Università
degli Studi
di Padova

Head Office: Università degli Studi di Padova
Department: Information Engineering

Ph.D. course in: Information Engineering
Curriculum: Information and Communication Technologies (ICT)
Series: XXXVI, 2023

Statistical Learning Techniques for Causal Structure Discovery and Effect Estimation

Coordinator: Ch.mo Prof. Fabio Vandin
Supervisor: Ch.mo Prof. Fabio Vandin

Ph.D. Candidate: Dario Simionato

Abstract

The increasing availability of data and the decreasing computational power cost sparked the data-revolution that we are in nowadays, with machine learning and artificial intelligence methods influencing our daily life more and more. Such trend also influenced several scientific fields, as now complex and massive data analyses are performed in multiple fields allowing researchers to test elaborate hypotheses and to speed-up discoveries. A significant drawback of traditional machine learning approaches, however, is their ability to discover *only correlations* between variables that do not always reflect the true *causal* mechanisms of the phenomenon under study, possibly leading to misleading conclusions.

In light of such obstacles, the field of *causality* has gained significant traction due to its natural ability to answer two fundamental questions for knowledge discovery from data. The first is to select the important variables among a pool of observed ones, as big datasets comprised of multiple and heterogeneous measurements are collected for subsequent analyses without any prior knowledge of the importance of each feature. The second is to understand how those variables influence each other, as this helps understanding the evolution of the scenario under study. Both questions can be answered using the causal framework as *causal discovery* algorithms aim to recover cause and effect relationships among variables, upon which it is possible to identify the important ones for the task under study, and *effect estimation* techniques quantify how modifying a feature (or treatment) in the real-world influences the other variables, allowing us to better understand the system under study.

One last common issue of data analysis on datasets is to report *false discoveries* in output, that are results that arise by chance without reflecting causal effects or other relationships in the data. This problem is especially important when performing large analyses comprised of multiple hypotheses, and this is critical in high-stake fields such as in financial or medical analyses. One way to address this problem is to adopt suitable techniques designed to bound the *Family-Wise Error Rate* (*FWER*), that is the probability of returning at least one false discovery in output, below an user-defined threshold. In this Thesis we develop two causality methods with rigorous guarantees on the FWER: the first focuses on a causal discovery problem and the second involves effect estimation and its application on cancer data.

In the first part of the Thesis we focus on a sub task of causal discovery, the *local* causal discovery task, that given a target variable and a candidate set of variables, aims at selecting a subset of the latter with specific causal or statistical properties with the target. In particular, local causal discovery focuses on inferring two sets of variables: the *Parent-Children (PC)* set, which is composed of variables that are direct causes or direct consequences of the target, and the *Markov boundary (MB)* of the target, which is the minimal set of variables with the highest target prediction performances. We present the first two algorithms for local causal discovery that bound the FWER of their output, as the inference of PC and MB sets requires performing multiple independence tests from data. We prove that state-of-the-art algorithms cannot be adapted for the task due to untestable and unrealistic assumptions on the statistical power of independence tests used for the discovery, while our algorithms come with provable guarantees on their results and require less assumptions. We successfully control the FWER either by exploiting the well-known *Bonferroni correction* for multiple hypotheses testing or by implementing data-dependent bounds based on *Rademacher averages*, a tool commonly used to measure the complexity of a family of functions. To the best of our knowledge, our work is the first one introducing the use of Rademacher averages in (local) causal discovery. We then introduce two test statistics to be used in independence testing with Rademacher averages. Finally, we analyse the performances of our algorithms and our proposed statistics both on synthetic and real-world data.

We then focus on the problem of inferring the effect of multiple treatments on a target variable using the syntax of *causal rules*, which are convenient ways of representing multiple variables taking specific values. Our aim is to discover the top-k (where k is a user-defined parameter) rules with highest effect from a dataset of observations by controlling the FWER, since to address this problem multiple hypotheses need to be tested. We develop a branch-and-bound algorithm with provable guarantees for such discovery task, and we also prove that the underlying problem is NP-hard. We then adapted our algorithm for the breast cancer context by adding two parameters: a first one that encodes the set of admissible rules to study, and a second one that controls the discovery non-overlapping results. These two features steer the rule-mining process towards a set of diverse biologically-informed rules that represent the combinations of treatments with the highest causal effect on the target variable. We extensively assess the performances of our tool on synthetic data and we run it on real-world breast cancer datasets, where it was able to identify both well-known mutational patterns that cause this malignancy and novel candidate relationships to test via follow-up studies.

Acknowledgments

Eccomi qui, alla fine di questo dottorato a scrivere i ringraziamenti nonostante sembri ieri il momento in cui scrivevo quelli di tesi magistrale. Mi piace paragonare il mio dottorato a un cammino: è un percorso che richiede impegno ma regala emozioni uniche sia nel mentre che una volta completato. E in questo momento mi sento come fossi in cima a un picco. Da qui in alto vedo orizzonti che prima mi erano occlusi, e voltandomi riconosco di avere fatto tanta strada e aver visto una miriade di nuovi paesaggi. Se sono arrivato sin qui è merito di molte persone, che spero di ringraziare adeguatamente in quanto segue.

Innanzitutto non posso che ringraziare Fabio: è grazie a te se ho avuto l'opportunità di vivere tutto questo. In questi anni mi hai guidato con il tuo distintivo ottimismo e la tua umile gentilezza dandomi un esempio che mi ha fatto crescere prima come persona e poi come studente. È anche grazie a te, infine, che ho vissuto 6 mesi magici nella città più pazza al mondo.

Talking about the craziest and most beautiful city in the world, I'd like to thank Iman for the amazing opportunity of staying in New York City and visiting his lab. It was an incredible experience that I never thought I would have been able to make, and of which I will always keep lots of memories with me. I'd also like to thank Eeshaan, Lauren, Salil, Suraj, William, and all the other guys in the lab for all the funny lunches and stimulating discussions that made my time in the City fly.

Rimanendo in tema amici, non posso non ricordare tutte le altre persone che mi sono state accanto in questi 3 anni: siete davvero tante, e non posso che essere orgoglioso e grato di avervi avuto al mio fianco. Grazie a Jacopo, Marco e Silvia per aver alleggerito i tragitti da e verso il DEI, ad Antonio per le sessioni di brain storming cestistico, a Dario e Fabrizio per l'umorismo e le ricariche di zuccheri, a Ilie per l'organizzazione della BBQ, a Paola per le discussioni profonde, ad Andrea, Davide, Diego e Leonardo per la disponibilità e gli scambi d'opinione. E come se ciò non bastasse a rendere il mio clima migliore, grazie anche a Erica, Francesca, Giacomo, Giulia, Isotta, Marco, Massimo, Matteo e Mikele per i caffè assieme e le sfide a pallavolo.

Last but not least (at all), un ringraziamento speciale va alla mia famiglia. Non solo mi avete permesso di arrivare dove sono, ma mi avete sempre appoggiato e amato incondizionatamente, specialmente in questi ultimi, sfidanti, anni.

Mai avrei immaginato di fare un dottorato, mai di fare *questo* dottorato. Che questo stupore mi accompagni e sia un buon auspicio per il futuro.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As humanity, we are now traversing the information age, an era in which we generate, retrieve and analyse an unprecedented quantity of data on which we ground our choices. Such revolution has allowed us to record and store data from various phenomena, such as industrial processes, medical analyses, or market trends, with the hope of extracting useful information from such data that will lead us to improve our productive systems, health statuses, and economies.

In order to get such value from data, however, we need to process it and extract meaningful information upon which to base our decisions. This can be accomplished via *hypothesis testing* that leverages observed data to validate a thesis while ensuring a certain level of statistical reliability within a specified margin of error. Another way is to automatically analyse data for discovering specific trends and relationships, as in *data mining* tasks. Finally, it is possible to combine the two worlds by designing data mining algorithms with statistical guarantees on their results. Achieving such goal is not straightforward, as it involves testing *multiple* hypotheses and the combination of their statistical errors often leads to an increased probability of reporting incorrect results in the output. In the context of *multiple hypothesis testing* we therefore have to adopt specific techniques to control the error achieved over the discoveries in output to the algorithms. A commonly adopted strategy is to bound the *Family-Wise Error Rate (FWER)*, that is the probability of returning in output at least one false positive, and algorithms with such guarantees play a key role in high-stakes decisions fields, as in the financial or in the medical areas.

Even if they provide statistical guarantees, however, most of data mining results are based on *correlations* and that may lead to incorrect (or even worse, harmful) decisions, especially in the just mentioned high-stakes field. Such problem is not related to correlation strength, as an high correlation between variables may not reflect a true cause-effect relationship, but it has a different nature. As a simple example, let us examine the relationship between the number of fires in a city and the size of its firefighting personnel. The number of firefighters is strongly and positively correlated with the amount of fires in a city, but that does not mean that reducing the former will decrease the latter (it would actually have an opposite effect). The key here is that the real-world model is more complex than the one under analysis: both variables are direct consequences of a third one, that is the fire hazard of the city under study, and not including such variable would lead us to erroneous conclusions. Such issues may arise in much more complex scenarios in which our knowledge of cause-effect relationships is not as deep as the previous example, and they justify the development of *causal* techniques that distinguish spurious correlations from real-world mechanisms using data.

The gold standard for inferring causal relationships is to perform a *randomized control trial (RCT)*, that is a real-world experiment in which the population under study is divided into two homogeneous groups, one of which performs a given action (e.g. ingesting a treatment pill) while the other does not (e.g. it takes a placebo), for then comparing the two cohorts outcomes. In the previous example context, that would translate into splitting the pool of cities under study in two groups with homogenous features (such as population, infrastructure status, and climate), and to decrease the amount of firefighters in one of the two. As in this example, it is clear that performing RCTs may be risky, unethical or unfeasible, therefore the causal community developed methods to answer causal questions from *observational data*, i.e. data gathered without performing any experiment.

Causal inference can be exploited to answer two fundamental scientific questions concerning the presence of interactions between different variables composing a system, and their strength. The first is a *causal discovery* question, whose aim is to infer cause and effect relationships among the variables under study and that is particularly useful in exploratory data analyses to understand each variable's role. As an example, causal discovery applied to biological data aims at inferring relationships (such as cooperation or competition mechanisms) among groups of organisms. The second is an *effect estimation* task that aims at quantifying the strength of each cause-effect relationship without accounting for confounding effects and spurious correlations. Such techniques are exploited in the medical field during preliminary phases of drug effect assessment, that is before performing a RCT which will empirically confirm the drug effectiveness.

In this Thesis we develop causal algorithms for the two fundamental problems introduced above with theoretical guarantees on the FWER of their output. That is especially important in high-stake scenarios, as our algorithms are able to address finite sample issues while distinguishing between spurious and genuine associations. Our algorithms work on observational data, as it is the most common type of data that does not require conducting any RCT. This Thesis work contributes to the state-of-the-art with the following results:

- In Chapter 3 we address the causal discovery problem, with guarantees on the returned outputs. In particular, we focus on the *local* causal discovery problem, which aims at inferring the causal structure in the proximity of a target variable, and it is a useful primitive for global causal discovery, which aims at the whole network inference. Given a target variable, local causal discovery aims at inferring two sets of elements: the *parent-children* set (or *PC*), that contains all the variables that are direct causes or consequences of such target, and the *Markov boundary* (or *MB*), that is the minimal set of variables that make the target independent from any other variable in the network upon conditioning on it. We prove, both analytically and experimentally, that state-of-the-art algorithms for local causal discovery do not provide guarantees on the FWER of their output and cannot be adapted for the task by simply correcting for multiple hypotheses testing due to their reliance on strong and untestable assumptions. We then present RAveL-PC and RAveL-MB, that are the first local causal discovery algorithms with statistical guarantees on the FWER of their outputs. Our algorithms rely on *Rademacher averages*, a key concept from statistical learning theory, to tackle the multiple hypotheses testing problem in scenarios where a large number of independence tests are performed. To the best of our knowledge, this is the first application of such concepts in (local) causal discovery. We then introduce two novel statistics for the task, and we evaluate them exhaustively both on synthetic and real-world datasets. This work appeared in [Simionato and Vandin, 2022, 2023].

- In Chapter 4 we focus on an effect estimation problem exploiting the syntax of *causal rules*. Causal rules represent specific combinations of treatments, that are the variables for which we want to estimate the causal effect, to which it is assigned a score representing their *average treatment effect* on the target variable. More specifically, we focus on the *reliable* causal effect estimation that takes into account estimation errors due to finite data samples providing confidence intervals for each estimate. We extend such framework taking into account the multiple hypotheses testing problem and develop ALLSTAR, a branch-and-bound algorithm for mining the top-k rules with the highest reliable effect and probabilistic guarantees on the FWER of its output. We prove that the underlying discovery problem is NP-hard, and we adapt the algorithm for applications in the medical domain. In such scenarios it is typical to collect plenty of measurements from a modest amount patients, which can complicate the discovery process. In order to deal with such issue, we study rules with up to a given number of distinct variables and we exploit a graph $\mathcal{G}$ for encoding the set of biologically-meaningful rules to study. Additionally, we present a novel and tighter bound to speed-up the discovery in single-core machines. We analyse ALLSTAR performances on synthetic data, and we run it on real-world breast cancer data, where it is able to retrieve both relations well established in the literature and novel biologically-sound rules to be verified via follow-up experiments. We finally tested the stability of the retrieved rules with respect to variations in $\mathcal{G}$ comparing such results with cancer literature.

The reminder of the Thesis is organized as follows. In Chapter 2 we introduce basic notions and notations used throughout this work. In Chapter 3 we introduce the problem of local causal discovery, and we present `RAveL-PC` and `RAveL-MB`, that are the first algorithms for PC and MB discovery with guarantees on the FWER of their results. In Chapter 4 we address the effect estimation task and we present `ALLSTAR`, our algorithm for causal rule reliable effect estimation, and its application on real-world breast cancer data. Finally, Chapter 5 discusses the contribution presented in this Thesis and highlights future research directions.

Chapter 2

Preliminaries

In this Chapter, we introduce preliminary concepts and notations that will be used throughout the Thesis. Section 2.1 introduces the concept of causal Bayesian network for representing cause and effect relationships among a set of variables. Section 2.2 and Section 2.3 present the structure discovery and the effect estimation problems, for which we developed ad-hoc algorithms in Chapter 3 and 4 respectively. Finally, Section 2.4 introduces the concept of statistical hypothesis testing and presents the challenges associated with conducting multiple tests.

2.1 Bayesian Networks

Bayesian Networks (BNs) are probabilistic graphical models useful to encode joint probabilities among a set of variables $\mathbf{V}$ by means of a *Directed Acyclic Graph (DAG)* $\mathcal{G}$ (one such example is shown in Figure 2.1) and a probability distribution function p over $\mathbf{V}$. Formally, they are defined as follows.

Definition 2.1 (Bayesian network [Neapolitan et al., 2004]). *Let p be a joint probability distribution over $\mathbf{V}$. Let $\mathcal{G} = (\mathbf{W}, \mathbf{A})$ be a DAG where the vertices $\mathbf{W}$ of $\mathcal{G}$ are in a one-to-one correspondence with members of $\mathbf{V}$, and such that $\forall X \in \mathbf{V}$, X is conditionally independent of all non-descendants of X, given the parents of X (i.e., the* Markov condition *holds). A* Bayesian Network (BN) *is defined as a triplet $\langle \mathbf{V}, \mathcal{G}, p \rangle$.*

As by definition, the Markov condition allows us to represent BNs compactly by modelling the probability distribution function of each variable $X \in \mathbf{V}$ as a function of its parents $pa(X)$ [Pearl, 2009] only, where the parents $pa(X)$ are those elements Y for which the arc $\{Y \longrightarrow X\} \in \mathbf{A}$. This allows us to decompose big (i.e. within a large set of variables) join probabilities into smaller ones. In a Bayesian network with $\mathcal{G}$ defined as in Figure 2.1, this would translate for example of having the probability distribution function of X_3 as a function of X_1 and X_2 only, that is $pa(X_3) = \{X_1, X_2\}$.

Statistical dependencies among variables of a BN can be inferred by studying the paths that link them, that is by studying $\mathcal{G}$ structure. In order to introduce such concept, let us define a *path* of any directionality from

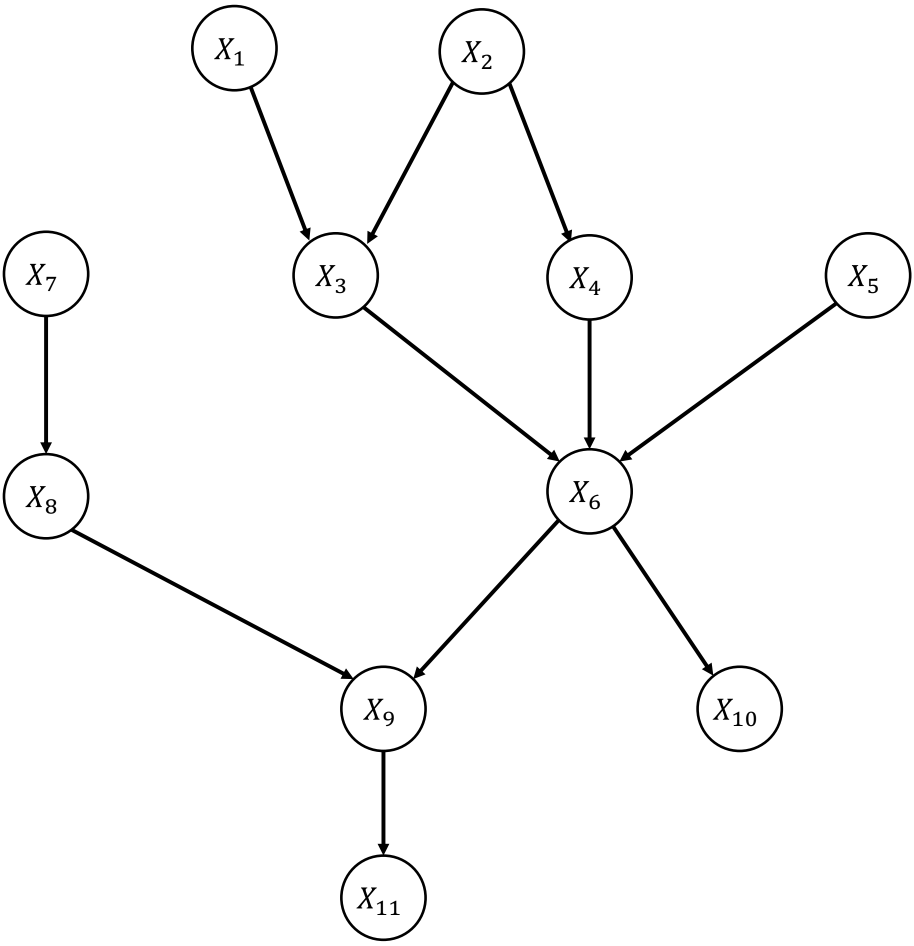


Figure 2.1: Example of DAG $\mathcal{G}$ associated to a Bayesian network.

$X \in \mathbf{V}$ to $Y \in \mathbf{V} \setminus \{X\}$ as a sequence of arcs among adjacent nodes that starts from X and ends on Y. The path is *directed* if each all the edges are oriented in the same direction. The *directional separation*, or *d-separation* [Pearl, 2009], criterion can be used to study the dependence between two subsets $\mathbf{X}$ and $\mathbf{Y}$ of variables conditional on another set $\mathbf{Z}$, such that $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ are disjoint. Informally, the criterion marks a path between any variable in $\mathbf{X}$ and any variable in $\mathbf{Y}$ as *blocked* by $\mathbf{Z}$ if the flow of dependency between the two sets is interrupted and therefore the two sets are *independent* conditioning on $\mathbf{Z}$, written $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_{\mathcal{G}}$. Viceversa, if the two sets $\mathbf{X}$ and $\mathbf{Y}$ are conditionally dependent given $\mathbf{Z}$, denoted with $(\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})_{\mathcal{G}}$, the path is marked as *open*. More formally, the definition of d-separated path is the following.

Definition 2.2 (d-separation [Pearl, 2009]). *A path q is* d-separated, *or* blocked, *by a set of nodes $\mathbf{Z}$ if and only if:*

1. *q contains a* chain $I \rightarrow M \rightarrow J$ *or a* fork $I \leftarrow M \rightarrow J$ *such that $M \in \mathbf{Z}$, or*

2. *q contains an* inverted fork *(or* collider*)* $I \rightarrow M \leftarrow J$ *such that $M \notin \mathbf{Z}$*

and no descendant of M is in $\mathbf{Z}$.

A set $\mathbf{Z}$ *is said to d-separate* $\mathbf{X}$ *from* $\mathbf{Y}$ *if and only if* $\mathbf{Z}$ *blocks every path from a node in* $\mathbf{X}$ *to a node in* $\mathbf{Y}$.

As an example on the structure of Figure 2.1, X_6 and X_7 are independent without conditioning on any variable (that is $(X_6 \perp\!\!\!\perp X_7|\emptyset)_{\mathcal{G}}$, or $(X_6 \perp\!\!\!\perp X_7)_{\mathcal{G}}$), as the collider passing through X_9 is blocked. If we condition on X_9, the two variables become dependent (written $(X_6 \not\perp\!\!\!\perp X_7|X_9)_{\mathcal{G}}$), but if we also condition on X_8 they are independent due to the chain $X_6 \longrightarrow X_8 \longrightarrow X_9$ being blocked (i.e. $(X_6 \perp\!\!\!\perp X_7|X_8, X_9)_{\mathcal{G}}$).

In most real-world scenarios, however, the graph $\mathcal{G}$ is not known a priori and therefore it should be inferred from data: this is possible only under the assumption of *faithfulness* that links the probabilistic properties of p with the graphical ones of $\mathcal{G}$, as defined below.

Definition 2.3 (Faithfulness [Spirtes et al., 2000]). *A directed acyclic graph* $\mathcal{G}$ *is* faithful *to a joint probability distribution* p *over a variable set* $\mathbf{V}$ *if and only if every independence present in* p *is entailed by* $\mathcal{G}$ *and the Markov Condition. A distribution* p *is* faithful *if and only if there exists a DAG* $\mathcal{G}$ *such that* $\mathcal{G}$ *is faithful to* p.

Under the faithfulness assumption, the independencies encoded by the d-separation criterion map one-to-one to statistical independencies, that is $(X \perp\!\!\!\perp Y|\mathbf{Z})_{\mathcal{G}} \iff (X \perp\!\!\!\perp Y|\mathbf{Z})_p$ [Pearl, 2009], where the latter notation represents statistical independence between X and Y conditional on $\mathbf{Z}$. In the following we will consider only faithful BNs, therefore we will use the notation $X \perp\!\!\!\perp Y|\mathbf{Z}$ to refer to either d-separation or statistical independence depending on the context. Faithfulness is an untestable assumption on the underlying data generative scenario that may not always hold, as in the case in which multiple different paths from X to Y (e.g. the two paths $X_2 \longrightarrow X_3 \longrightarrow X_6$ and $X_2 \longrightarrow X_4 \longrightarrow X_6$ from X_2 to X_6 of Figure 2.1) cancel each other out creating a statistical independence that is not implied in the graph structure (as in Example 6.34 of Peters et al. [2017]). Luckily, under mild conditions, such scenario happens with zero probability (see Theorem 3.2 of Spirtes et al. [2000]).

Lastly, a faithful BN can be *causal* if it encodes cause-effect relationships allowing us to reason about *interventions* on variables [Pearl, 2009]. An atomic intervention setting the value of X to x (written as $do(X = x)$) is a real-world action that forces the variable X to take the specified constant value x without imposing any additional constraint on the values of the other variables in the network. In the example of Figure 2.1, $P(X_6 = x_6|do(X_4 = x_4))$ measures the probability of observing $X_6 = x_6$ after forcing the variable X_4 to take the value x_4, action that may change the probability distribution of each descendant of X_4 in the graph, that are X_6, X_9, X_{10}, and X_{11}. Intervening on X makes it independent of its parents (as now we have $p(X = x) = 1$ and 0 otherwise, without considering $pa(X)$ values) and allows variables that are causally influenced by X to change as a consequence of the intervention. Interventional distributions (i.e. $p(Y = y|do(X = x))$) can be inferred from observational distributions (i.e. $p(Y = y|X = x)$) if there exists a set $\mathbf{Z}$ of variables that satisfies the *back-door* criterion, defined as follows.

Definition 2.4 (Back-door [Pearl, 2009]). *A set of variables $\mathbf{Z}$ satisfies the back-door criterion relative to an ordered pair of variables (X, Y) in a DAG $\mathcal{G}$ if:*

1. *no node in $\mathbf{Z}$ is a descendant of X; and*

2. *$\mathbf{Z}$ blocks all paths between X and Y that contain an arrow into X.*

Similarly, if $\mathbf{X}$ and $\mathbf{Y}$ are two disjoint set of nodes in $\mathcal{G}$, then $\mathbf{Z}$ is said to satisfy the back-door criterion relative to $(\mathbf{X}, \mathbf{Y})$ if it satisfies the criterion relative to any pair (X_i, Y_j) such that $X_i \in \mathbf{X}$ and $Y_j \in \mathbf{Y}$.

If $\mathbf{Z}$ satisfies the back-door criterion for (X, Y), then interventional distributions is identifiable by *back-door adjustment* as follows.

