Bounding the Family-Wise Error Rate in Local Causal Discovery using Rademacher Averages

Dario Simionato[0000−0001−5533−425X] and Fabio Vandin(✉)[0000−0003−2244−2320]

Department of Information Engineering, University of Padova, Italy
dario.simionato@phd.unipd.it,fabio.vandin@unipd.it

Abstract. Many algorithms have been proposed to learn local graphical structures around target variables of interest from observational data. The Markov boundary (MB) provides a complete picture of the local causal structure around a variable and is a theoretically optimal solution for the feature selection problem. Available algorithms for MB discovery have focused on various challenges such as scalability and data-efficiency. However, current approaches do not provide guarantees in terms of false discoveries in the MB.
In this paper we introduce a novel algorithm for the MB discovery problem with rigorous guarantees on the Family-Wise Error Rate (FWER), that is, the probability of reporting any false positive. Our algorithm uses Rademacher averages, a key concept from statistical learning theory, to properly account for the multiple-hypothesis testing problem arising in MB discovery. Our evaluation on simulated data shows that our algorithm properly controls for the FWER, while widely used algorithms do not provide guarantees on false discoveries even when correcting for multiple-hypothesis testing. Our experiments also show that our algorithm identifies meaningful relations in real-world data.

Keywords: Local causal discovery · Markov boundary · Rademacher averages · FWER

1 Introduction

One of the most fundamental and challenging problems in science is the discovery of causal relations from observational data [20]. Bayesian networks are a type of 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 [21, 27], medicine [36], and others [37, 13].

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 [30], 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 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 [32], that is, it is the minimal set of variables with optimal predictive performance for T.

In several real-world applications, such as biology [27] and neuroscience [8], 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 [17, 13], 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, currently 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).

Our contributions. In this paper we introduce a novel algorithm, Radamacher Averages for Local structure discovery, or `RAveL-MB`, for the MB discovery problem with rigorous guarantees on the FWER. Our `RAveL-MB` uses a novel algorithm, `RAveL-PC`, that we developed for the identification of the PC of a target variable while bounding the FWER. 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 [4], 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. 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 paper is organized as follows. Section 2 introduces the preliminary concepts used in the rest of the paper. Section 3 describes previous works related to our contribution. Section 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 5 describes our experimental evaluation on synthetic and real data. Finally, Section 6 offers some concluding remarks.

2 Preliminaries

In this section, we introduce basic notions and preliminary concepts used in the rest of the paper. More specifically, in Section 2.1 we formally define Bayesian networks (BNs) and the sets $PC(T)$ and $MB(T)$ for a target variable T. In Section 2.2 we describe the statistical testing procedure commonly used by algorithms for the identification of $PC(T)$ and $MB(T)$. In Section 2.3 we introduce the multiple hypotheses testing problem and the family-wise error rate (FWER). Finally, in Section 2.4 we introduce the concept of Rademacher averages for supremum deviation estimation.

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 1 (Bayesian network [19]). *Let p be a joint probability distribution over $\mathbf{V}$. Let $G = (\mathbf{W}, \mathbf{A})$ be a DAG where the vertices $\mathbf{W}$ of 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 $\langle \boldsymbol{V}, G, p \rangle$.*

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

Definition 2 (Faithfulness [30]). *A directed acyclic graph 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 G and the Markov Condition. A distribution p is* faithful *if and only if there exists a DAG G such that 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. $X \rightarrow Y$ or $X \leftarrow Y$) in G. In particular, the *directional separation*, or *d-separation* [20], 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 $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$. Viceversa, if the two sets $\mathbf{X}$ and $\mathbf{Y}$ are conditionally dependent given $\mathbf{Z}$, denoted with $\mathbf{X} \not\perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$, the path is marked as *open*. More formally, the definition of d-separated path is the following.

Definition 3 (d-separation [20]). *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 $\mathbf{Y}$.

A *causal* Bayesian network is a Bayesian network with causally relevant edge semantics [20, 16].

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* $PC(T)$.

