different phases:

1. Causal model's creation (in this approach, we suggest the usage of a modified version of PC);
2. Relevant variables' identification. These variables are directly related to the target variable:
    - They are its parents and children;
    - They belong to its Markov blanket (*i.e.* parents, children and spouses).
3. Inference of the probabilities associated with each pair *{target variable, associated variable}*;
4. Creation of the new features using these probabilities. The number of features should be $number\ of\ associated\ variables \times number\ of\ classes$.

The framework starts by creating a full causal model representing the causal associations between all the variables in the first step. This is done through the application of a modified version of PC [94]. In this modified version, the state of the art independence test (usually χ^2 or G^2) is replaced by the GCMH. This test has the advantage (over χ^2 and G^2) of adjusting for confounding factors [121].

It is important to note that, in some cases, PC cannot direct every edge. Hence it creates a Completed Partially Directed Acyclic Graph (CPDAG). In those cases, we apply a method to direct such edges. This method, proposed by Dor and Tarsi [122] searches recursively for possible ways to direct undirected edges.

In the second step, the framework selects the relevant variables. We propose two approaches to select these attributes: parents and children and the Markov blanket.

In the parents and children (P-C) approach, as the name says, the variables selected are the ones that, in the causal graph, have an edge directed to the target (parents) or from it (children).

In the Markov blanket (MB) approach, both the target's parents and children are selected, as well as the nodes that have edges directed to the child nodes (also called spouse nodes). It is important to note that the most common way to select the variables that influence the target is through Markov Blanket (often used in causal feature selection methods [123]). However, several authors proposed to use only parents and children, as these variables can be considered to be the ones that influence the target the most, within its Markov blanket [124–126].

In the third step, the framework infers a set of probabilities that representing the relevant variables' influence on the classes of the target: posterior probability distribution (5.1). In these probabilistic queries, the objective is to find what the influence that evidence (particular values of the relevant variable) has on the value of the target [127]. This is performed for all the values in each variable, and the resulting probability matrix is similar to Table 5.2.

$$P(Target = t|Attr = a) = \frac{occurrences_{t \cap a}}{ococcurrences_a} \tag{5.1}$$

Finally, the new features are created and added to the data set in the fourth step. Each new feature represents the probability of the relevant variables' influence on a specific

TABLE 5.2: Example of probabilities generated by the probability queries

		Attr		
		0	1	2
Target	0	0.63	0.53	0.13
	1	0.34	0.29	0.67
	2	0.14	0.25	0.56

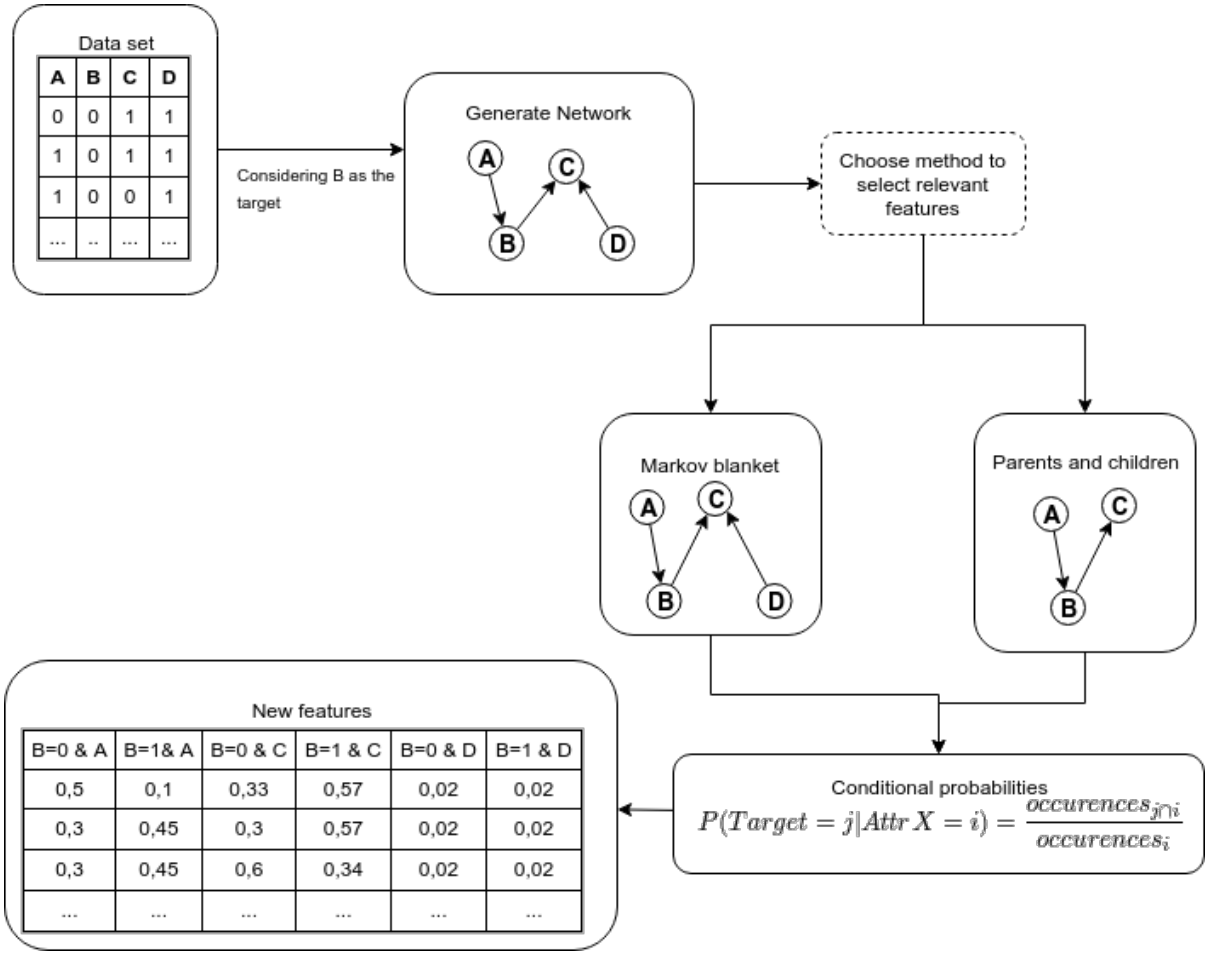


FIGURE 5.1: Example of the operation of the proposed framework

class, *i.e.*, if we have, for example, a target variable with two classes ($\{0, 1\}$) and a relevant variable *Attr*, two new features representing *Attr*'s influence in each class (each instance of the feature represents *Attr* values's influence on the class represented in that feature) will be created.

An overview of the framework can be seen in Figure 5.1.

An Illustrative Example

To explain in more detail how this approach works, we will use as an example a data set with six discrete variables (A, B, C, D, E and F) with 5000 instances*. The values for variables A,B,C, D, and E can be $\{0, 1, 2\}$, while F can have the values $\{0, 1\}$. For this example, we will use variable **B** as the target.

As it was explained in the first step, the approach starts by generating the full network with PC and GCMH. The generated network can be seen in Figure 5.2.

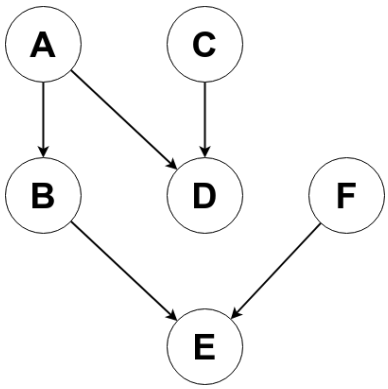


FIGURE 5.2: Example: network generated

After the full network's creation, the relevant variables are selected. These variables can be parents or children (P-C) ($\{A, E\}$) or the Markov blanket (MB) of **B** ($\{A, E, F\}$).

In the third step, the framework generates the chosen variables' inference probabilities (Table 5.3). Taking A=0 and B=0 as an example, the probabilities are obtained for each one of the target values are calculated by dividing the number of times both $A = 0$ and $B = 0$ occur by the number of times A=0 occurs, or in other words, $P(B = 0|A = 0) = 0.86$.

TABLE 5.3: Probabilities generated for the Markov blanket variables. In parents and children's case, the probabilities for F are not generated.

		A			E			F	
		0	1	2	0	1	2	0	1
Target	0	0.86	0.45	0.11	0.74	0.46	0.15	0.47	0.48
	1	0.03	0.22	0.09	0.08	0.11	0.16	0.11	0.12
	2	0.11	0.32	0.78	0.19	0.44	0.68	0.41	0.41

These probabilities are then added to the global data set. The resulting data set is similar to Table 5.4. There is a difference between the number of new features created since the

*https://www.bnlearn.com/documentation/man/learning-test.html

number of generated features is equal to the product between the number of values in the target and the number of relevant variables. Since the MB approach selects more variables than P-C, the number of generated features will be higher (in theory). So, in the case of P-C features, we have six new features, and in the case of MB, we generate nine new features.

TABLE 5.4: Features generated with the probabilities for Markov blanket variables. In parents and children's case, the features related with *F* are not generated.

A	B	C	D	E	F	A & B=0	A & B=2	A & B=2	E & B=0	E & B=1	E & B=2	F & B=0	F & B=1	F & B=2
1	2	1	0	1	1	0.44	0.22	0.35	0.45	0.10	0.44	0.48	0.12	0.41
1	0	2	0	1	1	0.44	0.22	0.35	0.45	0.10	0.44	0.48	0.12	0.41
0	0	0	0	0	0	0.87	0.02	0.11	0.73	0.08	0.19	0.47	0.11	0.41
0	0	0	0	1	1	0.87	0.02	0.11	0.45	0.10	0.44	0.48	0.12	0.41
0	0	1	2	0	0	0.87	0.02	0.11	0.73	0.08	0.19	0.47	0.11	0.41

5.3 Experimental Setup

To evaluate the proposed approaches and make a comparative study, the following configuration of experiments was designed: the performance of Random Forest, using the original data, and the versions generated by the two proposed approaches were compared.

This comparative analysis was made through 10-fold cross validation in several public data sets (Table 5.1). For each fold, the two approaches are applied to the train set. Then, the resulting conditional probabilities are used to create the new features for both the train and test set (this ensures that no information about the classes in the test set is added to the new features).

A sensitivity analysis was performed to choose the optimal parameters for the approaches presented in the following sections. This analysis consisted of obtaining the error (*1 - accuracy*) for the presented data sets (by dividing them into 70 % train, 30 % test). In the case of PC, this test was repeated for significance levels 1 % and 5 %. In these tests, we concluded that the error of the algorithms in the data sets did not change much when the parameters were changed. For this reason, for all the data sets, we select and present a significance level of 5 %.

5.4 Results

Next, we evaluate the proposed approaches by studying their application in several public data sets. As a classification algorithm, Random Forest was used.

The performance of this algorithm was compared in terms of error rate. This comparison was performed using the *No new features* as a reference. The classification algorithm performance, trained with causal features in each data set, compared to the reference using the Wilcoxon signed ranked test. The sign $+/-$ indicates that the algorithm is significantly better/worse than the reference with a p-value of less than 5 %. Besides this, the algorithms are also compared in terms of average and geometric mean of the errors, average ranks, average error ratio, win/losses, significant win/losses (number of times that the reference was better or worse than the algorithm, using signed ranked-test) and the Wilcoxon signed ranked-test. For the Wilcoxon signed ranked-test, we also consider a p-value of 5 %.