Theorem 2.1 (Back-door adjustment [Pearl, 2009]). *If a set of variables $\mathbf{Z}$ satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by the formula*

$$p(Y = y|do(X = x)) = \sum_{\mathbf{z}} p(Y = y|X = x, \mathbf{Z} = \mathbf{z})p(\mathbf{Z} = \mathbf{z}).$$

The back-door adjustment allows us to infer causal relationships from observational data if the back-door paths for the couple (X, Y) are properly blocked. This can be the case for calculating $P(X_6 = x_6|do(X_4 = x_4))$ by conditioning on $\mathbf{Z} = \{X_2\}$ given the structure of Figure 2.1, leading to

$$p(X_6 = x_6|do(X_4 = x_4)) = \sum_{x_2} p(X_6 = x_6|X_4 = x_4, X_2 = x_2)p(X_2 = x_2).$$

2.2 Structure Discovery

One of the main causal inference tasks is the *causal discovery task* which aims at inferring structural properties of the BN from observational data, proving to be useful in data exploration phases for understanding dependencies between variables. The discovery task may focus on inferring the complete graph $\mathcal{G}$, that is the *global discovery task*, or the region in proximity of a target variable $T \in \mathbf{V}$, as in the *local causal discovery* task. Algorithms for the latter can be used as useful primitives for global causal discovery and they focus on discovering two sets of variables, namely the *Parent-Children* set and the *Markov boundary* of T.

The *Parent-Children* (or *PC*) set of T is composed of variables that are direct causes or direct consequences of T, that are respectively the parents and the children of T in $\mathcal{G}$.

Definition 2.5 (Parent-children set of T [Ma and Tourani, 2020]). *The* parent-children set of T*, or* PC(T)*, is the set of all parents and all children of* T*, i.e., the elements directly connected to* T*, in the DAG* $\mathcal{G}$*.*

Statistically, the elements in $PC(T)$ are the only that cannot be d-separated by T (as a consequence of being directly connected to it) that is, by the Markov property, $PC(T) = \{X \in \mathbf{V} \setminus \{T\} : \forall \mathbf{Z} \subseteq \mathbf{V} \setminus \{T, X\}, T \not\perp\!\!\!\perp X|\mathbf{Z}\}$. As an example, in Figure 2.1 we have $PC(X_6) = \{X_3, X_4, X_5, X_9, X_{10}\}$ that is the union of the parents of X_6 (i.e. X_3, X_4, and X_5) and its children (i.e. X_9 and X_{10}).

The second set is the *Markov Boundary* (or *MB*) of T that is the minimal set of variables that make T independent of any other variable in the BN by conditioning on it. Such property makes it the optimal solution for the feature selection task when predicting the value of T [Ma and Tourani, 2020, Tsamardinos et al., 2003a], and its definition follows.

Definition 2.6 (Markov Boundary of T [Pearl, 2009, Tsamardinos et al., 2003a]). *The* Markov Boundary of T *or* MB(T) *is the smallest set of variables in* $\mathbf{V} \setminus \{T\}$ *conditioned on which all other variables are independent of* T*, that is*

$$MB(T) = \{X \in \mathbf{V} \setminus \{T\} : \forall Y \in \mathbf{V} \setminus \{MB(T) \cup \{T\}\}, T \perp\!\!\!\perp Y|MB(T)\}.$$

The Markov boundary of T is composed by parents, children, and *spouses* of T, that are parents of children of T that are not T. This implies that T and each spouse X form a collider structure with a common child Y, and it mathematically translates into the following set $spouses(T) = \{X \in \mathbf{V} \setminus \{T\} : \exists Y \in PC(T), X \in PC(Y) : \forall \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y, T\}, X \perp\!\!\!\perp T|\mathbf{Z} \wedge X \not\perp\!\!\!\perp T|\mathbf{Z} \cup \{Y\}\}$. The Markov Boundary of X_6 in Figure 2.1 is $MB(X_6) = PC(X_6) \cup spouses(X_6) = \{X_3, X_4, X_5, X_9, X_{10}\} \cup \{X_8\}$. X_8 is in fact, the only spouse of X_6 with X_9 as common child, that is the spouse definition holds by setting $T = X_6, X = X_8$, and $Y = X_9$.

2.3 Effect Estimation

Another main causal inference task is *effect estimation*, which goal is to infer how much a treatment X influences a specific target variable T taking into account the presence of confounder variables $\mathbf{Z}$[1]. Such tasks focuses on estimating post-intervention probabilities (i.e. $p(T = t|do(X = x))$) from observational data avoiding biases and finite sample data issues, for then combining them to calculate the effect value of interest.

When dealing with binary treatments, the most studied quantity is the *Average Treatment Effect* (*ATE*), or *Average Causal Effect* (*ACE*), defined as [Hol-

[1] This section definitions can be easily extended in the case of multiple treatments $\mathbf{X}$.

land, 1986, Rubin, 1974][2]

$$ATE(X \to T) = E[T|do(X=1)] - E[T|do(X=0)]$$

which measures the average increase of T when performing the action $do(X=1)$ w.r.t. $do(X=0)$. Such concept can be adapted for categorical treatment by considering two categories at a time, as suggested in [Wang et al., 2017], or for continuous data studying the *average causal derivative* [Chernozhukov et al., 2022] defined as follows

$$ACD(X \to T) = E[\delta/\delta_x T|do(X=x)].$$

2.4 Statistical and Multiple Hypothesis Testing

Causal discovery tasks assume to rely on the knowledge of the true probability distribution function p in order to test for independencies between variables but in most cases such function is unknown pushing us to infer its properties from data via *statistical hypothesis testing*. Such tests take as an input a null hypothesis H_0 and a set of observations $\mathcal{S}$, and they compute a test statistic that follows a specific distribution if H_0 is true. As an output, such tests return a p-value representing the probability of observing a test statistic as extreme as the one observed if H_0 holds.

In practice, while testing for the independence of two variables X and Y, they are considered dependent if the p-value of the corresponding test is below a threshold δ, that is if there is a very low probability (lower than δ) of observing such dependence by chance. It is easy to see that such procedure guarantees that if X and Y are independent, then the probability of a *false discovery*, that is *falsely rejecting* their independence, is at most δ. The situation is drastically different when a large number N of hypotheses are tested, as in the case of local causal discovery. In this case, if the same threshold δ is used for every test, the expected number of false discoveries can be as large as δN. Therefore, it is of fundamental importance to correct for multiple hypothesis testing, with the goal of providing guarantees on false discoveries.

A commonly used guarantee is provided by the *Family-Wise Error Rate (FWER)*, which is the probability of having at least one false discovery among all the tests. A common approach to control the FWER is the so called *Bonferroni correction* [Bonferroni, 1936], which performs each test with a corrected threshold $\delta_{corr} = \delta/N$ (a simple union bound shows that the resulting FWER is at most δ).

[2] The reader used to Rubin causal models may recall the same definition with counterfactuals in place of interventional probabilities. Despite being related to two different concepts, counterfactual and interventional probabilities are equivalent if we do not condition on any evidence, as the abduction phase (see Th. 7.1.7 of [Pearl, 2009]) does not update the probabilities over the unobserved variables **U**. For a more in-depth discussion, we point the interested reader to Chapter 7 of [Pearl, 2009], and its note at page 221.

A similar problem arises when we want to estimate confidence intervals around empirical probabilities $\hat{p}$ that contain the true ones p with probability $1 - \delta$, as in the task of effect estimation with guarantees. If multiple confidence intervals are estimated, then we have to account for the multiple hypothesis problem and perform each inference with a threshold $\delta_{corr} = \delta/N$ in order to bound the FWER of the family of estimations.

2.5 Notation

Finally, we summarize part of the notation we will use in this work in Table 2.1.

Table 2.1: Notation table.

Symbol	Description
X	Variable X
x	Value x of variable X
$\mathbf{X}$	Set of variables $\mathbf{X}$
$\mathbf{x}$	Set of values $\mathbf{x}$ of variables $\mathbf{X}$
$\mathcal{G} = (\mathbf{V}, \mathbf{A})$	Directed Acyclic Graph (DAG) $\mathcal{G}$ with nodes $\mathbf{V}$ and arcs $\mathbf{A}$
$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$	D-separation of $\mathbf{X}$ and $\mathbf{Y}$ given $\mathbf{Z}$ on graph $\mathcal{G}$
$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_p$	Statistical conditional independence w.r.t. probability distribution function p of $\mathbf{X}$ and $\mathbf{Y}$ conditioning on $\mathbf{Z}$
$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})$	Conditional independence of $\mathbf{X}$ and $\mathbf{Y}$ conditioning on $\mathbf{Z}$, equivalent both to statistical conditional independence and d-separation if faithfulness holds.

Chapter 3

RAveL

In this Chapter we address the local causal discovery problem with guarantees on the FWER. After proving that state-of-the-art algorithms cannot be adapted for the task, we introduce `RAveL-PC` and `RAveL-MB` for the *PC* and *MB* discovery, respectively. `RAveL-PC` and `RAveL-MB` come with provable guarantees on the FWER of their outputs, and they tackle the multiple hypothesis problem using data-dependent bounds based on Rademacher averages. We finally assessed `RAveL-PC` and `RAveL-MB` performances both on synthetic and real-world datasets.

3.1 Introduction

One of the most fundamental and challenging problems in science is the discovery of causal relations from observational data [Pearl, 2009]. Bayesian networks are graphical models that are widely used to represent causal relations and have been the focus of a large amount of research in data mining and machine learning. Bayesian networks represent random variables or events as vertices of graphical models, and encode conditional-independence relationships according to the (directed) Markov property among the variables or events as directed acyclic graphs (DAGs). They are a fundamental tool to represent causality relations among variables and events, and have been used to analyze data from several domains, including biology [Pe'er, 2005, Sachs et al., 2005], medicine [Velikova et al., 2014], and others [Yusuf et al., 2021, Kusner and Loftus, 2020].

One of the core tasks in learning Bayesian networks from observational data is the identification of local causal structures around a target variable T. In this work we focus on two related local structures. The first one is the set of parents and children (i.e., the neighbours) of T in the DAG, denoted as the parent-children set $PC(T)$. $PC(T)$ has a natural causal interpretation as the set of *direct* causes and effects of T [Spirtes et al., 2000], and the accurate identification of $PC(T)$ is a crucial step for the inference of Bayesian networks. The second structure is the Markov boundary of T, denoted as $MB(T)$. $MB(T)$ is a minimal set of variables that makes T conditionally independent of all the other variables, and comprises the elements of $PC(T)$

and the other parents of the children of T. Thus, $MB(T)$ includes all direct causes, effects, and causes of direct effects of T. Moreover, under certain assumptions, the Markov boundary is the solution of the variable selection problem [Tsamardinos and Aliferis, 2003], that is, it is the minimal set of variables with optimal predictive performance for T.

In several real-world applications, such as biology [Sachs et al., 2005] and neuroscience [Bielza and Larrañaga, 2014], the elements in $PC(T)$ and $MB(T)$ identified from observational data provide *candidate* causal relations explored in follow-up studies and experiments, which often require significant resources (e.g., time or chemical reagents). In other areas, such as algorithmic fairness [Mhasawade and Chunara, 2021, Kusner and Loftus, 2020], local causal discovery can help in identifying discriminatory relationships in data. In these scenarios, it is crucial to identify *reliable* causal relations between variables, ideally avoiding any false discovery.

While the stochastic nature of random sampling implies that false discoveries cannot be avoided with absolute certainty (when at least a relation is reported), a common approach from statistics to limit false discoveries is to develop methods that rigorously bound the Family-Wise Error Rate (FWER), that is, the probability of reporting one or more false discoveries. However, current approaches for local causal discovery do not provide guarantees on false discoveries in terms of FWER, and the study of causal discovery with false positive guarantees has received scant attention in general (see Section 3.3).

Our contributions In this Chapter we introduce two novel algorithms that exploit Rademacher Averages for Local structure discovery (RAveL) providing rigorous guarantees on the FWER: RAveL-MB for the MB discovery task and RAveL-PC for the PC identification task. To the best of our knowledge, our algorithms are the first ones to allow the discovery of the PC set and the MB of a target variable while providing provable guarantees on false discoveries in terms of the FWER. Our algorithms crucially rely on Rademacher averages, a key concept from statistical learning theory [Bartlett and Mendelson, 2002], to properly account for the multiple-hypothesis testing problem arising in local causal discovery, where a large number of statistical test for conditional independence are performed. To the best of our knowledge, this work is the first one to introduce the use of Rademacher averages in (local) causal discovery. We prove, both analytically and experimentally, that currently used approaches to discover the PC set and the MB of a target variable cannot be adapted to control the FWER simply by correcting for multiple-hypothesis testing. This is due to their additional requirement of conditional dependencies being correctly identified, which is an unreasonable assumption due to the stochastic nature of random sampling and finite sample sizes. We then introduce two test statistics to be used in independence testing with Rademacher averages. Our experimental evaluation shows that our algorithms do control the FWER while allowing for the discovery of elements in the PC set and in the MB of a target variable. On real data, our algorithms return a subset of variables that causally influences the target in agreement with prior knowledge.

The rest of the Chapter is organized as follows. Section 3.2 revisits the preliminary concepts used in the rest of the Chapter. Section 3.3 describes previous works related to our contribution. Section 3.4 describes our algorithms and their analysis, and the assumptions required by previously proposed algorithms in order to provide rigorous results in terms of the FWER. For clarity, we describe our algorithms focusing on the case of continuous variables, but our algorithms can be easily adapted to discrete and categorical variables. Section 3.5 describes our experimental evaluation on synthetic and real data. Finally, Section 3.6 offers some concluding remarks.

3.2 Preliminaries

In this section, we revisit basic notions and preliminary concepts used in the rest of the Chapter. More specifically, in Section 3.2.1 we formally define Bayesian networks (BNs) and the sets $PC(T)$ and $MB(T)$ for a target variable T. In Section 3.2.2 we describe the statistical testing procedure commonly used by algorithms for the identification of $PC(T)$ and $MB(T)$. In Section 3.2.3 we introduce the multiple hypotheses testing problem and the FWER. Finally, in Section 3.2.4 we introduce the concept of Rademacher averages for supremum deviation estimation.

3.2.1 Bayesian networks

Bayesian Networks (BNs) are convenient ways to model the influence among a set of variables $\mathbf{V}$. BNs represent interactions using a *Direct Acyclic Graph (DAG)*, and employ probability distributions to define the strength of the relations. More formally, they are defined as follows.

Definition 3.1 (Bayesian network [Neapolitan et al., 2004]). *Let p be a joint probability distribution over $\mathbf{V}$. Let $\mathcal{G} = (\mathbf{W}, \mathbf{A})$ be a DAG where the vertices $\mathbf{W}$ of $\mathcal{G}$ are in a one-to-one correspondence with members of $\mathbf{V}$, and such that $\forall X \in \mathbf{V}$, X is conditionally independent of all non-descendants of X, given the parents of X (i.e., the* Markov condition *holds). A* Bayesian Network (BN) *is defined as a triplet $\langle \mathbf{V}, \mathcal{G}, p \rangle$.*

A common assumption for the study of BNs is *faithfulness*, defined as follows.

Definition 3.2 (Faithfulness [Spirtes et al., 2000]). *A directed acyclic graph $\mathcal{G}$ is* faithful *to a joint probability distribution p over variable set $\mathbf{V}$ if and only if every independence present in p is entailed by $\mathcal{G}$ and the Markov Condition. A distribution p is* faithful *if and only if there exists a DAG $\mathcal{G}$ such that $\mathcal{G}$ is faithful to p.*

The dependencies between variables in a faithful BN can be analyzed through the study of *paths*, which are sequences of consecutive edges of any directionality (i.e. $X \longrightarrow Y$ or $X \longleftarrow Y$) in $\mathcal{G}$. In particular, the *directional separation*, or *d-separation* [Pearl, 2009], criterion can be used to study the dependence between two subsets $\mathbf{X}$ and $\mathbf{Y}$ of variables conditioning on another set $\mathbf{Z}$ of variables, such that $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ are disjoint. Informally, the criterion marks a path between any variable in $\mathbf{X}$ and any variable in $\mathbf{Y}$ as *blocked* by $\mathbf{Z}$ if the flow of dependency between the two sets is interrupted and therefore the two sets are *independent* conditioning on $\mathbf{Z}$, written $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$. Viceversa, if the two sets $\mathbf{X}$ and $\mathbf{Y}$ are conditionally dependent given $\mathbf{Z}$, denoted with $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$, the path is marked as *open*. More formally, the definition of d-separated path is the following.

Definition 3.3 (d-separation [Pearl, 2009])**.** *A path q is* d-separated*, or* blocked*, by a set of nodes $\mathbf{Z}$ if and only if:*

1. *q contains a* chain $I \longrightarrow M \longrightarrow J$ *or a* fork $I \longleftarrow M \longrightarrow J$ *such that $M \in \mathbf{Z}$, or*

2. *q contains an* inverted fork *(or* collider*) $I \longrightarrow M \longleftarrow J$ such that $M \notin \mathbf{Z}$ and no descendant of M is in $\mathbf{Z}$.*

A set $\mathbf{Z}$ is said to d-separate $\mathbf{X}$ from $\mathbf{Y}$ if and only if $\mathbf{Z}$ blocks every path from a node in $\mathbf{X}$ to a node in $\mathbf{Y}$.

A *causal* Bayesian network is a Bayesian network with causally relevant edge semantics [Pearl, 2009, Ma and Tourani, 2020].

Local causal discovery

The task of inferring the local region of a causal BN related to a target variable T from data is called *local causal discovery.* Two sets of variables are of major importance in local causal discovery. The first set is the *parent-children set $PC(T)$*, which contains the variables that are direct cause of T or that are its direct consequence.

Definition 3.4 (Parent-children set of T [Ma and Tourani, 2020])**.** *The* parent-children set of T*, or* PC(T)*, is the set of all parents and all children of T, i.e., the elements directly connected to T, in the DAG $\mathcal{G}$.*

The elements in $PC(T)$ are the only variables that cannot be d-separated from T, that is, by the Markov property, for each X in $PC(T) : X \not\perp\!\!\!\perp T \mid \mathbf{Z}, \forall \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, T\}$. The second set is the *Markov boundary $MB(T)$* of a target variable T, defined as follows.

Definition 3.5 (Markov boundary of T [Pearl, 2009, Tsamardinos et al., 2003a])**.** *The* Markov boundary of T *or* MB(T) *is the smallest set of variables in $\mathbf{V} \setminus \{T\}$ conditioned on which all other variables are independent of T, that is $\forall Y \in \mathbf{V} \setminus MB(T), Y \neq T, T \perp\!\!\!\perp Y \mid MB(T)$.*

Given its definition and the d-separation criteria, in a faithful BN $MB(T)$ is composed of all parents, children, and *spouses* (i.e., parents of children) of T [Ma and Tourani, 2020], that are those variables $X \in \mathbf{V} \setminus \{T\}$ for which $\exists Y \in PC(T)$ such that $X \perp\!\!\!\perp T \mid \mathbf{Z}$ and $X \not\perp\!\!\!\perp T \mid \mathbf{Z} \cup \{Y\}$ for all $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, T\}$. $MB(T)$ is the minimal subset $\mathbf{S} \subseteq \mathbf{V}$ for which $p(T \mid \mathbf{S})$ is estimated accurately [Ma and Tourani, 2020, Tsamardinos et al., 2003a], therefore is the optimal solution for feature selection tasks.

3.2.2 Statistical testing for independence

The identification of $PC(T)$ and $MB(T)$ is based on the definitions of conditional dependence and independence between two variables X and Y. In practice, given a dataset, the conditional dependencies between variables are assessed using statistical hypothesis testing. Since a universal independence test does not exist [Shah and Peters, 2020], a commonly used approach is to compute the *Pearson's linear correlation coefficient* r between two vectors $\mathbf{x}$ and $\mathbf{y}$ of k elements:

$$r_{\mathbf{x},\mathbf{y}} = \frac{\sum_{i=1}^{k} x_i y_i - k\bar{x}\bar{y}}{(k-1)s_{\mathbf{x}}s_{\mathbf{y}}} \tag{3.1}$$

where x_i and y_i are the i-th element $\mathbf{x}$ and $\mathbf{y}$, respectively, $\bar{x}$ and $\bar{y}$ are the sample mean of $\mathbf{x}$ and $\mathbf{y}$, respectively, and $s_{\mathbf{x}}$ and $s_{\mathbf{y}}$ are the sample standard deviations.

The vectors $\mathbf{x}$ and $\mathbf{y}$ correspond to the observations of X and Y in the data, but their definition depends on whether the test is unconditional, or conditional on a set $\mathbf{Z}$ of variables. In the first case, $\mathbf{x}$ and $\mathbf{y}$ are the vectors of observations for variables X and Y, respectively. In the second case, $\mathbf{x}$ and $\mathbf{y}$ represent the residuals of the linear regression of the observations of the variables in $\mathbf{Z}$ on the ones in X (respectively, for $\mathbf{y}$, the ones in Y). For sake of simplicity, in what follows we will use $r_{X,Y,\mathbf{Z}}$ to denote the value of $r_{\mathbf{x},\mathbf{y}}$ when $\mathbf{x}$ and $\mathbf{y}$ are obtained conditioning on the set $\mathbf{Z}$, potentially with $\mathbf{Z} = \emptyset$ (i.e., for unconditional testing), as we just described.

Under the *null hypothesis* of independence between X and Y conditional on $\mathbf{Z}$ (including the case $\mathbf{Z} = \emptyset$), the expected value of $r_{X,Y,\mathbf{Z}}$ is 0, and the statistic $t = \frac{r_{X,Y,\mathbf{Z}}}{\sqrt{(1-r^2_{X,Y,\mathbf{Z}})/(k-2)}}$ follows a *Student's t* distribution with $k-2$ degrees of freedom. The dependence between X and Y is then usually assessed by computing (with *Student's t* distribution) the *p-value* for the test statistic t, that is the probability that the statistic is greater or equal than t under the null hypothesis of independence. In practice, algorithms for local causal discovery (e.g., [Tsamardinos et al., 2003b, Pena et al., 2007]) consider X and Y as independent (unconditionally or conditional on $\mathbf{Z}$) if the p-value is greater than a threshold δ (common values for δ are 0.01 or 0.05), while X and Y are considered as dependent otherwise.

3.2.3 Multiple hypotheses testing

As described above, in testing for the independence of two variables X and Y, they are considered dependent if the p-value of the corresponding test is below a threshold δ. It is easy to see that such procedure guarantees that if X and Y are independent, then the probability of a *false discovery*, that is *falsely rejecting* their independence, is at most δ. The situation is drastically different when a large number N of hypotheses are tested, as in the case of local causal discovery. In this case, if the same threshold δ is used for every test, the expected number of false discoveries can be as large as δN. Therefore, it is necessary to correct for multiple hypothesis testing, with the goal of providing guarantees on false discoveries. A commonly used guarantee is provided by the *Family-Wise Error Rate (FWER)*, which is the probability of having at least one false discovery among all the tests. A common approach to control the FWER is the so called *Bonferroni correction* [Bonferroni, 1936], which performs each test with a corrected threshold $\delta_{test} = \delta/N$ (a simple union bound shows that the resulting FWER is at most δ).

3.2.4 Supremum deviation and Rademacher averages

While Bonferroni correction does control the FWER, it conservatively assumes the worst-case scenario (of independence) between *all* null hypotheses. This often leads to a high number of *false negatives* (i.e. false null hypotheses that are not rejected). We now describe Rademacher averages [Bartlett and Mendelson, 2002, Koltchinskii and Panchenko, 2000], which allow to compute *data-dependent* confidence intervals for *all hypotheses simultaneously*, leading to improved tests for multiple hypotheses testing scenarios [Pellegrina et al., 2022]. Rademacher averages are a concept from statistical learning theory commonly used to measure the complexity of a family of functions and that, in general, also provide a way to probabilistically bound the deviation of the empirical means of the functions in the family from their expected values.

Let $\mathcal{F}$ be a family of functions from a domain $\mathcal{D}$ to $[a, b] \subset \mathbb{R}$ and let $\mathcal{S}$ be a sample of m i.i.d. observations from an unknown data generative distribution $\mathcal{W}$ over $\mathcal{D}$. We define the *empirical sample mean* $\hat{\mathbb{E}}_{\mathcal{S}}[f]$ of a function $f \in \mathcal{F}$ and its *expectation* $\mathbb{E}[f]$ as

$$\hat{\mathbb{E}}_{\mathcal{S}}[f] \doteq \frac{1}{m} \sum_{s_i \in \mathcal{S}} f(s_i) \text{ and } \mathbb{E}[f] \doteq \mathbb{E}_{\mathcal{W}} \left[\frac{1}{m} \sum_{s_i \in \mathcal{S}} f(s_i) \right]. \tag{3.2}$$

Note that $\mathbb{E}[f] = \mathbb{E}_{\mathcal{W}}[f]$, that is, the expected value of the empirical mean corresponds to the expectation according to distribution $\mathcal{W}$. A measure of the maximum deviation of the empirical mean from the (unknown) expectation for every function $f \in \mathcal{F}$ is given by the *supremum deviation* (SD) $D(\mathcal{F}, \mathcal{S})$ that is defined as

$$D(\mathcal{F}, \mathcal{S}) = \sup_{f \in \mathcal{F}} |\hat{\mathbb{E}}_{\mathcal{S}}[f] - \mathbb{E}[f]|. \tag{3.3}$$

Computing $D(\mathcal{F}, \mathcal{S})$ exactly is not possible given the unknown nature of $\mathcal{W}$, therefore bounds are commonly used. An important quantity to estimate tight bounds on the SD is the *Empirical Rademacher Average* (ERA) $\hat{R}(\mathcal{F}, \mathcal{S})$ of $\mathcal{F}$ on $\mathcal{S}$, defined as

$$\hat{R}(\mathcal{F}, \mathcal{S}) \doteq \mathbb{E}_{\sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(s_i) \right] \tag{3.4}$$

where $\boldsymbol{\sigma}$ is a vector of m i.i.d. Rademacher random variables, i.e. for which each element σ_i equals 1 or -1 with equal probability. ERA is an alternative of *VC dimension* for computing the expressiveness of a set $\mathcal{S}$ over class function $\mathcal{F}$, whose main advantage is that it provides tight *data-dependent* bounds while the *VC dimension* provides *distribution-free* bounds that are usually fairly conservative ([Mitzenmacher and Upfal, 2017], chap. 14).