Definition 4 (Parent-children set of T [16]). *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 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 $PC(T)$: $X \not\perp\!\!\!\perp T|\mathbf{Z}, \forall \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, T\}$. The second set is the *Markov boundary* $MB(T)$ of a target variable T, defined as follows.

Definition 5 (Markov boundary of T [20, 33]). *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 $\forall Y \in \mathbf{V} \setminus MB(T), Y \neq T, T \perp\!\!\!\perp Y|MB(T)$.*

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 [16], that are those variables $X \in \mathbf{V} \setminus \{T\}$ for which $\exists Y \in PC(T)$ such that $X \perp\!\!\!\perp T|\mathbf{Z}$ and $X \not\perp\!\!\!\perp T|\mathbf{Z} \cup \{Y\}$ for all $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, T\}$. MB is the minimal subset $\mathbf{S} \subseteq \mathbf{V}$ for which $p(T|\mathbf{S})$ is estimated accurately [16, 33], therefore is the optimal solution for feature selection tasks.

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 [29], 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:

$$r_{\mathbf{x},\mathbf{y}} = \frac{\sum_{i=0}^{k} x_i y_i - k\bar{x}\bar{y}}{(k-1)s_{\mathbf{x}}s_{\mathbf{y}}} \tag{1}$$

where $\bar{x}$ and $\bar{y}$ are the sample mean of $\mathbf{x}$ and $\mathbf{y}$, respectively, while $s_{\mathbf{x}}$ 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 and $k-2$ 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., [34, 24]) 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.

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 (MHT), 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* [9], 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).

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 [4, 12], which allow to compute *data-dependent* confidence intervals for *all hypotheses simultaneously*, leading to improved tests for MHT scenarios [22]. 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 $\mathcal{X}$ to $[a, b] \subset \mathbb{R}$ and let S be a sample of m i.i.d. observations from an unknown data generative distribution μ over $\mathcal{X}$. We define the *empirical sample mean* of a function $f \in \mathcal{F}$, $\hat{\mathbb{E}}_S[f]$, and its *expectation* $\mathbb{E}[f]$ as

$$\hat{\mathbb{E}}_S[f] \doteq \frac{1}{m} \sum_{s_i \in S} f(s_i) \text{ and } \mathbb{E}[f] \doteq \mathbb{E}_\mu \left[\frac{1}{m} \sum_{s_i \in S} f(s_i) \right]. \tag{2}$$

Note that $\mathbb{E}[f] = \mathbb{E}_\mu[f]$, that is, the expected value of the empirical mean corresponds to the expectation according to distribution μ. 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) $D(\mathcal{F}, S)$ that is defined as

$$D(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} \left| \hat{\mathbb{E}}_S[f] - \mathbb{E}[f] \right|. \tag{3}$$

Computing $D(\mathcal{F}, S)$ exactly is not possible given the unknown nature of μ, 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

$$\hat{R}(\mathcal{F}, \mathcal{S}) \doteq \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(s_i) \right] \tag{4}$$

where σ 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 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 ([18], 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 σ. The n-samples Monte-Carlo Empirical Rademacher Average (n-MCERA) $\hat{R}^n_m(\mathcal{F},\mathcal{S},\sigma)$ is defined as

$$\hat{R}^n_m(\mathcal{F},\mathcal{S},\sigma)\doteq\frac{1}{n}\sum_{j=1}^{n}\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{s_i\in S}\sigma_{j,i}f(s_i) \tag{5}$$

with σ 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 1 (Th. 3.1 of [22]). *Let $\delta\in(0,1)$. For ease of notation let*

$$\tilde{R}=\hat{R}^n_m(\mathcal{F},\mathcal{S},\sigma)+2z\sqrt{\frac{\ln\frac{4}{\delta}}{2nm}} \tag{6}$$

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

$$D(\mathcal{F},S)\leq 2\tilde{R}+\frac{\sqrt{c(4m\tilde{R}+c\ln\frac{4}{\delta})\ln\frac{4}{\delta}}}{m}+\frac{c\ln\frac{4}{\delta}}{m}+c\sqrt{\frac{\ln\frac{4}{\delta}}{2m}} \tag{7}$$

