DRCFS: Doubly Robust Causal Feature Selection

Francesco Quinzan[*] [†] [1] **Ashkan Soleymani**[*] [2] **Patrick Jaillet**[2] **Cristian R. Rojas**[3] **Stefan Bauer**[4 5]

Abstract

Knowing the features of a complex system that are highly relevant to a particular target variable is of fundamental interest in many areas of science. Existing approaches are often limited to linear settings, sometimes lack guarantees, and in most cases, do not scale to the problem at hand, in particular to images. We propose DRCFS, a doubly robust feature selection method for identifying the causal features even in nonlinear and high dimensional settings. We provide theoretical guarantees, illustrate necessary conditions for our assumptions, and perform extensive experiments across a wide range of simulated and semi-synthetic datasets. DRCFS significantly outperforms existing state-of-the-art methods, selecting robust features even in challenging highly non-linear and high-dimensional problems.

1. Introduction

We study the fundamental problem of *causal* feature selection for non-linear models. That is, consider a set of features $\boldsymbol{X} = \{X_1, \ldots, X_m\}$, and an outcome Y specified with an additive-noise model on some of the features (Hoyer et al., 2008; Schölkopf et al., 2012; Peters et al., 2014):

$$\text{Axiom (A)} \quad Y = f(\mathsf{Pa}(Y)) + \varepsilon,$$

for a subset $\mathsf{Pa}(Y) \subseteq \boldsymbol{X}$ and posterior additive noise ε. Our goal is to identify the set of relevant features $\mathsf{Pa}(Y)$ from observations. Feature selection is an important cornerstone of high-dimensional data analysis (Liu & Motoda, 2007; Bolón-Canedo et al., 2015; Li et al., 2017; Butcher & Smith, 2020), especially in data-rich settings. By including only relevant variables and removing nuisance factors, feature selection allows us to build models that are simple, interpretable, and more robust (Yu et al., 2020; Janzing et al., 2020). Moreover, in many applications in science and industry, we are not only interested in predictive features but we also aim to identify causal relationships between them (Pearl, 2009; Spirtes et al., 2000; Murphy, 2001).

Knowing the causal structure allows one to understand the potential effects of interventions on a system of interest, spanning various fields such as economics (Varian, 2016), biology (Hu et al., 2018), medicine (Mehrjou et al., 2021; Lv et al., 2021), software engineering (Siebert, 2022), agriculture (Tsoumas et al., 2022), and climate research (Runge et al., 2019). In general, it is impractical to infer the underlying causal graph solely from observational data, and in some cases, it may even be infeasible (Pearl, 2009; Spirtes et al., 2000; Chickering et al., 2004; Maclaren & Nicholson, 2019). Often, the focus is on specific target variables, such as stock returns in trading, genes associated with a specific phenotype or disease, or reduction of CO_2 emissions in environmental studies, rather than all the interactions between the many variables affecting the underlying system.

Capturing the causal relationships between different features and a particular target variable is highly non-trivial in practice. Existing approaches for causal feature selection often make unrealistic assumptions on the data generating process (DGP), or they simply do not scale to the problem at hand in terms of computational efficiency (Yu et al., 2020) and/or statistical efficiency (Yu et al., 2021). Moreover, many of these approaches have only limited theoretical guarantees, and they have only been evaluated on noiseless data, far from the needs of practitioners in industry and science (Kira & Rendell, 1992; Guyon, 2008).

To overcome these limitations, we propose a novel, doubly robust causal feature selection method (DRCFS) that significantly outperforms existing state-of-the-art approaches, especially in non-linear, noisy, and data-sparse settings. Our contributions are as follows:

- We propose DRCFS, a novel method for doubly robust feature selection, even in non-linear and high dimensional settings. In particular, our approach has guarantees for realistic cases, when there are cycles or hidden confounders between the features (see Figure 1).

[*] Equal contribution [†] This work was initiated and partially done while the author was employed at KTH Royal Institute of Technology [1]Department of Computer Science, University of Oxford [2]Laboratory for Information & Decision Systems (LIDS), Massachusetts Institute of Technology [3]KTH Royal Institute of Technology [4]TU Munich [5]Helmholtz Munich. Correspondence to: Francesco Quinzan <quinzan@kth.se>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

- Based on our model description, we provide theoretical guarantees that substantiate our framework. In particular, we illustrate the necessity of our assumptions for causal feature selection and demonstrate that our approach is doubly robust while achieving $\sqrt{n}$-consistency.
- We provide a comprehensive experimental evaluation across various synthetic and semi-synthetic datasets, demonstrating that DRCFS significantly outperforms an extensive list of state-of-the-art baselines for feature selection.

Additional related work. Learning causal features has been investigated early on, and Guyon & Aliferis (2007) provides a broad review of these works. Since then, multiple other works have proposed different ideas on how to learn causal features from data (Cawley, 2008; Paul, 2017; Yu et al., 2021). However, none of these works recover all the direct causal parents asymptotically or non-asymptotically, and they do not provide guarantees, especially for the cyclic or confounded setting.

Under the expense of additional assumptions such as faithfulness, there has been an orthogonal line of research aiming to infer the Markov equivalence class through iterative testing of d-separation statements using conditional independence test as done in the PC-algorithm (Mastakouri et al., 2019; Pearl, 2009; Spirtes et al., 2000). For a recent overview of causal feature selection approaches and their evaluation, we refer to Yu et al. (2020). Another approach of interest is to consider strong assumption on the causal structure of the DGP to achieve identifiability (Hoyer et al., 2008; Peters et al., 2014). This approach includes the works by Shimizu et al. (2006b) and Peters et al. (2011).

2. Model Ddescription

2.1. Causal structure

We study causal feature selection for a model as in Axiom (A), using the language of probabilistic causality and structural causal models (SCMs, see Pearl (2000); Bongers et al. (2021)). Following Bongers et al. (2021), we define the causal structure of a model in terms of direct causal effects among the variables (see Appendix A for the precise concepts introduced in this section). Direct causal effects are defined by distribution changes due to *interventions* on the DGP. An intervention amounts to actively manipulating the generative process of a random variable X_j, without altering the remaining components of the DGP. For instance, in a randomized experiments an intervention could consist of giving patients a medical treatment. We specifically consider perfect interventions $X_j \leftarrow x_j$, by which the post-interventional variable is set to a constant $X_j \equiv x_j$. We denote with $Y \mid do(x_j)$ the post-interventional outcome Y. Using this notation, a variable X_j has a direct causal effect on the outcome Y if intervening on X_j can affect the outcome while keeping fixed the other variables $\boldsymbol{X}_j^c := \{X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_m\}$, i.e., there exists $x_j' \neq x_j$ such that

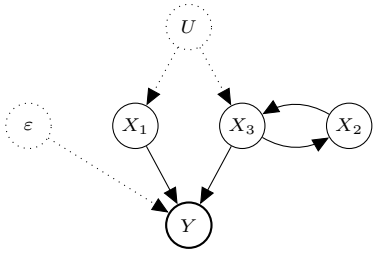


Figure 1: Causal structure of a DGP as in Axiom (A)-Axiom (C). The outcome is specified as $Y = f(X_1, X_3) + \varepsilon$, with f a deterministic non-linear function, and ε exogenous independent noise. U is a latent variable that acts as a confounder for X_1 and X_3.

$$\mathbb{P}(Y \mid do(x_j', \boldsymbol{x}_j^c)) \neq \mathbb{P}(Y \mid do(x_j, \boldsymbol{x}_j^c)) \tag{1}$$

for some array $\boldsymbol{x}_j^c$ in the range of $\boldsymbol{X}_j^c$. Following Bongers et al. (2021), we define the causal structure $\mathcal{G}$ of the DGP as a directed graph whose edges represent all direct causal effects among the variables.

We assume unique solvability of the SCM (see Definition 3.3 by Bongers et al. (2021) and Appendix B). By this assumption, the distribution on the observed variables $\boldsymbol{V} = \{X_1, \ldots, X_m, Y\}$ is specified by a mixing measurable function of the form $\boldsymbol{V} = \boldsymbol{g}(\boldsymbol{U})$. The function $\boldsymbol{g}$ is uniquely defined by the structural equations of the model (see Appendix B). Here, $\boldsymbol{U}$ is a random vector of latent sources. According to our model, there may be potential hidden confounders for the observed variables. Furthermore, the underlying graph $\mathcal{G}$ may contain cycles. An example of a causal graph for a uniquely solvable SCM is presented in Figure 1.

2.2. Notation

We always denote with Y the outcome, with X_j the features, for $j = 1, \ldots, m$, and with n, the number of observations. We use the bold capital script, e.g. $\boldsymbol{X}, \boldsymbol{Z}$, to denote a group of random variables. The set $\boldsymbol{X}$ always refers to the set of all the features. For a given index $j = 1, \ldots, m$, we denote with $\mathbf{X}_j^c$ the set consisting of all the features *except* X_j. We use the standard $Y \mid do(\cdot)$ notation, to denote the post-interventional outcome. The notation $\mathsf{Pa}(Y)$ denotes the subset of features that yield a direct causal effect on the outcome. We use the letter $\mathcal{D}$ to denote a generic dataset, and we denote with $\hat{\mathbb{E}}_{\mathcal{D}}[\,\cdot\,]$ the empirical expected value over this dataset.

2.3. Necessary Conditions for Causal Feature Selection

We assume that the outcome is specified by the formula in Axiom (A). Although very general, this model does not guarantee that we may identify $\mathsf{Pa}(Y)$ from samples. In fact, there may be multiple models as in Axiom (A) with different causal parents, that induce the same joint probability distribution on the observed variables. To overcome this problem, we consider the following additional assumptions on the data-generating process:

Axiom (B) ε is exogenous noise, independent of $\boldsymbol{X}$.

Axiom (C) Y has no direct causal effect on any of the features.

Axiom (B) is a standard assumption in the literature of continuous additive noise models (Peters et al., 2014). Axiom (C) requires that the outcome Y does not causally affect any of the features X_j. This requirement is aligned with previous related work (Soleymani et al., 2022). As we show in Section 3.2, under Axioms Axiom (A)-Axiom (C) we can identify the causal parents of the outcome from samples. However, if either Axiom (B) or Axiom (C) fail, then performing Causal Feature Selection from observations is impossible. We show this, by considering two counterexamples, given in Example 2.1 and Example 2.2 respectively.

Example 2.1 (Necessity of Axiom (B))**.** *Consider two datasets $\{X_i, Y_i\}$ for $i = 1, 2$. The respective DGPs are defined as follows:*

$$\begin{cases} X_1 = N_1 \\ Y_1 = \varepsilon_1 \end{cases} \quad \textit{and} \quad \begin{cases} X_2 = N_2 \\ Y_2 = X_2 + \varepsilon_2 \end{cases}$$

with $(N_1, \varepsilon_1) \sim \mathcal{N}(\mathbf{0}, \Sigma)$, a zero-mean joint Gaussian distribution with covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

while $N_2 \sim \mathcal{N}(0, 1)$ and $\varepsilon_2 \equiv 0$. Note that $\{X_1, Y_1\}$ satisfies Axiom (C) but violates Axiom (B), since the noise ε_1 is correlated with X_1. Both datasets entail the same joint probability distribution. However, Y_1 is exogenous to the model, whereas X_2 yields a direct causal effect on Y_2.

Example 2.1 shows that if Axiom (B) is violated, then it is impossible to perform Causal Feature Selection from observational data. We provide a second example, to show that Axiom (C) is also necessary.

Example 2.2 (Necessity of Axiom (C))**.** *Consider two datasets $\{X_i, Y_i\}$ for $i = 1, 2$, with DGPs defined as:*

$$\begin{cases} X_1 = N_1 \\ Y_1 = X_1 + \varepsilon_1 \end{cases} \quad \textit{and} \quad \begin{cases} X_2 = 0.5 \cdot Y_2 + N_2 \\ Y_2 = \varepsilon_2 \end{cases}$$

where $\varepsilon_1, N_1, \sim \mathcal{N}(0, 1)$, $\varepsilon_2 \sim \mathcal{N}(0, 2)$ and $N_2 \sim \mathcal{N}(0, 0.5)$ are independent. In this case, dataset $\{X_2, Y_2\}$ satisfies Axiom (B) but violates Axiom (C), since Y_2 causes X_2. Then, both datasets are jointly normal with zero mean and the same covariance matrix. However, Y_1 has as causal parent X_1, and Y_2 has no causal parents.

3. Methodology

3.1. Double Machine Learning (DoubleML)

We provide a statistical test for Causal Feature Selection based on DoubleML estimators. DoubleML is a general framework for parameter estimation, which uses debiased score functions and (double) cross-fitting to achieve $\sqrt{n}$-consistency guarantees. Following Rotnitzky et al. (2020); Chernozhukov et al. (2022), we provide a suitable debiased score function using the Riesz Representation Theorem and the Mixed Bias Property.