Let's analyse Table 5.5. It is possible to see that, in general, +*Causal features P-C*(the addition of features representing the conditional probability of parents and children features on the target) has a better performance than *No new features* since the value obtained in the Wilcoxon test is 0.0266 (less than the p-value of 5 %), which means that the difference between the performance is significant. This difference can also be seen in the values of the average and geometric ranks. More specifically, if we look at the average ranks, we can see that +*Causal features P-C* has lower ranks (in average) than *No new features* (1.436 against 2.538).

If we now compare the second approach proposed (+*Causal features MB*) with the reference, we can see that there is a positive difference in the results (although not significant). It is possible to see this difference, once again, in the average and geometric mean and the average rank (1.538).

In Table 5.6, it is possible to see the AUC values for the three analysed approaches for the lucas data set (this data set has been used in other causal-related tasks, and it's known for having causal relationships retrievable in the data). The results presented in this table were obtained by dividing this data set by train and test (70%/30%). The model scores were then obtained for the test data.

TABLE 5.5: Error rates of Random Forest for classification with causal features

Data set		No new features	+Causal features P-C	+Causal features MB
1	breast cancer	28.60 ± 9.88	28.60 ± 7.49	28.00 ± 8.39
2	cervical	6.88 ± 1.51	6.65 ± 1.66	6.53 ± 1.49
3	corral	5.62 ± 5.47	+0.01 ± 0.10	+0.01 ± 0.10
4	earthquake	0.26 ± 0.14	0.20 ± 0.14	0.20 ± 0.14
5	head injury	7.08 ± 1.23	7.43 ± 0.83	7.05 ± 0.69
6	lucas	15.20 ± 2.02	14.50 ± 2.12	14.50 ± 2.12
7	medpar	32.70 ± 4.29	33.00 ± 3.91	34.10 ± 3.23
8	mifem	20.10 ± 4.28	20.00 ± 4.30	19.90 ± 3.63
9	qualitative bankruptcy	0.40 ± 1.26	0.01 ± 0.10	0.80 ± 2.53
10	respiratory	40.90 ± 6.79	40.20 ± 6.20	41.20 ± 6.90
11	survey	44.60 ± 2.26	44.40 ± 2.05	44.40 ± 2.05
12	titanic	21.40 ± 2.52	20.20 ± 2.19	20.50 ± 1.83
13	xd6	0.41 ± 0.72	0.10 ± 0.10	0.10 ± 0.10
Average Mean		17.242	16.562	16.715
Geometric Mean		7.161	2.889	4.039
Average Ranks		2.538	1.462	1.538
Average Error Ratio		1	0.764	0.914
Wicoxon test			0.0266	0.1465
Win/Losses			10/2	10/3
Significant win/losses			1/0	1/0

TABLE 5.6: AUC for Lucas data set

	AUC
No new features	0.877
+Causal features P-C	0.887
+Causal features MB	0.889

In this table, it is possible to see that *+Causal features MB* has the highest area, meaning that, in the data set with the causal probabilistic features that represent the relations between the target and its Markov blanket, Random Forest can distinguish better the classes than with the data from the other approaches, thus having a better performance [104]. Although *+Causal features MB* was the best approach in terms of AUC, the other proposed

approach +*Causal features P-C* also obtained an AUC higher than the reference.

Finally, from these results, we can conclude that there is evidence that applying causality to the creation of new features can have a positive impact on the classification algorithm's performance.

5.5 Summary

The achievement of satisfactory results in a classification problem depends not only on the chosen classifier but also on the processed data. One possible way to improve the performance of classifiers is to apply feature engineering. In other words, use the original data to infer new information, create new attributes, and alter others to obtain more descriptive features. Furthermore, most of the proposed methodologies do not consider the possible causal relationships in the data. This information can help create more accurate models since we encode information about the interaction between variables in one variable, thus reinforcing their importance.

In this chapter, we proposed a framework that uses causal discovery to create new features based on posterior probabilistic analysis of the relationships between a target variable and the variables considered relevant, being these variables the parents and children of the Markov Blanket the target.

We compared the approaches with the original data in the experiments, using Random Forest in public data sets. From these results, we can conclude that there is evidence that the application of causality in the creation of new supposed probabilistic features may positively impact the overall performance of the classification algorithm.

Chapter 6

Temporal Nodes Causal Discovery for ICU Survival Analysis

In hospital and after discharge deaths in Intensive Care Units (ICUs) are unfortunate but usual, given the severity of the condition under which many of them are admitted to these wings. However, some recent studies show that one in five patients die even after being discharged from the ICU from complications related to the admission, with some deaths being called as "failure to rescue" [128]. Given this, it is crucial to promptly identify and follow these cases closely so that, if possible, the outcome can be changed.

The diagnosis of a medical problem can be seen as the relationship between a disease and the symptoms it induces. This notion of causal discovery (finding out what is causing a set of symptoms) is implemented regularly in medicine, although not consciously or through algorithms. The application of causal discovery in the medical field has been debated over the years [129] since the application of this type of technique can help in the fastest diagnosis of certain diseases.

However, working with medical data can be challenging since this type of data can be composed of thousands of variables, measured only one time or in regular and irregular intervals, depending on the exams performed. Primarily ICU data is characterised by a high flow of information measured in different intervals, usually accompanied by the length of patient's stay as well as their outcome [130]. Moreover, this type of data usually comprises patient data where, for each subject, there is a set of measurements, hence being

considered a panel or longitudinal data. This heterogeneity in the sampled data raises the need for specialised methodologies that:

1. Transform irregular multivariate time-series into stationary;

2. Somehow deal with them as irregular time-series.

In this chapter we aim at answering **RQ.4** and **RQ.5** (Section 1.3): *Can we create causal models from sequential data?* and *Are causal relationships helpful, and can they bring significant gains?* We started by studying the potential usage of causal discovery methodologies to generate models for irregular multivariate time-series data.

Causal Bayesian networks are a type of Bayesian Network that captures supposed causal relationships from observational data (data that represents a snapshot of a system) and are known to be an explicable method since their graph-like appearance mimics human decisions. This type of methodology can aid medical staff in performing simple decisions more easily, as the everyday user can easily understand it. PC algorithm [94] is an example of a Bayesian network specifically designed to ensure that every relationship can be assumed as causal. Although Bayesian networks are traditional methods designed for cross-sectional data, since they do not consider time, methods that deal with time have emerged in more recent years. This is the case of the Dynamic Bayesian Networks (DBNs) [131]. However, these methods have two significant restrictions: they can only be applied in stationary time-series data.

This work aims to address the problem of ICU patients' non-survival early detection while maximising the data usage by taking advantage of the timing irregularity. To do this, we propose the ItsPC, a causal Bayesian network-like approach that can model irregular multivariate time-series data. This method models time by incorporating it into the variables' values (instead of creating new variables representing the stages in a particular timestamp). This method combines the time stamp of every instance during the measured value for each temporal variable, thus creating instances that represent both stage and time. To obtain a more accurate depiction of reality, every interval-value is adjusted according to the variables' parents (obtained from the network). Hence it is based on the parents' delayed manifestation and not on the absolute time. In this method, every variable represents a temporal change and every edge a causal, temporal relationship between variables [132].

6.1 Temporal Bayesian Networks

The Temporal node Bayesian Network (TnBN) [132] are an extension of the Bayesian Networks, designed to deal with multivariate time-series data. Each node represents a temporal change (based on their relationships) in this method, and each edge represents a temporal relationship. This method first discretises all temporal variables, transforming them into time intervals. Next, it applies the K2 Bayesian network to this new discretised data set. After that, and using the information obtained from the model (such as the parents of each temporal variable), the algorithm adjusts the intervals present in the temporal variables and re-generates the model. However, this methodology has an issue: as temporal variables, it only accepts the value that represents the moment where the value was measured, for example, at what time the doctor saw dilated pupils, hence dealing with them as binary variables, where each measurement details if something was measured or not and at what moment was measured. This majorly restrains the number of potential applications, especially in the medical domain, where variables can represent continuous values, or discrete stages, always measured in different intervals of time, consequently representing "hybrid variables" *STAGE A [t1-t2]* or *[interval of continuous measure][t1-t2]*. Besides this, the usage of K2 to create the Bayesian model does not ensure the existence of causal relationships between variables (temporal or not), and that can be crucial to identify what is causing changes in the system, as it is not prepared for such a task.

Tawfik and Neufeld [133] proposed a different approach. The Temporal Bayesian Networks (TBNs) are a Bayesian network designed to represent time by expressing the probabilities as a function of time. This means that, arguably, if one variable depends on another, this dependency represents a time interval between them. Unfortunately, despite a simple and intuitive representation, this algorithm seems capable of representing time through binary choices (is the dog out or not) instead of a multiple state variable (the dog is in the garden, in the driveway or the house).

More similar to the proposed idea are the Irregular Time Bayesian Networks (ITBNs) [134], designed to deal with irregular time-series. This approach generalises the DBNs so that data can be sampled in different time intervals. Despite taking into account that variables may be measured in irregular intervals, thus generating smaller networks when compared with the DBNs, this method, like the DBNs, assumes that all variables are measured at these intervals.

Finally, it is important to note that none of these three approaches infers causal relationships from the data.

6.2 Problem

In hospital and after discharge deaths are a well-known problem by ICU practitioners [128]. Especially the death of discharged patients has been considered a problem, with studies showing that potentially one in five patients dies after being discharged from the hospital, with some of these deaths being considered preventable [135]. As some of these cases are considered as a failure of assistance or "failure to rescue", meaning that, given awareness, they could be addressed and prevented, being their timely prediction a key to saving lives [136].

This problem can also be seen from a cost perspective [137]. After the first discharge, patients who need more care signify more costs for the hospital and the patient. Besides this, the care needed may be more intensive than if the patient had been closely followed after discharge or not been discharged.

Machine learning algorithms can be applied to ensure:

1. The timely patient assessment;

2. Cost reduction, as this type of methodologies can be more affordable than traditional approaches.

This is the case of the work presented by Garcia-Gallo et al. [138], where the authors use a Stochastic Gradient Boosting methodology to model and predict one-year mortality in critical patients diagnosed with sepsis. A different approach was taken by Chia et al. [139], where the authors evaluated the usage of logistic regression, decision tree, and Cox-Proportional Hazards to identify the feature that better help predict the patients' outcomes.

Although there were significant advances in predicting ICU outcomes, none of these studies indeed considered the underlying supposed causes of hospital and after discharge deaths in ICU patients. Moreover, these studies do not consider the lack of regularity in hospital records.

Data

The physionet data set [140] is a subset of the MIMIC II, a data set with more than 25 000 patients admitted in the Boston's Beth Israel Deaconess Medical Center's ICU, from 2005 to 2008. This subset is composed by 12 000 patients (divided in train set [4000 patients], test set [4000 patients] and scoring set [4000 patients]), that were followed during the first 48 hours of their stay in the ICU.

The raw data was cleaned and pre-processed. First, every variable was discretised according to the literature. Next, all the variables whose stages were not defined in the literature (for example, the patient's height) were discretised using equal-frequency discretisation with three bins. Next, all the variables with single values and with at least 50 % missing data across all subjects were discarded. Finally, every patient with at least 50 % of missing values across all variables was also removed.