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

Theorem 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 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 [33, 2, 24, 1], 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. Note however that, as described in Section 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. [24] 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 $IAMB$ [34] 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. [35] introduced the problem of MHT in the context of local causal discovery, and proposed to use the Benjamini-Hochberg correction [6] 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 [3], the authors extended the analysis of [35] from the local discovery task to the BN inference while [14, 15, 31] re-implemented the PC algorithm for BN structure discovery using the Benjamini-Yekutieli [7] 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 [25, 26, 22, 28, 10, 23]). To the best of our knowledge, ours is the first work to introduce their use in (local) causal discovery.

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 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 5). We then present in Section 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 4.3 we describe how Rademacher averages are used by our algorithms for effective independence testing.

4.1 Analysis and limitations of $PCMB$ and $IAMB$

The algorithms presented in Section 3 are correct under the assumption that the independence tests result in no false positive *and* no false negative [24,

34]. In this section we determine milder sufficient conditions that allow $GetPC$ [24] to control the FWER for the PC discovery task, and $PCMB$ [24] and $IAMB$ [34] 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 be corrected for MHT 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*.

We start by proving sufficient conditions for bounding the FWER of the elements returned by $GetPC$ [24]. (All proofs are in the Appendix.)

Theorem 2. *$GetPC(T)$ 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.*

The following proves that similar requirements are needed for $PCMB$ [24] to have guarantees on the FWER.

Theorem 3. *$PCMB(T)$ 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.*

The following result proves analogous requirements for $IAMB$.

Theorem 4. *$IAMB(T)$ 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.*

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 present two such examples in the Appendix. Moreover, our experimental evaluation in Section 5 shows that this situation does happen in practice.

4.2 Algorithms `RAveL-PC` and `RAveL-MB`

As shown in Section 4.1, correcting for 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): 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 1. 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 1: RAveL-MB(T,$\mathbf{V}$,δ)

Input: target variable T, set of variables $\mathbf{V}$, threshold $\delta \in (0, 1]$
Output: A subset of $MB(T)$ with FWER lower than δ.

1 $MB \leftarrow$ RAveL-PC(T,$\mathbf{V}$,δ) ;
2 **foreach** $X \in MB$ **do**
3 **foreach** $Y \in$ *RAveL-PC(X,$\mathbf{V}$,δ)* ***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,$\mathbf{V}$,δ) (shown in Algorithm 2) 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,$\mathbf{V}$,δ) 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.

Both algorithms RAveL-MB and RAveL-PC employ a function, denoted as test_indep(X,Y,$\mathbf{Z}$,δ), 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. Prac-

Algorithm 2: `RAveL-PC(`T`,`$\mathbf{V}$`,`δ`)`

Input: target variable T, set of variables $\mathbf{V}$, threshold $\delta \in (0, 1]$
Output: A subset of $PC(T)$ with FWER lower than δ.

```
1 PC ← V \ {T};
2 foreach X ∈ V \ {T} do
3     foreach Z ⊆ V \ {X, T} do
4         if test_indep(T,X,Z,δ) then
5             PC ← PC \ {X};
6 return PC;
```

tical details on our implementation of `test_indep(`X`,`Y`,`$\mathbf{Z}$`,`δ`)` are provided in Section 4.3.

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

Theorem 5. *`RAveL-PC(`T`,`$\mathbf{V}$`,`δ`)` outputs a set of elements in $PC(T)$ with FWER $\leq \delta$.*

Theorem 6. *`RAveL-MB` outputs a set of elements in $MB(T)$ with FWER $\leq \delta$.*

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.

4.3 Rademacher averages for independence testing

Note that our algorithms `RAveL-PC(`X`,`$\mathbf{V}$`,`δ`)` and `RAveL-MB(`X`,`$\mathbf{V}$`,`δ`)` 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 2.2) and use Bonferroni correction (see Section 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[1] between the variables in $\mathbf{V}$, that is

[1] 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.