The Riesz Representation Theorem. Let $\boldsymbol{V}$ be the collection of observed random variables of an SCM as in Definition A.1. Let $\boldsymbol{X}$ be a random variable in $\boldsymbol{V}$, g any real-valued function of $\boldsymbol{X}$ such that $\mathbb{E}[g^2(\boldsymbol{X})] < \infty$, and consider a linear functional $m(\boldsymbol{V}; g)$ in g. The Riesz Representation Theorem ensures that, under certain conditions, there exists a function α_0 of $\boldsymbol{X}$ such that $\mathbb{E}[m(\boldsymbol{V}; g)] = \mathbb{E}[\alpha_0(\boldsymbol{X}) g(\boldsymbol{X})]$. The function α_0 is called the *Riesz Representer* (RR). Crucially, the RR only depends on the functional m, and not on the function g. Chernozhukov et al. (2021) shows that the Riesz representer can be estimated by solving the following optimization problem:

$$\alpha_0 = \arg\min_{\alpha} \mathbb{E}[\alpha(\boldsymbol{X})^2 - 2m(\boldsymbol{V}; \alpha)]. \tag{2}$$

Using the RR, we will derive a debiased score function for the parameter $\theta_0 := \mathbb{E}[m(\boldsymbol{V}; g_0)]$ with $g_0(\boldsymbol{X}) = \mathbb{E}[Y \mid \boldsymbol{X}]$. This score function is "debiased" in the sense that it fulfills the Mixed Bias Property.

The Mixed Bias Property. The RR is crucial when learning an estimate $\hat{\theta}_0$ of the parameter $\theta_0 := \mathbb{E}[m(\boldsymbol{V}; g_0)]$ as described above. In fact, as shown by Chernozhukov et al. (2022); Rotnitzky et al. (2020), one can achieve $\sqrt{n}$-consistency for $\hat{\theta}_0$ by combining (double) cross-fitting with a debiased learning objective of the form

$$\varphi(\theta, \boldsymbol{\eta}) := m(\boldsymbol{V}; g) + \alpha(\boldsymbol{X}) \cdot (Y - g(\boldsymbol{X})) - \theta. \tag{3}$$

Here, $\boldsymbol{\eta} := (\alpha, g)$ is a nuisance parameter consisting of a pair of square-integrable functions. Note that it holds $\mathbb{E}[\varphi(\theta_0, \boldsymbol{\eta}_0)] = 0$, with $\boldsymbol{\eta}_0 := (\alpha_0, g_0)$ consisting of the RR α_0 of the moment functional $m(\boldsymbol{V}; g)$ and the conditional expected value $g_0(\boldsymbol{X}) = \mathbb{E}[Y \mid \boldsymbol{X}]$. As shown by Chernozhukov et al. (2022), the score function Eq. (3) yields

$$\mathbb{E}[\varphi(\theta_0, \boldsymbol{\eta})] = -\mathbb{E}[(\alpha(\boldsymbol{X}) - \alpha_0(\boldsymbol{X}))(g(\boldsymbol{X}) - g_0(\boldsymbol{X}))].$$

This equation corresponds to the Mixed Bias Property as in Definition 1 of Rotnitzky et al. (2020). Hence, when cross-

fitting or double-cross is employed, the resulting estimator $\hat{\theta}$ has the *double robustness property*. That is, the quantity $\sqrt{n}(\hat{\theta} - \theta_0)$ converges in distribution to a zero-mean Normal distribution whenever the product of the mean-square convergence rates of α and g is faster than $\sqrt{n}$. By these guarantees, if an estimator of g_0 has a good convergence rate, then the rate requirement on the estimator of the RR is less strict, and vice versa.

3.2. The Average Controlled Direct Effect (ACDE)

Our approach to causal feature selection uses the ACDE. The ACDE is a concept introduced by Pearl (2001) to provide an operational description of the statement "the variable X_j directly influences the outcome Y". By this conceptualization, we ask whether the expected outcome would change under an intervention $X_j \leftarrow x_j$, while holding the remaining observed variables at a predetermined value. Following the notation given in Section 2, the ACDE is defined as

$$\mathrm{ACDE}(x_j, x_j' | \boldsymbol{x}_j^c) := \mathbb{E}[Y \mid do(x_j, \boldsymbol{x}_j^c)] - \mathbb{E}[Y \mid do(x_j', \boldsymbol{x}_j^c)]. \quad (4)$$

This quantity captures the difference between the distribution of the outcome Y under $X_j \leftarrow x_j$ and $X_j \leftarrow x_j'$, when we keep all other variables $\boldsymbol{X}_j^c$ fixed at some values $\boldsymbol{x}_j^c$. As we will show later, under Axioms Axiom (A)-Axiom (C) a variable X_j is a causal parent of Y if and only if the ACDE is non-zero for at least a triple $(x_j, x_j', \boldsymbol{x}_j^c)$ of possible interventional values. Hence, we can perform causal feature selection by testing whether the ACDE is identically zero or not.

Using the ACDE for Causal Feature Selection. Our approach to causal feature selection essentially consists of testing whether a feature X_j yields a non-zero average controlled direct effect on the outcome. This approach is justified by the following result.

Lemma 3.1. *Consider a causal model as in Axioms Axiom (A)-Axiom (C), and fix a feature X_j. Then, $\mathbb{E}[Y \mid do(x_j, \boldsymbol{x}_j^c)] - \mathbb{E}[Y \mid do(x_j', \boldsymbol{x}_j^c)] \neq 0$ for some interventional values x_j, x_j' and $\boldsymbol{x}_j^c$ if and only if $X_j \in \mathsf{Pa}(Y)$.*

The proof is deferred to Appendix C. Intuitively, this proof tells us that direct causal effects can be checked with the expected value of the post-interventional outcome, instead of the full distribution as in Eq. (1). This lemma crucially relies on Axioms Axiom (A)-Axiom (C) and it does not hold for general causal models.

Estimating the ACDE from samples. A major challenge in using the ACDE for Causal Feature Selection is that we do not assume knowledge of the post-interventional outcome distribution. However, it turns out that we can perform this test from observational data, by considering this quantity:

$$\chi_j := \mathbb{E}_{(x_j, \boldsymbol{x}_j^c) \sim (X_j, \boldsymbol{X}_j^c)} \left[\left(\mathbb{E}[Y | x_j, \boldsymbol{x}_j^c] - \mathbb{E}[Y | \boldsymbol{x}_j^c] \right)^2 \right]. \quad (5)$$

Under mild assumptions, testing whether $\chi_j = 0$ is equivalent to testing if there is no average direct effect.

Lemma 3.2. *Consider a causal model as in Axioms Axiom (A)-Axiom (C). Then, $\mathbb{E}[Y \mid do(x_j, \boldsymbol{x}_j^c)] - \mathbb{E}[Y \mid do(x_j', \boldsymbol{x}_j^c)] \neq 0$ almost surely if and only if $\chi_j \neq 0$. It follows that $\chi_j \neq 0$ if and only if $X_j \in \mathsf{Pa}(Y)$.*

The proof is deferred to Appendix D. From this lemma, we can use χ_j to estimate if a feature X_j is a causal parent of the outcome Y, i.e., perform Causal Feature Selection, by testing if $\chi_j = 0$. Crucially, χ_j can be estimated efficiently from samples.

3.3. Estimating the ACDE with DoubleML

We provide a debiased score function for χ_j, using the Riesz representation theorem, as described in Section 3.1. Crucially, we show that χ_j is the difference of the expected value of two linear moment functionals. The following lemma holds.

Lemma 3.3. *Let χ_j be as in Eq. (5), and $\boldsymbol{X} = \{X_j\} \cup \boldsymbol{X}_j^c$. For any square-integrable random variable $g(\boldsymbol{X})$, consider the moment functional $m_0(\boldsymbol{V}; g) := Yg(\boldsymbol{X})$. Similarly, for any square-integrable random variable $h(\boldsymbol{X}_j^c)$, consider the moment functional $m_j(\boldsymbol{V}; h) := Yh(\boldsymbol{X}_j^c)$.[1] Then, it holds that*

$$\chi_j = \mathbb{E}[m_0(\boldsymbol{V}; g_0)] - \mathbb{E}[m_j(\boldsymbol{V}; h_0)],$$

with $g_0(\boldsymbol{X}) = \mathbb{E}[Y \mid \boldsymbol{X}]$ and $h_0(\boldsymbol{X}_j^c) = \mathbb{E}[Y \mid \boldsymbol{X}_j^c]$.

The proof is deferred to Appendix E. By this lemma, we can estimate the parameter χ_j, by learning the parameters $\theta_0 := \mathbb{E}[m_0(\boldsymbol{V}; g_0)]$ and $\theta_j := \mathbb{E}[m_j(\boldsymbol{V}; h_0)]$ separately and them taking their difference. Both θ_0 and θ_j are expected values of linear moment functionals, so we can apply DoubleML as described in Section 3.1 to obtain $\sqrt{n}$-consistent estimates for them. The resulting estimators $\hat{\theta}_0$ and $\hat{\theta}_j$ have the double robustness property, and so does their difference $\chi_j = \hat{\theta}_0 - \hat{\theta}_j$. We can then take advantage of the fast convergence rate, to determine if $\hat{\theta}_0 \approx \hat{\theta}_j$ with a paired t-test.

4. The DRCFS Algorithm

4.1. Overview

Our approach to causal discovery essentially consists of testing whether a feature X_j yields a non-zero average controlled direct effect on the outcome, following Lemma 3.1.

[1] Note that $m_0(\boldsymbol{V}; g)$ and $m_j(\boldsymbol{V}; h)$ are distinct functionals, since they are defined over sets of functions with different domains.

Algorithm 1 Doubly Robust Causal Feature Selection Algorithm (DRCFS)

1: split the n samples $\mathcal{D} = \{(x_{1i}, \ldots, x_{mi}, y_i)\}_{i=1,\ldots,n}$ into k disjoint sets $\mathcal{D}_1, \ldots, \mathcal{D}_k$;
2: define $\mathcal{D}_l^c \leftarrow \mathcal{D} \setminus \mathcal{D}_l$ for all $l \in [k]$;
3: **for** each $l \in [k]$ **do**
4: construct an estimator $\hat{g}_l(\boldsymbol{X})$ for $\mathbb{E}[Y \mid \boldsymbol{X}]$ and an estimator $\hat{\alpha}_l(\boldsymbol{X})$ for the RR of m_0, using samples in $\mathcal{D}_l^c$;
5: $\hat{\theta}_{0,l} \leftarrow \hat{\mathbb{E}}_{\mathcal{D}_l}[Y \cdot \hat{g}_l(\boldsymbol{X}) - Y \cdot \hat{\alpha}_l(\boldsymbol{X}) - \hat{\alpha}_l(\boldsymbol{X}) \cdot \hat{g}_l(\boldsymbol{X})]$;
6: $\hat{\sigma}_{0,l}^2 \leftarrow \hat{\mathbb{E}}_{\mathcal{D}_l}[(Y \cdot \hat{g}_l(\boldsymbol{X}) - Y \cdot \hat{\alpha}_l(\boldsymbol{X}) - \hat{\alpha}_l(\boldsymbol{X}) \cdot \hat{g}_l(\boldsymbol{X}) - \hat{\theta}_{0,l})^2]$;
7: **end for**
8: $\hat{\theta}_0 \leftarrow \frac{1}{k}\sum_l \hat{\theta}_{0,l}$ and $\hat{\sigma}_0^2 \leftarrow \frac{1}{k}\sum_l \hat{\sigma}_{0,l}^2$;
9: **for** each feature $X_j \in \boldsymbol{X}$ **do**
10: **for** each $l \in [k]$ **do**
11: construct an estimator $\hat{h}_l^j(\boldsymbol{X}_j^c)$ for $\mathbb{E}[Y \mid \boldsymbol{X}_j^c]$ and an estimator $\hat{\alpha}_l^j(\boldsymbol{X}_j^c)$ for the RR of m_0^j, using samples in $\mathcal{D}_l^c$;
12: $\hat{\theta}_{j,l} \leftarrow \hat{\mathbb{E}}_{\mathcal{D}_l}[Y \cdot \hat{h}_l^j(\boldsymbol{X}_j^c) - Y \cdot \hat{\alpha}_l^j(\boldsymbol{X}_j^c) - \hat{\alpha}_l^j(\boldsymbol{X}_j^c) \cdot \hat{h}_l^j(\boldsymbol{X}_j^c)]$;
13: $\hat{\sigma}_{j,l}^2 \leftarrow \hat{\mathbb{E}}_{\mathcal{D}_l}[(Y \cdot \hat{h}_l^j(\boldsymbol{X}_j^c) - y \cdot \hat{\alpha}_l^j(\boldsymbol{X}_j^c) - \hat{\alpha}_l^j(\boldsymbol{X}_j^c) \cdot \hat{h}_l^j(\boldsymbol{X}_j^c) - \hat{\theta}_{j,l}^j)^2]$;
14: **end for**
15: $\hat{\theta}_j \leftarrow \frac{1}{k}\sum_l \hat{\theta}_{j,l}$ and $\hat{\sigma}_j^2 \leftarrow \frac{1}{k}\sum_l \hat{\sigma}_{j,l}^2$;
16: perform a paired t-test to determine if $\hat{\theta}_j \approx \hat{\theta}_0$ and select feature X_j if the null-hypotheses is rejected.
17: **end for**
18: **return** the selected features X_j;

We refer to our approach as the Doubly Robust Causal Feature Selection Algorithm (DRCFS, see Algorithm 1). This algorithm consists of the following steps:

- Select a variable X_j to test if $X_j \in \mathsf{Pa}(Y)$.
- Estimate the parameter χ_j using DoubleML, as described in Section 3.3. The resulting estimator $\hat{\chi}_j$ has the double-robustness property.
- By Lemmas 3.2 and 3.1, the variable X_j is not a parent of Y if and only if $\chi_j = 0$. Use a paired t-test for $\hat{\chi}_j$ to select or discard X_j as a parent of Y.