Computing the exact value of $\hat{R}(\mathcal{F}, \mathcal{S})$ is often infeasible since the expectation is taken over 2^m elements. A common approach is then to estimate $\hat{R}(\mathcal{F}, \mathcal{S})$ using a Monte-Carlo approach with n samples of $\boldsymbol{\sigma}$. The n-samples Monte-Carlo Empirical Rademacher Average (n-MCERA) $\hat{R}^n_m(\mathcal{F}, \mathcal{S}, \boldsymbol{\sigma})$ is defined as

$$\hat{R}^n_m(\mathcal{F}, \mathcal{S}, \boldsymbol{\sigma}) \doteq \frac{1}{n} \sum_{j=1}^{n} \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{s_i \in S} \sigma_{j,i} f(s_i) \tag{3.5}$$

with $\boldsymbol{\sigma}$ being a $m \times n$ matrix of i.i.d. Rademacher random variables. n-MCERA is useful to derive probabilistic upper bounds to the SD, as the following.

Theorem 3.1 (Th. 3.1 of [Pellegrina et al., 2022]). *Let $\delta \in (0, 1)$. For ease of notation let*

$$\tilde{R} = \hat{R}^n_m(\mathcal{F}, \mathcal{S}, \boldsymbol{\sigma}) + 2z\sqrt{\frac{\ln \frac{4}{\delta}}{2nm}} \tag{3.6}$$

With a probability of at least $1 - \delta$ over the choice of $\mathcal{S}$ and $\boldsymbol{\sigma}$, it holds

$$D(\mathcal{F}, \mathcal{S}) \leq 2\tilde{R} + \frac{\sqrt{c(4m\tilde{R} + c \ln \frac{4}{\delta}) \ln \frac{4}{\delta}}}{m} + \frac{c \ln \frac{4}{\delta}}{m} + c\sqrt{\frac{\ln \frac{4}{\delta}}{2m}} \tag{3.7}$$

where $z = \max\{a, b\}$ and $c = b - a$.

Theorem 3.1 allows us to obtain confidence intervals around the empirical mean containing the expectation with probability at least $1 - \delta$ for all functions in $\mathcal{F}$ simultaneously.

3.3 Related work

Given a target variable T, the task of finding $MB(T)$ is strictly related to the discovery of $PC(T)$. A common approach for MB discovery consists of creating a candidate set of elements in $MB(T)$ by running a PC discovery algorithm twice (first on T, and then on all the elements reported as member of $PC(T)$) to find the elements at distance at most 2 from T, and then to eliminate false positives, which are those elements that are not parents, children, or spouses of T. Various algorithms follow this general scheme [Tsamardinos et al., 2003a, Aliferis et al., 2003, Pena et al., 2007, Aliferis et al., 2010], each one with a different variant that aims at minimizing the number of independence tests *actually* performed and their degrees of freedom to reduce the amount of data required. However, as described in Section 3.4.3, this does not decrease the number of statistical tests to be considered for MHT correction, since *a priori* all tests could *potentially* be performed. Among such algorithms, Pena et al. [Pena et al., 2007] proposed *PCMB* and proved its correctness under the assumption of all statistical tests being correct, that is, not returning *any* false positive or false negative. A different approach has been proposed for *IAMB* [Tsamardinos et al., 2003b] that incrementally grows a candidate set of elements in $MB(T)$ without searching for $PC(T)$, and then performs a false positive removal phase. Both *PCMB* and *IAMB* do not report false positives only under the assumption of not having any false positive and any false negative. Such assumptions are unrealistic in real-world scenarios due to noise in the data, finite sample sizes, and probabilistic guarantees of statistical tests, especially in multiple hypotheses scenarios. Our algorithms `RAveL-PC` and `RAveL-MB` do not require such assumptions to identify $PC(T)$ and $MB(T)$ with guarantees on the FWER.

To the best of our knowledge, the study of local causal discovery with guarantees on false discoveries has received scant attention. Tsamardinos et al. [Tsamardinos and Brown, 2008] introduced the problem of MHT in the context of local causal discovery, and proposed to use the Benjamini-Hochberg correction [Benjamini and Hochberg, 1995] to estimate the False Discovery Rate (FDR) of elements retrieved by $PC(T)$ discovery algorithms. However, such work does not provide an algorithm with guarantees for $MB(T)$. To the best of our knowledge, no method has focused on local causal discovery while bounding the FWER, which is extremely important in domains where false positives are critical or where follow-up studies require significant resources (e.g., biology and medicine).

Additional works focused on the more general task of BN inference. In [Armen and Tsamardinos, 2014], the authors extended the analysis of [Tsamardinos and Brown, 2008] from the local discovery task to the BN inference while [Li and Wang, 2009, Liu et al., 2012, Strobl et al., 2019] re-implemented the PC algorithm for BN structure discovery using the Benjamini-Yekutieli [Benjamini and Yekutieli, 2001] correction for the FDR, the former focusing on the skeleton retrieving and the latter deriving bounds on edge orientation as well. Our work instead focuses on *local* causal discovery tasks.

Rademacher averages have been successfully used to speed-up data mining tasks (e.g., pattern mining [Riondato and Upfal, 2015, Pellegrina et al., 2022, Santoro et al., 2020, Pellegrina and Vandin, 2021]). To the best of our knowledge, ours is the first work to introduce their use in (local) causal discovery.

3.4 Algorithms for local causal discoveries with FWER guarantees

In this Section we describe algorithms to obtain $PC(T)$ and $MB(T)$ with guarantees on the FWER. First, we discuss in Section 3.4.1 the requirements for previously proposed algorithms *PCMB* and *IAMB* to obtain guarantees on the FWER. In particular, we show that they require unrealistic assumptions that are not met in practice, as confirmed by our experimental evaluation (see Section 3.5). We then present in Section 3.4.2 our algorithms `RAveL-PC` and `RAveL-MB` for the computation of $PC(T)$ and $MB(T)$ with guarantees on the FWER. Finally, in Section 3.4.3 we describe how our algorithms perform effective independence testing by combining a novel test statistic with Rademacher averages.

3.4.1 Analysis and limitations of *PCMB* and *IAMB*

The algorithms presented in Section 3.3 are correct under the assumption that the independence tests result in no false positive *and* no false negative [Pena et al., 2007, Tsamardinos et al., 2003b]. In this Section we determine milder sufficient conditions that allow *GetPC* [Pena et al., 2007] to control the FWER for the PC discovery task, and *PCMB* [Pena et al., 2007] and *IAMB* [Tsamardinos et al., 2003b] to control the FWER for the MB discovery task. In all cases, a first requirement is that the independence tests performed by the algorithms must account for multiple hypotheses testing in order to bound the FWER. However, we also show that an additional requirement on the ability to identify dependent variables (i.e., on the *power* of the tests) is needed. In particular, we refer to the situation where *all tests* on dependent variables correctly reject the null hypothesis of independence as the *infinite power assumption.* In some cases, we consider the infinite power assumption only for independence tests between pairs of variables that are directly connected in the underlying DAG. We refer to such situation as the *local infinite power assumption.*

PCMB

Both *GetPC* and *PCMB* make use of a subroutine called *GetPCD* [Pena et al., 2007] whose aim is to return a set with parents, children, and (eventually) other descendants of T by applying a sequence of independence tests. In this section we will study under which conditions each method does not output any false positive, and how each subroutine result may affect the output of other algorithms.

We first start by studying under which conditions *GetPCD* [Pena et al., 2007] returns a false positive in output.

Theorem 3.2 (Study of false positives in *GetPCD*)**.** *An element $X \notin PCD(T)$ is returned from GetPCD only if not all the parents of T are detected or the null hypotheses of some independence tests is wrongly rejected.*

Proof. Let us recall that an element $X \in \mathbf{V}$ returned by $GetPCD(T, \mathbf{V})$ is a false negative if and only if $X \notin Parents(T) \cup Descendants(T)$.

It is easy see that an element is returned by *GetPCD* only if it is not removed at lines 9 and 19 of Algorithm 1, which means that the null hypothesis of tests at lines 8 and 18 gets always rejected[1]. The independence test determines the dependence of T from X only if conditioning on $\mathbf{Z} = sep[X]$ there is an open path between X and T (i.e. $T \not\perp\!\!\!\perp X \mid \mathbf{Z}$), or if the null hypothesis gets wrongly rejected.

Let us now study the two topological cases of X being disconnected to T and of X being connected to T.

Disconnected case. Let X be disconnected from T. Since there are no paths from X to T (therefore no open paths from X to T), X may be returned by *GetPCD* only if independence tests at lines 8 and 18 is wrongly rejected.

Connected case. Let $X \notin PCD(T)$ be connected to T. X is returned in output only if in any iteration of the cycle the null hypothesis on tests at lines 8 and 18 is wrongly rejected or there is an open path conditioning on $Sep[X]$.

By assuming of not having wrong rejections of the null hypotheses, $\mathbf{Z} = Parents(T)$ d-separates X and T by definition of parents since X is not a descendant of T. This implies that if some parent of T is undetected, then it may not be possible to d-separate X from T. □

We can then determine under which conditions *GetPCD* is able to control the FWER.

Theorem 3.3. *$GetPCD(T, \mathbf{V})$ outputs a set of elements in $PCD(T)$ with FWER lower than δ if the FWER of every independence test performed by GetPCD is below δ and the local infinite power assumption holds.*

Proof. By analyzing *GetPCD* structure as in Th. 3.2, an element is returned only if both independence tests at lines 8 and 18 of Algorithm 1 reject the null hypothesis. Therefore the algorithm outputs a false positive if under infinite power assumption for elements directly connected at least one independence test returns a false positive. Let us define the events E =*"GetPCD(T, $\mathbf{V}$) outputs a false positive"* and E_i = *"the i-th independence test returns a false positive"*. We then have

$$FWER = P(E) \leq P\left(\cup_i E_i\right) \leq \delta$$

by definition of FWER. □

We now provide sufficient conditions for bounding the FWER of the elements returned by *GetPC* [Pena et al., 2007].

[1] The "if" clause does not hold since an element may be added and then subsequently removed leading to the end of the repeat cycle because *PCD* did not change, but there still are elements in *canPCD* i.e. unremoved elements.

Algorithm 1: *GetPCD*$(T, \mathbf{V})$ [Pena et al., 2007]

```
Input: target variable $T$, set $\mathbf{V}$ of variables
Output: $PCD(T) = \{X \in \mathbf{V} \mid X \in PC(T) \vee X \in descendants(T)\}$
$PCD \leftarrow \emptyset$;
$CanPCD \leftarrow \mathbf{V} \setminus \{T\}$;
repeat
    /* Remove false positives from CanPCD */ ;
    foreach $X \in CanPCD$ do
        $Sep[X] \leftarrow \arg\min_{\mathbf{Z} \subseteq PCD} dep(T, X \mid \mathbf{Z})$;
    foreach $X \in CanPCD$ do
        if $T \perp\!\!\!\perp X \mid Sep[X]$ then
            $CanPCD \leftarrow CanPCD \setminus \{X\}$;
    /* Add the best candidate to PCD */ ;
    $Y \leftarrow \arg\max_{X \in CanPCD} dep(T, X \mid Sep[X])$;
    $PCD \leftarrow PCD \cup \{Y\}$;
    $CanPCD \leftarrow CanPCD \setminus \{Y\}$;
    /* Remove false positives from PCD */;
    foreach $X \in PCD$ do
        $Sep[X] \leftarrow \arg\min_{\mathbf{Z} \subseteq PCD \setminus \{X\}} dep(T, X \mid \mathbf{Z})$;
    foreach $X \in PCD$ do
        if $T \perp\!\!\!\perp X \mid Sep[X]$ then
            $PCD \leftarrow PCD \setminus \{X\}$;
until PCD does not change;
return PCD;
```

Theorem 3.4. *GetPC*$(T, \mathbf{V})$ *outputs a set of elements in* $PC(T)$ *with FWER* $\leq \delta$ *if the independence tests performed by GetPC have FWER* $\leq \delta$ and *the local infinite power assumption holds.*

Proof. *GetPC* outputs a false positive only if at least one call to *GetPCD* at lines 2-3 of Algorithm 2 outputs a false positive and, under the infinite power assumption while testing the independence of elements directly connected, this happens only if at least one independence test outputs a false positive. Let us define the events E =“*GetPC*$(T, \mathbf{V})$ *outputs a false positive*” and E_i =“*the i-th independence test returns a false positive*”. We then have

$$FWER = P(E) \leq P\left(\cup_i E_i\right) \leq \delta$$

by definition of FWER. □

The following proves that similar requirements are needed for *PCMB* [Pena et al., 2007] to have guarantees on the FWER.

Theorem 3.5. *PCMB*$(T, \mathbf{V})$ *outputs a set of elements in* $MB(T)$ *with FWER* $\leq \delta$ *if the independence tests performed by PCMB have FWER* $\leq \delta$ *and the infinite power assumption holds.*

Algorithm 2: *GetPC*$(T, \mathbf{V})$ [Pena et al., 2007]

```
Input: target variable $T$, set $\mathbf{V}$ of variables
Output: $PC(T)$
$PC \leftarrow \emptyset$;
foreach $X \in GetPCD(T, \mathbf{V})$ do
    if $T \in GetPCD(X, \mathbf{V})$ then
        $PC \leftarrow PC \cup \{X\}$
return $PC$;
```

Proof. *PCMB* outputs a false positive only if there is a false positive in any independence test performed by *GetPC* calls at lines 2 and 6 of Algorithm 3, or if tests at lines 8 and 9 return a false negative or a false positive, respectively. Given the infinite power assumption and Theorem 3.4, *PCMB* outputs a false positive only if at least one independence test outputs a false positive and by defining the events E =*"PCMB*$(T, \mathbf{V})$ *outputs a false positive"* and E_i =*"the i-th independence test returns a false positive"* we have

$$FWER = P(E) \leq P\left(\cup_i E_i\right) \leq \delta$$

by definition of FWER. □

Algorithm 3: *PCMB*$(T, \mathbf{V})$ [Pena et al., 2007]

```
   Input: target variable $T$, set $\mathbf{V}$ of variables
   Output: $MB(T)$
1  /* Add true positives to MB */ ;
2  $PC \leftarrow GetPC(T, \mathbf{V})$;
3  $MB \leftarrow PC$;
4  /* Add more true positives to MB */ ;
5  foreach $Y \in PC$ do
6      foreach $X \in GetPC(Y, \mathbf{V})$ do
7          if $X \notin PC$ then
8              find $\mathbf{Z}$ such that $T \perp\!\!\!\perp X \mid \mathbf{Z}$ and $T, X \notin \mathbf{Z}$ ;
9              if $T \not\perp\!\!\!\perp X \mid \mathbf{Z} \cup Y$ then
10                 $MB \leftarrow MB \cup \{X\}$;
11 return $MB$;
```

IAMB

The following result proves analogous requirements of Section 3.4.1 for *IAMB*.

Theorem 3.6. *IAMB*$(T, \mathbf{V})$ *outputs a set of elements in* $MB(T)$ *with FWER* $\leq \delta$ *if the independence tests performed by IAMB have FWER* $\leq \delta$ *and the infinite power assumption holds.*

Proof. *IAMB* outputs a false positive only if an element $X \notin MB(T)$ gets added to MB at lines 5-6, and it does not get removed from MB at lines 10-11 of Algorithm 4. Under the infinite power assumption, all elements in $PC(T)$ get added at lines 5-6 by definition of PC, therefore X gets returned by *IAMB* only if independence tests at lines 10-11 output a false positive. Then, by defining the events $E =$"*$GetPC(T, \mathbf{V})$ outputs a false positive*" and $E_i =$"*the i-th independence test returns a false positive*", we have

$$FWER = P(E) \leq P\left(\cup_i E_i\right) \leq \delta$$

by definition of FWER. □

Algorithm 4: *$IAMB(T, \mathbf{V})$* [Tsamardinos et al., 2003b]

```
    Input: target variable $T$, set $\mathbf{V}$ of variables
    Output: $MB(T)$
 1  /* Add true positives to MB */ ;
 2  $MB \leftarrow \emptyset$;
 3  repeat
 4      $Y \leftarrow \arg\max_{X \in \mathbf{V} \setminus MB \setminus \{T\}} dep(T, X, MB)$;
 5      if $T \not\perp\!\!\!\perp Y \mid MB$ then
 6          $MB \leftarrow MB \cup \{Y\}$ ;
 7  until MB does not change;
 8  /* Remove false positives from MB */ ;
 9  foreach $X \in MB$ do
10      if $T \perp\!\!\!\perp X \mid MB \setminus \{X\}$ then
11          $MB \leftarrow MB \setminus \{X\}$ ;
12  return MB;
```

Relaxation of the infinite power assumption

Note that the results above require the (local) infinite power assumption to hold in order to have guarantees on the FWER of the output of previously proposed algorithms. In fact, if the (local) infinite power assumption does not hold, such algorithms may output false positives even when *all* independence tests do not return a single false positive. We now present three such examples by considering the subgraph of Figure 3.1 in Section 3.5 between variables $\mathbf{V} = \{C_1, A_2, B_2, C_2\}$ with edges $\mathbf{E} = \{C_1 \rightarrow A_2, C_1 \rightarrow B_2, A_2 \rightarrow C_2, B_2 \rightarrow C_2\}$. Moreover, our experimental evaluation in Section 3.5 shows that these situations do happen in practice.

Scenario 1: The infinite power assumption holds only for directly connected elements. Let us study the subgraph previously described under only local infinite power assumption. Let us suppose to run $PCMB(C_1, \mathbf{V})$ and that the call at line 2 correctly returned $GetPC(C_1, \mathbf{V}) = \{A_2, B_2\}$. Let us further suppose that $GetPC(A_2, \mathbf{V}) = \{C_1, C_2\}$ and that a false negative arises when testing the unconditional dependence between C_1 and C_2, leading to the choice of $\mathbf{Z} = \emptyset$ on line 8. If the conditional independence test at line 9 correctly assesses the conditional dependence of C_1 and C_2 conditioning on A_2, then C_2 is wrongly considered a spouse of C_1.

Scenario 2: No infinite power assumption. Consider as an example the calculus of $GetPC(C_2, \mathbf{V})$ in the subgraph previously described. Let us suppose that a false negative occurs when testing the unconditional independencies between C_2 and A_2 and between C_1 and A_2. Let us further suppose $\mathbf{Z} = \{A_2, B_2\}$ to be the only set for which the null hypothesis of independence between C_1 and C_2 is not rejected. Then $GetPCD(C_2, \mathbf{V})$ will contain C_1 (because the independence conditioning on $\mathbf{Z} = \{A_2, B_2\}$ is never tested), and similarly $GetPCD(C_1, \mathbf{V})$ will contain C_2 leading C_1 to be returned by $GetPC(C_2, \mathbf{V})$.

Scenario 3: No infinite power assumption and $GetPC$ **does not return false positives.** Let us finally consider a situation in which the infinite power assumption does not hold and $GetPC$ does not return any false positive, as this may be the case of a modification of the algorithms proposed by [Pena et al., 2007] using Bonferroni correction. Let us suppose $GetPC(C_1, \mathbf{V}) = \{A_2\}$, and $GetPC(C_2, \mathbf{V}) = \{A_2\}$. Let us suppose line 8 to return $\mathbf{Z} = \emptyset$, and the conditional independence test at line 9 to correctly assess the conditional dependence of C_1 and C_2 conditioning on A_2. Under these assumptions, C_2 is wrongly considered a spouse of C_1. Note that this scenario differs from the first because the local infinite power assumption does not hold, leading to a partial discovery of the variables in $PC(C_1)$ whose elements are not enough to d-separate C_1 and C_2.

3.4.2 Algorithms RAveL-PC and RAveL-MB

As shown in Section 3.4.1, controlling the FWER of every independence test is not sufficient for bounding the FWER of the variables returned by current state-of-the-art algorithms for PC and MB discovery. In addition, infinite statistical power is a strong assumption which is impossible to test and ensure in real-world scenarios. Motivated by these observations, we developed RAveL-PC and RAveL-MB, two algorithms for the discovery of elements in PC and MB, respectively, that control the FWER of their outputs without making any assumption on statistical power.

RAveL-MB follows the same overall approach used by previously proposed algorithms (e.g., *PCMB*, see Section 3.3): it first identifies elements in $PC(T)$ and adds them to $MB(T)$, and then tests the spouse condition on elements at distance 2 from T, that are variables $Y \in PC(X)$ with $X \in PC(T)$ and $Y \notin PC(T)$. The pseudocode of RAveL-MB is shown in Algorithm 5. RAveL-MB inizializes MB to the output of the function RAveL-PC(T,$\mathbf{V}$,δ) (line 1), which returns a subset of $PC(T)$. For each element $X \in MB$ (line 2), RAveL-MB computes RAveL-PC(X,$\mathbf{V}$,δ) and, for every returned element Y that is not already in MB (line 3), an independence test of T on Y conditioning on $\mathbf{V} \setminus \{Y, T\}$ using function test_indep(T,Y,$\mathbf{V} \setminus \{Y, T\}$,$\delta$) is performed to test whether Y is a spouse of T with respect to X (line 4). If such test determines the conditional dependence between T and Y, then Y is added to MB (line 5). Finally, after analyzing all variables originally in MB, `RAveL-MB` outputs the set of elements in the Markov Boundary (line 6).

Note that the spouse condition is tested by conditioning only on the set $\mathbf{V} \setminus \{Y, T\}$. This is sufficient, since it is a set conditioned on which T and Y are d-connected if and only if Y is directly connected or is a spouse of T. In fact, if Y does not belong to any of these elements, then Y is connected to T through paths that contain chains or forks whose middle element is in $\mathbf{V} \setminus \{Y, T\}$. That is, Y is connected to T only through d-blocked paths.

Algorithm 5: `RAveL-MB`$(T, \mathbf{V}, \delta)$

```
  Input: target variable $T$, set $\mathbf{V}$ of variables, threshold $\delta \in (0, 1]$
  Output: A subset of $MB(T)$ with FWER lower than $\delta$.
1 $MB \leftarrow$ RAveL-PC$(T, \mathbf{V}, \delta)$ ;
2 foreach $X \in MB$ do
3     foreach $Y \in$ RAveL-PC$(X, \mathbf{V}, \delta)$ and $Y \notin MB$ do
4         if not test_indep$(T, Y, \mathbf{V} \setminus \{Y, T\}, \delta)$ then
5             $MB \leftarrow MB \cup \{Y\}$;
6 return $MB$;
```

`RAveL-MB` uses algorithm `RAveL-PC`$(X, \mathbf{V}, \delta)$ (shown in Algorithm 6) for the discovery of variables of a set $\mathbf{V}$ that are in $PC(X)$. The parameter δ controls the overall FWER of the procedure. `RAveL-PC`$(X, \mathbf{V}, \delta)$ identifies $PC(X)$ by using the definition of parent-children set, that is, $Y \in PC(X)$ gets returned if only if all independence tests between X and Y reject the null hypothesis.

Algorithm 6: `RAveL-PC`$(T, \mathbf{V}, \delta)$

```
  Input: target variable $T$, set $\mathbf{V}$ of variables, threshold $\delta \in (0, 1]$
  Output: A subset of $PC(T)$ with FWER lower than $\delta$.
1 $PC \leftarrow \mathbf{V} \setminus \{T\}$;
2 foreach $X \in \mathbf{V} \setminus \{T\}$ do
3     foreach $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, T\}$ do
4         if test_indep$(T, X, \mathbf{Z}, \delta)$ then
5             $PC \leftarrow PC \setminus \{X\}$;
6 return $PC$;
```

Both algorithms `RAveL-MB` and `RAveL-PC` employ a function, denoted as `test_indep`$(X, Y, \mathbf{Z}, \delta)$, that performs the independence test between $X, Y \in \mathbf{V}$ conditioning on $\mathbf{Z} \subseteq \mathbf{V}$ while controlling the FWER of *all testable hypotheses* with threshold δ, and returns `true` only if the null hypothesis gets rejected. Practical details on our implementation of `test_indep`$(X, Y, \mathbf{Z}, \delta)$ are provided in Section 3.4.3.