$N = |\mathbf{V}|(|\mathbf{V}|-1)2^{|\mathbf{V}|-3}$. Note that the value of N grows 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). 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(`X`,`$\mathbf{V}$`,`δ`)` and `RAveL-MB(`X`,`$\mathbf{V}$`,`δ`)` 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 $r_{X,Y,\mathbf{Z}}$ 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 0, which is the expected value of $r_{X,Y,\mathbf{Z}}$ under the null hypothesis of independence.

To implement the idea described above, we express Eqn. 1 as an additive function on the samples as follows. First, we assume that all variables have been centered around 0 and then normalized by dividing for the maximum absolute value (i.e., $\bar{x} = 0$ and $\max(|\mathbf{x}|) = 1$ for all variables). Let $s_1, s_2, \ldots, s_k$ 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 that, given a sample s_i, provides an estimate of $r_{X,Y,\mathbf{Z}}$ using only s_i

$$r_{X,Y,\mathbf{Z}}(s_i) = k\frac{x_i y_i}{k-1}. \tag{8}$$

Note that 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 $\mathbf{y}$ (see Section 2.2).

Therefore, we have the following modified version of Pearson's coefficient, where $s_{\mathbf{x}}$ is replaced by $\max(|\mathbf{x}|) - \bar{\mathbf{x}}$ (similarly for $s_{\mathbf{y}}$):

$$r_{X,Y,\mathbf{Z}} = \frac{1}{k}\sum_{i=1}^{k} r_{X,Y,\mathbf{Z}}(s_i). \tag{9}$$

By considering the family $\mathcal{F}$ of functions defined by $r_{X,Y,\mathbf{Z}}$ 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. 5) is

$$\hat{R}_k^n(\mathcal{F}, \mathcal{S}, \sigma) \doteq \frac{1}{n}\sum_{j=1}^{n} \sup_{r_{X,Y,\mathbf{Z}} \in \mathcal{F}} \frac{1}{k}\sum_{i=1}^{k} \sigma_{j,i} r_{X,Y,\mathbf{Z}}(s_i). \tag{10}$$

After the n-MCERA has been computed as above, we compute the supremum deviation $D(\mathcal{F}, S)$ according to Theorem 1, which allows us to obtain confidence intervals around the empirical $r_{X,Y,\mathbf{Z}}$ as

$$CI_{X,Y,\mathbf{Z}} = [r_{X,Y,\mathbf{Z}} - D(\mathcal{F}, S), r_{X,Y,\mathbf{Z}} + D(\mathcal{F}, S)] \tag{11}$$

with the guarantee that, *simultaneously* for all $r_{X,Y,\mathbf{Z}} \in \mathcal{F}$, $CI_{X,Y,\mathbf{Z}}$ contains the expected value of $r_{X,Y,\mathbf{Z}}$ 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.

5 Experimental Evaluation

This section describes the experimental evaluation performed to empirically assess our algorithms. In Section 5.1 we compare `RAveL-PC` and `RAveL-MB` performances with other state-of-the-art methods on synthetic data. Section 5.2 presents the analysis on real world data. We implemented[2] `RAveL-PC`, `RAveL-MB`, and the other algorithms considered in this section in Python 3 and R.

5.1 Synthetic data

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 our synthetic data, each variable is a linear combination of its parents values plus a Gaussian noise term. The related structural model includes 13 variables and is shown in the Appendix. 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 4.3.)

For the PC discovery task, we compared two versions of *GetPC* [24], the original one (without any correction for MHT) and a modified version with Bonferroni correction, our algorithm `RAveL-PC` (that uses Rademacher averages as described in Section 4.3), and a variant of `RAveL-PC` that uses Bonferoni correction instead of Rademacher averages for MHT. Figure 1(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 4.2, and we observe that, for the specific BN we consider, algorithm *GetPC* 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* [24] (which uses *GetPC* as subroutine) and of *IAMB* [34], the original one (without any correction for MHT) and a modified version with Bonferroni correction, our algorithm

[2] Code and appendix available at https://github.com/VandinLab/RAveL .