This procedure can be iterated for each of the features. Crucially, the estimator $\hat{\chi}_j$ ought to have the double-robustness property. To this end, we resort to Lemma 3.3. We first estimate the parameters $\theta_0 := \mathbb{E}[m_0(\boldsymbol{V}, g_0)]$ and $\theta_j := \mathbb{E}[m_j(\boldsymbol{V}, h_0)]$ separately using DoubleML, and then obtain the desired estimator $\hat{\chi}_j$ by taking the difference.

4.2. Double-Robustness of $\hat{\chi}_j$

We now show that the estimand $\hat{\chi}_j$ as in Line 16 of Algorithm 1 has the double-robustness property. To this end, we show that $\hat{\theta}_0$ and $\hat{\theta}_j$ in Algorithm 1 have the double robustness property. We focus on $\hat{\theta}_0$ since the case for $\hat{\theta}_j$ is analogous. To this end, we can apply the Riesz representation theorem, as described in Section 3.1, to obtain that $\theta_0 = \mathbb{E}[\alpha_0(\boldsymbol{X}) \cdot m_0(\boldsymbol{V}; g_0)]$, with α_0 the Riesz representer of the moment functional, and g_0 the conditional expected value.

We can use DoubleML as in Section 3.1 to derive a debiased score function for the term θ_0. Consider a dataset of n samples $\mathcal{D} = \{(x_{1i}, \ldots, x_{mi}, y_i)\}_{i=1,\ldots,n}$ and let $\mathcal{D}_1, \ldots, \mathcal{D}_k$ be a disjoint k-partition of this dataset. Following Eq. (3), we provide an estimator $\hat{\theta}_{0,l}$ for θ_0 on the partition $\mathcal{D}_l$ as

$$\hat{\theta}_{0,l} = \hat{\mathbb{E}}_{\mathcal{D}_l}[Y \cdot \hat{g}_l(\boldsymbol{X}) - Y \cdot \hat{\alpha}_l(\boldsymbol{X}) - \hat{\alpha}_l(\boldsymbol{X}) \cdot \hat{g}_l(\boldsymbol{X})] \quad (6)$$

Here, $\hat{\mathbb{E}}_{\mathcal{D}_l}$ denotes the empirical expected value over $\mathcal{D}_l$. The function $\hat{\alpha}_l$ is an estimator for the Riesz representer obtained by Eq. (2) over the complementary sample set $\mathcal{D}_l^c = \mathcal{D} \setminus \mathcal{D}_l$. Note that Eq. (6) corresponds to Line 5 of Algorithm 1. Similarly, $\hat{g}_l$ is an estimator for the conditional expected value, which is obtained as a solution of the ℓ_1-regularized regression problem (Rotnitzky et al., 2020) over the sample set $\mathcal{D}_l^c = \mathcal{D} \setminus \mathcal{D}_l$. We can also estimate the variance over the samples as in Line 6 of Algorithm 1. Estimates for the parameters θ_0 and the variance are then

$$\hat{\theta}_0 = \frac{1}{k}\sum_{l=1}^{k} \hat{\theta}_{0,l} \quad \text{and} \quad \hat{\sigma}_0^2 = \frac{1}{k}\sum_{l=1}^{k} \hat{\sigma}_{0,l}^2.$$

These estimators are given in Line 8 of Algorithm 1, and they have the double-robustness property, i.e.,

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) \rightsquigarrow \mathcal{N}(0, \sigma_0^2).$$

Furthermore, the empirical variance $\hat{\sigma}_{\chi_j}^2$ is a $\sqrt{n}$-consistent estimator for σ_0^2. Similarly, the parameters $\hat{\theta}_j$ and $\hat{\sigma}_j^2$ as in Line 15 of Algorithm 1 are doubly-robust.

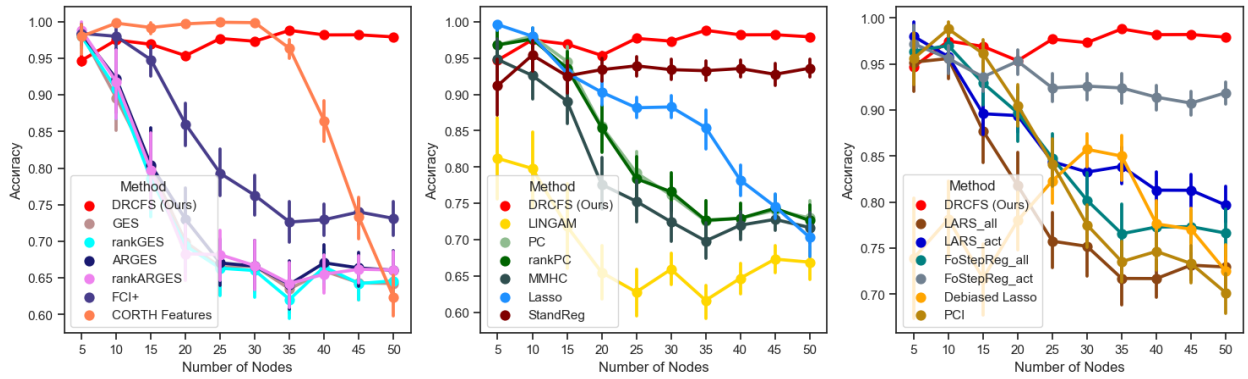


Figure 2: Performance (accuracy) of the algorithms w.r.t. number of nodes m for fully linear causal structures ($f = f_1$ with probability 1), where $p_s = 0.3$, $p_h = 0$, and $\epsilon \sim \mathcal{N}(0, 1)$. Each case is averaged over 50 simulations. We use ForestRiesz with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$. DRCFS's performance shows stability even for large graphs while the baselines suffer in high dimensions. Plots for other metrics are provided in Appendix G, Figures 7 and 8.

4.3. Statistical Test

For a given confidence interval, testing whether $\hat{\chi}_j \approx 0$, is equivalent to testing whether the estimates of $\hat{\theta}_j$ and $\hat{\theta}_0$ have the same mean. To accomplish this, we perform paired sample t-tests. Given the presence of multiple dependent tests, in case the data is highly dependent, it is important to control for false discovery rate (FDR) in order to accurately assess the results. For this purpose, we apply the Benjamini-Yekutieli procedure (Benjamini & Yekutieli, 2001).

5. Experiments

In this section, we evaluate the performance of our algorithm extensively. First, we show the superiority of our method compared to well-established algorithms, using synthetic data consisting of causal structures created by various DGPs. Second, we demonstrate a real-world application of DRCFS on microbiome abundance. Lastly, we further showcase the performance and scalability of our algorithm on bnlearn benchmarks (Scutari, 2010).

5.1. Causal Feature Selection for Synthetic Data

Here, we discuss the data generating process, comparison of the performance with the baselines, robustness of performance w.r.t. various characteristics of the underlying causal structure such as connectivity level, and quality of estimations and statistical tests.

Data Generating Process. The DGP generally follows the same procedure as Soleymani et al. (2022) to produce direct acyclic graphs (DAGs): (1) Nodes are randomly permutated to form a topological order. (2) For each pair of nodes X_i and X_j, where X_i precedes X_j in this order, an edge $X_i \to X_j$ is added to the graph with probability p_c (connectivity level). (3) The values are assigned to each variable X as a transformation f of the direct causes of that variable, plus posterior additive noise as in Axioms Axiom (A)-Axiom (C) for each variable $X' \in \boldsymbol{X} \cup \{Y\}$. (4) Each node $X \in \boldsymbol{X}$ is concealed to serve as a unseen confounder with probability p_h. The parameters of interest in the DGP are: number of nodes m, number of observations n, connectivity level p_c, transformation function f, random variable ϵ representing the additive noise, probability of hiding each node p_h to serve as unseen confounders. For convenience, different choices of function f are given in Appendix F.2. Let $\{f_1, f_2, \ldots, f_k\}$ and $\{\pi_1, \pi_2, \ldots, \pi_k\}$ be the potential choices of f and their corresponding probabilities, then f is chosen by $f \sim \sum_{i=1}^{k} \pi_i \delta_{f_i}$, for each variable $X' \in \boldsymbol{X} \cup \{Y\}$, where δ is the Dirac probability measure.

Baselines. We compare the performance of our method with a diverse set of causal structure learning and inference for regression algorithms: LiNGAM (Shimizu et al., 2006a), order-independent PC (Colombo & Maathuis, 2014), rankPC, rankFCI (Spirtes et al., 2000; Heinze-Deml et al., 2018), MMHC (Tsamardinos et al., 2006), GES (Chickering, 2003), rankGES, ARGES (adaptively restricted GES (Nandy et al., 2016)), rankARGES, FCI+ (Claassen et al., 2013), PCI (Shah & Peters, 2020), CORTH Features (Soleymani et al., 2022), Standard Linear Regression, Lasso with exact post-selection inference (Lee et al., 2016), Debiased Lasso (Javanmard et al., 2015), Forward Stepwise Regression for active variables (Loftus & Taylor, 2014; Tibshirani et al., 2016), Forward Stepwise Regression for all variables (Loftus & Taylor, 2014; Tibshirani et al., 2016), LARS for active variables (Efron et al., 2004; Tibshirani et al., 2016), and LARS for all variables (Efron et al., 2004; Tibshirani et al., 2016). R Packages "CompareCausalNet-

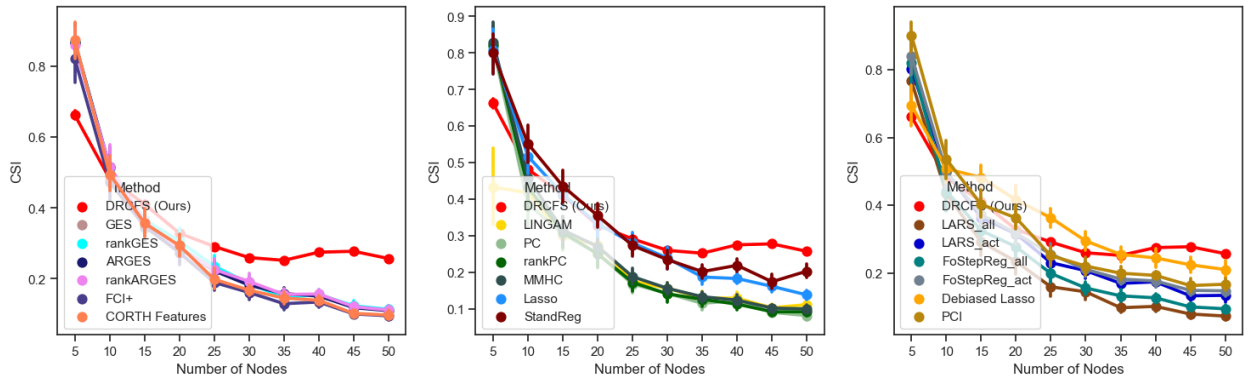


Figure 3: Performance (CSI) of the algorithms w.r.t. number of nodes m for fully Log-sum-exp causal structures ($f = f_6$ with probability 1), where $p_s = 0.5$, $p_h = 0.1$, and $\epsilon \sim \mathcal{N}(0, 1)$. Each case is averaged over 50 simulations. We use ForestRiesz with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$. DRCFS's performance shows stability even for large graphs while the baselines suffer in high dimensions. Plots for other metric are provided in Appendix G, Figure 12.