The following results prove that `RAveL-PC` and `RAveL-MB` control the FWER of PC and MB, respectively.

Theorem 3.7. `RAveL-PC(`T`,`$\mathbf{V}$`,`δ`)` *outputs a set of elements in* $PC(T)$ *with FWER* $\leq \delta$*.*

Proof. Note that the number of false positives of `RAveL-PC(`T`,`$\mathbf{V}$`,`δ`)` is greater than 0 if and only if there is at least one variable X of $\mathbf{V} \setminus \{T\}$ that is not in $PC(T)$ and is in the set PC reported by `RAveL-PC(`T`,`$\mathbf{V}$`,`δ`)`. A variable X is returned in PC if and only if all independence tests between T and X (conditioning on the various sets $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, T\}$) reject the null hypothesis. Therefore `RAveL-PC(`T`,`$\mathbf{V}$`,`δ`)` reports a false positive only if at least one independence test returns a false positive, which happens with probability at most δ by definition of `test_indep(`T`,`X`,`$\mathbf{Z}$`,`δ`)`. □

Theorem 3.8. `RAveL-MB(`T`,`$\mathbf{V}$`,`δ`)` *outputs a set of elements in* $MB(T)$ *with FWER* $\leq \delta$*.*

Proof. The set of `RAveL-MB(`T`,`$\mathbf{V}$`,`δ`)` output elements is the union of the set O_1 of variables returned by `RAveL-PC(`T`,`$\mathbf{V}$`,`δ`)`, and the set O_2 of candidate spouses Y for which `test_indep(`T`,`Y`,`$\mathbf{V}\setminus\{Y,T\}$`,`$\delta$`)` rejects the null hypothesis. Then, a necessary condition to return a false positive is that at least one between sets O_1 and O_2 contains a false positive. The last event happens if and only if all calls to `test_indep(`T`,`X`,`$\mathbf{Z}$`)` returns at least a false positive, which happens with probability at most δ. □

The choice of $\mathbf{V} \setminus \{Y, T\}$ as conditioning set for testing the spouse condition is a consequence of `RAveL-PC` returning, with probability at least $1 - \delta$, a subset of $PC(T)$, and of any superset of $PC(T)$ allowing the discovery of spouses by `RAveL-MB`. We note that prior knowledge may be incorporated in the algorithm, if available, by conditioning on smaller set of variables, therefore increasing the precision of independence tests.

3.4.3 Rademacher averages for independence testing

Note that our algorithms `RAveL-PC` and `RAveL-MB` both rely on the availability of function `test_indep(`X`,`Y`,`$\mathbf{Z}$`,`δ`)`, which assesses the independence between $X, Y \in \mathbf{V}$ conditioning on $\mathbf{Z} \subseteq \mathbf{V}$ and returns `true` only if the null hypothesis gets rejected, while controlling the FWER of *all testable hypotheses* below a threshold δ.

The naïve implementation of `test_indep(`X`,`Y`,`$\mathbf{Z}$`,`δ`)` would be to perform a standard statistical test (see Section 3.2.2) and use Bonferroni correction (see Section 3.2.3) to correct for multiple hypothesis testing. In particular, this requires to use a modified threshold δ/N for every hypothesis, where N is the maximum number of hypotheses that could be tested. Therefore, N is the maximum number of conditional independencies[2] between the variables in $\mathbf{V}$, that is $N = \mathbf{V}(\mathbf{V} - 1)2^{\mathbf{V}-3}$. Note that the value of N grows

[2] N counts, in fact, the total number of possible conditional independencies between any couple of variables by considering the symmetry property of independence tests, that is testing the (conditional) independence of X from Y is equivalent to testing the one of Y from X.

exponentially with $\mathbf{V}$, leading to a Bonferroni correction which is very conservative and, therefore, to a high number of false negatives (independence tests between dependent variables for which the null hypothesis does not get rejected).

The high number of tests is not a feature of our algorithms only, but it is, in essence, shared by other widely used algorithms such as IAMB and PCMB (see Section 3.3). In fact, for both algorithms, the potential number of independence tests they perform can be as high as $N = \mathbf{V}(\mathbf{V} - 1)2^{\mathbf{V}-3}$, even if a smaller number of tests may be considered in practice, depending on the output of the tests in previous steps, and a proper MHT correction depends on the maximum number of tests that could be performed.

Our solution to make our algorithms `RAveL-PC` and `RAveL-MB` practical is to implement `test_indep(`X`,`Y`,`$\mathbf{Z}$`,`δ`)` exploiting Rademacher averages to obtain data-dependent bounds and confidence intervals. The key idea is to estimate confidence intervals around the empirical test statistics so that they contain the true values *simultaneously* with probability $1 - \delta$. In this way, testing for independence corresponds to check whether a confidence interval contains the expected value of the test statistic under the null hypothesis of independence.

To implement the idea described above, we express Eqn. 3.1 as an additive function on the samples as follows. First, let us assume the observations $\mathbf{x}$ of each variable X to follow a probability distribution $\mathcal{X}$ with mean $\mu_{\mathcal{X}}$ and whose absolute value is bounded by $\max_{\mathcal{X}}$. Let us also assume that all variables have been centered around 0 (i.e. by subtracting $\mu_{\mathcal{X}}$) and then normalized by dividing for $\max_{\mathcal{X}} - \mu_{\mathcal{X}}$ (i.e. they take values in $[-1, 1]$).

Let $s_1, s_2, \ldots, s_k$ be the samples in the dataset $\mathcal{S} = \{s_1, s_2, \ldots, s_k\}$, where each s_i is a collection of observations $s_i = \{v_1^i, v_2^i, \ldots\}$ of variables in $\mathbf{V}$, where v_j^i is the observation of the j-th variable $V_j \in \mathbf{V}$ in sample s_i. Given two variables $X, Y \in \mathbf{V}$, and a set of variables $\mathbf{Z} \subset \mathbf{V}$, we define the following function $\tilde{r}_{X,Y,\mathbf{Z}}(s_i)$ on a sample s_i as

$$\tilde{r}_{X,Y,\mathbf{Z}}(s_i) = k\frac{x_i y_i}{k-1}, \tag{3.8}$$

where the conditioning set $\mathbf{Z}$ does not explicitly appear in the term $k\frac{x_i y_i}{k-1}$ but it is used in the definition of the values in $\mathbf{x}$ and $\mathbf{y}$ as in Section 3.2.2.

We then define the following modified version $\tilde{r}$ of Pearson's r coefficient, which we refer to as the *modified r statistic* (or *ModR*), where $s_{\mathbf{x}}$ is replaced by $\max_{\mathcal{X}} - \mu_{\mathcal{X}}$ (similarly for $s_{\mathbf{y}}$):

$$\tilde{r}_{X,Y,\mathbf{Z}} = \frac{1}{k}\sum_{i=1}^{k} \tilde{r}_{X,Y,\mathbf{Z}}(s_i). \tag{3.9}$$

By considering the family $\mathcal{F}$ of functions defined by $\tilde{r}_{X,Y,\mathbf{Z}}$ for each pair X, Y of variables and each set $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$, we have that the n-MCERA

(Eqn. 3.5) is

$$\hat{R}^n_k(\mathcal{F}, \mathcal{S}, \sigma) \doteq \frac{1}{n} \sum_{j=1}^{n} \sup_{\tilde{r}_{X,Y,\mathbf{Z}} \in \mathcal{F}} \frac{1}{k} \sum_{i=1}^{k} \sigma_{j,i} \tilde{r}_{X,Y,\mathbf{Z}}(s_i). \tag{3.10}$$

After the n-MCERA has been computed as above, we compute a bound $\mathcal{B}$ to the supremum deviation $D(\mathcal{F}, \mathcal{S})$ according to Theorem 3.1, which allows us to obtain confidence intervals around the empirical $\tilde{r}_{X,Y,\mathbf{Z}}$ as

$$CI_{X,Y,\mathbf{Z}} = \left[\tilde{r}_{X,Y,\mathbf{Z}} - \mathcal{B}, \tilde{r}_{X,Y,\mathbf{Z}} + \mathcal{B}\right] \tag{3.11}$$

with the guarantee that, *simultaneously* for all $\tilde{r}_{X,Y,\mathbf{Z}} \in \mathcal{F}$, $CI_{X,Y,\mathbf{Z}}$ contains the expected value of $\tilde{r}_{X,Y,\mathbf{Z}}$ with probability at least $1 - \delta$. Then, for a pair X, Y of variables and a set $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$, we reject the null hypothesis of independence between X, Y conditioning on $\mathbf{Z}$ (i.e., `test_indep(`X`,`Y`,`$\mathbf{Z}$`,`δ`)` returns `true`) if $CI_{X,Y,\mathbf{Z}}$ does not contain the value 0. In practice, we replace the unknown quantities $\mu_{\mathcal{X}}$ and $\max_{\mathcal{X}}$ with their empirical estimates, that is, we replace $\mu_{\mathcal{X}}$ with the empirical sample mean $\bar{\mathbf{x}}$ and $\max_{\mathcal{X}}$ with $\max_{\mathbf{x}}$.

We finally propose another test statistic on a sample s_i, which we refer to as the *r-centered statistic* (or $\tilde{r}^c$), defined as

$$\tilde{r}^c{}_{X,Y,\mathbf{Z}}(s_i) = \frac{x_i y_i}{(max\{x_i, y_i\})^2} \tag{3.12}$$

where $\mathbf{x}$ and $\mathbf{y}$ are defined as previously (see Section 3.2.2). The same independence testing procedure described for $\tilde{r}_{X,Y,\mathbf{Z}}$ applies for the empirical average of $\tilde{r}^c{}_{X,Y,\mathbf{Z}} = \frac{1}{k} \sum_{i=1}^{k} \tilde{r}^c{}_{X,Y,\mathbf{Z}}(s_i)$, since its expectation is zero under independence assumption and data centered around zero as follows.

Theorem 3.9. *Let $\mathcal{W}$ be the joint distribution of the variables X, Y, and $\mathbf{Z}$. If X and Y are independent, then $\mathbb{E}_{\mathcal{W}}[\tilde{r}^c{}_{X,Y,\mathbf{Z}}] = 0$.*

Proof. We have that

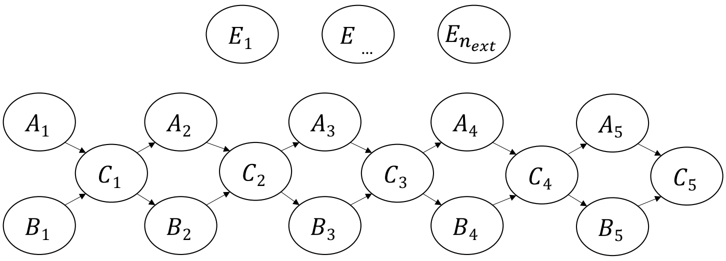
$$\mathbb{E}_{\mathcal{W}}\left[\tilde{r}^c{}_{X,Y,\mathbf{Z}}\right] = \mathbb{E}_{\mathcal{W}}\left[\frac{1}{k} \sum_{i=1}^{k} \frac{x_i y_i}{(max\{x_i, y_i\})^2}\right]$$

which is proportional to $\mathbb{E}_{\mathcal{W}}[\hat{\mathbb{E}}_{\mathcal{S}}[XY]]$ (see Section 3.2.4 for definitions of $\mathbb{E}_{\mathcal{W}}$ and $\hat{\mathbb{E}}_{\mathcal{S}}$). Under the independence assumption, we have $\mathbb{E}_{\mathcal{W}}[\hat{\mathbb{E}}_{\mathcal{S}}[XY]] = \mathbb{E}_{\mathcal{W}}[\hat{\mathbb{E}}_{\mathcal{S}}[X]] \times \mathbb{E}_{\mathcal{W}}[\hat{\mathbb{E}}_{\mathcal{S}}[Y]]$, and the result follows since $\mathbb{E}_{\mathcal{W}}[\hat{\mathbb{E}}_{\mathcal{S}}[X]] = \mathbb{E}_{\mathcal{X}}[\hat{\mathbb{E}}_{\mathcal{S}}[X]] = 0$. □

3.5 Experimental evaluation

This section describes the experimental evaluation performed to empirically assess our algorithms. In Section 3.5.1 we compare `RAveL-PC` and `RAveL-MB` performances with other state-of-the-art methods on synthetic data. Section 3.5.2 present the analysis on two real world datasets (see the Appendix for details). We implemented[3] RAveL-PC, RAveL-MB, and the other algorithms considered in this section in Python 3. On each run we assumed no prior knowledge of the data distributions values for each variable X.

3.5.1 Synthetic data



$$A_1 \sim N(0,1)$$
$$B_1 \sim N(0,1)$$
$$C_1 \sim 2A_1 + 3B_1 + N(0,\sigma^2)$$
$$\forall i \in [2,3,4,5]$$
$$A_i \sim 3C_{i-1} + N(0,\sigma^2)$$
$$B_i \sim -2C_{i-1} + N(0,\sigma^2)$$
$$C_i \sim 2A_i + 3B_i + N(0,\sigma^2)$$
$$\forall j \in [1,\ldots,n_{ext}]$$
$$E_j \sim N(0,1)$$

Figure 3.1: Bayesian Network used for synthetic data generation, parametrized by two values σ^2 and n_{ext}. After drawing all the observations $\mathbf{x}$ for a particular variable X, $\mathbf{x}$ is normalized such that $mean(\mathbf{x}) = 0$ and $var(\mathbf{x}) = 1$, then the values for the descendants of X are sampled.

We used synthetic data to evaluate RAveL-PC and RAveL-MB against state-of-the-art algorithms for the task of PC and MB discovery, respectively. In this scenario, each variable is a linear combination of its parents values plus Gaussian noise. The related structural model (shown in Figure 3.1) is composed of 15 connected variables and n_{ext} external variables, and it is specified by two parameters: σ^2 which controls the amount of noise in the estimations, and n_{ext} which sets the number of external variables.

In these experiments we set the rejection threshold $\delta = 0.05$, which is a common value in literature, and we run each algorithm on increasing size datasets. We repeated each trial 100 times and used $n = 1000$ for the n-MCERA. For each dataset, we considered all variables as target variable T in turn and run the algorithms for each choice of T. (Note that the number N of potential hypotheses tested is still the same as defined in Section 3.4.3.). Lastly, we limited our algorithms to consider only conditioning sets $\mathbf{Z}$ of at most 2 variables (except for the independence test at line 6 of RAveL-MB) for avoiding the analysis of all the exponential number N of hypotheses. We chose such value since each variable X is d-separated by each $Y \notin PC(X)$ by conditioning on a $\mathbf{Z}$ of size at most 2, and by running the algorithms on synthetic data allowing higher maximum sizes, we observed no differences in results w.r.t. the ones we are presenting.

In the first experiment, we compared different local causal discovery algorithms on the BN obtained setting $\sigma^2 = 1$ and $n_{ext} = 15$. For the PC discovery task, we compared two versions of *GetPC* [Pena et al., 2007], the original one (without any correction for MHT) and one adaptation that uses Bonferroni correction, with three versions of RAveL-PC: one that uses the modi-

[3] Code available at https://github.com/VandinLab/RAveL .

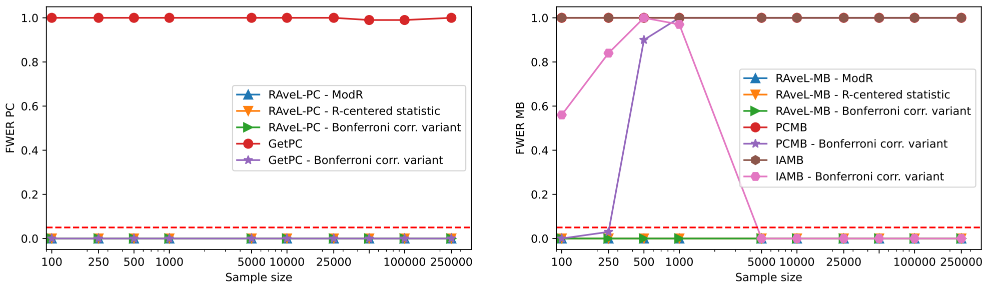


Figure 3.2: Empirical FWER of various PC discovery (a) and MB discovery (b) algorithms on synthetic data for different sample sizes. FWER is the fraction of 100 trials in which at least one false positive is reported. The dashed line represents the bound $\delta = 0.05$ to the FWER used in the experiments.

fied r statistic (or *ModR*) defined in Eqn. 3.9, another that exploits $\tilde{r}^c$, and a variant of RAveL-PC that uses Bonferroni correction instead of Rademacher averages for MHT. Figure 3.2(a) shows the estimated FWER of each method (that is, the fraction of trials in which at least a false positive is reported). The results confirm our analysis in Section 3.4.2, and we observe that, for the specific BN we consider, the adaptation of *GetPC* that uses Bonferroni correction has FWER below the threshold, even if this is not guaranteed from our theoretical analysis.

For the MB discovery task, we compared two versions of *PCMB* [Pena et al., 2007] and of *IAMB* [Tsamardinos et al., 2003b], the original ones (without any correction for MHT) and two adaptations that use Bonferroni correction, with three versions of RAveL-MB: one that uses the modified r statistic defined in Eqn. 3.9, another that exploits $\tilde{r}^c$, and a variant of RAveL-MB that uses Bonferroni correction instead of Rademacher averages for MHT. Figure 3.2(b) shows the FWER of each method. The results confirms RAveL-MB (with both statistics) and its variant to be the only algorithms with guarantees on the FWER at any sample size, that is without infinite power assumption. Moreover, note that *PCMB* reports false positives with high probability even if its PC discovery method *GetPC* does not. This is due to elements at distance 2 from T that are correctly identified as candidate spouses, but for which the spouse condition used by *PCMB* results in a false positive due to false negatives in $PC(T)$, as described in Section 12 (scenario 3).

We then assessed the fraction of false negatives for our algorithms, which are the only ones with guarantees on the FWER, on datasets with sample sizes up to 250000 elements by repeating each trial 100 times. Figure 3.3 summarizes (with solid lines) these results on a scenario with $\sigma^2 = 1$ (in Figure 3.3(a,b)) and another with $\sigma^2 = 5$ (in Figure 3.3(c,d)). For each setting, we run the algorithms by considering a different number of variables ($n_{ext} = 0$ and $n_{ext} = 15$), and we highlighted the difference in performances between the two cases. The results show how the approaches based on Rademacher averages do not suffer from the addition of external variables (i.e. their FN% are equivalent), as opposed to the versions of RAveL-PC and

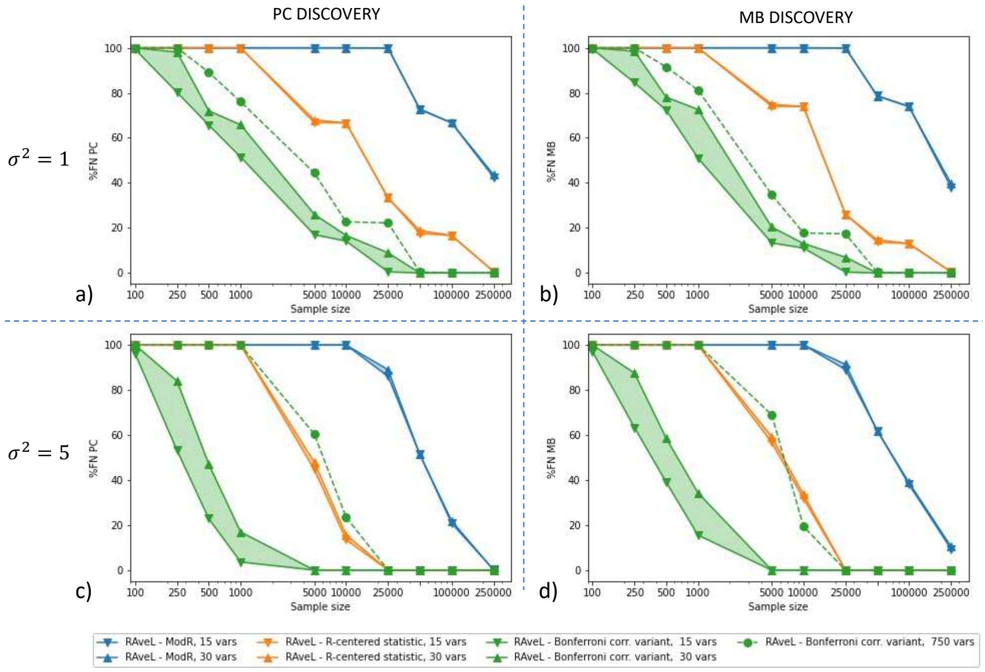


Figure 3.3: Empirical FN% of RAveL-PC (a,c) and RAveL-MB (b,d) on synthetic data for different sample sizes in two data generative scenarios. We sampled data from Figure 3.1 in two scenarios with different noise level: $\sigma^2 = 1$ for (a,b), and $\sigma^2 = 5$ for (c,d). FN% is the mean percentage of false negatives out of 100 trials. In each experiment we compared the approach that uses the Pearson's R test with Bonferroni correction, and two implementations that exploits Rademacher averages, one using the modified r statistic *ModR* defined in Eqn. 3.9, and another with $\tilde{r}^c$. Solid lines represent experiments on datasets with $n_{ext} = 0$ and $n_{ext} = 15$, and performance gaps between the two are highlighted. Dashed lines show simulated results on datasets with $n_{ext} = 750$.

RAveL-MB that exploit the Bonferroni correction, whose performances degrade by increasing the number of variables under analysis. Both behaviors are expected as the Bonferroni correction becomes stricter since the number N of hypotheses to test increases (see Section 3.2.3), while the bound to the supremum deviation remains stable as the complexity of the function class $\mathcal{F}$ does not increase[4]. Motivated by these observations, we simulated the performances of RAveL-PC and RAveL-MB variants that exploit Bonferroni correction in a high-dimensional scenario with 750 total variables, and we reported them as well in Figure 3.3 (dashed lines).

Figures 3.3(a,b) show differences between the approach that exploits Rademacher averages with the modified r statistic defined in Eqn. 3.9 and the one that exploits $\tilde{r}^c$, with the FN% of the first one decreasing for datasets with more than 10000 samples and the latter one just at 5000 samples. Such

[4] The most complex statistics in $\mathcal{F}$ are in fact the ones for which there is indepencence between $\mathbf{x}$ and $\mathbf{y}$, that are the ones with the highest variance.

difference is due to the normalization procedure employed by the former approach (see Sec. 3.4.3). Such procedure allows us to bound the test statistic (and therefore to use the Rademacher averages) but it also lowers the test statistic value as the sample size increases (since it will increase the chances of observing more extreme values) degrading the statistical power and requiring more accurate estimates of the bound $\mathcal{B}$ to the supremum $D(\mathcal{F}, S)$. $\tilde{r}^c$ instead is not affected by such issue and shows higher statistical power, highlighting the importance of the choice of the test statistic. From Figure 3.3(a,b) we also observe that the use of Bonferroni correction leads to a high statistical power, even with a high number of variables, in the $\sigma^2 = 1$ scenario. Such trend does not hold when $\sigma^2 = 5$ and the dimensionality is high (Figure 3.3(c,d)), for which `RAveL-PC` and `RAveL-MB` that exploit $\tilde{r}^c$ have more statistical power than algorithmic variants with Bonferroni correction.

3.5.2 Real datasets

We tested our algorithms on the Boston housing dataset [Harrison Jr and Rubinfeld, 1978] (see Appendix 6.1), which contains data about house prices in Boston suburbs, considering the median price of homes in each suburb as target T. Since the number of variables for such dataset is small, we used the Bonferroni variant of our algorithms `RAveL-PC` and `RAveL-MB`, with $\delta = 0.01$. Given the small number of observations (506 samples), we limited our analysis to conditioning sets $\mathbf{Z}$ of size at most 2 for maintaining a high statistical power in the independence testing. Both algorithms reported in output two variables, one related to the number of rooms per house, and the other to the median income of the suburb residents, that clearly influence the median price of the houses in the neighborhood. The first variable is a common indicator of the price of a house, while the second confirms the intuition that between two identical houses, the one built in a wealthier neighborhood has a higher price.

We finally tested our algorithms on the Framingham dataset (see Appendix 6.2), that provides information about the development of coronary heart disease (CHD) in 10 years for 3656 citizens of the city of Framingham, with 16 features describing health status and lifestyle. Given the relatively small number of samples, we limited our analysis to conditioning sets $\mathbf{Z}$ of size at most 2 for maintaining an high statistical power in the independence testing. We preprocessed the dataset by removing samples with missing data and binary features that were highly unbalanced, for which therefore we would not have had enough statistical power to test our assumptions[5]. We tested `RAveL-PC` and `RAveL-MB` variants using Bonferroni correction with $\delta = 0.05$ and got in output, for both discovery tasks, three variables: *Age*, *Systolic Blood Pressure*, and *Glucose*. Such results are supported by the World Health Organization guidelines[6]. Overall, our results on real data provide

[5] Dataset information on the Appendix.

[6] More information available on the official site `https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)` [Accessed: March 2023]

empirical evidence that our algorithms identify meaningful causal relations while avoiding false positives.