The new data set is composed of 11 657 patients, split into train and test sets. The following variables are present in the data:

- Patient identifier
- Measurement's time
- Height
- Age
- Weight
- Alanine transaminase (ALP)
- Aspartate transaminase (AST)
- Alkaline phosphatase (ALP)
- Lactate
- Biliburin
- Respiration rate
- O_2 saturation in haemoglobin (SaO2)
- Blood urea nitrogen (BUN)
- Creatinine
- Fractional inspired oxygen(FiO2)
- Glasgow Coma Scale (GCS)
- Glucose
- Bicarbonate (HCO3)
- Hematocrit (HCT)
- Heart rate (HR)
- Potassium (K)
- Magnesium (Mg)
- Mean blood pressure (invasive and non-invasive) (MAP, NIMAP)
- Mechanical ventilation

- Sodium (Na)
- Arterial blood gas (PaCO2, PaO2)
- Urine
- Temperature
- Blood pressure (invasive and non-invasive diastolic blood pressure and invasive and non-invasive systolic blood pressure)(DiasABP, NIDiasABP, NISysABP, SysABP)
- Haemoglobin saturation
- Platelets and Arterial pH
- Length of stay
- Survival

The target variable of this processed data set is the patient's survival. It is a merge between the *Survival* and *In Hospital Death* variables present in the original data set. This new variable represents whether a patient died in the hospital after the discharge with a problem related to his first hospitalisation (DEATH AFTER DISCHARGE), if he died in the hospital, whether he is still in the ICU or another hospital inpatient unit (IN HOSPITAL DEATH) or is still alive (ALIVE) and is measured after the patients leave the ICU. These classes are distributed as follows: 61.20 % (*ALIVE*), 15.51 % (*IN HOSPITAL DEATH*) and 23.28 % (*DEATH AFTER DISCHARGE*).

Running example

A running example representing a patient admitted to the hospital ICU will be used to explain better the proposed methodology and how it is applied to the data. Patient 134432 is a 70 years old male admitted to the surgical ICU. For easier identification, we will call this patient John Doe. Mr Doe was followed for 48h during his stay in the ICU. During this period, several tests were performed and recorded. Table 6.1 represents part of these tests. This patient was hospitalised for three days and died in the hospital.

6.3 Methodology

In this section, we present the *Irregular time-series PC* or ItsPC, a causal Bayesian network for irregular multivariate time-series data, designed to deal with data that represents measurements done at specific moments and repeated several times, hence being represented by a value and a timestamp. The model creation process applied by this method can be divided into six steps: state/timestamp conjunction, model generation, first redefinition

TABLE 6.1: Mr. Doe's medical tests

Time	Age	Gender	Height	ICUType	BUN	Creatinine	GCS	SaO2	Weight		Length of stay	Survival
0	SENIOR	MALE	80	Surgical ICU	NA	NA	NA	NA	[171.5, 189.7)		[1, 9)	IN HOSPITAL DEATH
2.88	SENIOR	MALE	80	Surgical ICU	NA	NA	NA	NA	[171.5, 189.7)		[1, 9)	IN HOSPITAL DEATH
3.02	SENIOR	MALE	80	Surgical ICU	NA	NA	NA	NA	[171.5, 189.7)		[1, 9)	IN HOSPITAL DEATH
3.18	SENIOR	MALE	80	Surgical ICU	NA	NA	SEVERE	NA	[171.5, 189.7)		[1, 9)	IN HOSPITAL DEATH
4.18	SENIOR	MALE	80	Surgical ICU	NA	NA	SEVERE	NA	[171.5, 189.7)		[1, 9)	IN HOSPITAL DEATH
5.18	SENIOR	MALE	80	Surgical ICU	NA	NA	SEVERE	NA	[171.5, 189.7)		[1, 9)	IN HOSPITAL DEATH
5.85	SENIOR	MALE	80	Surgical ICU	NA	NA	SEVERE	NORMAL	[171.5, 189.7)		[1, 9)	IN HOSPITAL DEATH
6.18	SENIOR	MALE	80	Surgical ICU	NA	NA	SEVERE	NORMAL	[171.5, 189.7)		[1, 9)	IN HOSPITAL DEATH
6.43	SENIOR	MALE	80	Surgical ICU	NA	NA	SEVERE	NORMAL	[171.5, 89.7)		[1, 9)	IN HOSPITAL DEATH
6.85	SENIOR	MALE	80	Surgical ICU	NA	NA	SEVERE	NORMAL	[171.5, 189.7)		[1, 9)	IN HOSPITAL DEATH
7.18	SENIOR	MALE	80	Surgical ICU	NA	NA	MILD	NORMAL	[171.5, 189.7)		[1, 9)	IN HOSPITAL DEATH
8.18	SENIOR	MALE	80	Surgical ICU	NA	NA	MILD	NORMAL	[171.5, 189.7)		[1, 9)	IN HOSPITAL DEATH
8.53	SENIOR	MALE	80	Surgical ICU	HIGH	NORMAL	MILD	NORMAL	[171.5, 189.7)		[1, 9)	IN HOSPITAL DEATH
9.18	SENIOR	MALE	80	Surgical ICU	HIGH	NORMAL	MILD	NORMAL	[171.5, 189.7)		[1, 9)	IN HOSPITAL DEATH
							...					
47.18	SENIOR	MALE	80	Surgical ICU	HIGH	NORMAL	SEVERE	NORMAL	[171.5, 189.7)		[1, 9)	IN HOSPITAL DEATH

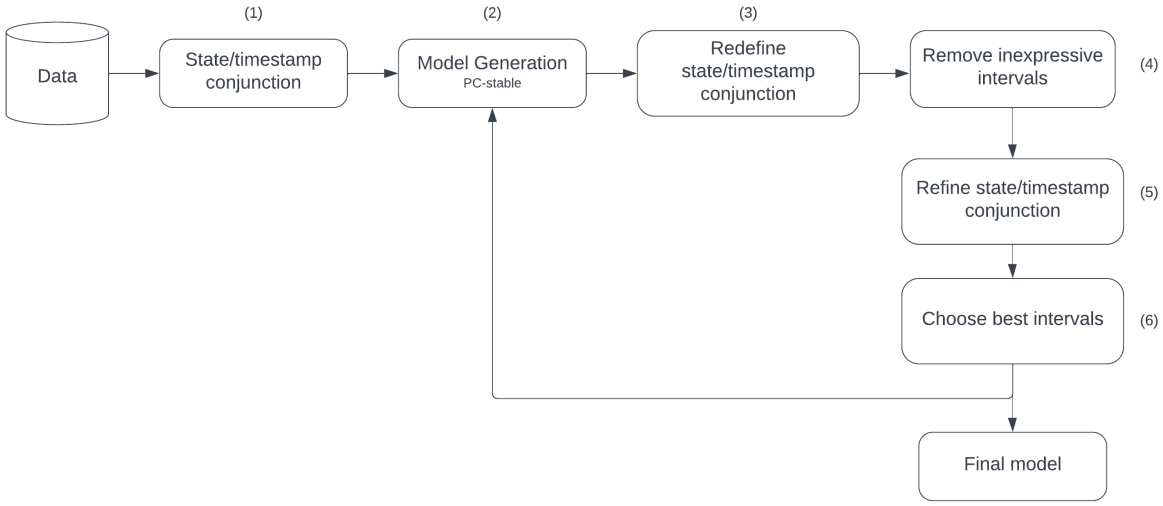


FIGURE 6.1: Its PC pipeline

of state/timestamp conjunction, inexpressive intervals removal, second redefinition of state/timestamp conjunction and optimal interval selection (Figure 6.1).

Initially, the algorithm starts by merging the temporal states with the respective timestamps (Figure 6.2, (1)). As shown in the running example (Table 6.1), the timestamp is saved separately from the measured correspondent value. For each discrete variable marked as varying over time, the algorithm divides the corresponding timestamps according to their categories (Figure 6.2, (2)). For each category, the method discretises the timestamps (Figure 6.2, (3)). In the presence of missing data, the unknown value is replaced by the state *UNK[min, max]*, where *min* and *max* represent the minimum and maximum timestamp found in that specific variable. In the running example, Mr Doe is

grouped with other patients to create the merged states (to ensure (1) the states are meaningfully for the majority of the subjects and (2) the number of generated merged states is low). The result for the GCS (Glasgow Coma Scale) is shown in Table 6.2.

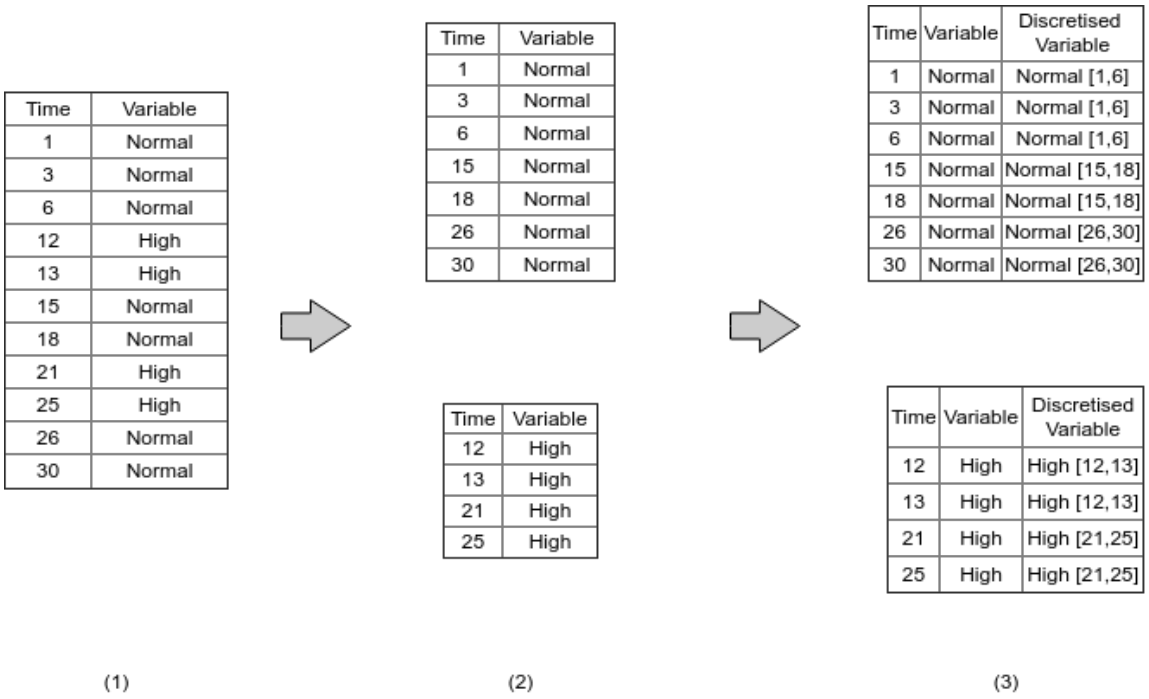


FIGURE 6.2: First discretisation (example)

TABLE 6.2: Discretisation for Mr. Doe's GCS measure

Time	GCS
0	UNK[0,48]
2.88	UNK[0,48]
3.02	UNK[0,48]
3.18	SEVERE [3,23]
4.18	SEVERE [3,23]
5.18	SEVERE [3,23]
5.85	SEVERE [3,23]
6.18	SEVERE [3,23]
6.43	SEVERE [3,23]
6.85	SEVERE[3,23]
7.18	MILD [7,35]
8.18	MILD [7,35]
8.53	MILD [7,35]
9.18	MILD [7,35]
	...
47.18	SEVERE [25,48]

At this moment, and before we introduce the next step, it is essential to introduce the definition of initial and final cross-sectional variables. As initial variables, we perceive them as variables measured at the study's beginning and not changing over time (for example, age). In contrast, final variables are understood as variables measured only once, but after all, the temporal variables (for example, if a patient survived or not).