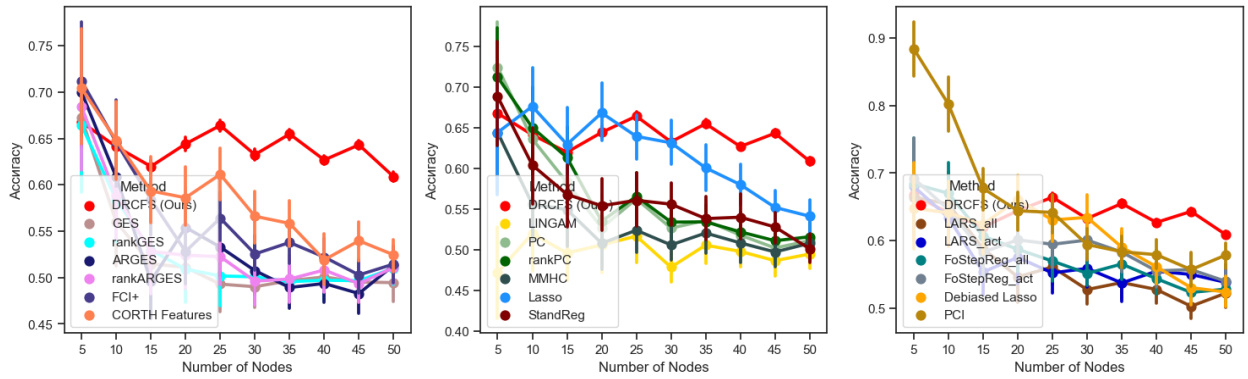


Figure 4: Performance (accuracy) of the algorithms w.r.t. number of nodes m for causal structures with geometric mean relationship ($f = f_5$ with probability 0.8, $f = f_1$ with probability 0.2), where $p_s = 0.5$, $p_h = 0$, and $\epsilon \sim \mathcal{N}(0, 1)$. Each case is averaged over 50 simulations. We use ForestRiesz with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$. DRCFS's performance shows stability even for large graphs while the baselines suffer in high dimensions. Plots for other metric are provided in Appendix G, Figures 13 and 14.

works"[2] and "selectiveInference: Tools for Post-Selection Inference"[3] are used for a great number of the baselines.

Evaluation. We use accuracy, F1 score, and CSI (discussed in Appendix F.1) as our evaluation metrics. We consider F1 score and CSI because they put emphasis on the number of true positives.

Estimation Technique. In this section, we leverage a special case of Generalized Random Forests (Athey et al., 2019), called ForestRiesz (Chernozhukov et al., 2021) to estimate the conditional expected value and RR as in Lemma 3.3.[4] We illustrate this technique, by focusing on g_0 and α_0, since the case for g_j, α_j is analogous. Assume that g_0 and α_0 are locally linear with respect to a smooth feature map ϕ, i.e., $g_0(\boldsymbol{X}) = \langle \phi(\boldsymbol{X}), \beta(\boldsymbol{X}) \rangle$ and $\alpha_0(\boldsymbol{X}) = \langle \phi(\boldsymbol{X}), \gamma(\boldsymbol{X}) \rangle$ with $\beta(\boldsymbol{X})$ and $\gamma(\boldsymbol{X})$ non-parametric estimators derived by the tree. Then, in order to learn g_0, we minimize the square loss $\mathcal{R}(\beta) = \mathbb{E}_n[(\langle \phi(\boldsymbol{X}), \beta(\boldsymbol{X}) \rangle - Y)^2]$. This is equivalent to searching for the solution of the equation

$$\mathbb{E}_n[(\langle \phi(\boldsymbol{X}), \beta(\boldsymbol{X}) \rangle - Y)\phi(\boldsymbol{X}) \mid \boldsymbol{X}] = 0$$

in the variable β. Similarly, we learn the RR by minimizing the score $\mathcal{R}(\gamma) = \mathbb{E}_n[\langle \phi(\boldsymbol{X}), \gamma(\boldsymbol{X}) \rangle^2 - 2\gamma(\boldsymbol{X})^T m(\boldsymbol{V}; \phi(\boldsymbol{X}))]$. This amounts to solving

$$\mathbb{E}_n[\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T \gamma(\boldsymbol{X}) - m(\boldsymbol{V}; \phi(\boldsymbol{X})) \mid \boldsymbol{X}] = 0.$$

[2] https://cran.r-project.org/web/packages/CompareCausalNetworks/index.html

[3] https://cran.r-project.org/web/packages/selectiveInference/

[4] Another candidate estimator can be utilized to learn conditional expected value and RR under the presence of complicated structures is RieszNet introduced by Chernozhukov et al. (2021).

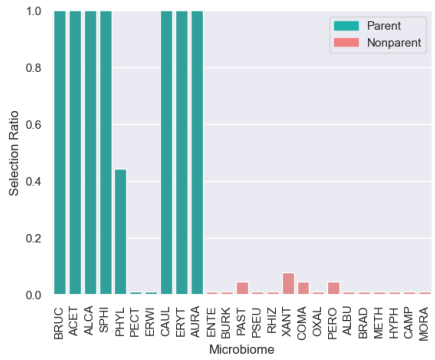


Figure 5: The ratio of simulations that each variable is selected by DRCFS to the total number of simulations (30) for linear target $f = f_1$. The ground truth about parental status is given in the legend. DRCDS captures most of the direct causes only with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$.

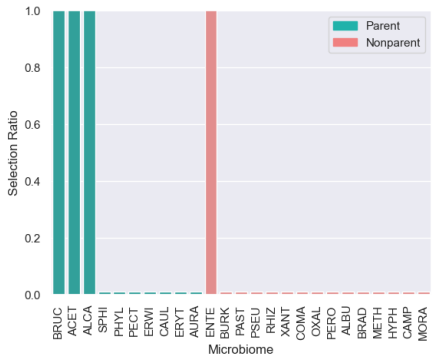


Figure 6: The ratio of simulations that each variable is selected by DRCFS to the total number of simulations (30) for Log-sum-exp target $f = f_6$. The ground truth about parental status is given in the legend. Despite the highly complicated structure and limited number of observations, DRCDS captures some of the direct causes only with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$.

Thus, our approach is in line with the works by Athey et al. (2019); Chernozhukov et al. (2021).

In order to effectively ensure fair evaluations against baselines, and to account for efficiency, the feature map used in our evaluations within this section is set to the identity function $\phi(\boldsymbol{X}) = \boldsymbol{X}$. However, it should be noted that alternative feature maps may be employed, to cater to specific requirements and prior domain knowledge of the dataset.

Results. Evaluation of the performance of DRCFS and the baselines for linear models ($f = f_1$) are illustrated w,r,t number of nodes in Figure 2. DRCFS significantly outperforms the baselines in high dimensions. The same plot for fully nonlinear models (Log-sum-exp $f = f_6$ and geometric mean $f = f_5$) are shown in Figures 3 and 4. Note that geometric mean has a highly nonlinear structure that most of the existing models cannot support theoretically. Similar plots for different settings, e.g., hide probability p_h, connectivity level p_c, linear and nonlinear transformation functions f, random variable ϵ are represented in Figures 7 to 14 (See Appendix G). DRCFS establishes reasonably good performance w.r.t these settings, even though identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$ is used,

5.2. Causal Feature Selection for Semi-synthetic Data on Microbiome Abundance

In this part, to assess the performance of our algorithm with a taste of real-world application, we conduct a semi-synthetic experiment based on microbiome abundance data in plant leaves from Regalado et al. (2020). The dataset contains shotgun sequencing of 275 wild *Arabidopsis Thaliana* leaf microbiomes. This shotgun data provides the ratio of microbial load to plant DNA on the host planet. Microbiome abundance datasets are known to have highly complicated underlying structures with hidden confounders such as the environment, host genetics, and other biological interactions (Lim et al., 2021; Regalado et al., 2020; Yang & Chen, 2022).

Dataset. Following the preprocessing, the dataset comprises 625 observations for 25 microorganisms (to serve as confounders $\boldsymbol{X}$) that exhibit the highest level of variations within the dataset. These microorganisms are given in Table 2. The natural target variables of interest for study often include the abundance of bacterial, eukaryotic, and total pathogenic organisms. Guided by this intuition, a subset of 10 confounders is randomly selected as direct causes from the entire set of confounders. Subsequently, target variables are constructed in adherence to Axioms Axiom (A)-Axiom (C) with different choices of function f.

Results. The selection ratio by DRCFS for parent and non-parent nodes is illustrated in Figures 5 and 6 for both linear and nonlinear target functions. Despite a limited number of observations, DRCFS is able to accurately identify most of the direct causes in the linear case and many in the nonlinear case, while maintaining a low false positive rate. Additional plots for other target functions f can be found in Figure 15 (See Appendix G).

5.3. Causal Feature Selection for Bnlearn Benchmarks

To further demonstrate the performance, scalability, and adaptability of our method across various domains, we consider four additional bnlearn (Scutari, 2010)[5] benchmarks with different properties. The results and details

[5] https://cran.r-project.org/web/packages/bnlearn/index.html

Table 1: Report on Performance of DRCFS on bnlearn benchmarks. The characteristics of the included networks are provided. To see the performance of DRCFS thoroughly on the underlying causal structure, please refer to Appendix G, Figures 16 to 19. In the table, BN tands for "Bayesian Network".

Category\Dataset	**ANDES**	**MEHRA**	**ALARM**	**ARTH150**
Type	Discrete BN	Conditional Linear Gaussian BN	Discrete BN	Gaussian BN
Number of nodes	223 (very large network)	24 (medium network)	37 (medium network)	107 (very large network)
Number of arcs	338	71	46	150
Number of parameters	1157	324423	509	364
Average Markov blanket size	5.61	13.75	3.51	3.35
Average degree	3.03	5.92	2.49	2.8
Maximum in-degree	6	13	4	6
Chosen target name	SNode_65	pm2.5	PRSS	786
Accuracy	**0.973**	**0.74**	**0.833**	**0.915**

on these networks are reported in Table 1. The semi-synthetic datasets that we include are ANDES (Conati et al., 1997) (a very large size discrete Bayesian network), MEHRA (Vitolo et al., 2018) (a medium size conditional linear Gaussian Bayesian Network), ALARM (Beinlich et al., 1989) (a medium size discrete Bayesian network) and ARTH150 (Opgen-Rhein & Strimmer, 2007) (a very large size Gaussian Bayesian network). The diagrams of these networks with the inferred direct causes by DRCFS are depicted in Appendix G, Figures 16 to 19. Notably, our method shows good performance in all instances, with very few false positives and no false negatives. In essence, it effectively captures the entirety of direct causes while maintaining a substantially low incidence of wrongly selected variables.

6. Discussion

Limitations. In this work we have focused on the case of learning causal features from observational data rather than full causal discovery. Moreover, while we have shown $\sqrt{n}$-consistency, the expectations in Algorithm 1 might in practice only be approximated well if enough samples are available to perform the required sample splitting efficiently and without sacrificing approximation quality. Finally, since our method is inspired by double machine learning-based approaches (Rotnitzky et al., 2020), the focus is on bias rather than variance reduction of estimates.

Conclusion and future work. We presented DRCFS, a doubly robust feature selection method for identifying causal features even in high-dimensional and nonlinear settings. In extensive experiments including state-of-the-art baselines, we demonstrated that our approach is significantly more accurate across various metrics and across a wide range of settings in synthetic and semi-synthetic datasets. Future work could see the extension of our work, especially to biomedical settings where the evaluation can not be conducted with respect to the ground truth or to representation learning approaches where features first have to be learned before they can be selected. Another potential avenue for future research is the extension to time series data. In the long-term, we hope that these future algorithms as well as the provided feature selection method will overall enable users to better understand, interpret and explain the results of machine learning-based decision-making.

7. Acknowledgments

This work was partly supported by Digital Futures project EXTREMUM. This project received partial support from ELSA: European Lighthouse on Secure and Safe AI project (grant agreement No. 101070617 under UK guarantee) and the ERC under the European Union's Horizon 2020 research and innovation programme (FUN2MODEL, grant agreement No. 834115). This work was partly supported by the Swedish Research Council under contract number 2016-06079 (NewLEADS). The research was partially supported by a grant from the MIT-IBM Watson AI Lab.

References

Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Beinlich, I. A., Suermondt, H. J., Chavez, R. M., and Cooper, G. F. The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In *AIME 89: Second European Conference on Artificial Intelligence in Medicine, London, August 29th–31st 1989. Proceedings*, pp. 247–256. Springer, 1989.

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

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. *Feature Selection for High-Dimensional Data*. Springer, 2015.

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. Foundations of structural causal models with cycles and latent

variables. *The Annals of Statistics*, 49(5):2885 – 2915, 2021.

Butcher, B. and Smith, B. J. Feature engineering and selection: A practical approach for predictive models: by max kuhn and kjell johnson. boca raton, fl: Chapman & hall/crc press, 2019, xv+ 297 pp., 79.95 (h), isbn: 978-1-13-807922-9., 2020.

Cawley, G. C. Causal & non-causal feature selection for ridge regression. In *Causation and Prediction Challenge*, 2008.

Chernozhukov, V., Newey, W. K., Quintas-Martinez, V., and Syrgkanis, V. Automatic debiased machine learning via neural nets for generalized linear regression. *arXiv:2104.14737*, 2021.

Chernozhukov, V., Newey, W., Quintas-Martinez, V., and Syrgkanis, V. RieszNet and ForestRiesz: Automatic debiased machine learning with neural nets and random forests. In *Proc. of ICML*, pp. 3901–3914, 2022.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3: 507–554, 2003.

Chickering, M., Heckerman, D., and Meek, C. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

Claassen, T., Mooij, J. M., and Heskes, T. Learning sparse causal models is not np-hard. In *Proc. of UAI*, pp. 172–181, 2013.

Colombo, D. and Maathuis, M. H. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(116):3921–3962, 2014.