3.6 Conclusions

In this Chapter we presented two algorithms, `RAveL-PC` and `RAveL-MB`, for the task of local causal discovery. In contrast to state-of-the-art approaches, our algorithms provide guarantees on false discoveries in terms of bounding the FWER. Our algorithms use Rademacher averages to to properly account for multiple hypothesis testing, and our experimental evaluation shows that our algorithms properly control for false discoveries. Our algorithms can be extended to other (e.g., non-linear) test statistics and to other tests. In particular, Rademacher averages provide appealing time-effective alternatives for independence testing with test statistics whose distributions are unknown, since in such scenarios a typical solution is to rely on permutation testing, which require to analyze a large number of permuted datasets in order to achieve high statistical power. Interesting research directions include the application of our framework to recently proposed independence tests [Bellot and van der Schaar, 2019], improving the efficiency of our algorithms, and exploiting them for structure discovery.

Chapter 4

ALLSTAR

In this Chapter we study one instance of the effect estimation with statistical guarantees problem on observational data. We use the semantics of causal rules to describe combinations of treatment values and we propose `ALLSTAR` for the discovery of the rule with the highest reliable causal effect. Our algorithm has been adapted to discovery sound biological rules, and it has been applied on a breast cancer dataset to discover which combination of somatic genomic alteration is causally associated to a specific cancer type. `ALLSTAR` behaviour and performances have been assessed on synthetic datasets, and discovery results from real-world datasets have been validated against the literature.

4.1 Introduction

In the last ten years, the advances in DNA sequencing technologies have allowed to precisely depict the landscape of somatic alterations in large cohorts of tumours for various cancer types [Mardis, 2019, Weinstein et al., 2013, The International Cancer Genome Consortium, 2010]. The study of these data has shown that cancer is characterized by an extreme inter-tumour *heterogeneity*, with the alterations observed in different tumours being almost entirely different for any pairs of tumours. A number of computational tools have been designed to try to identify the alterations that drive the insurgence and development of tumours, while tackling inter-tumour heterogeneity [Cortés-Ciriano et al., 2022]. These tools are able to detect various types of signals [Cibulskis et al., 2013, Vandin et al., 2012, Mularoni et al., 2016, Arnedo-Pac et al., 2019] and integrate different prior and/or clinical information [Cowen et al., 2017, Reyna et al., 2020, Sarto Basso et al., 2019], but a common feature of such tools is that they detect alterations *correlated* with cancer phenotypes. That is, they identify alterations, or groups of alterations, significantly enriched in a group of patients or significantly associated with a (clinical) phenotype.

While the identification of alterations correlated with cancer phenotypes provides interesting insights into cancer initiation and progression, it does not guarantee that causal relations between somatic mutations and cancer are reported. While experimental and clinical validation is a necessary step to demonstrate the significance of alterations, tools reporting causal relations with guarantees on the quality of their findings would greatly reduce the resources needed to identify relevant alterations in follow-up experimental and clinical studies.

Randomized control trials are the gold standard in observational studies [Concato et al., 2000, Rosenbaum et al., 2010], and, in recent years, a lot of attention has been devoted towards mining *causal* rules [Silverstein et al., 2000] from observational data. Recently, Budhathoki et al. [2021] proposed a novel estimator of a rule's effect, taking into account the uncertainty of the estimates derived from data, and developed a branch and bound algorithm for the discovery task. Similarly, our work aims at finding reliable causal rules, but properly implements a correction for controlling the Family-Wise Error Rate (FWER) in a multiple hypothesis testing scenario, which is a fundamental feature of cancer studies given the high number of alterations found in tumours.

In this Chapter, we describe ALLSTAR, a novel tool to identify reliable causal relations between somatic mutations and cancer phenotypes. ALLSTAR identifies causal relations in the form of *rules* highlighting combinations of alterations with the highest average effect on the phenotype. Our contributions are fourfold. Firstly, we prove that the underlying computational problem is NP-hard. Secondly, we show that one needs to properly correct for multiple hypothesis testing when identifying *reliable* causal rules. Thirdly, we design ALLSTAR, an effective branch-and-bound algorithm to identify the *k* rules with the highest reliable average effect on the phenotype, with guarantees on the *Family-Wise Error Rate* (FWER) of the output. ALLSTAR identifies rules where genes are connected in a large interaction graph provided in input, and employs an iterative procedure leading to the identification of diverse rules, which highlight different causal relations potentially linked to cancer heterogeneity. Fourthly, we perform an extensive evaluation of ALLSTAR on both synthetic and cancer data, showing the statistical robustness of ALLSTAR and its ability to report well-supported as well as potentially novel causal relations between somatic mutations and cancer phenotypes.

ALLSTAR focuses on estimating the impact of genomic alterations on a tumour subtype, unlike Bayesian approaches such as Zhang et al. [2014] that learned a causal graph from The Cancer Genome Atlas (TCGA) mutation data to identify alterations relevant to ovarian cancer but without considering their effect on a target variable. A step towards the identification of causal relations between multi-omics data and a target variable (e.g., phenotype) has been made by the tools Aristotle [Mansouri et al., 2022] and CauMu [Liu et al., 2022], both identifying single features (i.e., alterations, or genes) linked to the phenotype. Our tool provides an efficient approach to identify rules comprising multiple features, which is an important characteristic given the high inter-tumour heterogeneity. Moreover, Aristotle focuses on the significance of the relation (by computing a corresponding

p-value), rather than their effect as done by ALLSTAR. Other causal tools, instead, leverage the increasing availability of single-cell RNAseq data (e.g., Cifuentes-Bernal et al. [2022]) and the estimates of pseudo-time derived from such data to identify causal relations at the transcriptomic level.

4.2 Methods

4.2.1 Causal Rules

Causal rules study the influence that a subset of actionable variables $\mathbf{X} = \{X_1, \dots, X_n\}$ exert on a target variable Y accounting for the possible confounding influence of a set of control variables $\mathbf{Z} = \{Z_1, \dots, Z_m\}$, that are common causes of at least one $X_i \in \mathbf{X}$ and Y. More specifically, a rule $\sigma = \pi_1 \wedge \pi_2 \wedge \cdots \wedge \pi_\ell$ is a conjunction of boolean propositions π_i defined on an actionable variable (e.g., $\pi_i \equiv X_5 = 1$), and which evaluates as *true* ($\top$) under an assignment $\mathbf{x} = \{x_1, \dots, x_n\}$ if all its propositions are verified by setting each X_j to the value x_j (otherwise the rule σ is *false*, or $\perp$). The *causal effect* of a rule σ [Budhathoki et al., 2021] on the target variable Y taking value y is defined as

$$e^y(\sigma) = \sum_{\mathbf{z}} (p(Y|\sigma(\mathbf{x}) = \top, \mathbf{Z} = \mathbf{z}) - p(Y|\sigma(\mathbf{x}) = \perp, \mathbf{Z} = \mathbf{z}))p(\mathbf{Z} = \mathbf{z})$$

where $\sigma(\mathbf{x})$ represents the value of σ under assignment $\mathbf{x}$. $e^y(\sigma)$ takes value in $[-1, 1]$ and it measures the increase in the probability that the target Y takes value y when the rule σ is true w.r.t. when σ is false. Despite being defined on conditional probabilities, $e^y(\sigma)$ measures the *causal* influence that the variables composing the propositions in σ exert on the event $Y = y$ if the *admissible input structure* assumptions [Budhathoki et al., 2021] are met, that are:

1. the target variable Y is not a cause of any $X_i \in \mathbf{X}$;

2. none of the variables $X_i \in \mathbf{X}$ is a cause of any $Z_j \in \mathbf{Z}$;

3. none of the variables $X_i \in \mathbf{X}$ is a cause of any other $X_j \in \mathbf{X}$; and

4. there is no unobserved variable U that directly cause $X_i \in \mathbf{X}$.

In other words, if the admissible input assumptions are met, $e^y(\sigma)$ measures the average treatment effect that the variables in σ exert on the event Y taking value y without including any spurious (i.e. non-causal) statistical correlation.

In this work we focus on applying the framework above to somatic mutations in cancer datasets, defining $\mathbf{X}$, $\mathbf{Z}$, and Y as follows:

1. the set $\mathbf{X}$ of features includes somatic alterations (i.e., SNVs, loss of heterozygosity, hypermethylation) in a set of genes, and the observations are provided by a binary matrix describing the status (present or not) of such alterations in a cohort of patients;

2. the set $\mathbf{Z}$ of confounders includes relevant germline mutations and clinical information (i.e. race, age, etc.), and the observations are provided by a corresponding matrix of relevant clinical variables;

3. the target Y is a phenotype of interest, such as histological or molecular marker-derived cancer subtypes.

In our setting, the constraints required by an admissible input structure for causal rule discovery translate as follows:

1. the target variable Y does not cause somatic alterations;

2. there is no somatic alteration that is a cause of any confounder;

3. there are no causal relations between somatic alterations;

4. there are no causal relations between somatic alterations and relevant unobserved variables.

Assumptions 1, 2, and 4 are satisfied by a proper choice of target variable Y, of confounders $\mathbf{Z}$, and the features $\mathbf{X}$ to include in the study. Assumption 3 is instead supported by the fact that somatic alterations arise as independent observations in the genome (even in normal cells), even if specific somatic alterations may modify the overall distribution of alterations in the genome (e.g., due to their impact on processes involved in mutagenesis). In such setup, each rule represents the observation of a specific set of gene alterations that occur simultaneously, and the rule effect is a measure of the influence of such pattern on having a specific cancer type.

4.2.2 Rule Effect Estimation

The estimation of probabilities from data is challenging when sample sizes are small, as the estimates obtained with naïve empirical estimators have high variance. As a consequence, rules discovered by data using such naïve empirical estimators have effects whose estimates are far from their true effects. To mitigate this phenomenon, which may lead to overfitting, Budhathoki et al. [2021] proposes a *reliable* estimator for the effect of causal rules.

Let us start the analysis by defining $\hat{p}(Y = y|\sigma = \top) = \frac{n_{Y=y,\sigma=\top}}{n_{\sigma=\top}}$ where $n_{\sigma=\top}$ is the number of instances for which $\sigma = \top$ (i.e, σ is true), and $n_{Y=y,\sigma=\top}$ is the number of instances for which $Y = y$ and $\sigma = \top$. Analogously we have $\hat{p}(Y = y|\sigma =\perp) = \frac{n_{Y=y,\sigma=\perp}}{n_{\sigma=\perp}}$. By considering all samples such that $\sigma = \top$ (resp. $\sigma =\perp$), the value y follows a binomial distribution with success probability $\hat{p}(Y = y|\sigma = \top)$. For a given confidence level $\alpha \in (0, 1)$, by defining $\beta(\alpha)$ as the $1-\alpha/2$ quantile of a standard normal distribution, the confidence bound for $\hat{p}_c(Y = y|\sigma = \top)$ proposed by Budhathoki et al. [2021] is then

$$\left[\hat{p}_c(Y = y|\sigma = \top) - \frac{\beta(\alpha)}{2\sqrt{n_{\sigma=\top}}}, \hat{p}_c(Y = y|\sigma = \top) + \frac{\beta(\alpha)}{2\sqrt{n_{\sigma=\top}}}\right].$$

Such bound allows us to compute the effect of **reliable causal rules**, defined as the lower bound of the effect of causal rules. That is, the estimated reliable effect $\hat{e}^{y}_{rel}(\sigma)$ of a causal rule σ on Y taking value y with confidence α is defined as:

$$\begin{aligned}\hat{e}^{y}_{rel}(\sigma,\alpha) =& \hat{p}_c(Y=y|do(Q_\sigma)) - \hat{p}_c(Y=y|do(Q_{\bar{\sigma}}))+\\ &-\frac{\beta(\alpha)}{2\sqrt{n_{\sigma=\top}}}-\frac{\beta(\alpha)}{2\sqrt{n_{\sigma=\bot}}}.\end{aligned}$$

where Q_σ is a *stochastic policy* i.e., a probability distribution over the interventions (see Budhathoki et al. [2021] for more details), which combines *atomic interventions* $do(X=x)$ [Pearl, 2009], i.e. changes the value of the variable X_i to x_i while keeping the values of all the other variables fixed.

4.2.3 Probability Estimation from Data

The estimator introduced by Budhathoki et al. [2021] is correct for the effect estimation of just one rule, but it may lead to false positives if multiple hypotheses (i.e. multiple rules) are analyzed, as in our case of discovering the top-k rules with the largest effect. As we focus on discovering rules bounding the FWER, that is the probability of returning in output at least one false positive, we correct the estimator proposed in Budhathoki et al. [2021] for multiple hypothesis testing using Bonferroni correction [Bonferroni, 1936], that is, we consider a corrected threshold $\alpha_c = \alpha/N$ for each hypothesis where N is the number of (potential) hypotheses tested.

4.3 ALLSTAR: Inferring Reliable Causal Rules between Somatic Mutations and Cancer Phenotypes

In this Section, we present our algorithm ALLSTAR (reliable cAusaL ruLe discovery between Somatic muTations and cAnceR phenotypes) for causal rule discovery with guarantees on its results. We start our analysis by proving that the underlying problem of estimating the rule with highest causal effect is NP-hard even if the probability distributions are known a priori (Section 4.3.1). In Section 4.3.2 we describe the algorithm and its subfunctions, and we prove ALLSTAR correctness. In Section 4.3.3 we present a novel bound that speeds-up the discovery in single-core machines, and Section 4.3.4 describes the cleanup procedure employed by ALLSTAR.

4.3.1 Computational Problem Definition and NP-Hardness

We now define the computational problem at the core of finding causal rules. In particular, we consider the problem of finding the rule with the largest positive effect on a target variable, defined as follows.

Definition 4.1. `Max Positive CRD` *problem.* *Consider variables* $\mathbf{Z} \cup \mathbf{X}$ *and a target variable* Y. *Find the rule* σ^* *with i)* $e(\sigma^*) > 0$ *and ii)* $\sigma^* = \arg\max_\sigma e(\sigma)$.

The *Max Positive CRD* problem is a simplified version of the problem of finding the rule with largest positive effect from data, since it assumes that one has access to the *exact* probabilities for the events of interests, while, in practice, such probabilities are estimated from an observational dataset (see Section 4.2). Nonetheless, we prove that the problem above is computationally difficult. In particular, we prove that finding the causal rule with the maximum effect is NP-hard, even when no confounder is considered (i.e., when $\mathbf{Z} = \emptyset$) and the true probabilities are described by a Bayesian Network.

We now define the aforementioned problem, that we call the `MaxCRD` problem.

Definition 4.2. `MaxCRD` *Problem.* *Given a Bayesian Network* B, *output* $\top$ *if the rule* $\sigma^* = \arg\max_\sigma |e(\sigma)|$ *with the highest absolute effect has a non-zero effect.*

The following theorem proves that the `MaxCRD` problem is computationally difficult.

Theorem 4.1. `MaxCRD` *is NP-hard.*

Proof. We prove that `MaxCRD` is NP-hard by reducing from `SAT`. The proof is divided in two steps: first we show a polynomial-time reduction of an input of `SAT` to an input of `MaxCRD`, and then we show that solving `MaxCRD` on such input allows to derive a solution to `SAT` in polynomial time on the original instance.

We start by describing the reduction from `SAT`. Let $\psi(\mathbf{X})$ be a boolean formula over variables in $\mathbf{X}$. Let us define $\mathcal{G} =< \mathbf{V}, \mathbf{E} >$ with $\mathbf{V} = \mathbf{X} \cup \{Y\}$ and $\mathbf{E} = \{X_i \rightarrow Y | X_i \in \mathbf{X}\}$. Let us define each $X_i \sim \mathcal{B}(0.5)$ be a Bernoulli distribution with probability $p(X_i = 0) = p(X_i = 1) = 0.5$. Let Y take values in $\{0, 1\}$ and let $p(Y = 1 | X_1 = x'_1, ..., X_n = x'_n) = 1$ if and only if $\psi((x'_1, ..., x'_n)) = \top$ else $p(Y = 1 | X_1 = x'_1, ..., X_n = x'_n) = 0$[1]. We then define the BN $B =< \mathcal{G}, p >$ as the reduced input for `MaxCRD`.

We now prove that solving `MaxCRD` on the reduced input leads to solving `SAT` in polynomial time on the original instance by proving that (i) if `MaxCRD`$(B) = \top$ then $\psi(\mathbf{X}) = \top$ and (ii) if `MaxCRD`$(B) = \bot$ then we can build a polynomial-time algorithm that solves `SAT`.

Let us prove (i). If `MaxCRD`$(B) = \top$ then $\exists \sigma | e_{corr}(\sigma) \neq 0$ that is $p(Y = y | \sigma = \top) - p(Y = y | \sigma =\bot) \neq 0$. By construction, we have two cases: $y = 1$ or $y = 0$. If $y = 1$ then $\psi(\mathbf{X})$ is satisfiable by construction since at least one between $p(Y = 1 | \sigma = \top)$ and $p(Y = 1 | \sigma =\bot)$ is positive. (Note that $\sigma =\bot$ corresponds to all assignments of variables $\mathbf{X}$ for which rule σ is not satisfied, and $p(Y = 1 | \sigma =\bot) > 0$ if and only if at least one such

[1] Note that the probability distribution function is fully specified since $p(Y = 0 | X_1 = x'_1, ..., X_n = x'_n) = 1 - p(Y = 1 | X_1 = x'_1, ..., X_n = x'_n)$.

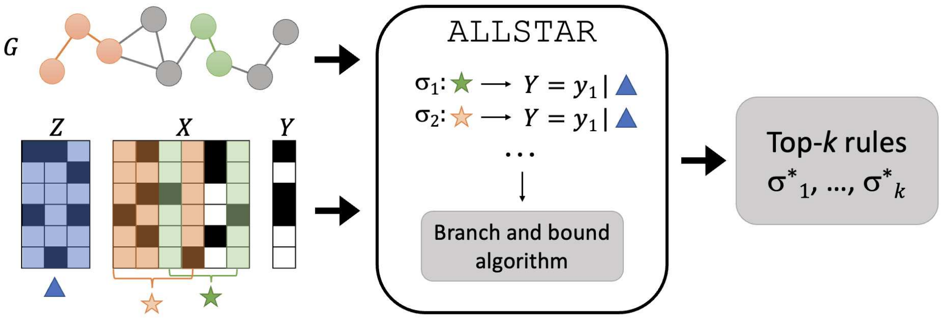


Figure 4.1: An illustration of ALLSTAR framework. From a dataset comprising a set of confounders $\mathbf{Z}$, treatments $\mathbf{X}$, and a target Y, ALLSTAR uses a branch and bound approach to discover the top-k rules $\sigma_1^*, ..., \sigma_k^*$ with the highest reliable causal effect. ALLSTAR exploits a gene-gene interaction network $\mathcal{G}$ to focus on biologically meaningful rules.

assignment lead to $Y = 1$, that by definition implies that such assignment satisfies $\psi(\mathbf{X})$.) If $y = 0$ then we notice that the same rule evaluated on $y = 1$ has a non-zero effect given that $y = 1$ is $y = 0$'s complementary event therefore $p(Y = 0|\sigma =\perp) = 1 - p(Y = 1|\sigma =\perp)$ (and the same holds for $\sigma =\perp$).

Let us prove (ii). If $\texttt{MaxCRD}(B) = \perp$ then $\forall\sigma$ we have $p(Y = y|\sigma = \top) = p(Y = y|\sigma =\perp) = p(Y = y)$ that is the value of Y is independent on $\mathbf{X}$ assignments. This means that $\psi(\mathbf{X})$ is either a tautology or a contradiction[2] and by evaluating $\psi(\mathbf{X})$ on any assignment we can distinguish between the two cases. □

As stated before, in practice we do not have access to the exact probabilities and, therefore, to the exact effect $e(\sigma)$ for a rule σ. We are therefore interested in finding the rule with the largest positive *reliable* effect from an observational dataset, which we formalize in the problem below.

Definition 4.3. `Max Reliable Positive CRD` *Problem.* *Consider an observational dataset $\mathcal{D}$ on variables $\mathbf{Z} \cup \mathbf{X} \cup \{Y\}$ and a confidence level $\alpha \in (0, 1)$. Find the rule σ^* such that i) $\hat{e}_{rel}(\sigma^*) > 0$ and ii) $\sigma^* = \arg\max_\sigma \hat{e}_{rel}(\sigma)$.*

4.3.2 ALLSTAR Algorithm

At its core, ALLSTAR (see Figure 4.1) employs the branch-and-bound approach proposed in Budhathoki et al. [2021] to discover the rule with the highest causal effect, while limiting to rules with at most ℓ alterations. Moreover, since in practice we are interested in finding multiple and diverse

[2] If not, then it would be possible to discover a rule with non-zero effect $\sigma : X_1 = x_1' \wedge ... \wedge X_n = x_n'$ on all elements of $\mathbf{X}$. By construction, in fact, $p(Y = 1|\sigma = \top) \in \{0, 1\}$ since it evaluates on just one element, and $p(Y = 1|\sigma =\perp) \neq p(Y = 1|\sigma = \top)$ otherwise the value of Y would be constant.

rules with positive reliable effect and with functionally related alterations, ALLSTAR uses an iterative approach to identify at most k rules, where k is a parameter provided by the user, and an interaction graph $\mathcal{G}$ to consider only rules with functionally related alterations.

```
Algorithm 7: ALLSTAR
   Input: alterations $\mathbf{X}$, confounders $\mathbf{Z}$, value $y$ of target $Y$, max. rule
          length $\ell$, confidence $\alpha$, graph $\mathcal{G} = (\mathbf{X}, E)$, integer $k$, clean-up
          threshold $t$
   Output: top-$k$ reliable causal rules
1  $N \leftarrow$ calculateRulesNumber($\mathcal{G}, \ell$); $\alpha_c \leftarrow \alpha/N$;
2  output $\leftarrow \emptyset$; Q $\leftarrow$ empty FIFO queue;
3  for $i \leftarrow 1$ to $k$ do
4      $\hat{e}_{\max} \leftarrow -\infty$; $\sigma_{\max} \leftarrow \emptyset$;
5      for $X_j \in \mathbf{X}$ do Q.enqueue("$X_j = 1$");
6      while |Q| > 0 do
7          $\sigma \leftarrow$ Q.dequeue();
8          if upperBoundRelATE($\sigma$,$y$,$\mathbf{Z}$,$\alpha_c$) > $\hat{e}_{\max}$ then
9              $\hat{e}_\sigma \leftarrow$ computeRelATE($\sigma, y, \mathbf{Z}, \alpha_c$);
10             if $\hat{e}_\sigma > \hat{e}_{\max}$ then $\hat{e}_{\max} \leftarrow \hat{e}_\sigma$; $\sigma_{\max} \leftarrow \sigma$;
11             for $\sigma' \in$expand($\sigma, G, \ell$) do
12                 Q.enqueue($\sigma'$)
13     if $\hat{e}_{\max} > 0$ then output $\leftarrow$ output $\cup \{\sigma_{\max}\}$;
14     update($\mathbf{X}$,$\sigma_{\max}$,$t$);
15 return output;
```

Specifically, ALLSTAR takes in input a set $\mathbf{X}$ of alterations, a set $\mathbf{Z}$ of confounders, a value y of interest for the target variable Y, the maximum length ℓ of rules, a confidence level α, a graph $\mathcal{G}$ whose vertices are the alterations in $\mathbf{X}$ and whose edges represent some relation between alterations (e.g., an edge represents the interaction between the proteins where the alterations are found), the maximum number k of rules to be reported in output, and a clean-up threshold $t \in [0, 1]$ that controls the diversity of the rules reported in output. In output, ALLSTAR produces at most k rules containing up to ℓ alterations, with the highest reliable effect and where each rule consists of alterations that form a connected subgraph of $\mathcal{G}$. In addition, each reported rule comprises alterations that appear in a set of patients different from the alterations in other reported rules, where the difference is controlled by the parameter t.

ALLSTAR starts by computing the total number of *candidate* rules of length at most ℓ (that is the number of connected subgraphs in $\mathcal{G}$ of length at most ℓ) and then calculates the correct threshold α_c for each confidence bound (see Section 4.2.3) using Bonferroni correction (line 1). The rule discovery is then performed in k iterations (line 3). In each iteration, a breadth-first search (BFS) of the lattice defined by set of all possible rules with at most

ℓ alterations is performed by using a FIFO queue `Q` and its (standard) operations `enqueue` and `dequeue`. During the BFS, the best rule $\sigma_{\max}$, and its maximum reliable estimated effect $\hat{e}_{\max}$, discovered during the exploration are maintained. After the initialization of $\sigma_{\max}$ and $\hat{e}_{\max}$ (line 4), the queue `Q` is initialized by inserting the rules containing a single alteration (line 5). (Note that `ALLSTAR` can also consider the *absence* of an alteration as part of a rule (i.e. $X_i = 0$); for clarity's sake, this is not reported in Algorithm 7.) The BFS then proceeds by extracting the current rule σ (line 7) until `Q` is not empty (line 6). When a rule σ is extracted from `Q`, an upper bound to its reliable effect is computed with the function `computeRelATE`$(\sigma, y, \mathbf{Z}, \alpha_c)$. If such upper bound is greater than $\hat{e}_{\max}$ (line 8) then the (exact) reliable effect estimate $\hat{e}_\sigma$ of σ is computed (line 9), and the values $\hat{e}_{\max}, \sigma_{\max}$ are updated if $\hat{e}_\sigma > \hat{e}_{\max}$ (line 10). Then, the rules that are obtained by expanding σ, obtained with the function `expand`$(\sigma, \mathcal{G}, \ell)$, are added to the queue (lines 11-12). `expand`$(\sigma, \mathcal{G}, \ell)$ returns all rules (with at most ℓ alterations) that are obtained by adding to σ one alteration that must be connected in $\mathcal{G}$ to at least one alteration of σ. When the BFS completes, the best rule $\sigma_{\max}$ is added to the output set if its estimated reliable effect is positive (line 13), and the set $\mathbf{X}$ of alterations is updated (line 14) to avoid discovering highly-overlapping, redundant, rules (see below). At the end, the set of at most top-k rules is reported in output (line 15).