After generating the data set, PC (Figure 6.3) is applied to create the first model. In this case, the method treats all variables as cross-sectional. To ensure precedence in the model (thus generating a model that genuinely represents time), no temporal variable can cause initial variables, and no final variable can cause initial or temporal variables.

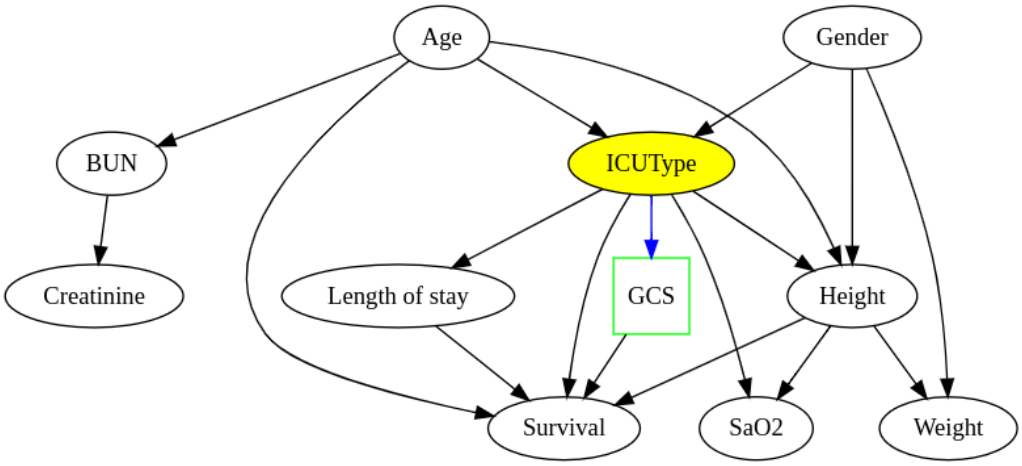


FIGURE 6.3: PC example model

After the network's creation, the model is analysed to discover the parents of the temporal variables. This information is later used to redefine the temporal variables. In Step 3, the Gaussian Mixture Models (GMM) creates new intervals for each partition based on the parent's information. These intervals are created by defining a n number of maximum intervals by partition. With this n value, the algorithm creates 1 to n different time intervals. Therefore, each partition, which represents a configuration of the parent nodes, has n different intervals. It is important to note that the minimum and maximum timestamps in each interval are given by each cluster's minimum and maximum timestamps.

Returning to Mr Doe's example, this patient (grouped with the other patients) has its temporal variables redefined (it is important to note that for this redefinition, we used the original states with no time associated), using the model generated in the previous step,

more specifically the parents' information. Using the GCS variable and its respective parent ICUType (this variable is a initial variable that takes the values *Surgical ICU*, *Medical ICU*, *Cardiac Surgery Recovery Unit* and *Coronary Care Unit*), the method splits GCS' values and timestamps taking into account ICUType's values. To these subsets, the GMM is applied, resulting in the discretised timestamps presented in Table 6.3.

Before we move for the next step, it is important to note that from now on a partition is considered as a combination between a value of the parent and a value of the child. In Table 6.3 we have 4 different partitions (ICUType= SurgicalICU; GCS=SEVERE, ICUType= SurgicalICU; GCS=MILD, ICUType= MedicalICU; GCS=SEVERE and ICUType= MedicalICU; GCS=MILD).

TABLE 6.3: Redefinition of GCS values using the parent's information (for simplicity, only two states of GCS and ICUType are used)

ICUType SurgicalICU	GCS SEVERE	[3-48]
		[3-25][29-48]
		[3-16][17-24][29-48]
	GCS MILD	[7-24]
		[7-12][14-24]
		[7-8][10-12][14-24]
ICUType MedicalICU	GCS SEVERE	[18-35]
		[18-21][23-35]
		[18-21][23-26][30-35]
	GCS MILD	[7-44]
		[7-30][31-44]
		[7-24][26-29][30-44]

As the number of intervals in each partition can be high but not expressive, depending on the number of parents and their values, a pruning method for removing inexpressive intervals is applied (all intervals with less than β instances are removed):

$$\beta = \frac{number\ of\ instances\ in\ the\ interval}{number\ of\ parent\ nodes \times 2} \tag{6.1}$$

The partitions are then combined based on common child values (before discretisation) For example, for *GCS= SEVERE*, the method combines the interval set from *ICUType= MedicalICU* and *ICUType= SurgicalICU*. From these combination nine intervals set result:

- $[3-48][18-35]$
- $[3-48][18-21][23-35]$

- $[3-48][18-21][23-26][30-35]$
- $[3-25][29-48][18-35]$
- $[3-25][29-48][18-21][23-35]$
- $[3-25][29-48][18-21][23-26][30-35]$
- $[3-16][17-24][29-48][18-35]$
- $[3-16][17-24][18-21][23-35]$
- $[3-16][17-24][18-21][23-26][30-35]$

Since many of these intervals overlap, a set of rules is used to combine them:

1. If one interval is contained in another (e.g. [18-35] is contained in [3-48]), the new interval will be *[minimum of the two, maximum of the two]* ([3-48]);
2. If two intervals partially overlap, ([3-25] and [18-35]), two new intervals are created: *[first interval minimum, average of contained values)[average of contained values+unit, second interval maximum]* ([3-21.5][21.6-35]). This process is continuously updating the intervals until all the intervals are adjusted. With this step, the method ultimately tries to instantiate a child node as a delayed occurrence of the parent node and not in absolute time.

After this, another pruning is performed: all partitions that have only one interval or more than n intervals (user-defined) are removed. For example, this means that if we have an adjusted set of intervals for variable V, with the intervals [1-12][12.1-43][45-56][67-70][70.1-90], and n has the value 3, these will be discarded. This pruning ensures that all accepted intervals have a broad representation in the data and are not the representation of only a few examples. Finally, in Step 6, the method chooses the optimal intervals for each temporal variable's value. This selection is made by combining each of the potential intervals' sets for each variable's values. Then, a model is created (with configurations identical to the first model). These models are then evaluated using the Brier Skill Score, a measure that calculates how precise a probability prediction in a model is (when compared to a reference) and is given by (6.2) [141] (the higher the value, the more precise is

the prediction).

$$BSS = 1 - \frac{BS}{BS^{ref}} \qquad BS = \frac{1}{n}\sum_{i=1}^{n}(1 - P_i)^2 \tag{6.2}$$

In this equation, n represents the number of unmeasured variables in the set, P_i represents the probability obtained from the unseen variables and BS_{ref} represents the reference Brier Score (this value is obtained by calculating the probabilities for the same unseen variables, with the model used in the previous steps). To determine P_i, a random subset of nodes is selected and instantiated with random values based on the original data distribution. With these values, we predict the P_i, probability of the unmeasured variable i, with the measured variables. It is important to note that the Brier Score formula used in this methodology is not the original version (designed for any discrete data) but the binary version instead. We use this equation instead because the algorithm studies, for each event, the probability of a particular value and not all values that the unseen variable can take, hence being a binary *true* and *false* problem. Subsequently, the set intervals chosen is the one that maximises the Brier Skill Score.

Steps 2 to 6 of Figure 6.1 are continually repeated until there are no changes to the model or data set.

6.4 Experimental Setup

To evaluate the proposed approach and make a comparative study, the following configuration of experiments was designed: we compare the model generated by our approach with a model generated by a DBN [131], in terms of performance (accuracy and F1-score). To do this, we derived ten data sets from the original one, presented in Section 6.2, by randomly sampling 70% of the patients for the train set and 30% for the test set. The results were also compared using the Wilcoxon signed ranked-test.

Since the DBNs are a type of model that only deals with regular intervals of time, the data set presented in Section 6.2 was transformed. To create this new data set, the mean timestamp interval (t_{mean}), mean minimum timestamp (t_{min}) and mean maximum timestamp (t_{max}) were calculated.

With this information, the original data set was transformed. This transformation was done following a set of rules. These rules were:

1. For each subject, the first and last time stamps were t_{min} and t_{max};
2. All the timestamps are distanced exactly t_{mean};
3. Each new timestamp (for each subject) is filled with the nearest timestamp from the original data. Suppose a particular subject does not have timestamps before the new timestamp is measured. In that case, the values in the new timestamp entrance are filled with missing values;
4. All timestamps from the original data set that is higher than the t_{max} are discarded.

The resulting data set is composed of 75 regular timestamps.

6.5 Results

To better understand how the algorithms perform in a general hospital situation, where the system encapsulates patients from different services, thus with distinct diseases and symptoms, we compared the DBNs and ItsPC (Table 6.4). If we analyse Table 6.4, which represents the mean accuracy and F1-score by class and overall accuracy and F1-score, it is possible to see that, in general, the proposed methodology has a better performance than the baseline (DBN).

TABLE 6.4: Results comparison

	Accuracy		F1-score	
	Dynamic Bayesian Network	ItsPC	Dynamic Bayesian Network	ItsPC
ALIVE	59.02 ± 0.70	+ 68.10 ± 0.45	71.98 ± 0.17	+ 78.31 ± 0.36
DEATH AFTER DISCHARGE	50.00 ± 0.03	+ 75.27 ± 0.45	0.02 ± 0.06	+ 24.08 ± 1.61
IN HOSPITAL DEATH	85.64 ± 0.21	86.43 ± 0.32	10.91 ± 11.71	+ 33.65 ± 1.61
Overall	53.65 ± 0.40	+ 64.82 ± 0.51	27.64 ± 3.91	+ 45.35 ± 0.62

To further assess the significance of these discrepancies, the performance of ItsPC in each test set was compared to the reference (DBN) using the Wilcoxon signed ranked-test. The sign $+/-$ indicates that the algorithm is significantly better/worse than the reference with a p-value of less than 5 %. As it is possible to observe in table Table 6.4, the difference between the two methods is significant. Furthermore, considering the proposed practical problem to assess if the patient will parish in the hospital or after being discharged, it is possible to notice that ItsPC is more successful in detecting future dead cases than the

baseline. Despite this, both methods demonstrate difficulty in accessing death after discharge patients, which is expectable since, in theory, these patients are not that different from those who survive.