Conati, C., Gertner, A. S., VanLehn, K., and Druzdzel, M. J. On-line student modeling for coached problem solving using bayesian networks. In *User Modeling: Proceedings of the Sixth International Conference UM97 Chia Laguna, Sardinia, Italy June 2–5 1997*, pp. 231–242. Springer, 1997.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Guyon, I. Practical feature selection: from correlation to causality. *Mining massive data sets for security: advances in data mining, search, social networks and text mining, and their applications to security*, pp. 27–43, 2008.

Guyon, I. and Aliferis, C. Causal feature selection. In *Computational methods of feature selection*. Chapman and Hall/CRC, 2007.

Heinze-Deml, C., Maathuis, M. H., and Meinshausen, N. Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391, 2018.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Proc. of NIPS*, pp. 689–696, 2008.

Hu, P., Jiao, R., Jin, L., and Xiong, M. Application of causal inference to genomic analysis: advances in methodology. *Frontiers in Genetics*, 9:238, 2018.

Janzing, D., Minorics, L., and Blöbaum, P. Feature relevance quantification in explainable ai: A causal problem. In *Proc. of AISTATS*, pp. 2907–2916, 2020.

Javanmard, A. et al. De-biasing the lasso: Optimal sample size for gaussian designs. arxiv, 2015.

Kira, K. and Rendell, L. A. A practical approach to feature selection. In *Machine Learning Proceedings 1992*, pp. 249–256. Elsevier, 1992.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. Feature selection: A data perspective. *ACM Computing Surveys*, 50(6):1–45, 2017.

Lim, M. Y., Hong, S., Bang, S.-J., Chung, W.-H., Shin, J.-H., Kim, J.-H., and Nam, Y.-D. Gut microbiome structure and association with host factors in a korean population. *Msystems*, 6(4):e00179–21, 2021.

Liu, H. and Motoda, H. *Computational Methods of Feature Selection*. CRC press, 2007.

Loftus, J. R. and Taylor, J. E. A significance test for forward stepwise model selection. *arXiv:1405.3920*, 2014.

Lv, B.-M., Quan, Y., and Zhang, H.-Y. Causal inference in microbiome medicine: principles and applications. *Trends in microbiology*, 29(8):736–746, 2021.

Maclaren, O. J. and Nicholson, R. What can be estimated? identifiability, estimability, causal inference and ill-posed inverse problems. *arXiv:1904.02826*, 2019.

Mastakouri, A., Schölkopf, B., and Janzing, D. Selecting causal brain features with a single conditional independence test per feature. In *Proc. of NeurIPS*, pp. 12532–12543, 2019.

Mehrjou, A., Soleymani, A., Jesson, A., Notin, P., Gal, Y., Bauer, S., and Schwab, P. Genedisco: A benchmark for experimental design in drug discovery. *arXiv:2110.11875*, 2021.

Murphy, K. P. Active learning of causal bayes net structure. Technical report, technical report, UC Berkeley, 2001.

Nandy, P., Hauser, A., and Maathuis, M. High-dimensional consistency in score-based and hybrid structure learning. *Annals of Statistics*, 46, 03 2016.

Opgen-Rhein, R. and Strimmer, K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC systems biology*, 1(1):1–10, 2007.

Paul, M. Feature selection as causal inference: Experiments with text classification. In *Proc. of CoNLL*, pp. 163–172, 2017.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

Pearl, J. Direct and indirect effects. In *Proc. of UAI*, pp. 411–420, 2001.

Pearl, J. *Causality: Models, Reasoning and Inference, 2nd Ed.* Cambridge University Press, 2009.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Identifiability of causal graphs using functional models. In *Proc. of UAI*, pp. 589–598, 2011.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Regalado, J., Lundberg, D. S., Deusch, O., Kersten, S., Karasov, T., Poersch, K., Shirsekar, G., and Weigel, D. Combining whole-genome shotgun sequencing and rrna gene amplicon analyses to improve detection of microbe–microbe interaction networks in plant leaves. *The ISME Journal*, 14(8):2116–2130, 2020.

Rotnitzky, A., Smucler, E., and Robins, J. M. Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238, 08 2020.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, 2019.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. M. On causal and anticausal learning. In *Proc. of ICML*, 2012.

Scutari, M. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010. doi: 10.18637/jss.v035.i03.

Shah, R. D. and Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct): 2003–2030, 2006a.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. J. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7: 2003–2030, 2006b.

Siebert, J. Applications of statistical causal inference in software engineering. *arXiv:2211.11482*, 2022.

Soleymani, A., Raj, A., Bauer, S., Schölkopf, B., and Besserve, M. Causal feature selection via orthogonal search. *Transactions on Machine Learning Research*, 2022.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search, 2nd Ed.* MIT Press, 2000.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

Tsamardinos, I., Brown, L., and Aliferis, C. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 10 2006.

Tsoumas, I., Giannarakis, G., Sitokonstantinou, V., Koukos, A., Loka, D., Bartsotas, N., Kontoes, C., and Athanasiadis, I. Evaluating digital tools for sustainable agriculture using causal inference. *arXiv:2211.03195*, 2022.

Varian, H. R. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113 (27):7310–7315, 2016.

Vitolo, C., Scutari, M., Ghalaieny, M., Tucker, A., and Russell, A. Modeling air pollution, climate, and health data using bayesian networks: A case study of the english regions. *Earth and Space Science*, 5(4):76–88, 2018.

Williams, D. *Probability with Martingales*. Cambridge University Press, 1991.

Yang, L. and Chen, J. A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. *Microbiome*, 10(1):130, 2022.

Yu, K., Guo, X., Liu, L., Li, J., Wang, H., Ling, Z., and Wu, X. Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys*, 53(5):1–36, 2020.

Yu, K., Liu, L., and Li, J. A unified view of causal and non-causal feature selection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(4):1–46, 2021.

A. Structural Causal Models

Definition A.1 (Structural Causal Model (SCM), Definition 2.1 by Bongers et al. (2021))**.** A structural causal model (SCM) is a tuple $\langle \mathtt{I}, \mathtt{J}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{U}} \rangle$ where (i) $\mathtt{I}$ is a finite index set of endogenous variables; (ii) $\mathtt{J}$ is a disjoint finite index set of exogenous variables; (iii) $\boldsymbol{V} = \prod_{j \in \mathtt{I}} V_j$ is the product of the domains of the endogenous variables, where each V_j is a standard measurable space; (iv) $\boldsymbol{U} = \prod_{j \in \mathtt{J}} U_j$ is the product of the domains of the exogenous variables, where each U_j is a standard measurable space; (v) $\boldsymbol{f} \colon \boldsymbol{V} \times \boldsymbol{U} \to \boldsymbol{V}$ is a measurable function that specifies the causal mechanism; (vi) $\mathbb{P}_{\boldsymbol{U}} = \prod_{j \in \mathtt{J}} \mathbb{P}_{\boldsymbol{U}_j}$ is a product measure, where $\mathbb{P}_{\boldsymbol{U}_j}$ is a probability measure on $\boldsymbol{U}_j$ for each $j \in \mathtt{J}$.

In SCMs, the functional relationships between variables are expressed in terms of deterministic equations. This feature allows us to model the cause-effect relationships of the data-generating process (DGP) using *structural equations*. For a given SCM $\langle \mathtt{I}, \mathtt{J}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{U}} \rangle$ a structural equation specifies an endogenous random variable V_l via a measurable function of the form $V_l = f_{V_l}(\boldsymbol{V}, \boldsymbol{U})$ for all $l \in \mathtt{I}$. A *parent* $i \in \mathtt{I} \cup \mathtt{J}$ of l is any index for which there is no measurable function $g \colon \prod_{j \in \mathtt{I} \setminus \{i\}} V_j \times \boldsymbol{U} \to V_l$ with $f_{V_l} = g$ almost surely. Intuitively, each endogenous variables V_j is specified by its parents together with the exogenous variables, via the structural equations. A structural equations model as in Definition A.1 can be conveniently described with the *causal graph*, a directed graph of the form $\mathcal{G} = (\mathtt{I} \cup \mathtt{J}, \mathcal{E})$. The nodes of the causal graph consist of the entire set of indices for the variables, and the edges are specified by the structural equations, i.e., $\{j \to l\} \in \mathcal{E}$ iff j is a parent of l. Note that the variables in the set $\mathsf{Pa}(V_l)$ are indexed by the parent nodes of l in the corresponding graph $\mathcal{G}$.

A.1. Interventions

We define the causal semantics of SCMs, by considering perfect interventions (Pearl, 2000). For a given a SCM as in Definition A.1, consider a variable $\boldsymbol{W} \coloneqq \prod_{j \in \mathtt{I}'} V_j$ for a set $\mathtt{I}' \subseteq \mathtt{I}$, and let $\boldsymbol{w} \coloneqq \prod_{j \in \mathtt{I}'} v_j$ be a point of its domain. The perfect intervention $\boldsymbol{W} \leftarrow \boldsymbol{w}$ amounts to replacing the structural equations $V_j = f_{V_j}(\boldsymbol{V}, \boldsymbol{U})$ with the constant functions $V_j \equiv v_j$ for all $j \in \mathtt{I}'$. We denote with $V_l \mid do(\boldsymbol{w})$ the variable V_l after performing the intervention. This procedure define a new probability distribution $\mathbb{P}(v_l \mid do(\boldsymbol{w}))$, which we refer to as interventional distribution. This distribution entails the following information: "Given that we have observed $\boldsymbol{W} = \boldsymbol{w}$, what would V_l have been had we set $do(\boldsymbol{w})$, instead of the value $\boldsymbol{W}$ had actually taken?".

B. Unique Solvability

We introduce the notions of solvability and unique solvability as defined by (). These notions describe the existence and uniqueness of measurable solution functions for the structural equations of a given subset of the endogenous variables. Solvability of an SCM is a sufficient and necessary condition for the existence of a solution of an SCM, and unique solvability implies the uniqueness of the induced observational distribution.

The notion of solvability is defined as follows.

Definition B.1 (Solvability, following Definition 3.1 by (Bongers et al., 2021))**.** Consider an SCM $\langle \mathtt{I}, \mathtt{J}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{U}} \rangle$. We say that the SCM is solvable if there exists a measurable mapping $g \colon \mathbf{V} \to \mathbf{U}$ such that $\boldsymbol{v} = g(\boldsymbol{u}) \Rightarrow \boldsymbol{v} = \boldsymbol{f}(\boldsymbol{v}, \boldsymbol{u})$ almost surely.

The unique solvability of an SCM is a stronger notion than mere solvability, and is defined as follows.

Definition B.2 (Unique Solvability, following Definition 3.3 by (Bongers et al., 2021))**.** Consider an SCM $\langle \mathtt{I}, \mathtt{J}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{U}} \rangle$. We say that the SCM is uniquely solvable if there exists a measurable mapping $g \colon \mathbf{V} \to \mathbf{U}$ such that

$$\boldsymbol{v} = g(\boldsymbol{u}) \Leftrightarrow \boldsymbol{v} = \boldsymbol{f}(\boldsymbol{v}, \boldsymbol{u})$$

almost surely.

The unique solvability condition essentially ensures that there exists a measurable solution for the structural equations, and that any possible solution induces the same observational distribution. Bongers et al. (2021) provide the following necessary and sufficient conditions for the unique solvability of an SCM.