Algorithm 8: `calculateRulesNumber`

```
Input: Graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, maximum rule length $\ell$
Output: Number $N$ of connected subgraphs of length at most $\ell$
        between elements in $\mathcal{G}$
$P \leftarrow \emptyset$;
$Q \leftarrow \emptyset$;
for $X \in \mathbf{V}$ do
    $P \leftarrow P \cup \{X\}$;
    $Q \leftarrow Q \cup \{X\}$;
for $i \leftarrow 1$ to $\ell - 1$ do
    $L \leftarrow \emptyset$;
    for $q \in Q$ do
        for $X \in q$ do
            for $e \in \mathbf{E}$ do
                if $X \in e$ & $e \setminus \{X\} \notin q$ then
                    $L \leftarrow L \cup \{q \cup \{e \setminus \{X\}\}\}$
    Remove duplicates from $L$;
    $P \leftarrow P \cup \{L\}$;
    $Q \leftarrow L$;
return size($P$);
```

`ALLSTAR` exploits three subroutines `calculateRulesNumber`, `upperBoundRelATE`, and `computeRelATE` that will be briefly explained in the following:

`calculateRulesNumber` takes as input a graph $\mathcal{G}$ and the maximum rule length ℓ and outputs the number of connected subgraphs of length at most ℓ between elements in $\mathcal{G}$. It is used to calculate the total number of possible rules under study, which is the amount of test performed in the worst case, and the pseudocode is described in Algorithm 8.

`upperBoundRelATE` takes as an input a rule σ, the value y for target Y, a set of confounders $\mathbf{Z}$, and a threshold α_c corrected for multiple hypotheses testing, and it outputs the tight optimistic upper bound to the effect for the rule proposed by Budhathoki et al. [2021]. It is used by the branch-and-bound algorithm for deciding whether to compute the values a specific branch (i.e. all children of a specific rule) or to avoid the computation because the best solution found in such branch would never improve the current best solution (i.e., the incumbent) $\hat{e}_{max}$.

More specifically, let us consider a rule σ and a more specific rule $\sigma' = \sigma \wedge \pi_j$. Let us define the quantity $\tilde{\tau}_{\sigma'}(\sigma, \mathbf{z})$ on the elements for which $\mathbf{Z} = \mathbf{z}$ holds as

$$\tilde{\tau}_{\sigma'}(\sigma, \mathbf{z}) = \max_{a'_\sigma \in \{0,1,\dots,a_\sigma\}} \frac{a'_\sigma + 1}{a'_\sigma + 2} - \frac{n_1 - a'_\sigma + 1}{n - a'_\sigma + 2} + \\ - \frac{\beta(\alpha_c)}{2\sqrt{a'_\sigma + 2}} - \frac{\beta(\alpha_c)}{2\sqrt{n - a'_\sigma + 2}}$$

where $\beta(\alpha_c)$ is the $1 - \alpha_c/2$ quartile of the standard normal distribution, n is the number of instances taken into account (i.e. with $\mathbf{Z} = \mathbf{z}$), n_1 of which have $Y = y$, and a_σ is the number of instances for which σ holds, $\mathbf{Z} = \mathbf{z}$ and $Y = y$. The upper bound is then defined as

$$U(\sigma') = \sum_{\mathbf{z}} (\tilde{\tau}_{\sigma'}(\sigma, \mathbf{z}) \hat{p}(\mathbf{Z} = \mathbf{z}))$$

where $\hat{p}(\mathbf{Z} = \mathbf{z})$ is the empirical probability of $\mathbf{Z}$ taking value $\mathbf{z}$. Differently from Budhathoki et al. [2021] our bound uses a confidence level $\alpha_c = \alpha/N$, where N is the total number of rules considered by the algorithm, to account for the multiple hypothesis testing problem.

`computeRelATE` takes in input a rule σ, the value y for target Y, a set of confounders $\mathbf{Z}$, and a threshold α_c and calculates the reliable effect of the rule $\hat{e}_{rel}(\sigma)$ as described in Section 4.2.3. Let us recall from 4.2.2 the definition of $\hat{p}(Y = y|\sigma = \top) = \frac{n_{Y=y,\sigma=\top}}{n_{\sigma=\top}}$ where $n_{\sigma=\top}$ is the number of instances for which $\sigma = \top$ (i.e, σ is true), and $n_{Y=y,\sigma=\top}$ is the number of instances for which $Y = y$ and $\sigma = \top$. Analogously we have $\hat{p}(Y = y|\sigma =\perp) = \frac{n_{Y=y,\sigma=\perp}}{n_{\sigma=\perp}}$. In extreme cases (e.g. $\sigma =\perp$ for all instances) such quantities are ill-defined, therefore the Laplace correction is applied to the estimated conditional probability, which becomes $\hat{p}_c(Y = y|\sigma = \top) = \frac{n_{Y=y,\sigma=\top}+1}{n_{\sigma=\top}+2}$. The returned value $\hat{e}^y_{rel}(\sigma)$ is then defined as

$$\begin{aligned}\hat{e}^{y}_{rel}(\sigma) = \sum_{\mathbf{z}} \Bigg[\Bigg(&\hat{p}_c(Y = y|\mathbf{Z} = \mathbf{z}, \sigma = \top)+ \\ &- \hat{p}_c(Y = y|\mathbf{Z} = \mathbf{z}, \sigma =\bot)+ \\ &- \frac{\beta(\alpha_c)}{2\sqrt{n_{\mathbf{Z}=\mathbf{z},\sigma=\top}}} - \frac{\beta(\alpha_c)}{2\sqrt{n_{\mathbf{Z}=\mathbf{z},\sigma=\bot}}}\Bigg)\hat{p}(\mathbf{Z} = \mathbf{z})\Bigg].\end{aligned}$$

Theoretical guarantees. The following theorem proves that ALLSTAR produces in output a set of rules with a rigorous bound on its FWER, where a false positive is defined as a rule σ reported in output but with effect $e(\sigma) \leq 0$.

Theorem 4.2. ALLSTAR*($\mathbf{X}$, $\mathbf{Z}$, y, ℓ, α, $\mathcal{G} = (\mathbf{X}, \mathbf{E})$, k, t) outputs a set of rules with FWER $\leq \alpha$.*

Proof. [*Sketch*] Let us notice that each iteration of the for loop at line 3 of Algorithm 7 considers an increasingly small subset of $\mathbf{X}$ and therefore the total amount N of candidate causal rules that may be evaluated by ALLSTAR (i.e. the total number of hypotheses tested in the worst scenario) is equal to the total number of rules that can be evaluated on the first iteration of the loop. In particular, the number of all the different rules of max length ℓ (i.e. N, line 1) is equivalent to the number of distinct connected subgraphs in $\mathcal{G}$ of length at most ℓ since ALLSTAR exploits $\mathcal{G}$ to expand a rule σ to a more specific $\sigma' \supset \sigma$ by adding a proposition $X_i = 1$ only if X_i is not already present in σ and it is connected to at least one treatment of σ.

We now prove that, by setting $\alpha_c = \alpha/N$ (line 1), ALLSTAR returns a false positive with probability at most α. Let us suppose that a false positive rule σ_{FP} (i.e. such that $e(\sigma_{FP}) \leq 0$) is returned in output by ALLSTAR. A necessary condition for this to happen is to add σ_{FP} to the top-k rules found (line 13) which in turn happens only if its estimated effect $\hat{e}_\sigma$ (calculated in line 9) is greater than 0 (line 13). By construction of the confidence intervals with confidence α_c, a rule with $e(\sigma_{FP}) \leq 0$ may have its estimated effect $\hat{e}_\sigma > 0$ with probability at most α_c. Since there are at most N rules under study, in the worst case the probability of having at least a false positive estimate is $N \times \alpha_c = \alpha$ which implies that the algorithm does not output any false positive with probability of at least $1 - \alpha$. □

4.3.3 Improved Bound Description

While the parallel implementation of ALLSTAR employs the branch-and-bound approach proposed in Budhathoki et al. [2021], we also develop an improved (i.e., tighter) upper bound on the reliable causal effect of a rule that is best suited for single-core runs since it requires a data structure shared among cores. Such bound relies on the key observation that one rule σ' is more specific of every rule in $\Omega_p = \{\sigma' \setminus \{\wedge\pi_k\} | \forall \pi_k \in \sigma'\}$. Consider a rule $\sigma = \pi_1 \wedge ... \wedge \pi_i$ and a more specific one $\sigma' = \sigma \wedge \pi_j$. Budhathoki et al. [2021]

defined the upper bound $\tilde{\tau}_{\sigma'}(\sigma, \mathbf{z})$ to the reliable effect estimate $\hat{e}_{rel}(\sigma')$ of σ' as a function of the number of instances n (in the $\mathbf{Z}$ strata), the number n_1 of instances with $Y = y$, and the number a_σ of instances for which σ holds and $Y = y$, as

$$\tilde{\tau}_{\sigma'}(\sigma, \mathbf{z}) = \max_{a'_\sigma \in \{0,1,\ldots,a_\sigma\}} \frac{a'_\sigma + 1}{a'_\sigma + 2} - \frac{n_1 - a'_\sigma + 1}{n - a'_\sigma + 2} + \\ - \frac{\beta(\alpha)}{2\sqrt{a'_\sigma + 2}} - \frac{\beta(\alpha)}{2\sqrt{n - a'_\sigma + 2}}$$

which upper bounds the effect of σ' by exploiting the fact that a_σ will upper bound the number $a_{\sigma'}$ of instances for which the σ' holds and $Y = y$, given that σ' is more specific than σ. We argue that σ' not only is more specific than σ, but also than every rule in the set $\Omega_p = \{\sigma' \setminus \{\wedge \pi_k\} | \forall \pi_k \in \sigma'\}$ of all possible rules chosen from σ' removing the proposition π_k. The proposed estimator must hold for each rule in Ω_p therefore we propose a tighter optimistic estimator that considers $a_{min} = \min_{\sigma_j \in \Omega_p} a_{\sigma_j}$ as

$$\bar{\tau}_{\sigma'}(\sigma, \mathbf{z}) = \max_{a'_{\Omega_p} \in \{0,1,\ldots,a_{min}\}} \frac{a'_{\Omega_p} + 1}{a'_{\Omega_p} + 2} - \frac{n_1 - a'_{\Omega_p} + 1}{n - a'_{\Omega_p} + 2} + \\ - \frac{\beta(\alpha)}{2\sqrt{a'_{\Omega_p} + 2}} - \frac{\beta(\alpha)}{2\sqrt{n - a'_{\Omega_p} + 2}}$$

Notice that if a rule $\sigma_{rem} \in \Omega_p$ has been pruned by the breadth-first branch and bound algorithm, then we can set $\bar{\tau}_{\sigma'}(\sigma, \mathbf{z}) = -\infty$ since the condition in line 8 does not hold for any such σ', given that it is more specific than σ_{rem}.

4.3.4 Cleanup Threshold

As stated above, in order to identify a diverse and more informative set of rules, the set $\mathbf{X}$ of alterations is updated after each rule is extracted. This is done with function update($\mathbf{X}, \sigma_{\max}, t$) (line 14), which we now describe. Such function removes from the set of alterations $\mathbf{X}$ the ones that either appear in the rule $\sigma_{\max}$ or are very similar to at least one alteration in $\sigma_{\max}$. The similarity is defined according to the normalized city-block Manhattan distance, defined for two vectors $\mathbf{a}$ and $\mathbf{b}$ in n dimensions as $d_M(\mathbf{a}, \mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n} |a_i - b_i|$. In particular, update($\mathbf{X}$,$\sigma_{\max}$,$t$) removes from $\mathbf{X}$ all alterations in $\sigma_{\max}$ and the ones with distance d_M less than t from at least one alteration in $\sigma_{\max}$, where the distance between the vectors describing the appearance of alterations in patients is considered and t is a user-defined threshold. This function therefore allows to recover non-overlapping rules over the whole alterations' search space.

Implementation. We implemented ALLSTAR in Python 3. Our implementation exploits multicore parallelism, when available. Code, data and

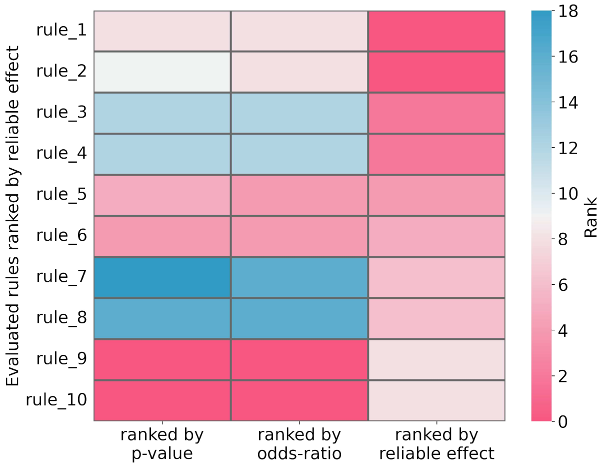


Figure 4.2: Ranking comparison of the top-10 binary rules with the highest effect computed on a dataset of 100 samples w.r.t. three different metrics (p-value, odds-ratio and reliable effect). Each row corresponds to a rule and each column corresponds to its ranking (with 1st scores being the highest) w.r.t. p-value, odds-ratio and ALLSTAR reliable effect, respectively. Colorscale representing ranking position on the right.

scripts to reproduce the experiments described below are available at https://github.com/VandinLab/ALLSTAR.

4.4 ALLSTAR Performances

In this Section, we assess ALLSTAR's performances on synthetic data. We start the analysis by comparing ALLSTAR with standard correlational approaches[3] (Section 4.4.1). We then evaluate the impact of the multiple hypotheses testing correction employed by ALLSTAR (Section 4.4.2), and the role of input parameters $\mathcal{G}$ (Section 4.4.3) and t (Section 4.4.4). We then assessed the stability of our algorithm on different combinations of ℓ and t (Section 4.4.5). Lastly, we studied ALLSTAR computational performances by comparing it to a brute-force algorithm (Section 4.4.6).

4.4.1 Comparison with Statistical Methods

Firstly we compared ALLSTAR with standard correlational approaches, to understand whether the results obtained are the same. In particular, we generated all possible binary rules comprising one gene and one target on a dataset of 100 samples, and for each one we computed i) the reliable effect as computed by ALLSTAR, ii) the p-value from the Fisher exact test, and

[3] We could not compare with Aristotle due to issues with its implementation, available at this link.

iii) the odds ratio[4]. We then sorted the results according to each computed value and compared the three rankings. In Figure 4.2 we show the rankings of the top-10 rules with the highest reliable effect: the top rules obtained by ALLSTAR have a much lower ranking as if they were ranked by p-value or odds ratio. For example, 4 of the top-10 rules according to the reliable effect are not in the top-10 by p-value or by odds ratio, with one rule appearing in the 18th position of the ranking by p-value. In general, while there is a concordance in terms of Kendall-tau coefficient [Kendall, 1938] between the ranking by reliable effect and the other measures (Kendall-tau coefficient 0.79 correlation between the odds ratio and effects; Kendall-tau coefficient 0.9 between p-values and effects), the reliable effect provides different top rules (which are the most interesting ones for any practical purpose) than standard correlation approaches.

4.4.2 Algorithm Correctness

We then performed multiple experiments to assess ALLSTAR's efficiency and correctness, using synthetic datasets. Every synthetic dataset resembles the structure of real cancer data, with mutated genes as treatments $\mathbf{X}$ and a binary outcome Y. (For simplicity we set $\mathbf{Z} = \emptyset$ in these analyses.) For each experiment, we sampled 10 datasets for every tested sample size (25, 50, 75, 100, 250, 500, 1000, 5000, 10000, and 25000). In each dataset, most alterations are drawn randomly with probability 0.5 and independently of the outcome Y. In some datasets, we planted alterations with a causal relation to the target Y; such alterations constitute the rules of interest to assess ALLSTAR's performance. Their relationships with Y are described in the related Section below.

In the first experiment, we assessed the impact of correcting for multiple hypothesis testing on false positives. In this experiment we considered only random alterations in each sample, hence no causal rule (i.e. any rule with a positive effect) with respect to the outcome was planted. We considered three different estimates of the (reliable) effect: the version based on the naïve estimate of probabilities, the reliable approach proposed in Budhathoki et al. [2021][5], and the one used by ALLSTAR (see Section 4.2.3). For the last two estimates, the value $\alpha = 0.05$ was considered. In particular, the naïve approach estimates the effect $\hat{e}(\sigma)$ as the difference $\hat{p}(Y = y|\sigma = \top) - \hat{p}(Y = y|\sigma = \perp)$ (i.e., empirical probabilities estimated from data and without any correction), while the reliable approach proposed in Budhathoki et al. [2021] considers $\hat{e}_{rel}(\sigma)$ (i.e., adding confidence bounds) but without correcting for multiple hypothesis testing, as it is done instead in our approach (see Sections 4.2.2-4.2.3). Our findings show that both the naïve and reliable approach incorrectly return at least one rule

[4] In the two latter cases, we computed the values taking into consideration the contingency table associated to each rule as in Budhathoki et al. [2021].

[5] The code available at the bitbucket repository does not run properly, therefore we implemented our own, equivalent, version.

with a positive effect for *every* dataset (i.e., corresponding to a FWER of 1), while ALLSTAR is the only one correctly returning zero false positives. These results show that the multiple hypothesis correction on ALLSTAR's reliable effect is a crucial component to avoid false discoveries.

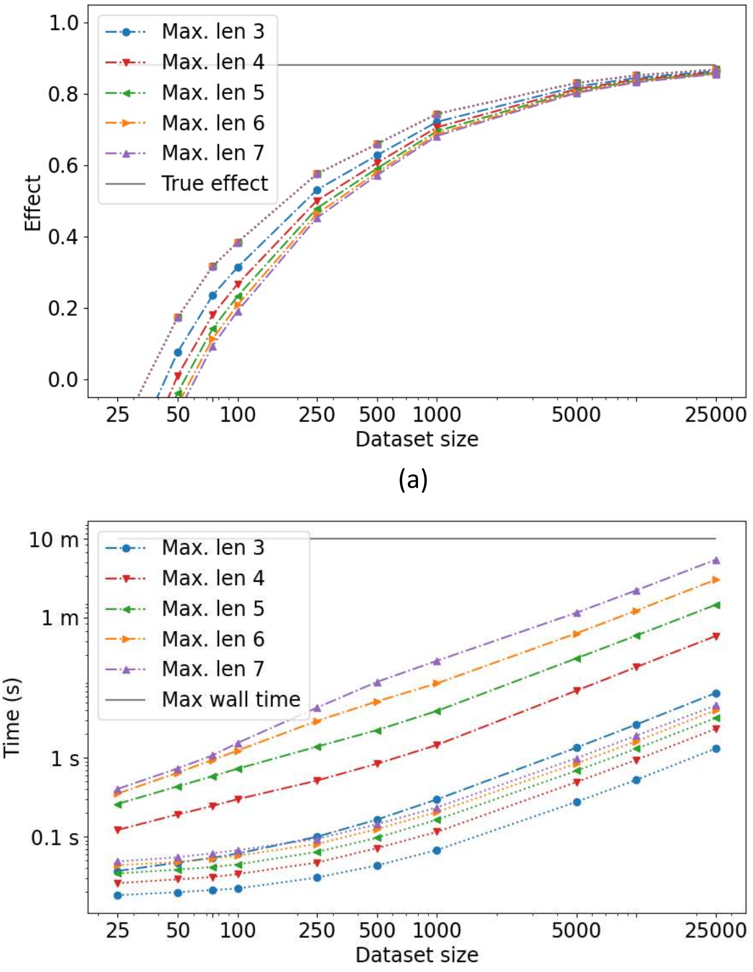
4.4.3 Role of $\mathcal{G}$

We then assess the effectiveness of using the interaction graph $\mathcal{G}$ in ALLSTAR when identifying causal rules composed of multiple alterations by sampling multiple datasets with a total of 22 alterations, of which 7 are part of a rule causally related to the target Y and constitute a connected subgraph of $\mathcal{G}$. Figure 4.3 shows the results obtained passing $\mathcal{G}$ in input (dotted line) and the results obtained when no prior knowledge on gene interaction is considered (dash-dotted line), obtained by passing a fully connected graph in input to ALLSTAR. In particular, we considered both the effect estimation of the implanted rule and the runtime, and we ran ALLSTAR for various values of the maximum rule length ℓ. As expected, the estimate of the effect converges to the true effect for all values of ℓ, and the estimate obtained using the interaction graph $\mathcal{G}$ is significantly better than the one when no prior knowledge is considered. Moreover, the use of $\mathcal{G}$ drastically reduces the runtime (due to a reduction in the number of candidate rules). For example, with 25000 samples and $\ell = 7$, the runtime using $\mathcal{G}$ is of few seconds, while almost 3 minutes are required when no prior knowledge is considered. This shows that the interaction graph leads to significant improvements in terms of the estimate of the true effect and of runtime.

4.4.4 Role of t

We then ran an experiment to assess the ability of ALLSTAR (i) to recover planted rules that cover a wide spectrum of diverse functional processes, a key feature given the high inter-tumor heterogeneity that characterizes cancer, (ii) even when some admissible input structure assumptions (see Sec. 4.2.1) are not satisfied.

To empirically test the first hypothesis, we compare ALLSTAR with a naïve greedy selection of the top-k rules by effect, without considering any cleanup threshold t. Let us consider the scenario of Section 4.4.3 experiment, in which we consider datasets with a rule σ implanted composed of 7 elements. By considering rules length $\ell \in \{2, 3, 4, 5\}$, all the elements of the implanted rule are returned by ALLSTAR by just setting $k = \lceil \frac{7}{\ell} \rceil$. In other words, all elements of σ are in the top-4 rules returned by ALLSTAR for $\ell = 2$, in the top-3 for $\ell = 3$, and in the top-2 for $\ell \in \{4, 5\}$. A greedy algorithm that ranks per effect all the rules of length ℓ and selects the top-k without exploiting a cleanup threshold, however, was never able to discover all the important genes in σ and always returned rules with repeated genes. Such results are sound w.r.t. different dataset sizes (as we tested datasets with 100, 1000, and 10000 samples) and statistical noises (as we tested 10 datasets per



(a)

(b)

Figure 4.3: Mean planted reliable rule effect (a) and mean runtimes (b) over multiple dataset sizes on 10 runs. In each plot, the dotted lines represent ALLSTAR results passing a protein-protein interaction $\mathcal{G}$ in input, and dash-dotted lines represent the approach with a fully connected graph (i.e., no prior knowledge).

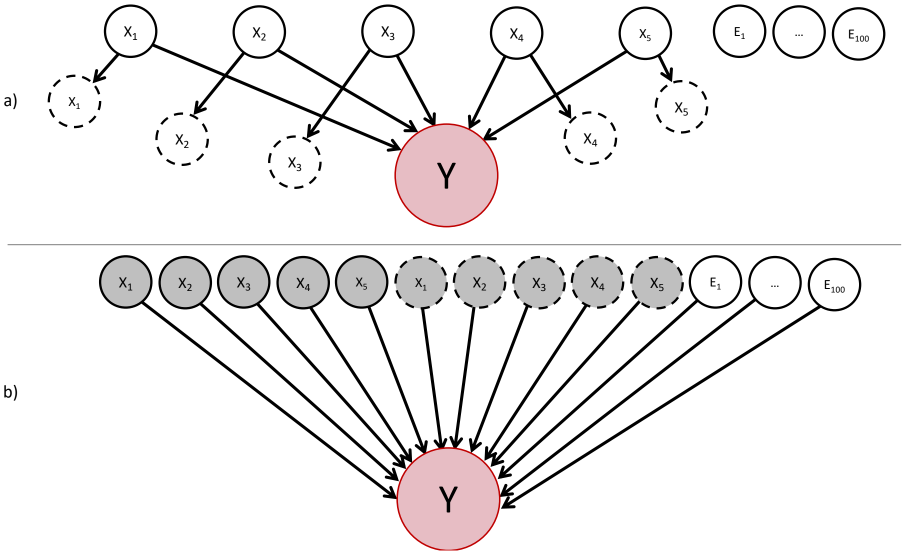


Figure 4.4: Data generative BN (a) and assumed graph (b) of last synthetic experiment. In the second plot, clone variables (i.e. those which definition depends on other variables) are shown with a dashed border, and variables that output rules with a positive effect by ALLSTAR without cleaning procedure are represented in grey.

sample size).