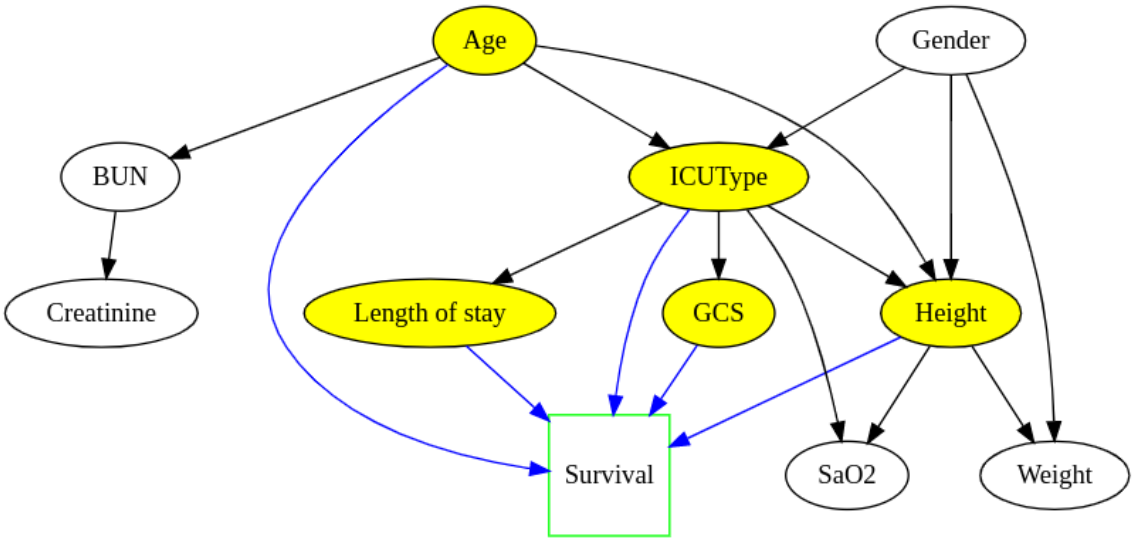


FIGURE 6.4: Simplified model generated be ItsPC

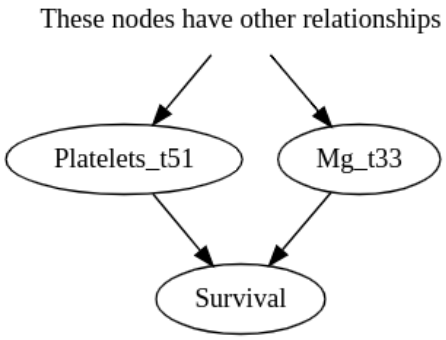


FIGURE 6.5: Simplified model generated by DBN (all the other 2435 nodes not related with *Survival* are omitted)

As a final note, it is essential to grasp that ItsPC generates significantly smaller models than DBN (an average of 37 nodes and 30 edges versus 2438 nodes and 4873 edges). Suppose we analyse Figure 6.4 and Figure 6.5, which represent the simplified versions of the models generated by ItsPC and DBN (with only nodes around the *Survival* node), respectively, we can see that there is a significant difference in size between the models. This happens due to the fact of how the algorithms deal with temporal variables: while DBN encapsulates time through the creation of new variables that represent each of the temporal variables in each timestamp, ItsPC encodes time in the nodes themselves, thus not creating more variables, instead of creating more states in each variable. This leads

to smaller and more interpretable models. Moreover, ItsPC finds relationships between a patient's survival and non-temporal variables, for example, ICU type and age, while DBN only finds relationships with temporal variables. This means that ItsPC's model can partially access the initial potential outcome with non-temporal information from the first moment. From a partitioner's perspective, having a clue right away about the future outcome of a patient, as well as knowing what exams to be more focused on, means that, for example, it is possible to recommend specific treatments that slow or even prevent the potential outcome.

6.6 Summary

In hospital and after ICU discharge, deaths are usual, given the severity of the condition under which many of them are admitted to these wings. Because of this, there is an urge to identify and follow these cases closely.

Given their interpretable properties, as they mimic human decision-making, Bayesian Networks, especially methods like PC, can aid in this problem. They can model the supposed causal relationships present in the data.

As ICU data is usually composed of variables measured in varying time intervals, there is a need for a method that can capture causal relationships in this type of data.

To solve this problem, we propose ItsPC, a causal discovery methodology that can model causal relationships in irregular multivariate time-series data. The results found that ItsPC creates smaller and more concise networks while maintaining the temporal properties that more accurately predict these cases.

Chapter 7

Conclusion

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

The main objective of this thesis is to try to identify how we can extract causal data relationships. To achieve this objective, several research hypotheses were formulated: first, we hypothesise that the usage of observational data, rather than experimental data, is possible since recent studies show that causal models can be created from observational data. Second, we also hypothesise that causal discovery can be applied through non-Bayesian causal algorithms, such as Causal Decision Trees, with better results. Third, we also hypothesise that applying causality to select and create variables can be advantageous since we are selecting and creating variables from the information of the relationships present in the data. Finally, it is possible to create accurate causal models for sequential data.

We started by answering **RQ.1**: *Is it possible to extract causal relationships from data? How?*. for this, we studied the potential usage of association rule mining to generate causal rules. We implemented CRPA-UC, a causal association rules method that generates causal rules for discrete cross-sectional observational data. This method uses the GCMH and χ^2 as independence tests and the UC as an orientation method. With this algorithm we were able to extract causal relationships that are fundamental to generate more accurate models. We explored several cases and obtained gains of 79.40 %.

After this, we studied the potential usage of a hybrid approach to improve the interpretability of decision trees, trying to answer **RQ.2**: *Is it possible to obtain more interpretable methods by using causal discovery?* Our objective in this research question was to identify a mean to maintain the decision tree's prediction properties (obtained through correlation) but generate a model that represented the supposed causal relationships found in the data. To do so, we proposed the SC Tree. This hybrid approach uses a custom information gain ratio equation that evaluates both the correlation between the target and the other variables and identifies the causal relationships between variables. With this algorithm we were able to extract causal relationships that are fundamental to generate interpretable models, while maintaining a performance similar to a tradicional decision trees. We explored a case and obtained gains of 80.06 % in binary data sets and 83.33 % in overall discrete datasets.

To try to answer **RQ.3**: *In what other situations can we apply causality beyond causal discovery?*, we proposed a framework that employs a causal discovery method (PC) to generate new supposedly causal features (based on posterior probabilistic analysis) that represent the causal relationships between a target variable and the variables considered relevant, being these variables the parents and children of the Markov Blanket the target. With this algorithm we were able to extract causal useful knowledge to generate more accurate models. We explored several discrete data sets and obtained gains of 84 %.

Next, to analyse **RQ.4**: *Can we create causal models from sequential data?*, we studied the practical case of ICU patients' survival. As patient data is characterised by its irregularity in measurements (mixture of static and temporal variables measures in regular or irregular time intervals), we proposed ItsPC, a causal discovery methodology that deals with this issue by modelling the variables as delayed occurrences instead of absolute ones, thus finding causal relationships in irregular multivariate time-series data. With this algorithm we were able to extract causal relationships that are fundamental to create models that represent relationships as perceived by humans. We explored a case and obtained gains of 64.82 %.

Finally, **RQ.5** (*Are causal relationships helpful, and can they bring significant gains?*) was studied and developed through the proposed methodologies. In Chapter 3, CRPA-UC had significantly better results than PC in several classification problems. In Chapter 4, SC

Tree proved to have a performance similar to J48 while creating smaller and more interpretable trees. Moreover, this methodology had significantly better results than CDT-PS, CDT-SPS and PC in several classification problems. In Chapter 5, the causal relationships found in the data helped improve the performance of Random Forest, and in Chapter 6, ItsPC had significantly better results than the DBNs in predicting the patient's survival.

As further research, we made available several resources used and created during the elaboration of this thesis: a practical guide for researchers and practitioners that are just entering the causality domain(https://github.com/AnaRitaNogueira/Methods-and-Tools-for-Causal-Discovery-and-Causal-Inference), data sets commonly used in causal related tasks (https://github.com/AnaRitaNogueira/Causality-Repository-data-sets-), list of currently available software (https://github.com/AnaRitaNogueira/Causality-Repository-software) and list of current causal related surveys(https://github.com/AnaRitaNogueira/-Causality-Repository-research-papers).

7.1 Limitations and Future Work

The proposed methodologies have some limitations despite answering all the proposed research questions. For example, these methodologies were designed to deal specifically with discrete data. Nevertheless, they can arguably be altered to deal with continuous or mixed data, as these solutions depend on the statistical tests developed for that specific type of data.

In the future, in studying further potential adaptations of traditional prediction methodologies so that they can generate causal models. Moreover, we will also analyse these methods and their potential modification to create models that represent both causality and correlation. Finally, address the irregular time-series data problem by implementing the same data transformation methodology in other causal discovery algorithms.

Bibliography

[1] T. M. Schmaltz, *Descartes on causation*. Oxford University Press, 2007. [Cited on page 1.]

[2] A. Falcon, "Aristotle on causality," in *The Stanford Encyclopedia of Philosophy*, spring 2015 ed., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2015. [Cited on pages 1 and 13.]

[3] C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969. [Cited on pages 1 and 32.]

[4] J. Pearl, *Causality*, 2nd ed. Cambridge University Press, 2009. [Cited on pages 1, 18, and 28.]

[5] S. Kleinberg, *Why: a guide to finding and using causes*. Sebastopol, CA: O'Reilly, 2015. [Cited on pages 1 and 78.]

[6] P. Illari, F. Russo, and J. Williamson, "Why look at causality in the sciences? A manifesto," in *Causality in the Sciences*. Oxford University Press, mar 2011, pp. 3–22. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.232.9105{&}rep=rep1{&}type=pdf [Cited on page 3.]

[7] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search, Second Edition*, ser. Adaptive computation and machine learning. MIT Press, 2000. [Cited on pages 4, 17, 18, 19, 20, and 21.]

[8] J. Li, S. Ma, T. D. Le, L. Liu, and J. Liu, "Causal decision trees," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 2, pp. 257–271, 2017. [Online]. Available: https://doi.org/10.1109/TKDE.2016.2619350 [Cited on pages 4, 27, 52, and 66.]

[9] A. R. Nogueira, C. Ferreira, J. a. Gama, and A. Pinto, "Generalised partial association in causal rules discovery," in *Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2021, p. 485–497. [Online]. Available: https://doi.org/10.1007/978-3-030-86230-5_38 [Cited on page 6.]

[10] A. R. Nogueira, C. A. Ferreira, and J. Gama, "Semi-causal decision trees," *Progress in Artificial Intelligence*, vol. 11, no. 1, pp. 105–119, 2022. [Cited on page 6.]

[11] A. R. Nogueira, J. Gama, and C. A. Ferreira, "Improving prediction with causal probabilistic variables," in *Advances in Intelligent Data Analysis XVIII - 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27-29, 2020, Proceedings*, ser. Lecture Notes in Computer Science, M. R. Berthold, A. Feelders, and G. Krempl, Eds., vol. 12080. Springer, 2020, pp. 379–390. [Online]. Available: https://doi.org/10.1007/978-3-030-44584-3_30 [Cited on page 7.]

[12] A. R. Nogueira, A. Pugnana, S. Ruggieri, D. Pedreschi, and J. Gama, "Methods and tools for causal discovery and causal inference," *WIREs Data Mining Knowl. Discov.*, vol. 12, no. 2, 2022. [Online]. Available: https://doi.org/10.1002/widm.1449 [Cited on page 8.]

[13] E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation*, 2009, vol. 19, no. 6. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/10543400903244270 [Cited on page 11.]

[14] 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*, 1985, pp. 329–334. [Cited on pages 11, 14, and 16.]

[15] M. Bunge, *Causality and modern science*. Routledge, 2017. [Cited on page 12.]

[16] F. E. Croxton and D. J. Cowden, "Applied general statistics." 1939. [Cited on page 12.]