Theorem B.3 (Following Theorem 3.6 by Bongers et al. (2021))**.** *Consider an SCM $\langle \mathtt{I}, \mathtt{J}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{U}} \rangle$. Then, the system of structural equations $\boldsymbol{V} = \boldsymbol{f}(\boldsymbol{V}, \boldsymbol{U})$ has a unique solution almost surely, if and only if the SCM is uniquely solvable. Furthermore, if the SCM is uniquely solvable, then there exists a solution, and all solutions have the same observational distribution.*

C. Proof of Lemma 3.1

Lemma 3.1. *Consider a causal model as in Axioms Axiom (A)-Axiom (C), and fix a feature X_j. Then, $\mathbb{E}[Y \mid do(x_j, \boldsymbol{x}_j^c)] - \mathbb{E}[Y \mid do(x_j', \boldsymbol{x}_j^c)] \neq 0$ for some interventional values x_j, x_j' and $\boldsymbol{x}_j^c$ if and only if $X_j \in \mathsf{Pa}(Y)$.*

Proof. Note that $X_j \notin \mathsf{Pa}(Y)$ if and only if $Y \mid do(x_j, \boldsymbol{x}_j^c) = Y \mid do(x_j', \boldsymbol{x}_j^c)$ for all possible interventional values $x_j, \boldsymbol{x}_j^c$. Hence, the first part of the claim follows by showing that $Y \mid do(x_j, \boldsymbol{x}_j^c) = Y \mid do(x_j', \boldsymbol{x}_j^c)$ if and only if $\mathrm{ACDE}(x_j, x_j' | \boldsymbol{x}_j^c) = 0$ almost surely. If $Y \mid do(x_j, \boldsymbol{x}_j^c) = Y \mid do(x_j', \boldsymbol{x}_j^c)$, it directly follows that $\mathrm{ACDE}(x_j, x_j' | \boldsymbol{x}_j^c) = 0$ almost surely, so it only remains to establish the converse. To this end, suppose that $\mathrm{ACDE}(x_j, x_j' | \boldsymbol{x}_j^c) = 0$, and define the group $\boldsymbol{Z} = \mathsf{Pa}(Y)$ consisting of all the parents of the outcome. Note that $\boldsymbol{Z} \subseteq \{X_j\} \cup \boldsymbol{X}_j^c$, since $\{X_j\} \cup \boldsymbol{X}_j^c$ consists of all observed variables of the model. Hence, the intervention $\{X_j, \boldsymbol{X}_j^c\} \leftarrow \{x_j, \boldsymbol{x}_j^c\}$ defines an intervention on the parents $\boldsymbol{Z} \leftarrow \boldsymbol{z}$, with $\boldsymbol{z}$ a sub-vector of $\{x_j, \boldsymbol{x}_j^c\}$. Further, we can write the potential outcome as $Y \mid do(x_j, \boldsymbol{x}_j^c) = f(\boldsymbol{z}) + \varepsilon$. Similarly, the intervention $\{X_j, \boldsymbol{X}_j^c\} \leftarrow \{x_j', \boldsymbol{x}_j^c\}$ defines an intervention of the form $\boldsymbol{Z} \leftarrow \boldsymbol{z}'$, and it follows that $Y \mid do(x_j', \boldsymbol{x}_j^c) = f(\boldsymbol{z}') + \varepsilon$. Therefore,

$$f(\boldsymbol{z}) + \mathbb{E}[\varepsilon] = \mathbb{E}[Y \mid do(x_j, \boldsymbol{x}_j^c)] = \mathbb{E}[Y \mid do(x_j', \boldsymbol{x}_j^c)] = f(\boldsymbol{z}') + \mathbb{E}[\varepsilon]. \tag{7}$$

where the first and the third equalities follow since ε is exogenous independent noise, and the second equality follows since $\mathrm{ACDE}(x_j, x_j' | \boldsymbol{x}_j^c) = 0$. From Eq. (7) we conclude that $Y \mid do(x_j, \boldsymbol{x}_j^c) = Y \mid do(x_j', \boldsymbol{x}_j^c)$, as claimed. □

D. Proof of Lemma 3.2

Lemma 3.2. *Consider a causal model as in Axioms Axiom (A)-Axiom (C). Then, $\mathbb{E}[Y \mid do(x_j, \boldsymbol{x}_j^c)] - \mathbb{E}[Y \mid do(x_j', \boldsymbol{x}_j^c)] \neq 0$ almost surely if and only if $\chi_j \neq 0$. It follows that $\chi_j \neq 0$ if and only if $X_j \in \mathsf{Pa}(Y)$.*

Proof. We first show that the claim follows if

$$\mathbb{E}[Y \mid do(x_j, \boldsymbol{x}_j^c)] = \mathbb{E}[Y \mid x_j, \boldsymbol{x}_j^c] \quad \text{a.s.}, \tag{8}$$

and then we will prove Eq. (8). To this end, assume that Eq. (8) holds and suppose that $\chi_j = 0$, i.e., $\mathbb{E}[Y | x_j, \boldsymbol{x}_j^c] = \mathbb{E}[Y | \boldsymbol{x}_j^c] = \mathbb{E}[Y | x_j', \boldsymbol{x}_j^c]$ a.s. Then,

$$\mathbb{E}[Y \mid do(x_j, \boldsymbol{x}_j^c)] = \mathbb{E}[Y \mid x_j, \boldsymbol{x}_j^c] = \mathbb{E}[Y \mid x_j', \boldsymbol{x}_j^c] = \mathbb{E}[Y \mid do(x_j', \boldsymbol{x}_j^c)] \quad \text{a.s.}$$

Here, the first and third equalities follow from Eq. (8). It follows that $\mathrm{ACDE}(x_j, x_j' | \boldsymbol{z}) = 0$ a.s. Conversely, suppose that Eq. (8) holds, and that $\mathrm{ACDE}(x_j, x_j' | \boldsymbol{z}) = 0$. Then, it holds that $\mathbb{E}[Y \mid do(x_j, \boldsymbol{x}_j^c)] = \mathbb{E}[Y \mid do(x_j', \boldsymbol{x}_j^c)]$ a.s., that is,

$$\mathbb{E}[Y \mid do(x_j, \boldsymbol{x}_j^c)] = \mathbb{E}\left[\mathbb{E}[Y \mid do(x_j, \boldsymbol{x}_j^c)] \mid \boldsymbol{x}_j^c\right]. \tag{9}$$

Hence,

$$\begin{aligned}
\mathbb{E}[Y \mid x_j, \boldsymbol{x}_j^c] &= \mathbb{E}[Y \mid do(x_j, \boldsymbol{x}_j^c)] && \text{[by Eq. (8)]} \\
&= \mathbb{E}[\mathbb{E}[Y \mid do(x_j, \boldsymbol{x}_j^c)] \mid \boldsymbol{x}_j^c] && \text{[by Eq. (9)]} \\
&= \mathbb{E}\left[\mathbb{E}[Y \mid x_j, \boldsymbol{x}_j^c] \mid \boldsymbol{x}_j^c\right] && \text{[by Eq. (8)]} \\
&= \mathbb{E}[Y \mid \boldsymbol{x}_j^c],
\end{aligned}$$

and the claim follows.

We conclude the proof by showing that Eq. (8) holds. To this end, define the group $\boldsymbol{Z} = \mathsf{Pa}(Y)$ consisting of all the parents of the outcome. Note that $\boldsymbol{Z} \subseteq \{X_j\} \cup \boldsymbol{X}_j^c$, since $\{X_j\} \cup \boldsymbol{X}_j^c$ consists of all observed variables of the model. By Axioms (A)-(C), the outcome can be described as $Y = f(\boldsymbol{Z}) + \varepsilon$, where ε is independent of $\{X_j\} \cup \boldsymbol{X}_j^c$. Hence by Rule 2 of the do-calculus (Pearl, 2000, page 85), $Y \mid do(x_j, \boldsymbol{x}_j^c) \sim Y \mid x_j, \boldsymbol{x}_j^c$, because Y becomes independent of $\{X_j\} \cup \boldsymbol{X}_j^c$ once all arrows from Z to Y are removed from the graph of the DGP. Therefore, $\mathbb{E}[Y \mid do(x_j, \boldsymbol{x}_j^c)] = \mathbb{E}[Y \mid x_j, \boldsymbol{x}_j^c]$. □

E. Proof of Lemma 3.3

Lemma 3.3. *Let χ_j be as in Eq. (5), and $\boldsymbol{X} = \{X_j\} \cup \boldsymbol{X}_j^c$. For any square-integrable random variable $g(\boldsymbol{X})$, consider the moment functional $m_0(\boldsymbol{V}; g) := Yg(\boldsymbol{X})$. Similarly, for any square-integrable random variable $h(\boldsymbol{X}_j^c)$, consider the moment functional $m_j(\boldsymbol{V}; h) := Yh(\boldsymbol{X}_j^c)$.[6] Then, it holds that*

$$\chi_j = \mathbb{E}[m_0(\boldsymbol{V}; g_0)] - \mathbb{E}[m_j(\boldsymbol{V}; h_0)],$$

with $g_0(\boldsymbol{X}) = \mathbb{E}[Y \mid \boldsymbol{X}]$ and $h_0(\boldsymbol{X}_j^c) = \mathbb{E}[Y \mid \boldsymbol{X}_j^c]$.

Proof. Since $\boldsymbol{X}_j^c \subset \boldsymbol{X}$, we have by the tower property of the expectation (Williams, 1991) that

$$\begin{aligned}
\chi_j &= \mathbb{E}[\big(\mathbb{E}[Y \mid \boldsymbol{X}] - \mathbb{E}[Y \mid \boldsymbol{X}_j^c]\big)^2] \\
&= \mathbb{E}[\mathbb{E}[\big(\mathbb{E}[Y \mid \boldsymbol{X}] - \mathbb{E}[Y \mid \boldsymbol{X}_j^c]\big)^2 \mid \boldsymbol{X}_j^c]] \\
&= \mathbb{E}[\mathbb{E}[(\mathbb{E}[Y \mid \boldsymbol{X}]^2 - \mathbb{E}[Y \mid \boldsymbol{X}]\mathbb{E}[Y \mid \boldsymbol{X}_j^c]) \mid \boldsymbol{X}_j^c]] \\
&= \mathbb{E}[\mathbb{E}[Y \mid \boldsymbol{X}]^2] - \mathbb{E}[\mathbb{E}[(\mathbb{E}[Y \mid \boldsymbol{X}]\mathbb{E}[Y \mid \boldsymbol{X}_j^c]) \mid \boldsymbol{X}_j^c]] \\
&= \mathbb{E}[\mathbb{E}[Y \mid \boldsymbol{X}]^2] - \mathbb{E}[\mathbb{E}[Y \mid \boldsymbol{X}_j^c]^2] \\
&= \mathbb{E}[Y\mathbb{E}[Y \mid \boldsymbol{X}]] - \mathbb{E}[Y\mathbb{E}[Y \mid \boldsymbol{X}_j^c]].
\end{aligned}$$

The claim follows since $m_0(\boldsymbol{V}; g_0) = Y\mathbb{E}[Y \mid \boldsymbol{X}]$ and $m_j(\boldsymbol{V}; h_0) = Y\mathbb{E}[Y \mid \boldsymbol{X}_j^c]$. □

F. Experimental Setup

F.1. Evaluation Metrics

We use Accuracy, F1 Score (harmonic mean of precision and sensitivity) and Critical Success Index (CSI) as metrics to assess the performance of the algorithms. Given the number of true positives TP, true negatives TN, false positives FP, and false negatives FN, these metrics are defined as,

- $\text{ACC} = \dfrac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$
- $\text{F1} = \dfrac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$
- $\text{CSI} = \dfrac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$

The reason that we consider F1 score, and CSI is the emphasis placed on the number of true positives within their calculations.

F.2. Designs for Transformation Function f

Distributions over the different choices of the transformation function f are used in the experiments to generate the variables based on $X = f(\mathsf{Pa}(X)) + \varepsilon$, for a subset $\mathsf{Pa}(X) \subseteq \boldsymbol{X}$:

- **Linear**:

$$f_1(\mathsf{Pa}(X)) = a \sum_{X' \in \mathsf{Pa}(X)} X' + b,$$

with $a = 0.5$ and $b = 0$, unless stated exactly otherwise.

- **Sum-sqrt**:

$$f_2(\mathsf{Pa}(X)) = a \sum_{X' \in \mathsf{Pa}(X)} \sqrt{|X'|} + b,$$

with $a = 0.5$ and $b = 0$, unless stated exactly otherwise.

- **Sum-sine**:

$$f_3(\mathsf{Pa}(X)) = a \sum_{X' \in \mathsf{Pa}(X)} \sin(c.X') + b,$$

with $a = 1, b = 0$ and $c = 0.5$, unless stated exactly otherwise.

- **Sum-tanh**:

$$f_4(\mathsf{Pa}(X)) = a \sum_{X' \in \mathsf{Pa}(X)} \tanh(c.X') + b,$$

with $a = 1, b = 0$ and $c = 2$, unless stated exactly otherwise.

[6] Note that $m_0(\boldsymbol{V}; g)$ and $m_j(\boldsymbol{V}; h)$ are distinct functionals, since they are defined over sets of functions with different domains.

- **Geometric mean**:

$$f_5(\mathsf{Pa}(X)) = a \prod_{X' \in \mathsf{Pa}(X)} |X'|^{\frac{1}{\mathbf{card}(\mathsf{Pa}(X))}} + b,$$

with $a = 3$ and $b = 0.1$, unless stated exactly otherwise.

- **Log-sum-exp**:

$$f_6(\mathsf{Pa}(X)) = a \log \Big(\sum_{X' \in \mathsf{Pa}(X)} e^{X'} \Big) + b,$$

with $a = 1$ and $b = \log 2$, unless stated exactly otherwise.

- **Sqrt-sum**:

$$f_7(\mathsf{Pa}(X)) = a \sqrt{\Big| \sum_{X' \in \mathsf{Pa}(X)} X' \Big|} + b,$$

with $a = 1$ and $b = 0$, unless stated exactly otherwise.

F.3. Semi-synthetic Data on Microbiome Abundance

The 25 covariates that demonstrate the highest variations in microbiome abundance within the dataset provided by Regalado et al. (2020) are listed in Table 2. These covariates subsume a diverse set of Bacteria/Eukaryote groups that are common in leaves, soil, and water.

Table 2: 25 covariates of the microbiome abundance with highest variations within the dataset provided by Regalado et al. (2020).