We then tested the second hypothesis by simulating datasets with 3 planted rules (of 5 genes in total) and 100 random alterations. For each alteration in a planted rule, we also planted a correlated alteration with 97.5% of values identical to the planted alteration (see equations in Section 6.3 the Appendix). This allows us to assess whether ALLSTAR reports the correct causal alterations and not the correlated (but non-causal) ones. We obtained the top-3 rules running both a variant of ALLSTAR obtained without using the Manhattan distance-based updating procedure (see function $\texttt{update}(\mathbf{X}, \sigma_{\max}, t)$ in Algorithm 7), and ALLSTAR with $t = 0.05$. ALLSTAR reports the planted rules and correctly disregards the rules comprising the correlated alterations. The variant of ALLSTAR that does not use the Manhattan distance-based updating procedure, instead, produces, among the top-3 rules, rules containing the correlated alterations. In the latter case, ALLSTAR variant returning correlations instead of causal relations is a consequence of the failure of assumption 3 (see Section 4.2.1) for the admissible causal structure, while the Manhattan distance-based updating procedure allowed us to remove the spuriously-linked variables and to report only causal relations. In particular, the graph of the data generative BN was the one shown in Figure 4.4(a), while the one assumed by ALLSTAR was Figure 4.4(b). The main difference between the two BNs are the d-separations between X_i, $1 \leq i \leq 5$ and their clones $X_{i(clone)}$. In particular, X_i always blocks spurious correlation paths (more on this and *d-separation* in Pearl [2009]) from $X_{i(clone)}$ in Figure 4.4(a) but not in Figure 4.4(b), therefore if we (incorrectly) assume the underlying graph to be as the latter, in order to still have correct results we

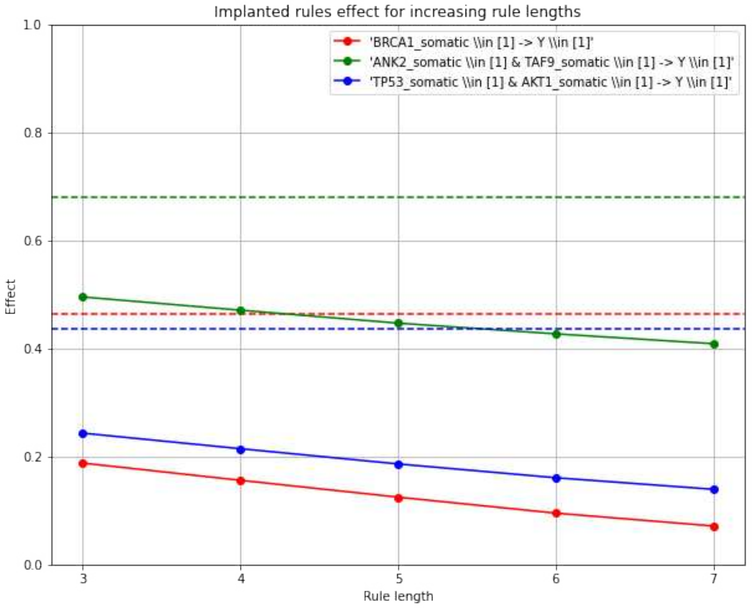


Figure 4.5: Average effect returned by ALLSTAR (solid lines) and theoretical value (dashed lines) for the 3 implanted rules of Section 4.4.3. Results have been averaged on datasets with 1000 samples, and the variability in each run results is negligible.

should be able to have some other heuristic mechanism (the threshold-based cleaning procedure) that removes the clones in order not to return them. Let us also notice that another difference between Figure 4.4(a) and Figure 4.4(b) relies on the links between the external variables ($E_i, 1 \leq i \leq 100$) and Y. Such links imply some form of (possible) dependence whose strength is defined by the probability distribution functions inferred by the observational dataset. ALLSTAR however, is able to confidently ignore such spurious correlations due to the use of the *reliable* effect estimator and its ability to deal with multiple hypotheses testing (more on that on Theorem 4.2 proof). Analysis of relaxation of the other assumptions (and the consequent development of new methods) is still an open research task, for which we point the reader to the discussion on Budhathoki et al. [2021].

4.4.5 Stability Analysis

We experimentally assessed the stability of ALLSTAR results with respect to the user-defined parameters ℓ and t.

In the first experiment, we run ALLSTAR on the datasets of Section 4.4.3, with ℓ taking values from 3 to 7, and $t = 0.05$. Figure 4.5 shows the average effect returned by ALLSTAR (solid lines) for the three implanted rules by varying ℓ, as well as their theoretical value (dashed lines). Results have been averaged for all datasets of 1000 samples, and their variability across the runs is negligible. As expected, the rule effect returned by ALLSTAR decreases as the rule length increases because the number of hypotheses to test increases and therefore the Bonferroni correction becomes stricter. Moreover, despite increasing ℓ allows `ALLSTAR` to evaluate more rules, our algorithm did not return any false positive.

We then analyzed results variability w.r.t. changes of t by running `ALLSTAR` on the same setup of Section 4.4.4, and setting t to $0.01, 0.025, 0.05, 0.075$, and 0.1. We finally set $k = 4$ to assess the ability of `ALLSTAR` to avoid returning duplicated rules. `ALLSTAR` returned duplicate rules consistently among all the runs for $t = 0.01$, among 6 of 10 runs for $t = 0.025$, and did not return any duplicated rule for all the other values of t tested. This is an expected behavior in this data generative scenario since each rule differs from its clone on 2.5% of samples on average (see Section 4.4.4).

4.4.6 Computational Performances

We finally compared the computational performances of `ALLSTAR` against a brute-force algorithm that exploits $\mathcal{G}$ to select the candidate rules to study, but calculates them all without exploiting the branch-and-bound. In this experiment we created 10 synthetic datasets with 1000 samples from the following distributions

$$\begin{aligned} X_1 &\sim \mathcal{B}(0.15) \\ E_i &\sim \mathcal{B}(0.1), 1 \leq i \leq 600 \\ Y &\sim X_1 \vee \mathcal{B}(0.05) \end{aligned}$$

and we searched for the rule with the highest effect ($k = 1$) by setting the target value $Y = 1$. We run both algorithms on 60 cores of our cluster and we tracked the runtimes without considering the time required to calculate the Bonferroni correction (i.e. function `calculateRulesNumber` of `ALLSTAR`) as our focus is to compare the performances of the two rule discovery approaches only[6]. Figure 4.6 compares the average runtimes in seconds of both approaches (y axis is log-scaled) over increasing maximum rule lengths ℓ. As expected, `ALLSTAR` is faster than the brute force approach due to the speedup given by its branch-and-bound, and such difference increases with the number of rules under study, therefore it increases monotonically with ℓ. As a reference, the brute force algorithm is more than 3 times slower than `ALLSTAR` when discovering rules setting $\ell = 4$, and nearly 20 times slower for $\ell = 5$.

4.5 Rule Discovery on Breast Cancer Data

In this Section, we describe the discovery results on real-world breast cancer data. We start in Section 4.5.1 by introducing the datasets, and in Sec-

[6] We remind that such procedure would be a prerequisite for both algorithms, therefore it would just add a bias term to both runtimes under analysis.

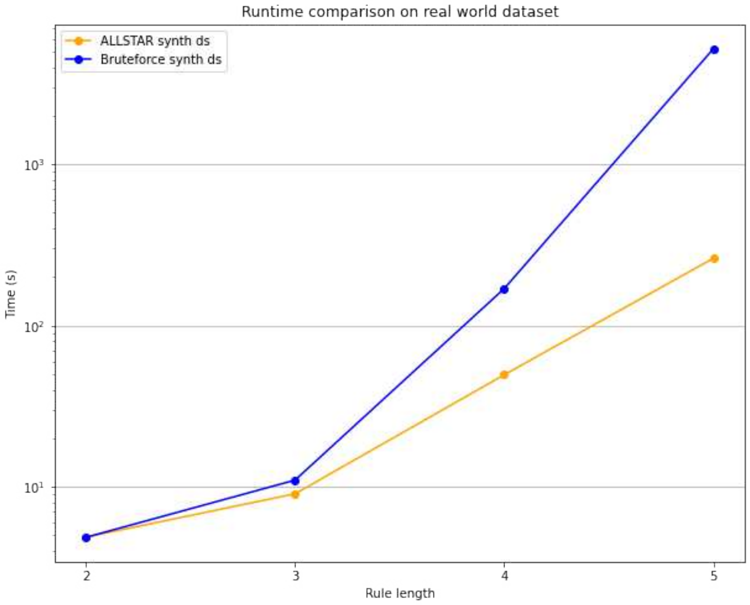


Figure 4.6: Average runtime comparison between ALLSTAR and a brute-force algorithm on 10 synthetic datasets over different rule lengths ℓ. Y-axis is logarithmically scaled, and variability across runs with the same ℓ is negligible (and therefore not plotted).

tion 4.5.2 we compare our method with other statistical methods. Section 4.5.3 describes our results on real-wold data and validates their biological relevance. Finally, Section 4.5.4 provides an assessment of results stability.

4.5.1 Cancer Data and Interaction Network

We tested ALLSTAR on publicly available breast cancer (BRCA) data from TCGA. In particular, we downloaded public clinical and somatic mutational data from the TCGA-BRCA repository, for a total of 1096 samples. We also included the subtype classification of TCGA-BRCA based on the 50-gene PAM50 model [Parker et al., 2009]. We also retrieved germline mutational patterns for TCGA patients in BRCA1 and BRCA2 from Kraya et al. [2019]. We integrated two additional alteration types that play a significant role in cancer: loss of heterozygosity (LOH) information from Riaz et al. [2017] and reported by Bodily et al. [2020], and hypermethylation from Xena Functional Genomics Explorer data [Goldman et al., 2015] and reported in Bodily et al. [2020]. The final datasets comprised a number of samples ranging from 898 to 935, depending on the target variable of interest. As an input graph $\mathcal{G}$ for ALLSTAR we considered the most recent Functional Interaction [Wu et al., 2010] gene network from Reactome[7], which comprises almost 14,000 genes and more than 250,000 edges. Note that patients in our dataset are

[7] FIsInGene (version 2021), available at this link.

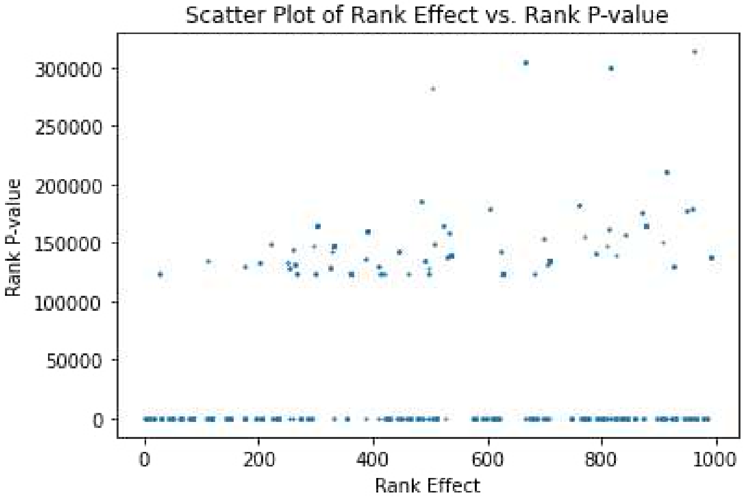


Figure 4.7: Comparison of the rankings in terms of reliable effect (ALLSTAR output, x-axis) and p-value (CMH test output, y-axis) for real-world data. Each dot corresponds to one of the top-1000 rules ranked by reliable effect.

all affected by cancer, therefore every reported rule implicitly conditions on such event.

4.5.2 Comparison with Statistical Methods

Initially, we emulated the analysis of Section 4.4.5 on real-world cancer data, that presents an even larger difference between rankings than synthetic data. For this test we selected the dataset presented in Section 4.5.1 with the 300 most frequent somatic mutations, the 300 most frequent LOHs, and the profiles of 22 frequently hypermethylated genes, as $\mathbf{X}$, and the Triple-Negative binary molecular classification, as target Y. We then ran both ALLSTAR and a python implementation of the Cochran–Mantel–Haenszel test (CMH), ranking reliable effects and p-values for rules built on every combination of one confounder, two alterations and the outcome. The scatter plot describing the two rankings' comparison for the first 1000 rules sorted by reliable effect, is shown in Figure 4.7. It seems clear that a considerable amount of rules ranked among the top-1000 in the effect ranking, when assessed using the correlation-based method CMH, are placed well beyond the 100000$^{\text{th}}$ position. Additionally, the p-value rankings compressed to the bottom of the plot are actually all valued 1. CMH is not able to differentiate all these rules, giving them a p-value of zero, which hints at their possible significance, but fails at prioritizing the combinations of genes that may be relevant to the Triple-Negative phenotype.

4.5.3 Rule Discovery on Breast Cancer Data

In this Section we provide more details on how we built the datasets, the parameters we used in our analyses with ALLSTAR, the results we obtained, and their biological relevance. We ran ALLSTAR on breast cancer data described in Section 4.5.1, split into treatments **X**, confounders **Z**, and outcomes Y. In each run, we considered a common set of confounders, while **X** and Y are combined in different ways in each run to focus on certain cancer mechanisms. In particular:

- We considered 7 confounder variables **Z**: gender, race, age at diagnosis, menopause status, history of another previous malignancy, and the presence of a germline mutation in genes BRCA1 and BRCA2.
- We selected a total of 622 alterations, which include the 300 most frequently somatically mutated genes, the 300 most frequent LOHs, and 22 frequently hypermethylated genes. We performed two types of analyses: one where each alteration corresponds to a treatment (element of **X**), and one where we considered as treatments **X** the 300 most frequently altered genes by considering a gene mutated if any of the 3 alterations above is present.
- As target Y, we considered three sub-typing classifications: an histological categorization (Ductal, Lobular, and Other carcinoma), an expanded molecular one, based on gene expression (Basal, HER2E, Luminal-A, Luminal-B, and Normal-like), and a specific binary molecular classification (Triple-Negative, or not).

We tested ALLSTAR under multiple settings on the differently combined datasets: we set the maximum rule length ℓ from 2 to 4, $k = 3$, and $t = 0.01$. Data requirements increase exponentially with the size of **Z**, and, therefore, for each dataset we run ALLSTAR multiple times each time passing a different subset of **Z** of cardinality at most 1. On each run, we set $\alpha = 0.05/(|\mathbf{Z}| + 1)$ to bound the FWER of all the tests on the same dataset below 0.05. Finally, we took into consideration both the presence and the absence of treatment. Table 4.1 shows the best rules with no confounders, at the top, and when conditioning on confounders, at the bottom.

The first three rules by effect include gene CDH1, which is a recurrently mutated gene in breast cancer and whose impact has been recognized as substantial [Pereira et al., 2016] in lobular histological subtype [McCart Reed et al., 2021, Erber and Hartmann, 2020], consistent with rules a and b; consistently, rule c states that the absence of an alteration in CDH1, given a breast cancer diagnosis, increases the chances of developing a ductal subtype, antagonist to the lobular one. Moreover, the combination of mutated CDH1 with unaltered ANK2 and SCN5A (rule b) provides an additional perspective on the mechanisms regulating the lobular subtype: ANK2 is typically downregulated in breast cancer, while SCN5A is upregulated in almost every neoplastic process. However, SCN5A is known to mediate the epithelial–mesenchymal transition (EMT), a biological trait underpinning cancer aggressiveness: the absence of a mutation in this gene can be interpreted as a normal state for EMT, aligned with the mild characteristics of the lobular subtype [Gradek et al., 2019, Luo et al., 2020]. Additionally, the rule including

ID	Rule	Effect
a	CDH1$_{som}$ = 1 ⟶ Lobular	0.470
b	CDH1$_{som}$ = 1 ∧ ANK2$_{som}$ = 0 ∧ SCN5A$_{som}$ = 0 ⟶ Lobular	0.430
c	CDH1$_{som}$ = 0 ⟶ Ductal Carcinoma	0.401
d	ITGB3$_{alt}$ = 1 ∧ RHOA$_{alt}$ = 1 ∧ MAP3K1$_{alt}$ = 1 ⟶ Basal	0.342
e	TP53$_{som}$ = 1 ∧ ATRIP$_{loh}$ = 1 ∧ ERBB2$_{loh}$ = 1 ⟶ Basal	0.300
f	ITGB3$_{alt}$ = 1 ∧ MAP3K1$_{alt}$ = 1 ⟶ Basal	0.297
g	TP53$_{som}$ = 0 ⟶ Luminal-A	0.289
h	RB1$_{loh}$ = 1 ∧ PHB$_{loh}$ = 1 ∧ LIMD1$_{loh}$ = 1 ⟶ Basal	0.271
i	TP53$_{som}$ = 0 ∧ BRCA1$_{meth}$ = 0 ⟶ Luminal-A	0.268
j	ERBB2$_{alt}$ = 1 ∧ MST1$_{alt}$ = 1 ⟶ Basal	0.245
k	MST1$_{loh}$ = 1 ∧ ERBB2$_{loh}$ = 1 ⟶ Basal	0.243
l	RB1$_{loh}$ = 1 ∧ PHB$_{loh}$ = 1 ⟶ Basal	0.242
m	STAT3$_{alt}$ = 1 ∧ ERBB2$_{alt}$ = 1 ∧ WNT5A$_{alt}$ = 1 ⟶ Basal	0.241
n	TP53$_{alt}$ = 1 ∧ RB1$_{alt}$ = 1 ∧ NGFR$_{alt}$ = 1 ⟶ Basal	0.241
o	PIK3CA$_{som}$ = 0 ∧ RHOA$_{loh}$ = 1 ∧ NGFR$_{loh}$ = 1 ⟶ Basal	0.240
p	TP53$_{som}$ = 1 ∧ NME1$_{loh}$ = 1 ⟶ Basal	0.230
q	TP53$_{loh}$ = 1 ∧ PRKCD$_{loh}$ = 1 ∧ NME1$_{loh}$ = 1 ⟶ Basal	0.226
r	PDX1$_{alt}$ = 1 ∧ SPOP$_{alt}$ = 1 ⟶ Basal	0.203
s	TP53$_{som}$ = 1 ∧ ERBB2$_{loh}$ = 1 ∧ PRKCD$_{loh}$ = 1 ⟶ TripleN	0.195
t	ITGB3$_{alt}$ = 1 ∧ RHOA$_{alt}$ = 1 ∧ MAP3K1$_{alt}$ = 1 ⟶ TripleN	0.184
u	TP53$_{som}$ = 0 ∧ BRCA2$_{som}$ = 0 ⟶ Luminal-A | gender	0.243
v	CDH1$_{som}$ = 1 ∧ AKT1$_{som}$ = 0 ⟶ Lobular | age_at_diagnosis	0.229
w	ERBB2$_{alt}$ = 1 ∧ RHOA$_{alt}$ = 1 ⟶ Basal | BRCA2$_{germ}$	0.202
x	TP53$_{som}$ = 0 ∧ BRCA2$_{som}$ = 0 ∧ BRCA1$_{meth}$ = 0 ⟶ Luminal-A | gender	0.197
y	TP53$_{som}$ = 0 ∧ RB1$_{som}$ = 0 ∧ BRCA1$_{meth}$ = 0 ⟶ Luminal-A | BRCA2$_{germ}$	0.191
z	TP53$_{som}$ = 1 ∧ ERBB2$_{loh}$ = 1 ⟶ Basal | history_other_malignancy	0.175

Table 4.1: Best rules, with (bottom) and without (top) confounder's conditioning, ordered by descending effect. Rules' description is as follows: GENE1$_{alterationtype}$ = [0,1] ∧ GENE2$_{alterationtype}$ = [0,1]∧ . . . ⟶ Target subtype | Confounder.

the absence of mutation in AKT1 in lobular carcinoma (rule *v*) is coherent, since this gene is strongly associated with ductal differentiation [Hinz and Jücker, 2019]. As a plus, this rule is strengthened by the conditioning on the confounder "age at diagnosis", which removes spurious correlations.

When considering a gene altered in the presence of either a somatic, LOH or hypermethylation, strong effects are linked to the molecular basal-like subtype. ALLSTAR reports the combination of aberrations occurring in ITGB3 and MAP3K (rule *f*) as strongly causal of the aforementioned subtype, in agreement with literature: Fuentes et al. [2020] and Li et al. [2022] converge on this conclusion due to their cancer-promoting activity and inclusion in the metastatic process. ITGB3 and MAP3K have recently gained attention relatively to basal-like breast cancer, but their combination is yet to be investigated. Even more interesting is the extension of this causal rule with the alteration of RHOA (rule *d*): the higher score of this expanded rule in association with basal-like subtype can be explained by the association of the outcome with precocious metastasization in accordance to RHOA's anti-metastatic function [Kalpana et al., 2021].

Even more specific mechanisms are retrieved by considering the decomposed treatments. Besides the strong positive effect of mutated TP53, which is well ascertained in non-luminal breast cancer [Abubakar et al., 2019], even more relevant is the causal effect increase in combination with the LOH event in ATRIP (rule *e*). When stable, this gene is responsible for anti-proliferative signal mediation [Venere et al., 2007], but its impairment's effect is not well established in the literature. The interaction with mutated TP53 is interesting but it needs further investigation.

Another combination strongly rooted in literature is LOH in RB1 and PHB (rule *l*), as explained by Wang et al. [1999]: RB1 is an important tumour suppressor gene [Herschkowitz et al., 2008], while PHB mediates anti-proliferation signalling [Jupe et al., 1996, Sato et al., 1992], therefore their combined action, if altered, is easily explainable in basal-like tumours. The addition of LOH in LIMD1 (rule *h*) is less established in breast cancer, being more associated with lung carcinoma, but its oncosuppressive role, and the correlation between LOH and mitosis, make it a potential key player in basal subtype [Huggins and Andrulis, 2008].

HER2-positive, basal-like, and triple-negative breast cancer are consistently determined by aberrations occurring in MST1 [Jin et al., 2021]. Our findings (rules j, k) coherently overlap this knowledge, extending it by pairing MST1 and ERBB2 within the same positively-scored rule. ERBB2 is a member of the epidermal growth factor (EGF) receptor family and its overexpression in 20-30% of invasive breast carcinomas leads to increased chemoresistance to certain chemotherapeutic agents [Tan and Yu, 2007]. Its mutational impact is undefined in literature, as only ERBB2's expression abnormalities have been encountered in breast malignancies, especially in triple-negative/basal-like. Our result in this particular case is partially coherent but can enable further studies into the MST1-ERBB2 interaction in terms of mammalian carcinoma profiling. Conversely, the joint action between ERBB2, STAT3, and WNT5A (rule *m*) is more explainable. STAT3 has a pivotal role in the initiation, progression, metastasis, and immune evasion of triple-negative breast cancer [Qin et al., 2019], while WNT5A reduces the clonogenicity, invasiveness, migration, and proliferation of carcinoma cells, and it is also considered a therapeutic target [Kobayashi et al., 2018]. The rule that ALLSTAR returned is not specific, as it emerged from the aggregated dataset, but it suggests a strong mutational involvement of these three genes in basal breast cancer. This being said, ERBB2 is recalled in rule *w* with RHOA: both genes offer potential reasons to be partnering in the determination of Basal subtype, but there is no clinical evidence of their combination, let alone an involvement of BRCA2 germline mutation as a confounder to condition over. This rule is a clear example of potential relations that need to be evaluated in future studies. It is not a surprise that various rules with a high conditional effect, which is one of the main contributions of choosing a causal approach, are related to one of the most debated genes, ERBB2, suggesting its direct involvement in breast cancer carcinogenesis (see also rule z in combination with TP53).

An additional point favourable to our methodology consists in the rules we have *not* found: long genes, such as TTN, HMCN1, or DMD, usually harbor several mutations simply due to their size. ALLSTAR seems robust to this drawback, even if those genes are in the top-20 of the most somatically mutated ones in TCGA data. As a term of comparison, Saravia et al. [2019] perform a chi-square test to detect meaningful mutations in triple-negative breast cancer, identifying TTN, HMCN1, and DMD, among others, as statistically significant players in recurrent patterns of genomic alterations with a potential contribution to tumour evolution. The authors themselves acknowledge the possibility their findings may be false positives and our results support this hypothesis.

As a further functional evaluation, we considered, for each analysis, the set of genes obtained by merging the alterations reported in any of the rules from ALLSTAR, and performed pathway enrichment analysis with DAVID [Huang et al., 2009] to find statistically overrepresented biological functions (encoded in KEGG database, Kanehisa and Goto [2000]) in each of these sets of genes. We selected 0.05 as the significance cut-off for pathways' p-value. We then counted the occurrence of each pathway, when significantly enriched, over all the results of the analyses we ran. The most represented pathway, occurring in 65% of the sets, is neurotrophin signaling pathway, whose relevance as a potential therapeutic target for breast cancer has been previously ascertained in preclinical studies [Hondermarck, 2012]. Interestingly, the breast cancer pathway (*KEGG: hsa05224*) occupies one of the top spots with a 57% of occurrence, alongside other known relevant pathways such as Rap1 and PI3K-Akt signaling [Zhang et al., 2017]. Additionally, the fluid shear stress and atherosclerosis (*KEGG: hsa05418*) scored an occurrence of 60%: the impact of this process in breast cancer, and in oncogenesis in general, is still unclear. However, this result seems to endorse some preliminary findings: according to Choi et al. [2019], in addition to promoting hematopoietic growth, biomechanical forces seem to be significant microenvironmental variables in the generation of cancer stem-like cells (CSLCs) or tumour-initiating cells (TICs) in cancer metastasis.