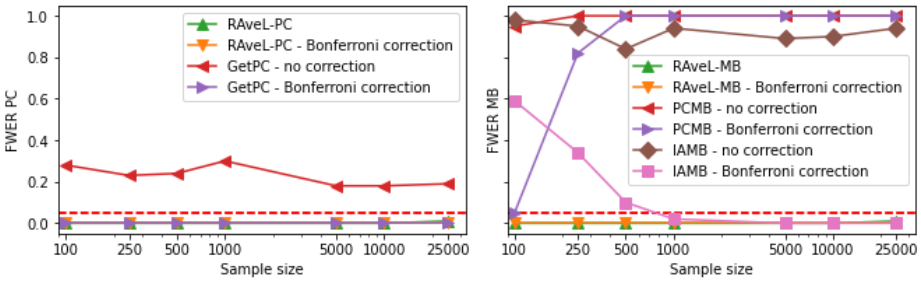

Fig. 1. Empirical FWER of various PC discovery (a) and MB discovery (b) algorithms on synthetic data for different sample sizes. FWER is the fraction of 100 trials in which at least one false positive is reported. The dashed line represents the bound $\delta = 0.05$ to the FWER used in the experiments.

`RAveL-MB`, and a variant of `RAveL-MB` that uses Bonferoni correction instead of Rademacher averages. Figure 1(b) shows the FWER of each method. The results confirms `RAveL-MB` 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 $GetPC$ 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 $PC(T)$.

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 500000 elements by repeating each trial 5 times. We analyzed only the standard version of our algorithms, which is the only practical option in scenarios with a large number of variables (i.e., when the number N of tests is very large). The percentage of false negatives returned by `RAveL-PC` and `RAveL-MB` starts decreasing for datasets with more than 25000 samples, but a simple modification of the test statistic (to be described in the journal version of this paper) greatly improves the performances lowering the data requirement to just 1000 samples. In all such tests our algorithms did not return any false positive.

5.2 Real-world dataset

We tested our algorithms on the Boston housing dataset [11], which contains data about 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 $\delta = 0.01$. 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 neighbourhood. 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. These results provide empirical evidence that our algorithms identify meaningful causal relations while avoiding false positives.

6 Conclusions

In this paper 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 (e.g., based on permutation testing). Interesting research directions include the application of our framework to recently proposed independence tests [5], improving the efficiency of our algorithms, and exploiting them for structure discovery.

Acknowledgements This work is supported, in part, by the Italian Ministry of Education, University and Research (MIUR), under PRIN Project n. 20174LF3T8 "AHeAD" (efficient Algorithms for HArnessing networked Data) and the initiative "Departments of Excellence" (Law 232/2016), and by University of Padova under project "SID 2020: RATED-X".

References

1. Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. JMLR (2010)
2. Aliferis, C.F., Tsamardinos, I., Statnikov, A.: Hiton: a novel markov blanket algorithm for optimal variable selection. In: Proc. of AMIA (2003)
3. Armen, A.P., Tsamardinos, I.: Estimation and control of the false discovery rate of bayesian network skeleton identification. Tech. rep., TR-441. U. of Crete (2014)
4. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. JMLR (2002)
5. Bellot A, van der Schaar M.: Conditional independence testing using generative adversarial networks. Advances in Neural Information Processing Systems. (2019)
6. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. of Royal stat. soc. (1995)
7. Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Annals of statistics (2001)
8. Bielza, C., Larranaga, P.: Bayesian networks in neuroscience: a survey. Frontiers in computational neuroscience (2014)
9. Bonferroni, C.: Teoria statistica delle classi e calcolo delle probabilita. Istituto Superiore di Scienze Economiche e Commericiali di Firenze (1936)
10. Cousins, C., Wohlgemuth, C., Riondato, M.: Bavarian: Betweenness centrality approximation with variance-aware rademacher averages. ACM SIGKDD (2021)