[17] N. Barrowman, "Correlation, causation, and confusion," *The New Atlantis*, pp. 23–44, 2014. [Cited on page 12.]

[18] Bruce W. Carlson, "Simpson's paradox | Definition, Example, and Explanation | Britannica.com," 2016. [Online]. Available: https://www.britannica.com/topic/Simpsons-paradox [Cited on page 12.]

[19] H. D. P. Lee *et al.*, *Timaeus and Critias*. Penguin, 1971. [Cited on page 13.]

[20] J. Barnes *et al.*, *Complete Works of Aristotle, Volume 1: The Revised Oxford Translation*. Princeton University Press, 2014, vol. 1. [Cited on page 13.]

[21] J. Huyssteen, *Encyclopedia of Science and Religion.* Gale Group, Inc, 2003. [Cited on page 13.]

[22] B. Stroud, "Hume and the idea of causal necessity," *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, vol. 33, no. 1, pp. 39–59, 1978. [Cited on page 13.]

[23] H. P. Grice and A. R. White, "Symposium: The causal theory of perception," *Proceedings of the Aristotelian Society, Supplementary Volumes*, vol. 35, pp. 121–168, 1961. [Cited on page 13.]

[24] A. Janiak, "Three concepts of causation in newton," *Studies in History and Philosophy of Science Part A*, vol. 44, no. 3, pp. 396 – 407, 2013. [Cited on page 13.]

[25] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Comput. Surv.*, vol. 53, no. 4, pp. 75:1–75:37, 2020. [Online]. Available: https://doi.org/10.1145/3397269 [Cited on pages 13 and 51.]

[26] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Trans. Knowl. Discov. Data*, vol. 15, no. 5, pp. 74:1–74:46, 2021. [Online]. Available: https://doi.org/10.1145/3444944 [Cited on page 13.]

[27] C. Hitchcock, "Causal Models," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2020. [Online]. Available: https://plato.stanford.edu/archives/sum2020/entries/causal-models/ [Cited on page 14.]

[28] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974. [Cited on page 14.]

[29] R. E. Neapolitan *et al.*, *Learning bayesian networks*. Pearson Prentice Hall Upper Saddle River, NJ, 2004, vol. 38. [Cited on page 14.]

[30] U. S. Kesmodel, "Cross-sectional studies–what are they good for?" *Acta obstetricia et gynecologica Scandinavica*, vol. 97, no. 4, pp. 388–393, 2018. [Cited on page 15.]

[31] S. Chiappa and W. S. Isaac, *A Causal Bayesian Networks Viewpoint on Fairness*. Cham: Springer International Publishing, 2019, pp. 3–20. [Online]. Available: https://doi.org/10.1007/978-3-030-16744-8_1 [Cited on page 15.]

[32] J. Pearl, *Probabilistic reasoning in intelligent systems - networks of plausible inference*, ser. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1989. [Cited on pages 17 and 18.]

[33] D. Janzing and B. Schölkopf, "Causal inference using the algorithmic markov condition," *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 5168–5194, 2010. [Cited on page 18.]

[34] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995. [Online]. Available: https://doi.org/10.1093/biomet/82.4.669 [Cited on page 19.]

[35] "A review on algorithms for constraint-based causal discovery," 2016, withdrawn. [Online]. Available: http://arxiv.org/abs/1611.03977 [Cited on pages 19 and 20.]

[36] D. Margaritis and S. Thrun, "Bayesian network induction via local neighborhoods," pp. 505–511, 1999. [Online]. Available: http://papers.nips.cc/paper/1685-bayesian-network-induction-via-local-neighborhoods [Cited on page 19.]

[37] D. Colombo and M. H. Maathuis, "Order-independent constraint-based causal structure learning," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3741–3782, 2014. [Online]. Available: https://dl.acm.org/doi/10.5555/2627435.2750365 [Cited on pages 22 and 36.]

[38] P. Bühlmann, M. Kalisch, and M. H. Maathuis, "Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm," *Biometrika*, vol. 97, no. 2, pp. 261–278, 2010. [Online]. Available: http://doi.org/10.1093/biomet/asq008 [Cited on page 22.]

[39] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Mach. Learn.*, vol. 65, no. 1, pp. 31–78, 2006. [Online]. Available: https://doi.org/10.1007/s10994-006-6889-7 [Cited on page 22.]

[40] P. Spirtes, C. Meek, and T. S. Richardson, "Causal inference in the presence of latent variables and selection bias," vol. abs/1302.4983, 2013. [Online]. Available: http://arxiv.org/abs/1302.4983 [Cited on page 22.]

[41] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in Genetics*, vol. 10, p. 524, jun 2019. [Cited on pages 22, 51, 57, and 78.]

[42] P. Spirtes, "An anytime algorithm for causal inference," 2001. [Online]. Available: http://www.gatsby.ucl.ac.uk/aistats/aistats2001/files/spirtes156.ps [Cited on page 23.]

[43] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson, "Learning high-dimensional directed acyclic graphs with latent and selection variables," *The Annals of Statistics*, vol. 40, no. 1, pp. 294 – 321, 2012. [Online]. Available: https://doi.org/10.1214/11-AOS940 [Cited on page 23.]

[44] J. M. Mooij, S. Magliacane, and T. Claassen, "Joint causal inference from multiple contexts," *J. Mach. Learn. Res.*, vol. 21, pp. 99:1–99:108, 2020. [Online]. Available: http://jmlr.org/papers/v21/17-123.html [Cited on page 23.]

[45] F. Jabbari, J. D. Ramsey, P. Spirtes, and G. F. Cooper, "Discovery of causal models that contain latent variables through bayesian scoring of independence constraints," in *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2017)*, ser. LNCS, vol. 10535. Springer, 2017, pp. 142–157. [Online]. Available: https://doi.org/10.1007/978-3-319-71246-8_9 [Cited on page 24.]

[46] D. M. Chickering, "Learning equivalence classes of bayesian networks structures," Tech. Rep., 2013. [Online]. Available: http://arxiv.org/abs/1302.3566 [Cited on page 24.]

[47] A. Hauser and P. Bühlmann, "Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs," *J. Mach. Learn. Res.*, vol. 13, pp. 2409–2464, 2012. [Online]. Available: https://dl.acm.org/doi/10.5555/2503308.2503320 [Cited on page 24.]

[48] J. D. 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.*, vol. 3, no. 2, pp. 121–129, 2017. [Online]. Available: https://doi.org/10.1007/s41060-016-0032-z [Cited on page 24.]

[49] J. M. Ogarrio, P. Spirtes, and J. Ramsey, "A hybrid causal search algorithm for latent variable models," in *International Conference on Probabilistic Graphical Models (PGM 2016)*, A. Antonucci, G. Corani, and C. P. de Campos, Eds., vol. 52. JMLR.org, 2016, pp. 368–379. [Online]. Available: http://proceedings.mlr.press/v52/ogarrio16.html [Cited on page 24.]

[50] M. A. Wiering, "Evolving causal neural networks," in *Benelearn'02: Proceedings of the Twelfth Belgian-Dutch Conference on Machine Learning*, 2002, pp. 103–108. [Cited on pages xi and 25.]

[51] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985. [Cited on page 25.]

[52] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, J. B. Bocca, M. Jarke, and C. Zaniolo, Eds. Morgan Kaufmann, 1994, pp. 487–499. [Online]. Available: http://www.vldb.org/conf/1994/P487.PDF [Cited on page 26.]

[53] J. Li, L. Liu, and T. D. Le, "Causal rule discovery with partial association test," in *Practical approaches to causal relationship exploration*. Springer, 2015, pp. 33–50. [Cited on pages 26 and 39.]

[54] Z. Jin, J. Li, L. Liu, T. D. Le, B. Sun, and R. Wang, "Discovery of causal rules using partial association," pp. 309–318, 2012. [Online]. Available: https://doi.org/10.1109/ICDM.2012.36 [Cited on pages 26, 39, and 57.]

[55] W. G. Cochran, "Some methods for strengthening the common χ 2 tests," *Biometrics*, vol. 10, no. 4, pp. 417–451, 1954. [Cited on pages 26 and 127.]

[56] J. Li, T. D. Le, L. Liu, J. Liu, Z. Jin, B. Sun, and S. Ma, "From observational studies to causal rule mining," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 2, pp. 14:1–14:27, 2016. [Online]. Available: https://doi.org/10.1145/2746410 [Cited on pages 26 and 39.]

[57] J. Pearl and T. S. Verma, "A theory of inferred causation," in *Studies in Logic and the Foundations of Mathematics*. Elsevier, 1995, vol. 134, pp. 789–811. [Cited on page 27.]

[58] V. K. Raghu, A. Poon, and P. V. Benos, "Evaluation of causal structure learning methods on mixed data types," vol. 92, pp. 48–65, 2018. [Online]. Available: http://proceedings.mlr.press/v92/raghu18a.html [Cited on page 28.]

[59] L. Cheng, R. Guo, R. Moraffah, P. Sheth, K. S. Candan, and H. Liu, "Evaluation methods and measures for causal learning algorithms," *IEEE Trans. Artif. Intell.*, vol. 3, no. 6, pp. 924–943, 2022. [Online]. Available: https://doi.org/10.1109/TAI.2022.3150264 [Cited on page 28.]

[60] M. Hossin and M. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, p. 1, 2015. [Online]. Available: https://doi.org/10.5121/ijdkp.2015.5201 [Cited on page 28.]

[61] M. Kalisch, M. Mächler, and D. Colombo, "Causal Inference with Graphical Models in R Package pcalg," Tech. Rep. 11, 2012. [Cited on page 29.]

[62] M. Scutari, "Learning bayesian networks with the bnlearn r package," *Journal of Statistical Software*, vol. 35, no. 3, pp. 1–22, 2010. [Online]. Available: https://doi.org/10.18637/jss.v035.i03 [Cited on page 29.]

[63] J. Hausser and K. Strimmer, "Entropy inference and the james-stein estimator, with application to nonlinear gene association networks," *J. Mach. Learn. Res.*, vol. 10, pp. 1469–1484, 2009. [Online]. Available: https://dl.acm.org/doi/10.5555/1577069.1755833 [Cited on page 29.]

[64] O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of Empirical Finance*, vol. 10, no. 5, pp. 603–621, dec 2003. [Online]. Available: https://doi.org/10.1016/S0927-5398(03)00007-0 [Cited on page 29.]

[65] J. D. Ramsey, K. Zhang, M. Glymour, R. S. Romero, B. Huang, I. Ebert-Uphoff, S. Samarasinghe, E. A. Barnes, and C. Glymour, "Tetrad—a toolbox for causal discovery," in *8th International Workshop on Climate Informatics*, 2018. [Cited on page 29.]

[66] K. Miley, P. Meyer-Kalos, S. Ma, D. J. Bond, E. Kummerfeld, and S. Vinogradov, "Causal pathways to social and occupational functioning in the first episode of schizophrenia: uncovering unmet treatment needs," *Psychological Medicine*, p. 1–9, 2021. [Cited on page 31.]

[67] X. Shen, S. Ma, P. Vemuri, and G. Simon, "Challenges and opportunities with causal discovery algorithms: application to alzheimer's pathophysiology," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020. [Cited on page 31.]