4.5.4 Results Stability

Lastly, we evaluated the stability of our discoveries by using a high confidence subnetwork of Reactome's Protein-Protein Interaction (PPI) (see Section 4.5.1), as input knowledge. Conveniently, the original PPI from Reactome is featured with a score ($s \in [0, 1]$) for each pair of genes, representing the confidence of their edge in the interaction network. To build the experiment, we removed every pair with a score lower than 1, thus keeping only high confidence links, and run ALLSTAR with the same data inputs and parameters as described in Section 4.5.3, with the exception of the PPI. We then compared the results obtained running our algorithm with the two PPIs. Keeping Table 4.1 as reference, a total of 7 rules out of 26 (27% of the reference), and specifically, rules *b*, *h*, *l*, *m*, *r*, *u*, and *v*, were not retrieved in this analysis. These results show that most of the rules found by ALLSTAR including lower confidence interactions, are still reported using only high-confidence interactions. Moreover, the excluded rules *r* and *u* were not extensively characterized by our oncologist due to lack of literature support. Conversely, rules *b*, *h*, *l*, *m*, and *v* were labelled as potentially novel discoveries: as motivated in Section 4.5.3, these rules refer to proteins whose impact on breast cancer is debated. Even if their role in breast cancer physiology is not specifically supported by sufficient literature, the underlying biological mechanisms are explainable, either because of their genetic properties and functionalities, or the existence of an analogous biological process in other cancer types. Overall, these results show that ALLSTAR can focus on well-characterized mechanisms by including only high-confidence interactions, but also that ALLSTAR can be used in to pinpoint potential novel discoveries by including lower-confidence interactions.

4.6 Conclusions

We introduced ALLSTAR, a novel tool to identify causal relations between somatic alterations and cancer phenotypes from mutational data measured in large cohorts of cancer patients, in contrast to previous approaches focusing on correlations. Our tool reports rules defined on several interactions and integrates prior information in the form of a graph to focus on functionally related alterations. It also uses an iterative procedure to identify diverse rules to tackle inter-tumour heterogeneity. Our extensive experimental evaluation shows that ALLSTAR is an efficient and effective tool and that it identifies well-supported causal relations from cancer data.

Chapter 5

Conclusions

Within this Section, we provide an overview of the contributions made in this Thesis, highlighting for each one possible uses, extensions, and research directions.

In Chapter 3 we studied the problem of local causal discovery with statistical guarantees on the returned output. In particular, we focused on controlling the Family-Wise Error Rate, that is the probability of returning at least one false positive in output, as it is a relevant measure that received scant attention in the causal discovery community. Initially, we proved that state-of-the-art algorithms cannot be adapted for such task due to unfeasible and untestable assumptions on the power of the independence tests used in the discovery processes. We then introduced two novel algorithms with provable guarantees, `RAveL-PC` and `RAveL-MB`, for the discovery of the Parent-Children set and the Markov boundary, respectively. Our algorithms rely on Rademacher averages to provide such guarantees on their outputs, which is a novelty in the causal discovery field. Additionally, we introduced two new test statistics for the task. We then evaluated the performances of our algorithms on synthetic datasets, empirically showing the correctness of our claims and comparing our algorithms with two modified versions that exploit the Bonferroni correction for multiple hypotheses testing. Our analysis shown how these latter versions are to be preferred in scenarios with few samples and few variables, while it is preferable to use the standard versions when dealing with high-dimensional problems and lot of samples. Lastly, we ran our algorithms on two real-world datasets obtaining results that are sound with scientific literature and prior knowledge.

`RAveL` algorithms can be exploited in dataset exploration phases in which false discoveries have a significant impact. One such example is clinical trials design, given that a false discovery in this field might lead to the development of useless or even harmful drugs. Our algorithms rely on Rademacher averages which may provide speed-ups in the causal discovery when the test statistic distribution is unknown, as typical solutions to overcome this issue rely on permutation testing which is a computationally intensive task. We therefore advise the use of `RAveL` algorithms in such contexts. Finally, other interesting research directions include exploiting

RAveL for the global causal discovery task with guarantees, as they rely on local causal discovery functions as primitives to perform the whole network inference.

In Chapter 4 we then studied the effect estimation problem with guarantees on the Family-Wise Error Rate on the returned output, as estimating multiple probabilities from datasets may result in false positives due to finite sample issues and the underlying multiple hypotheses problem. In our work we used the syntax of causal rules to represent combination of treatments taking specific values and influencing a target variable, and for each rule we assigned a score that is the average treatment effect that the specific combination of assignments exerts on the target w.r.t. any other combination. Firstly, we proved that the underlying problem of mining the rule with the highest effect is NP-hard even if the probability distributions are known. We then developed ALLSTAR, a branch-and-bound algorithm to discover the top-k causal rules with the highest effect by controlling the Family-Wise Error Rate of the rules returned in output. The algorithm takes in input a graph $\mathcal{G}$ that encodes the set of feasible rules to study, as we may want to use our prior knowledge to exclude certain combinations of treatments from the discovery. ALLSTAR also accepts in input a parameter ℓ used to study only rules composed of up to ℓ treatments, and another t that controls the diversity of the rules returned in output. All these input parameters make ALLSTAR a flexible tool for exploratory data analysis in multiple scenarios, especially biological ones. We evaluated the performances of our algorithm on synthetic datasets, and we show its ability to retrieve results that are different w.r.t. statistical (but not causal) methods commonly used in the biological field. We confirmed such discovery with experiments on real-world breast cancer data, where classical methods are not able to differentiate treatment effects as effectively as ALLSTAR. Lastly, we compared the discovery results with breast cancer literature, assessing ALLSTAR ability to retrieve both well-supported results and novel biologically sound rules.

Our algorithm can be exploited in multiple scenarios for which we want to assess the causal impact of multiple treatments on a specific target variable, not only on breast cancer data. Additionally, while ALLSTAR reports in output rules that are *conjunctions* of boolean propositions (i.e. treatment taking specific values), it would be interesting to design an algorithm that is able to report also *disjunctions* of propositions. Alternatively, future research may focus on improving ALLSTAR performances. In particular, our tool exploits the Bonferroni correction to properly control the Family-Wise Error Rate, which may be overly conservative in scenarios with lot of hypotheses (i.e. rules). An alternative approach for providing the same guarantees may exploit Rademacher averages, as described in Chapter 3. Lastly, ALLSTAR returns significant results only when analysing just a few confounders, as data requirements increase exponentially with the number of confounders under study. An interesting research direction is to study suitable techniques to tackle such issue, possibly by exploiting clustering techniques or by encoding confounders information in a lower dimensional space, in a similar fashion of autoencoders or other dimensionality reduction algorithms.

Chapter 6

Appendix

6.1 Variables in Boston housing dataset

Variables description follows from the paper describing the dataset Harrison Jr and Rubinfeld [1978].

Variable name	Explanation
CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	Proportion of non-retail business acres per town.
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX	Nitric oxides concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centres
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate per \$10,000
PTRATIO	Pupil-teacher ratio by town
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in \$1000's

6.2 Framingham dataset

Dataset and variable description are taken from `https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression`. Variables "CurrentSmoker", "PrevalentStroke", "PrevalentHyp", and "Diabetes" were removed in the data preprocessing phase.

Variable name	Explanation
Age	Age of the patient (Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
Current Smoker	Whether or not the patient is a current smoker (Nominal)
Cigs Per Day	The number of cigarettes that the person smoked on average in one day. (can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
BP Meds	Whether or not the patient was on blood pressure medication (Nominal)
Prevalent Stroke	Whether or not the patient had previously had a stroke (Nominal)
Prevalent Hyp	Whether or not the patient was hypertensive (Nominal)
Diabetes	Whether or not the patient had diabetes (Nominal)
Tot Chol	Total cholesterol level (Continuous)
Sys BP	Systolic blood pressure (Continuous)
Dia BP	Siastolic blood pressure (Continuous)
BMI	Body Mass Index (Continuous)
Heart Rate	Heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
Glucose	Glucose level (Continuous)
10 year risk of coronary heart disease CHD	Binary: “1”, means “Yes”, “0” means “No”

6.3 Equations of Section 4.4.4’s Experiment

The equations for sampling data from the graph of Figure 4.4(a) follows:

$$\begin{aligned}
X_1 &\sim \mathcal{B}(0.5) \\
X_2 &\sim \mathcal{B}(0.4) \\
X_3 &\sim \mathcal{B}(0.7) \\
X_4 &\sim \mathcal{B}(0.65) \\
X_5 &\sim \mathcal{B}(0.15) \\
X_{i(clone)} &\sim X_i \oplus \mathcal{B}(0.025) \\
E_i &\sim \mathcal{B}(0.5) \\
Y &\sim (X_1 \wedge X_2) \vee (X_3 \wedge X_4) \vee X_5
\end{aligned}$$

Bibliography

Mustapha Abubakar, Changyuan Guo, Hela Koka, Hyuna Sung, Nan Shao, Jennifer Guida, Joseph Deng, Mengjie Li, Nan Hu, Bin Zhou, et al. Clinicopathological and epidemiological significance of breast cancer subtype reclassification based on p53 immunohistochemical expression. *NPJ breast cancer*, 5(1):20, 2019.

Constantin F Aliferis, Ioannis Tsamardinos, and Alexander Statnikov. Hiton: a novel markov blanket algorithm for optimal variable selection. In *AMIA annual symposium proceedings*, volume 2003, page 21. American Medical Informatics Association, 2003.

Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(1), 2010.

Angelos P Armen and Ioannis Tsamardinos. Estimation and control of the false discovery rate of bayesian network skeleton identification. Technical report, Technical report, TR-441. University of Crete, 2014.

Claudia Arnedo-Pac, Loris Mularoni, Ferran Muiños, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. Oncodriveclustl: a sequence-based clustering method to identify cancer drivers. *Bioinformatics*, 35(22):4788–4790, 2019.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Alexis Bellot and Mihaela van der Schaar. Conditional independence testing using generative adversarial networks. *Advances in neural information processing systems*, 32, 2019.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

Concha Bielza and Pedro Larrañaga. Bayesian networks in neuroscience: a survey. *Frontiers in computational neuroscience*, 8:131, 2014.

Weston R Bodily, Brian H Shirts, Tom Walsh, Suleyman Gulsuner, Mary-Claire King, Alyssa Parker, Moom Roosan, and Stephen R Piccolo. Effects of germline and somatic events in candidate brca-like genes on breast-tumor signatures. *PLoS One*, 15(9):e0239197, 2020.

Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

Kailash Budhathoki, Mario Boley, and Jilles Vreeken. Discovering reliable causal rules. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 1–9. SIAM, 2021.

Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. Long story short: Omitted variable bias in causal machine learning. Technical report, National Bureau of Economic Research, 2022.

Hye Yeon Choi, Gwang-Mo Yang, Ahmed Abdal Dayem, Subbroto Kumar Saha, Kyeongseok Kim, Youngbum Yoo, Kwonho Hong, Jin-Hoi Kim, Cassian Yee, Kyung-Mi Lee, et al. Hydrodynamic shear stress promotes epithelial-mesenchymal transition by downregulating erk and gsk3β activities. *Breast Cancer Research*, 21(1):1–20, 2019.

Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3): 213–219, 2013.

Andres M Cifuentes-Bernal, Vu VH Pham, Xiaomei Li, Lin Liu, Jiuyong Li, and Thuc Duy Le. Dynamic cancer drivers: a causal approach for cancer driver discovery based on bio-pathological trajectories. *Briefings in Functional Genomics*, 21(6):455–465, 2022.

John Concato, Nirav Shah, and Ralph I Horwitz. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England journal of medicine*, 342(25):1887–1892, 2000.

Isidro Cortés-Ciriano, Doga C Gulhan, Jake June-Koo Lee, Giorgio EM Melloni, and Peter J Park. Computational analysis of cancer genome sequencing data. *Nature Reviews Genetics*, 23(5):298–314, 2022.

Lenore Cowen, Trey Ideker, Benjamin J Raphael, and Roded Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18(9):551–562, 2017.

Ramona Erber and Arndt Hartmann. Histology of luminal breast cancer. *Breast Care*, 15(4):327–336, 2020.

Pedro Fuentes, Marta Sesé, Pedro J Guijarro, Marta Emperador, Sara Sánchez-Redondo, Héctor Peinado, Stefan Hümmer, and Santiago Ramón y Cajal. Itgb3-mediated uptake of small extracellular vesicles facilitates intercellular communication in breast cancer cells. *Nature communications*, 11(1):1–15, 2020.

Mary Goldman, Brian Craft, Teresa Swatloski, Melissa Cline, Olena Morozova, Mark Diekhans, David Haussler, and Jingchun Zhu. The ucsc cancer genomics browser: update 2015. *Nucleic acids research*, 43(D1):D812–D817, 2015.

Frédéric Gradek, Osbaldo Lopez-Charcas, Stéphanie Chadet, Lucile Poisson, Lobna Ouldamer, Caroline Goupille, Marie-Lise Jourdan, Stéphan Chevalier, Driffa Moussata, Pierre Besson, et al. Sodium channel na v 1.5 controls epithelial-to-mesenchymal transition and invasiveness in breast cancer cells through its regulation by the salt-inducible kinase-1. *Scientific reports*, 9(1):18652, 2019.

David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

Jason I Herschkowitz, Xiaping He, Cheng Fan, and Charles M Perou. The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal b breast carcinomas. *Breast Cancer Research*, 10(5):1–13, 2008.

Nico Hinz and Manfred Jücker. Distinct functions of akt isoforms in breast cancer: a comprehensive review. *Cell Communication and Signaling*, 17: 1–29, 2019.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

Hubert Hondermarck. Neurotrophins and their receptors in breast cancer. *Cytokine & growth factor reviews*, 23(6):357–365, 2012.

Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57, 2009.

Christopher Jack Huggins and Irene L Andrulis. Cell cycle regulated phosphorylation of limd1 in cell lines and expression in human breast cancers. *Cancer letters*, 267(1):55–66, 2008.

Xiang Jin, Lihua Zhu, Sheng Xiao, Zhuhong Cui, Jing Tang, Jiangyong Yu, and Mingjun Xie. Mst1 inhibits the progression of breast cancer by regulating the hippo signaling pathway and may serve as a prognostic biomarker. *Molecular Medicine Reports*, 23(5):1–12, 2021.

Eldon R Jupe, Xiao-Tie Liu, Julie L Kiehlbauch, J Keith McClung, and RT Dell'Orco. Prohibitin in breast cancer cell lines: loss of antiproliferative activity is linked to 3'untranslated region mutations. *Cell Growth Differ.*, 3:4, 1996.

Gardiyawasam Kalpana, Christopher Figy, Jingwei Feng, Claire Tipton, Julius N De Castro, Vu N Bach, Clariza Borile, Alexandria LaSalla, Hussain N Odeh, Miranda Yeung, et al. The rhoa dependent anti-metastatic function of rkip in breast cancer. *Scientific reports*, 11(1):1–14, 2021.

Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2): 81–93, 1938.

Yoshie Kobayashi, Takayuki Kadoya, Ai Amioka, Hideaki Hanaki, Shinsuke Sasada, Norio Masumoto, Hideki Yamamoto, Koji Arihiro, Akira Kikuchi, and Morihito Okada. Wnt5a-induced cell migration is associated with the aggressiveness of estrogen receptor-positive breast cancer. *Oncotarget*, 9 (30), 2018.

Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer, 2000.

Adam A Kraya, Kara N Maxwell, Bradley Wubbenhorst, Brandon M Wenz, John Pluta, Andrew J Rech, Liza M Dorfman, Nicole Lunceford, Amanda Barrett, Nandita Mitra, et al. Genomic signatures predict the immunogenicity of brca-deficient breast cancerimmunogenetic signatures of brca1/2 breast cancer. *Clinical Cancer Research*, 25(14):4363–4374, 2019.

Matt J Kusner and Joshua R Loftus. The long road to fairer algorithms. *Nature*, 578(7793):34–36, 2020.

Cheukfai Li, Guochun Zhang, Yulei Wang, Bo Chen, Kai Li, Li Cao, Chongyang Ren, Lingzhu Wen, Minghan Jia, Hsiaopei Mok, et al. Spectrum of map3k1 mutations in breast cancer is luminal subtype-predominant and related to prognosis. *Oncology Letters*, 23(2):1–12, 2022.

Junning Li and Z Jane Wang. Controlling the false discovery rate of the association/causality structure learned with the pc algorithm. *Journal of Machine Learning Research*, 10(2), 2009.

Aiping Liu, Junning Li, Z Jane Wang, Martin J McKeown, et al. A computationally efficient, exploratory approach to brain connectivity incorporating false discovery rate control, a priori knowledge, and group inference. *Computational and mathematical methods in medicine*, 2012, 2012.

Yijun Liu, Ji Sun, Huiyan Sun, and Yi Chang. Identification of key somatic oncogenic mutation based on a confounder-free causal inference model. *PLoS Computational Biology*, 18(9):e1010529, 2022.

Qianxuan Luo, Ting Wu, Wenfang Wu, Gong Chen, Xuan Luo, Liping Jiang, Huai Tao, Mingqiang Rong, Shuntong Kang, and Meichun Deng. The functional role of voltage-gated sodium channel nav1. 5 in metastatic breast cancer. *Frontiers in Pharmacology*, 11:1111, 2020.

Sisi Ma and Roshan Tourani. Predictive and causal implications of using shapley value for model interpretation. In *Proceedings of the 2020 KDD workshop on causal discovery*, pages 23–38. PMLR, 2020.

Mehrdad Mansouri, Sahand Khakabimamaghani, Leonid Chindelevitch, and Martin Ester. Aristotle: stratified causal discovery for omics data. *BMC bioinformatics*, 23(1):1–18, 2022.

Elaine R Mardis. The impact of next-generation sequencing on cancer genomics: from discovery to clinic. *Cold Spring Harbor Perspectives in Medicine*, 9(9), 2019.

Amy E McCart Reed, Samuel Foong, Jamie R Kutasovic, Katia Nones, Nicola Waddell, Sunil R Lakhani, and Peter T Simpson. The genomic landscape of lobular breast cancer. *Cancers*, 13(8):1950, 2021.

Vishwali Mhasawade and Rumi Chunara. Causal multi-level fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 784–794, 2021.

Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

Loris Mularoni, Radhakrishnan Sabarinathan, Jordi Deu-Pons, Abel Gonzalez-Perez, and Núria López-Bigas. Oncodrivefml: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome biology*, 17(1):1–13, 2016.

Richard E Neapolitan et al. *Learning bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, 2004.

Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160, 2009.

Judea Pearl. *Causality: models, reasoning and inference.* Cambridge University Press, 2 edition, 2009.

Dana Pe'er. Bayesian network analysis of signaling networks: a primer. *Science's STKE*, 2005(281):pl4–pl4, 2005.

Leonardo Pellegrina and Fabio Vandin. Silvan: estimating betweenness centralities with progressive sampling and non-uniform rademacher bounds. *arXiv preprint arXiv:2106.03462*, 2021.

Leonardo Pellegrina, Cyrus Cousins, Fabio Vandin, and Matteo Riondato. Mcrapper: Monte-carlo rademacher averages for poset families and approximate pattern mining. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–29, 2022.

Jose M Pena, Roland Nilsson, Johan Björkegren, and Jesper Tegnér. Towards scalable and data efficient learning of markov boundaries. *International Journal of Approximate Reasoning*, 45(2):211–232, 2007.

Bernard Pereira, Suet-Feung Chin, Oscar M Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A Bardwell, Michelle Pugh, Linda Jones, Roslin Russell, Stephen-John Sammut, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature communications*, 7(1):1–16, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms.* The MIT Press, 2017.

Jiang-Jiang Qin, Li Yan, Jia Zhang, and Wei-Dong Zhang. Stat3 as a potential therapeutic target in triple negative breast cancer: a systematic review. *Journal of Experimental & Clinical Cancer Research*, 38(1):1–16, 2019.

Matthew A Reyna, David Haan, Marta Paczkowska, Lieven PC Verbeke, Miguel Vazquez, Abdullah Kahraman, Sergio Pulido-Tamayo, Jonathan Barenboim, Lina Wadi, Priyanka Dhingra, et al. Pathway and network analysis of more than 2500 whole cancer genomes. *Nature communications*, 11(1):1–17, 2020.

Nadeem Riaz, Pedro Blecua, Raymond S Lim, Ronglai Shen, Daniel S Higginson, Nils Weinhold, Larry Norton, Britta Weigelt, Simon N Powell, and Jorge S Reis-Filho. Pan-cancer analysis of bi-allelic alterations in homologous recombination dna repair genes. *Nature communications*, 8(1):857, 2017.

Matteo Riondato and Eli Upfal. Mining frequent itemsets through progressive sampling with rademacher averages. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1005–1014, 2015.

Paul R Rosenbaum, PR Rosenbaum, and Briskman. *Design of observational studies*, volume 10. Springer, 2010.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Diego Santoro, Andrea Tonon, and Fabio Vandin. Mining sequential patterns with vc-dimension and rademacher complexity. *Algorithms*, 13(5): 123, 2020.

César H Saravia, Claudio Flores, Luis J Schwarz, Leny Bravo, Jenny Zavaleta, Jhajaira Araujo, Silvia Neciosup, and Joseph A Pinto. Patterns of mutation enrichment in metastatic triple-negative breast cancer. *Clinical Medicine Insights: Oncology*, 13, 2019.

Rebecca Sarto Basso, Dorit S Hochbaum, and Fabio Vandin. Efficient algorithms to discover alterations with complementary functional association in cancer. *PLoS computational biology*, 15(5), 2019.

Takaaki Sato, Hiroko Saito, Jeff Swensen, Arnold Olifant, Carla Wood, David Danner, Takashi Sakamoto, Kenichi Takita, Fujio Kasumi, Yoshio Miki, et al. The human prohibitin gene located on chromosome 17q21 is mutated in sporadic breast cancer. *Cancer research*, 52(6), 1992.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. 2020.

Craig Silverstein, Sergey Brin, Rajeev Motwani, and Jeff Ullman. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, 4(2):163–192, 2000.

Dario Simionato and Fabio Vandin. Bounding the family-wise error rate in local causal discovery using rademacher averages. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 255–271. Springer, 2022.

Dario Simionato and Fabio Vandin. Bounding the family-wise error rate in local causal discovery using rademacher averages (extended abstract). In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search.* MIT press, 2000.

Eric V Strobl, Peter L Spirtes, and Shyam Visweswaran. Estimating and controlling the false discovery rate of the pc algorithm using edge-specific p-values. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–37, 2019.

Ming Tan and Dihua Yu. Molecular mechanisms of erbb2-mediated breast cancer chemoresistance. *Breast Cancer Chemosensitivity*, pages 119–129, 2007.

The International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010.

Ioannis Tsamardinos and Constantin F Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In *International Workshop on Artificial Intelligence and Statistics*, pages 300–307. PMLR, 2003.

Ioannis Tsamardinos and Laura E Brown. Bounding the false discovery rate in local bayesian network learning. In *AAAI*, pages 1100–1105, 2008.

Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–678, 2003a.

Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov. Algorithms for large scale markov blanket discovery. In *FLAIRS conference*, volume 2, pages 376–380. St. Augustine, FL, 2003b.

Fabio Vandin, Eli Upfal, and Benjamin J Raphael. De novo discovery of mutated driver pathways in cancer. *Genome research*, 22(2):375–385, 2012.

Marina Velikova, Josien Terwisscha Van Scheltinga, Peter JF Lucas, and Marc Spaanderman. Exploiting causal functional relationships in bayesian network modelling for personalised healthcare. *International Journal of Approximate Reasoning*, 55(1):59–73, 2014.

Monica Venere, Andrew Snyder, Omar Zgheib, and Thanos D Halazonetis. Phosphorylation of atr-interacting protein on ser239 mediates an interaction with breast-ovarian cancer susceptibility 1 and checkpoint function. *Cancer research*, 67(13):6100–6105, 2007.

Linbo Wang, Thomas S Richardson, and Xiao-Hua Zhou. Causal analysis of ordinal treatments and binary outcomes under truncation by death. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79 (3):719–735, 2017.

Sheng Wang, Niharika Nath, Matthew Adlam, and Srikumar Chellappan. Prohibitin, a potential tumor suppressor, interacts with rb and regulates e2f function. *Oncogene*, 18(23):3501–3510, 1999.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Guanming Wu, Xin Feng, and Lincoln Stein. A human functional protein interaction network and its application to cancer data analysis. *Genome biology*, 11(5):1–23, 2010.

Farzana Yusuf, Shaoming Cheng, Sukumar Ganapati, and Giri Narasimhan. Causal inference methods and their challenges: the case of 311 data. In *DG. O2021: The 22nd Annual International Conference on Digital Government Research*, pages 49–59, 2021.

Qingyang Zhang, Joanna E Burdette, and Ji-Ping Wang. Integrative network analysis of tcga data for ovarian cancer. *BMC systems biology*, 8:1–18, 2014.

Yi-Lei Zhang, Ruo-Chen Wang, Ken Cheng, Brian Z Ring, and Li Su. Roles of rap1 signaling in tumor cell migration and invasion. *Cancer biology & medicine*, 14(1):90, 2017.