11. Harrison Jr, D., Rubinfeld, D.L.: Hedonic housing prices and the demand for clean air. J. environmental economics and management (1978)
12. Koltchinskii, V., Panchenko, D.: Rademacher processes and bounding the risk of function learning. In: High dimensional probability II (2000)
13. Kusner, M.J., Loftus, J.R.: The long road to fairer algorithms. Nature (2020)
14. Li, J., Wang, Z.J.: Controlling the false discovery rate of the association/causality structure learned with the pc algorithm. JMLR (2009)
15. Liu, A., Li, J., Wang, Z.J., McKeown, M.J.: A computationally efficient, exploratory approach to brain connectivity incorporating false discovery rate control, a priori knowledge, and group inference. Comp. and math. met. in medicine (2012)
16. Ma, S., Tourani, R.: Predictive and causal implications of using shapley value for model interpretation. KDD Workshop on Causal Discovery. PMLR (2020)
17. Mhasawade, V., Chunara, R.: Causal multi-level fairness. Proc. AAAI/ACM Conference on AI, Ethics, and Society (2021)
18. Mitzenmacher, M., Upfal, E.: Probability and computing. Cambridge university press (2017)
19. Neapolitan, R.E., et al.: Learning bayesian networks. Pearson Prentice Hall (2004)
20. Pearl, J.: Causality. Cambridge University Press, 2 edn. (2009)
21. Pe'er, D.: Bayesian network analysis of signaling networks: a primer. Science's STKE (2005)
22. Pellegrina, L., Cousins, C., Vandin, F., Riondato, M.: Mcrapper: Monte-carlo rademacher averages for poset families and approximate pattern mining. ACM SIGKDD (2020)
23. Pellegrina, L., Vandin, F.: Silvan: Estimating betweenness centralities with progressive sampling and non-uniform rademacher bounds. arXiv:2106.03462 (2021)
24. Pena, J.M., Nilsson, R., Björkegren, J., Tegnér, J.: Towards scalable and data efficient learning of markov boundaries. Int. J. of Approximate Reasoning (2007)
25. Riondato, M., Upfal, E.: Mining frequent itemsets through progressive sampling with rademacher averages. ACM SIGKDD (2015)
26. Riondato, M., Upfal, E.: Abra: Approximating betweenness centrality in static and dynamic graphs with rademacher averages. ACM TKDD (2018)
27. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P.: Causal protein-signaling networks derived from multiparameter single-cell data. Science (2005)
28. Santoro, D., Tonon, A., Vandin, F.: Mining sequential patterns with vc-dimension and rademacher complexity. Algorithms (2020)
29. Shah, R.D. and Peters, J.: The hardness of conditional independence testing and the generalised covariance measure. The Annals of Statistics (2020)
30. Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: Causation, prediction, and search. MIT press (2000)
31. Strobl, E.V., Spirtes, P.L., Visweswaran, S.: Estimating and controlling the false discovery rate of the pc algorithm using edge-specific p-values. ACM TIST (2019)
32. Tsamardinos, I., Aliferis, C.F.: Towards principled feature selection: Relevancy, filters and wrappers. Int. Workshop on AI and Statistics. PMLR (2003)
33. Tsamardinos, I., Aliferis, C.F., Statnikov, A.: Time and sample efficient discovery of markov blankets and direct causal relations. ACM SIGKDD (2003)
34. Tsamardinos, I., Aliferis, C.F., Statnikov, A.R., Statnikov, E.: Algorithms for large scale markov blanket discovery. FLAIRS conference (2003)
35. Tsamardinos, I., Brown, L.E.: Bounding the false discovery rate in local bayesian network learning. AAAI (2008)

36. Velikova, M., van Scheltinga, J.T., Lucas, P.J., Spaanderman, M.: Exploiting causal functional relationships in bayesian network modelling for personalised healthcare. Int. J. of Approximate Reasoning (2014)
37. Yusuf, F., Cheng, S., Ganapati, S., Narasimhan, G.: Causal inference methods and their challenges: the case of 311 data. Int. Conf. on Digital Government Res. (2021)