[68] N. Afrianto, Y. Azzani, Y. Sa'adati, N. Tou, P. M. Endraswari, Y. S. R. Nur, N. Annisa, R. N. Widyanara, and R. Rahmadi, "Applying PC algorithm and GES to three clinical data sets: Heart disease, diabetes, and hepatitis," *IOP Conference Series: Materials Science and Engineering*, vol. 1077, no. 1, p. 012067, feb 2021. [Online]. Available: https://doi.org/10.1088/1757-899x/1077/1/012067 [Cited on page 31.]

[69] P. Esling and C. Agon, "Time-series data mining," *ACM Comput. Surv.*, vol. 45, no. 1, dec 2012. [Online]. Available: https://doi.org/10.1145/2379776.2379788 [Cited on page 31.]

[70] D. Entner and P. O. Hoyer, "On causal discovery from time series data using fci," *Probabilistic graphical models*, pp. 121–128, 2010. [Cited on page 32.]

[71] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, p. eaau4996, nov 2019. [Online]. Available: https://doi.org/10.1126/sciadv.aau4996 [Cited on pages 32 and 33.]

[72] J. Runge, "Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets," in *Conference on Uncertainty in Artificial Intelligence (UAI 2020)*, R. P. Adams and V. Gogate, Eds., vol. 124. AUAI Press, 2020, pp. 1388–1397. [Online]. Available: http://proceedings.mlr.press/v124/runge20a.html [Cited on page 33.]

[73] A. Gerhardus and J. Runge, "High-recall causal discovery for autocorrelated time series with latent confounders," in *Advances in Neural Information Processing (NeurIPS 2020)*, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/94e70705efae423efda1088614128d0b-Abstract.html [Cited on page 33.]

[74] J. J. McArdle and J. R. Nesselroade, *Longitudinal Data Analysis Using Structural Equation Models*. American Psychological Association, 2021/07/23/ 2014. [Online]. Available: http://doi.org/10.1037/14440-000 [Cited on page 33.]

[75] L. M. Collins, "Analysis of longitudinal data," *Annual review of psychology*, vol. 57, pp. 505–528, 2006. [Cited on page 33.]

[76] R. Moraffah, P. Sheth, M. Karami, A. Bhattacharya, Q. Wang, A. Tahir, A. Raglin, and H. Liu, "Causal inference for time series analysis: Problems, methods and evaluation," *CoRR*, vol. abs/2102.05829, 2021. [Online]. Available: https://arxiv.org/abs/2102.05829 [Cited on page 34.]

[77] F. Iglesias and W. Kastner, "Analysis of similarity measures in times series clustering for the discovery of building energy patterns," *Energies*, vol. 6, no. 2, pp. 579–597, jan 2013. [Online]. Available: https://doi.org/10.3390/en6020579 [Cited on page 34.]

[78] P. Barrett, "Euclidean distance: Raw, normalised, and double-scaled coefficients," *The technical whitepaper series*, vol. 6, pp. 1–26, 2005. [Cited on page 34.]

[79] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances," *Data Mining and Knowledge Discovery*, vol. 31, no. 3, pp. 606–660, nov 2016. [Online]. Available: https:/doi.org/10.1007/s10618-016-0483-9 [Cited on page 34.]

[80] L. Chen and R. Ng, "On the marriage of lp-norms and edit distance," in *Proceedings 2004 VLDB Conference*. Elsevier, 2004, pp. 792–803. [Online]. Available: https://doi.org/10.1016/b978-012088469-8.50070-x [Cited on page 34.]

[81] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Knowledge Discovery in Databases Workshop*. AAAI Press, 1994, pp. 359–370. [Online]. Available: https://dl.acm.org/doi/10.5555/3000850.3000887 [Cited on page 34.]

[82] A. Zeileis and T. Hothorn, "Diagnostic checking in regression relationships," *R News*, vol. 2, no. 3, pp. 7–10, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/ [Cited on page 35.]

[83] Y. Hmamouche, "Nlints: An R package for causality detection in time series," *R J.*, vol. 12, no. 1, p. 21, 2020. [Online]. Available: https://doi.org/10.32614/rj-2020-016 [Cited on page 35.]

[84] J. Runge, "Tigramite – causal discovery for time series datasets," https://tocsy.pik-potsdam.de/tigramite.php/, 2004–2021. [Cited on page 35.]

[85] D. E. Yagoubi, R. Akbarinia, B. Kolev, O. Levchenko, F. Masseglia, P. Valduriez, and D. E. Shasha, "Parcorr: efficient parallel methods to identify similar time series pairs across sliding windows," *Data Min. Knowl. Discov.*, vol. 32, no. 5, pp. 1481–1507, 2018. [Online]. Available: https://doi.org/10.1007/s10618-018-0580-z [Cited on page 36.]

[86] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769 – 2794, 2007. [Online]. Available: https://doi.org/10.1214/009053607000000505 [Cited on page 36.]

[87] J. Runge, "Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information," in *International Conference on Artificial Intelligence and Statistics (AISTATS 2018)*, vol. 84. PMLR, 2018, pp. 938–947. [Online]. Available: http://proceedings.mlr.press/v84/runge18a.html [Cited on page 36.]

[88] E. V. Strobl, "Improved causal discovery from longitudinal data using a mixture of dags," in *ACM SIGKDD Workshop on Causal Discovery, (CD@KDD 2019)*, vol. 104. PMLR, 2019, pp. 100–133. [Online]. Available: http://proceedings.mlr.press/v104/strobl19a.html [Cited on page 36.]

[89] B. Lindner, L. Auret, M. Bauer, and J. Groenewald, "Comparative analysis of granger causality and transfer entropy to present a decision flow for the application of oscillation diagnosis," *Journal of Process Control*, vol. 79, pp. 72–84, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095915241830516X [Cited on page 36.]

[90] V. Troster, M. Shahbaz, and G. S. Uddin, "Renewable energy, oil prices, and economic activity: A granger-causality in quantiles analysis," *Energy Economics*, vol. 70, pp. 440–452, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140988318300379 [Cited on page 36.]

[91] C. Krich, M. D. Mahecha, M. Migliavacca, M. G. D. Kauwe, A. Griebel, J. Runge, and D. G. Miralles, "Decoupling between ecosystem photosynthesis and transpiration: a last resort against overheating," *Environmental Research Letters*, vol. 17, no. 4, p. 044013, mar 2022. [Online]. Available: https://doi.org/10.1088/1748-9326/ac583e [Cited on page 37.]

[92] J. Li, L. Liu, and T. D. Le, "Practical approaches to causal relationship exploration," 2015. [Online]. Available: https://doi.org/10.1007/978-3-319-14433-7 [Cited on page 39.]

[93] H. Theil, "On the estimation of relationships involving qualitative variables," *American Journal of Sociology*, vol. 76, no. 1, pp. 103–154, 1970. [Cited on page 40.]

[94] P. Spirtes, C. Glymour, and R. Scheines, "Causation, prediction, and search, second edition," 2000. [Cited on pages 40, 47, 57, 61, 78, 81, and 90.]

[95] K. B. Korb and A. E. Nicholson, *Bayesian artificial intelligence*. CRC press, 2010. [Cited on page 40.]

[96] D. Ehring, *Causation and persistence: A theory of causation*. Oxford University Press, 1997. [Cited on page 40.]

[97] D. Dor and M. Tarsi, "A simple algorithm to construct a consistent extension of a partially oriented graph," *Technicial Report R-185, Cognitive Systems Laboratory, UCLA*, 1992. [Cited on page 45.]

[98] N. Mantel, "Chi-square tests with one degree of freedom; extensions of the mantel-haenszel procedure," *Journal of the American Statistical Association*, vol. 58, no. 303, pp. 690–700, 1963. [Cited on page 48.]

[99] A. Agresti and M. Kateri, "Categorical data analysis," in *International Encyclopedia of Statistical Science*, M. Lovric, Ed. Springer, 2011, pp. 206–208. [Online]. Available: https://doi.org/10.1007/978-3-642-04898-2_161 [Cited on page 48.]

[100] A. T. D. of Product, M. B. S. C. M. Manager, and P. P. G. P. M. Manager, "Correlation vs causation: Understand the difference for your product," May 2022. [Online]. Available: https://amplitude.com/blog/causation-correlation [Cited on page 52.]

[101] J. M. Mooij, J. Cremers, and Others, "An empirical study of one of the simplest causal prediction algorithms," in *UAI 2015 Workshop on Advances in Causal Inference*, no. 1504, 2015, pp. 30–39. [Cited on page 52.]

[102] R. DeFries, M. Agarwala, S. Baquie, P. Choksi, S. Khanwilkar, P. Mondal, H. Nagendra, and J. Uperlainen, "Improved household living standards can restore dry tropical forests," *Biotropica*, 2021. [Cited on page 52.]

[103] S. Tangirala, "Evaluating the impact of gini index and information gain on classification using decision tree classifier algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, pp. 612–619, 2020. [Cited on page 55.]

[104] J. Gama, A. C. P. d. L. Carvalho, K. Faceli, A. C. Lorena, M. Oliveira *et al.*, *Extração de conhecimento de dados: data mining*. Sílabo, 2015. [Cited on pages 55 and 86.]

[105] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar 1986. [Online]. Available: https://doi.org/10.1007/BF00116251 [Cited on page 55.]

[106] J. T. KENT, "Information gain and a general measure of correlation," *Biometrika*, vol. 70, no. 1, pp. 163–173, 04 1983. [Online]. Available: https://doi.org/10.1093/biomet/70.1.163 [Cited on page 56.]

[107] S. Ma and A. Statnikov, "Methods for computational causal discovery in biomedicine," *Behaviormetrika*, vol. 44, no. 1, pp. 165–191, Jan 2017. [Online]. Available: https://doi.org/10.1007/s41237-016-0013-5 [Cited on page 57.]

[108] S. Luma-Osmani, F. Ismaili, X. Zenuni, and B. Raufi, "A systematic literature review in causal association rules mining," in *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2020, pp. 0048–0054. [Cited on page 57.]

[109] A. Marx and J. Vreeken, "Testing conditional independence on discrete data using stochastic complexity," vol. 89, pp. 496–505, 2019. [Online]. Available: http://proceedings.mlr.press/v89/marx19a.html [Cited on page 58.]

[110] A. Agresti, *An introduction to categorical data analysis*. John Wiley & Sons, 2018. [Cited on page 58.]

[111] M. Birch, "The detection of partial association, i: the 2× 2 case," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 26, no. 2, pp. 313–324, 1964. [Cited on page 58.]

[112] C. J. Mantas and J. Abellán, "Credal-c4.5: Decision tree based on imprecise probabilities to classify noisy data," *Expert Systems with Applications*, vol. 41, no. 10, pp. 4625 – 4637, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417414000384 [Cited on page 61.]

[113] R. Piltaver, M. Lustrek, M. Gams, and S. Martincic-Ipsic, "What makes classification trees comprehensible?" *Expert Syst. Appl.*, vol. 62, pp. 333–346, 2016. [Online]. Available: https://doi.org/10.1016/j.eswa.2016.06.009 [Cited on page 65.]