Microbiome	Abbreviation
Enterobacteriaceae	ENTE
Burkholderiaceae	BURK
Pasteurellaceae	PAST
Brucellaceae	BRUC
Pseudomonadaceae	PSEU
Acetobacteraceae	ACET
Alcaligenaceae	ALCA
Rhizobiaceae	RHIZ
Sphingomonadaceae	SPHI

Microbiome	Abbreviation
Xanthomonadaceae	XANT
Comamonadaceae	COMA
Phyllobacteriaceae	PHYL
Oxalobacteraceae	OXAL
Peronosporaceae	PERO
Albuginaceae	ALBU
Pectobacteriaceae	PECT
Bradyrhizobiaceae	BRAD

Microbiome	Abbreviation
Erwiniaceae	ERWI
Caulobacteraceae	CAUL
Methylobacteriaceae	METH
Hyphomicrobiaceae	HYPH
Erythrobacteraceae	ERYT
Campylobacteraceae	CAMP
Aurantimonadaceae	AURA
Moraxellaceae	MORA

G. Additional Plots

In this section, additional plots for the experiments are provided.

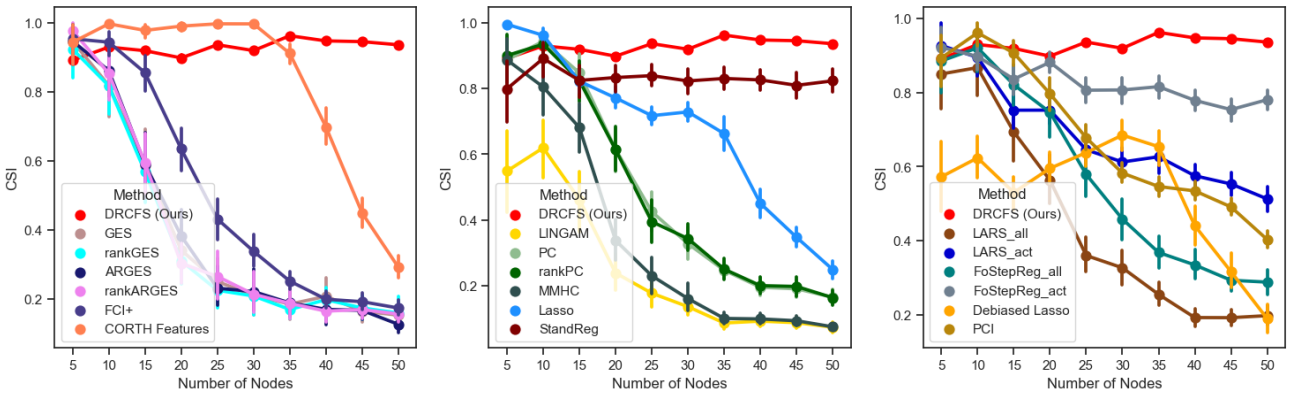


Figure 7: Performance (CSI) of the algorithms w.r.t. number of nodes m for fully linear causal structures ($f = f_1$ with probability 1), where $p_s = 0.3$, $p_h = 0$, and $\epsilon \sim \mathcal{N}(0, 1)$. Each case is averaged over 50 simulations. ForestRiesz with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$ has been used in these experiments. DRCFS's performance shows stability even for large graphs while the baselines suffer in high dimensions.

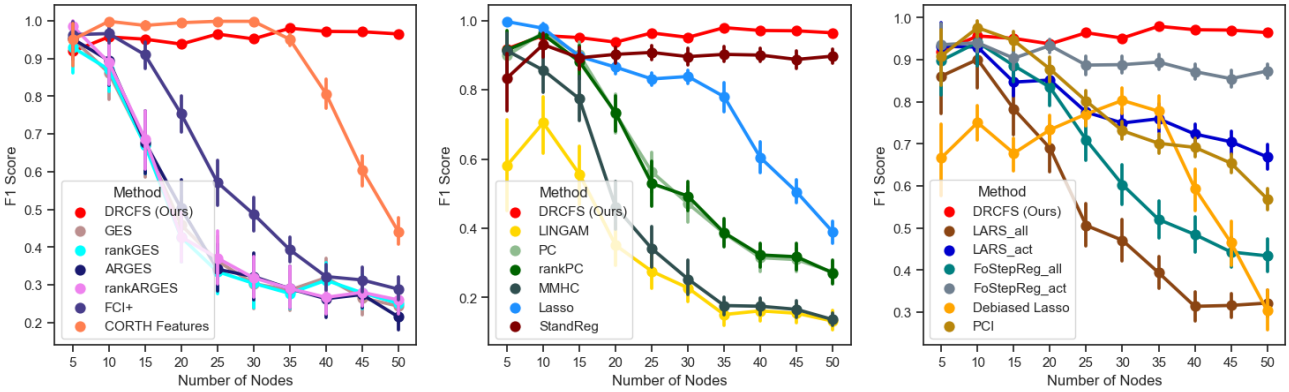


Figure 8: Performance (F1 score) of the algorithms w.r.t. number of nodes m for fully linear causal structures ($f = f_1$ with probability 1), where $p_s = 0.3$, $p_h = 0$, and $\epsilon \sim \mathcal{N}(0, 1)$. Each case is averaged over 50 simulations. ForestRiesz with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$ has been used in these experiments. DRCFS's performance shows stability even for large graphs while the baselines suffer in high dimensions.

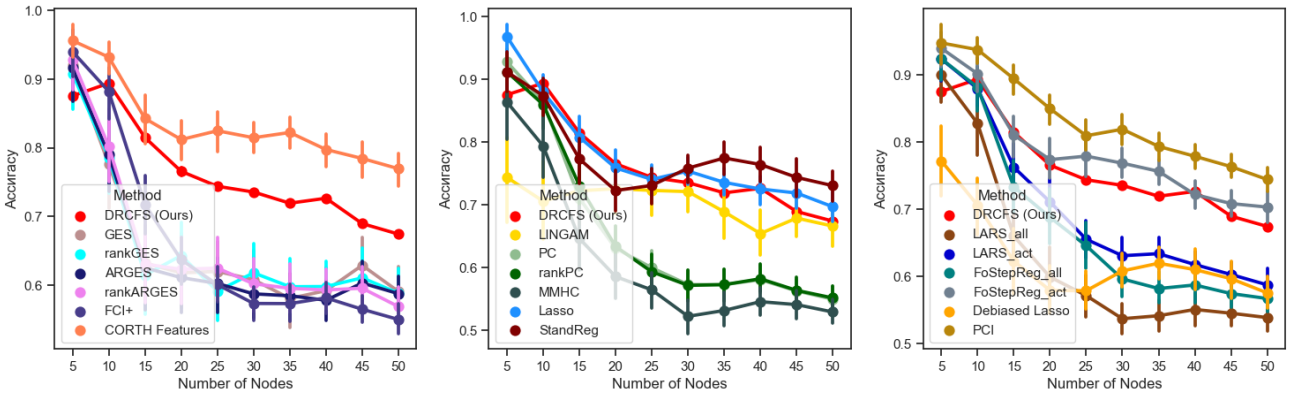


Figure 9: Performance (Accuracy) of the algorithms w.r.t. number of nodes m for causal structures defined as linear combinations of nonlinear functions ($f = f_2$ with probability 0.5, $f = f_3$ with probability 0.25, $f = f_4$ with probability 0.25), where $p_s = 0.5$, $p_h = 0$, and $\epsilon \sim \beta(2, 5)$. Each case is averaged over 50 simulations. ForestRiesz with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$ has been used in these experiments. DRCFS shows reasonably good performance, even though the identity feature map is used. Nevertheless, alternative feature maps taking into account prior domain knowledge of the dataset could be used. CORTH Features dominates others because in the regime that the noise is defined, summations of non-linear terms in DGP fro the target variable act approximately linear, hence, these results are consistent with Soleymani et al. (2022).

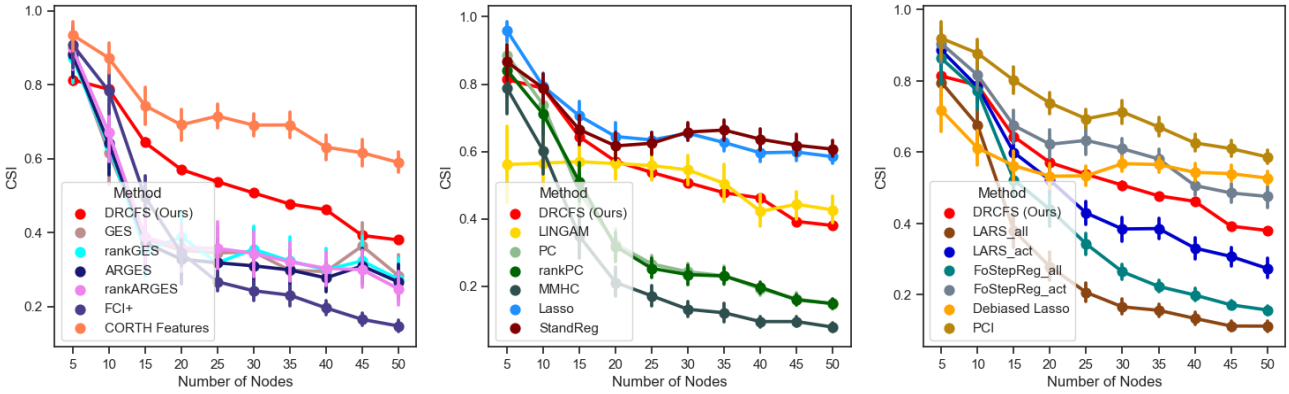


Figure 10: Performance (CSI) of the algorithms w.r.t. number of nodes m for causal structures defined as linear combinations of nonlinear functions ($f = f_2$ with probability 0.5, $f = f_3$ with probability 0.25, $f = f_4$ with probability 0.25), where $p_s = 0.5$, $p_h = 0$, and $\epsilon \sim \beta(2, 5)$. Each case is averaged over 50 simulations. ForestRiesz with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$ has been used in these experiments. DRCFS shows reasonably good performance, even though the identity feature map is used. Nevertheless, alternative feature maps taking into account prior domain knowledge of the dataset could be used. CORTH Features dominates others because in the regime that the noise is defined, summations of non-linear terms in DGP fro the target variable act approximately linear, hence, these results are consistent with Soleymani et al. (2022).

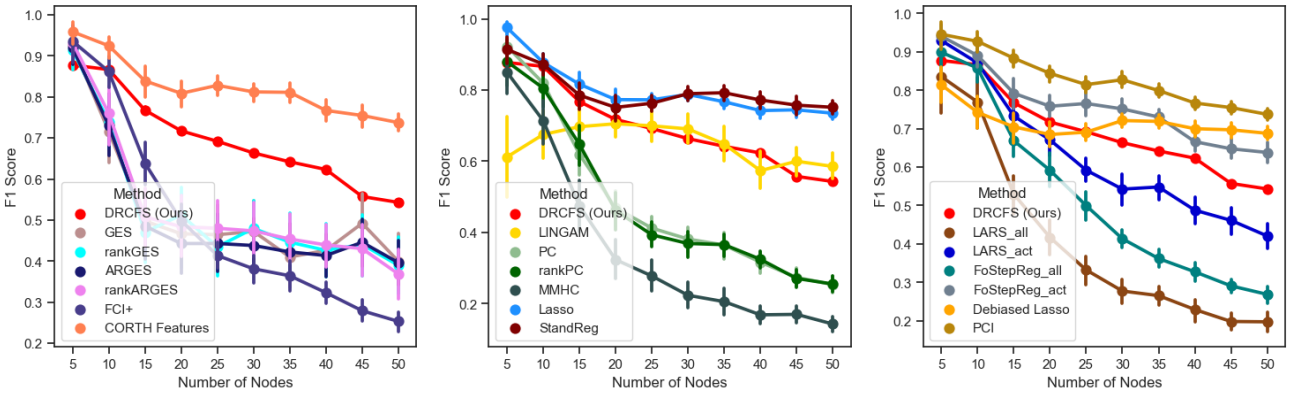


Figure 11: Performance (F1 score) of the algorithms w.r.t. number of nodes m for causal structures defined as linear combinations of nonlinear functions ($f = f_2$ with probability 0.5, $f = f_3$ with probability 0.25, $f = f_4$ with probability 0.25), where $p_s = 0.5$, $p_h = 0$, and $\epsilon \sim \beta(2, 5)$. Each case is averaged over 50 simulations. ForestRiesz with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$ has been used in these experiments. DRCFS shows reasonably good performance, even though the identity feature map is used. Nevertheless, alternative feature maps taking into account prior domain knowledge of the dataset could be used. CORTH Features dominates others because in the regime that the noise is defined, summations of non-linear terms in DGP fro the target variable act approximately linear, hence, these results are consistent with Soleymani et al. (2022).