[114] Q. Zhou, F. Liao, C. Mou, and P. Wang, "Measuring interpretability for different types of machine learning models," in *Trends and Applications in Knowledge Discovery and Data Mining - PAKDD 2018 Workshops, BDASC, BDM, ML4Cyber, PAISI, DaMEMO, Melbourne, VIC, Australia, June 3, 2018, Revised Selected Papers*, ser. Lecture Notes in Computer Science, M. Ganji, L. Rashidi, B. C. M. Fung, and C. Wang, Eds., vol. 11154. Springer, 2018, pp. 295–308. [Online]. Available: https://doi.org/10.1007/978-3-030-04503-6_29 [Cited on page 65.]

[115] P. M. Domingos, "The role of occam's razor in knowledge discovery," *Data Min. Knowl. Discov.*, vol. 3, no. 4, pp. 409–425, 1999. [Online]. Available: https://doi.org/10.1023/A:1009868929893 [Cited on page 65.]

[116] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, Oct. 2012. [Cited on pages 77 and 80.]

[117] W. Martin, "Making valid causal inferences from observational data," *Prev Vet Med*, vol. 113, no. 3, pp. 281–297, Sep. 2013. [Cited on page 77.]

[118] S. Listl, H. Jürges, and R. G. Watt, "Causal inference from observational data." *Community dentistry and oral epidemiology*, vol. 44 5, pp. 409–15, 2016. [Cited on page 77.]

[119] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic Books, 2018. [Cited on page 78.]

[120] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014. [Cited on page 78.]

[121] G. Tripepi, K. J. Jager, F. W. Dekker, and C. Zoccali, "Stratification for confounding–part 1: the mantel-haenszel formula," *Nephron Clinical Practice*, vol. 116, no. 4, pp. c317–c321, 2010. [Cited on page 81.]

[122] D. Dor and M. Tarsi, "A simple algorithm to construct a consistent extension of a partially oriented graph," *R-185*, no. October, pp. 1–4, 1992. [Cited on page 81.]

[123] I. Guyon, A. Elisseeff, and C. Aliferis, "Causal feature selection," *Training*, vol. 32, pp. 1–40, 2007. [Cited on page 81.]

[124] I. Tsamardinos, C. F. Aliferis, and A. R. Statnikov, "Algorithms for large scale markov blanket discovery," in *Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference, May 12-14, 2003, St. Augustine, Florida, USA*, I. Russell and S. M. Haller, Eds. AAAI Press, 2003, pp. 376–381. [Online]. Available: http://www.aaai.org/Library/FLAIRS/2003/flairs03-073.php [Cited on page 81.]

[125] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local causal and markov blanket induction for causal discovery and feature selection for classification part ii: Analysis and extensions," *Journal of Machine Learning Research*, vol. 11, no. Jan, pp. 235–284, 2010.

[126] P. Bühlmann, M. Kalisch, and M. H. Maathuis, "Variable selection in high-dimensional linear models: Partially faithful distributions and the pc-simple algorithm," *Biometrika*, vol. 97, no. 2, pp. 261–278, jun 2010. [Cited on page 81.]

[127] D. Koller and N. Friedman, *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 2009. [Online]. Available: http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=11886 [Cited on page 81.]

[128] J. Lee, Y. J. Cho, S. J. Kim, H. I. Yoon, J. S. Park, C. T. Lee, J. H. Lee, and Y. J. Lee, "Who dies after icu discharge? retrospective analysis of prognostic factors for in-hospital mortality of icu survivors," *Journal of Korean medical science*, vol. 32, no. 3, pp. 528–533, Mar 2017, 28145659[pmid]. [Cited on pages 89 and 92.]

[129] G. Giorello, "Causality in medicine," pp. 1–4, 2008. [Cited on page 89.]

[130] D. Shillan, J. A. C. Sterne, A. Champneys, and B. Gibbison, "Use of machine learning to analyse routinely collected intensive care unit data: a systematic review," *Critical Care*, vol. 23, no. 1, p. 284, Aug 2019. [Cited on page 89.]

[131] Z. Ghahramani, *Learning dynamic Bayesian networks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 168–197. [Cited on pages 90 and 100.]

[132] G. Arroyo-Figueroa and L. E. Sucar, "A temporal bayesian network for diagnosis and prediction," vol. abs/1301.6675, 2013. [Online]. Available: http://arxiv.org/abs/1301.6675 [Cited on pages 90 and 91.]

[133] A. Y. Tawfik and E. Neufeld, "Temporal bayesian networks," in *Proceedings of the TIME-94–International Workshop on Temporal Representation and Reasoning*. Citeseer, 1994. [Cited on page 91.]

[134] M. Ramati and Y. Shahar, "Irregular-time bayesian networks," *CoRR*, vol. abs/1203.3510, 2012. [Online]. Available: http://arxiv.org/abs/1203.3510 [Cited on page 91.]

[135] U. Hamsen, N. Drotleff, R. Lefering, J. Gerstmeyer, T. A. Schildhauer, C. Waydhas, and T. DGU, "Mortality in severely injured patients: nearly one of five non-survivors have been already discharged alive from icu," *BMC Anesthesiology*, vol. 20, no. 1, p. 243, Sep 2020. [Cited on page 92.]

[136] J. Sparling and E. A. Bittner, "Mortality risk after icu discharge: It's not over until it's over*," *Critical Care Medicine*, vol. 48, no. 1, 2020. [Cited on page 92.]

[137] M. Katsiari, K. Ntorlis, C. Mathas, and C. Nikolaou, "Predictors of adverse outcome early after icu discharge," *Int J Crit Care Emerg Med*, vol. 5, no. 1, pp. 1–6, 2018. [Cited on page 92.]

[138] J. E. García-Gallo, N. Fonseca-Ruiz, L. Celi, and J. Duitama-Muñoz, "A machine learning-based model for 1-year mortality prediction in patients admitted to an intensive care unit with a diagnosis of sepsis," *Medicina intensiva*, vol. 44, no. 3, pp. 160–170, 2020. [Cited on page 92.]

[139] A. H. T. Chia, M. S. Khoo, A. Z. Lim, K. E. Ong, Y. Sun, B. P. Nguyen, M. C. H. Chua, and J. Pang, "Explainable machine learning prediction of icu mortality," *Informatics in Medicine Unlocked*, vol. 25, p. 100674, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914821001593 [Cited on page 92.]

[140] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. E215–20, Jun. 2000. [Cited on page 93.]

[141] G. W. BRIER, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1 – 3, 1950. [Cited on page 99.]

[142] J. R. Landis, E. R. Heyman, and G. G. Koch, "Average Partial Association in Three-Way Contingency Tables: A Review and Discussion of Alternative Tests," *International Statistical Review*, vol. 46, no. 3, p. 237, 2006. [Cited on page 127.]

[143] H. Theil, "Statistical decomposition analysis; with applications in the social and administrative sciences," Tech. Rep., 1972. [Cited on page 129.]

[144] X. Zhang, C. Baral, and S. Kim, "An algorithm to learn causal relations between genes from steady state data: Simulation and its application to melanoma dataset," in *Artificial Intelligence in Medicine, 10th Conference on Artificial Intelligence in Medicine, AIME 2005, Aberdeen, UK, July 23-27, 2005, Proceedings*, ser. Lecture Notes in Computer Science, S. Miksch, J. Hunter, and E. T. Keravnou, Eds., vol. 3581. Springer, 2005, pp. 524–534. [Online]. Available: https://doi.org/10.1007/11527770_69 [Cited on page 129.]

[145] S. Samothrakis, D. P. Liebana, and S. M. Lucas, *Training Gradient Boosting Machines Using Curve-Fitting and Information-Theoretic Features for Causal Direction Detection*. Springer, 2019, pp. 331–338. [Online]. Available: https://doi.org/10.1007/978-3-030-21810-2_11 [Cited on page 129.]

Appendices

A Cochran-Mantel-Haenszel test

The Cochran-Mantel-Haenszel (CMH) test [55] is a test of independence, which differs from others like χ^2, because it tests if the relationship between two variables is maintained when influenced by the remaining variables, instead of only testing if two variables are related. There are two distinct versions of this test: the binary version and its generalised version [142], which can be used in every categorical data. The binary version is given by (1).

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

In the previous equation, the values n represent the cells of contingency tables identical to Table 1 (each cell of this table represent how many cases there are given the values of the studied variables and their supposed confounders), 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 cell in the first row, the sum of the cell in the second row, the sum of the cell in the first column, the sum of the cell in the second column and the sum of all the cells, of a table k.

As explained previously, this version of the CMH test can only be applied to binary data. However, other categorical non-binary data in which the application of this type of algorithms can be relevant. To those cases, the **GCMH test** is applied instead. This variant was designed to be used in contingency tables of size $I \times J \times K$ (instead of $2 \times 2 \times K$, as in the binary version) and is given by (2) [142]. In the equations previously presented, B_h

represents the product of Kronecker between C_h and R_h (these values are obtained from the partial contingency table, as showed in Table 2), Var the co-variance matrix, $(nh - mh)$ the difference between the observed and the expected and H_0 as the null hypothesis.

$$Q_{CMH} = G'Var\{G|H_0\}^{-1}G \qquad G_h = B_h(n_h - m_h) \qquad G = \sum_h G_h$$
$$Var\{G|H_0\} = \sum_h Var\{G_h|H_0\} \qquad B_h = C_h \bigotimes R_h. \tag{2}$$

TABLE 1: Example of a partial contingency table used in CMH test (in which $c_k = \{A = a1, B = b1\}$)

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

TABLE 2: Example of a partial contingency table used in GCMH test (in which $c_h = \{A = a1, B = b1\}$

$c_h = \{A, B\}$	$C = c_1$	$C = c_2$	$C = c_3$	...	$C = c_n$	Total
$R = r_1$	n_{11h}	n_{12h}	n_{13h}	...	n_{1nh}	$n_{1.h}$
$R = r_2$	n_{21h}	n_{22h}	n_{23h}	...	n_{2nh}	$n_{2.h}$
$R = r_3$	n_{31h}	n_{32h}	n_{3nh}	...	n_{3nh}	$n_{3.h}$
...	...	...	...	...	...	...
$R = r_n$	n_{n1h}	n_{n2h}	n_{n3h}	...	n_{nnh}	$n_{n.h}$
Total	$n_{.1h}$	$n_{.2h}$	$n_{.3h}$	...	$n_{.nh}$	$n_{..h}$

B Uncertainty Coefficient

The Uncertainty Coefficient (UC) is as measure of entropy used for discrete variables, that measures how much a variable x can explain a variable y and is given by the first formula in (3) [143].

$$U(y|x) = \frac{H(y) - H(y|x)}{H(y)} \tag{3}$$

This dependence is obtained by combining the entropy from y (4) ($H(y)$) and the entropy of y given x ($H(y|x)$). This coefficient has values comprehended between 0 and 1, being 0 the representation of no relation between the variables (x does not explain y) and 1 a full relation (x completely explains y).

$$H(y) = -\sum_{j} p_j \ln p_j \quad H(y|x) = -\sum_{i,j} p_{ij} \ln \frac{p_{ij}}{p_i} \tag{4}$$

This measure has already been used in causal discovery-related tasks. For example, Zhang et al. [144] applied this dependence measure to test the conditional independence of variables in the IC algorithm. In the work of Samothrakiset al. [145], the uncertainty coefficient is used as an asymmetric dependence feature that is used to train two Gradient Boosting Machines to detect a causal relationship between a pair of variables.

Given its asymmetric property, the uncertainty coefficient is a candidate to be used as an orientation method since it can assign the dependence of the relationship orientation and states its degree.