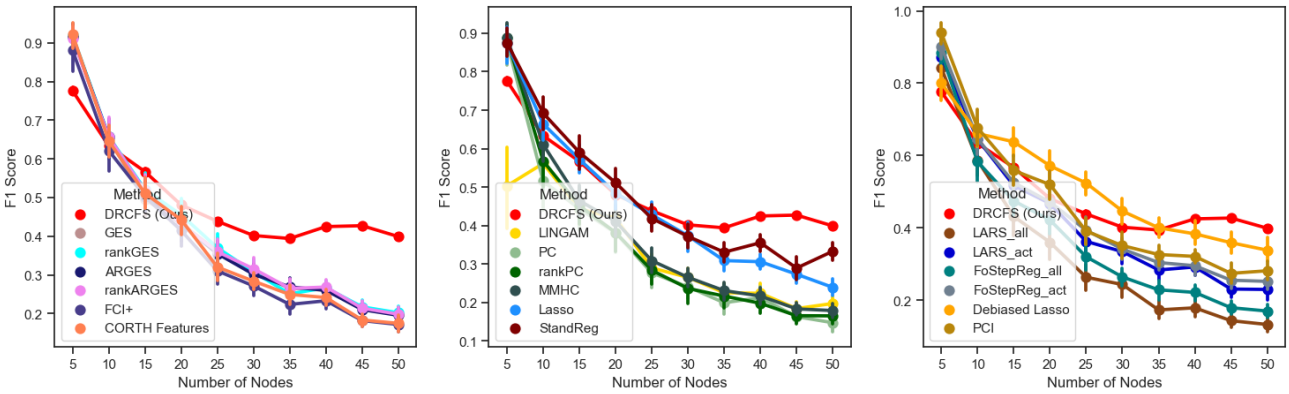


Figure 12: Performance (F1 score) of the algorithms w.r.t. number of nodes m for fully Log-sum-exp causal structures ($f = f_6$ with probability 1), where $p_s = 0.5$, $p_h = 0.1$, and $\epsilon \sim \mathcal{N}(0, 1)$. Each case is averaged over 50 simulations. We use ForestRiesz with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$. DRCFS's performance shows stability even for large graphs while the baselines suffer in high dimensions.

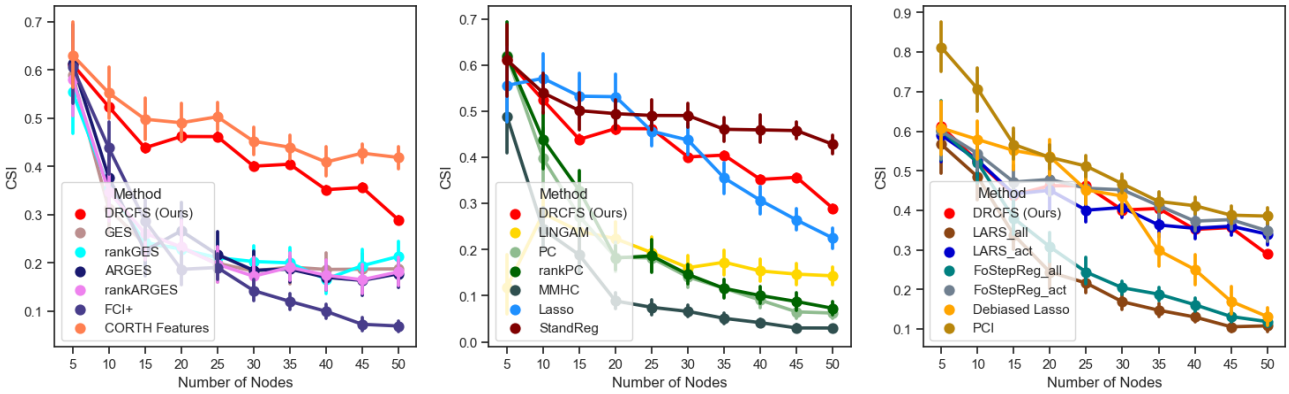


Figure 13: Performance (CSI) of the algorithms w.r.t. number of nodes m for causal structures with geometric mean relationship ($f = f_5$ with probability 0.8, $f = f_1$ with probability 0.2), where $p_s = 0.5$, $p_h = 0$, and $\epsilon \sim \mathcal{N}(0, 1)$. Each case is averaged over 50 simulations. We use ForestRiesz with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$. DRCFS's performance shows stability even for large graphs while the baselines suffer in high dimensions.

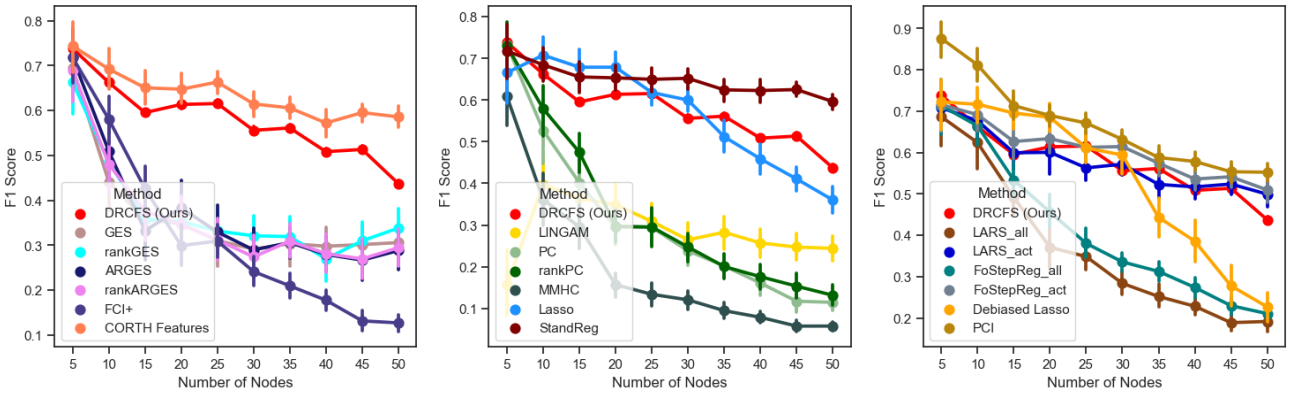
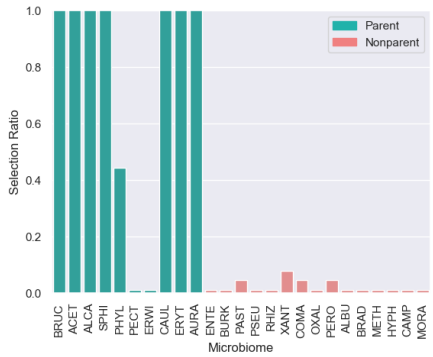
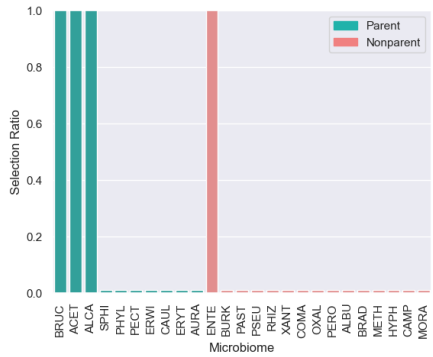


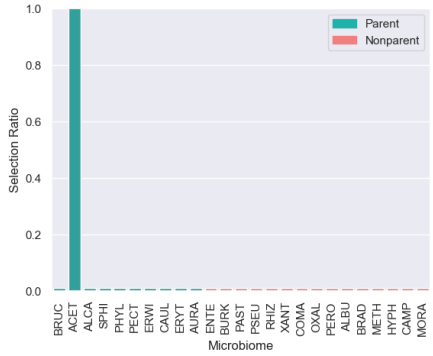
Figure 14: Performance (F1 Score) of the algorithms w.r.t. number of nodes m for causal structures with geometric mean relationship ($f = f_5$ with probability 0.8, $f = f_1$ with probability 0.2), where $p_s = 0.5$, $p_h = 0$, and $\epsilon \sim \mathcal{N}(0, 1)$. Each case is averaged over 50 simulations. We use ForestRiesz with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$. DRCFS's performance shows stability even for large graphs while the baselines suffer in high dimensions.



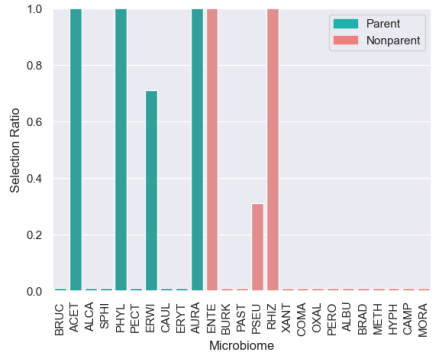
(a) Linear target $f = f_1$. The ground truth about parental status is given in the legend. Despite the highly complicated structure, DRCDS captures most of the direct causes only with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$.



(b) Log-sum-exp target $f = f_6$. The ground truth about parental status is given in the legend. Despite the highly complicated structure and limited number of observations, DRCDS captures some of the direct causes only with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$.



(c) Sqrt-sum target $f = f_7$. The ground truth about parental status is given in the legend. Despite the highly complicated structure and limited number of observations, DRCDS captures some of the direct causes only with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$.



(d) Sum-tanh target $f = f_4$. The ground truth about parental status is given in the legend. Despite the highly complicated structure and limited number of observations, DRCDS captures some of the direct causes only with identity feature map $\phi(\boldsymbol{X}) = \boldsymbol{X}$.

Figure 15: The ratio of simulations that each variable is selected by DRCFS to the total number of simulations (30) for different target functions f.

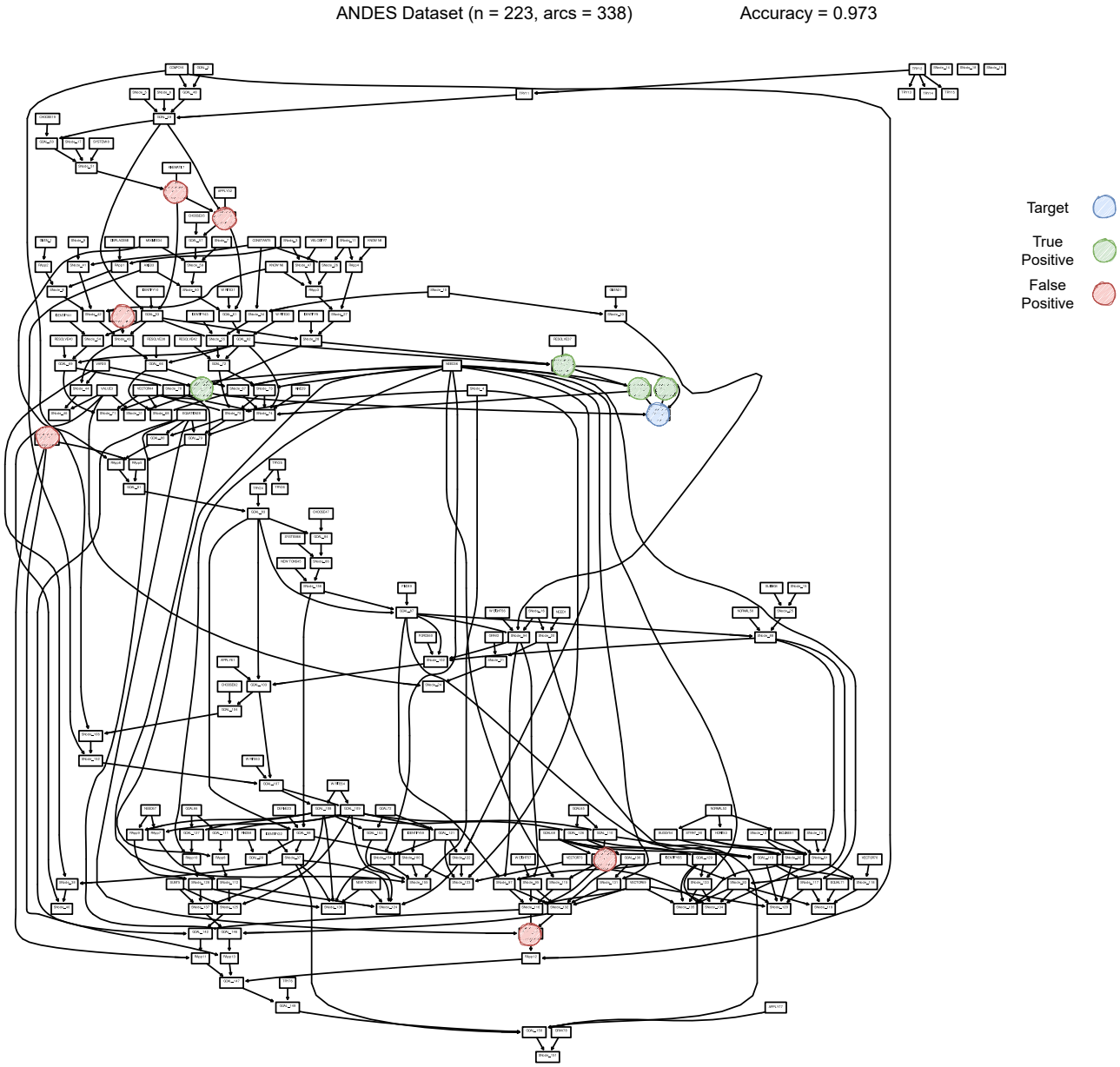


Figure 16: Causal Structure of ANDES benchmark (Scutari, 2010) and DRCFS's inferred causes. ANDES is a very large size discrete Bayesian network. DRCFS has good performance with very few false positives and no false negatives.

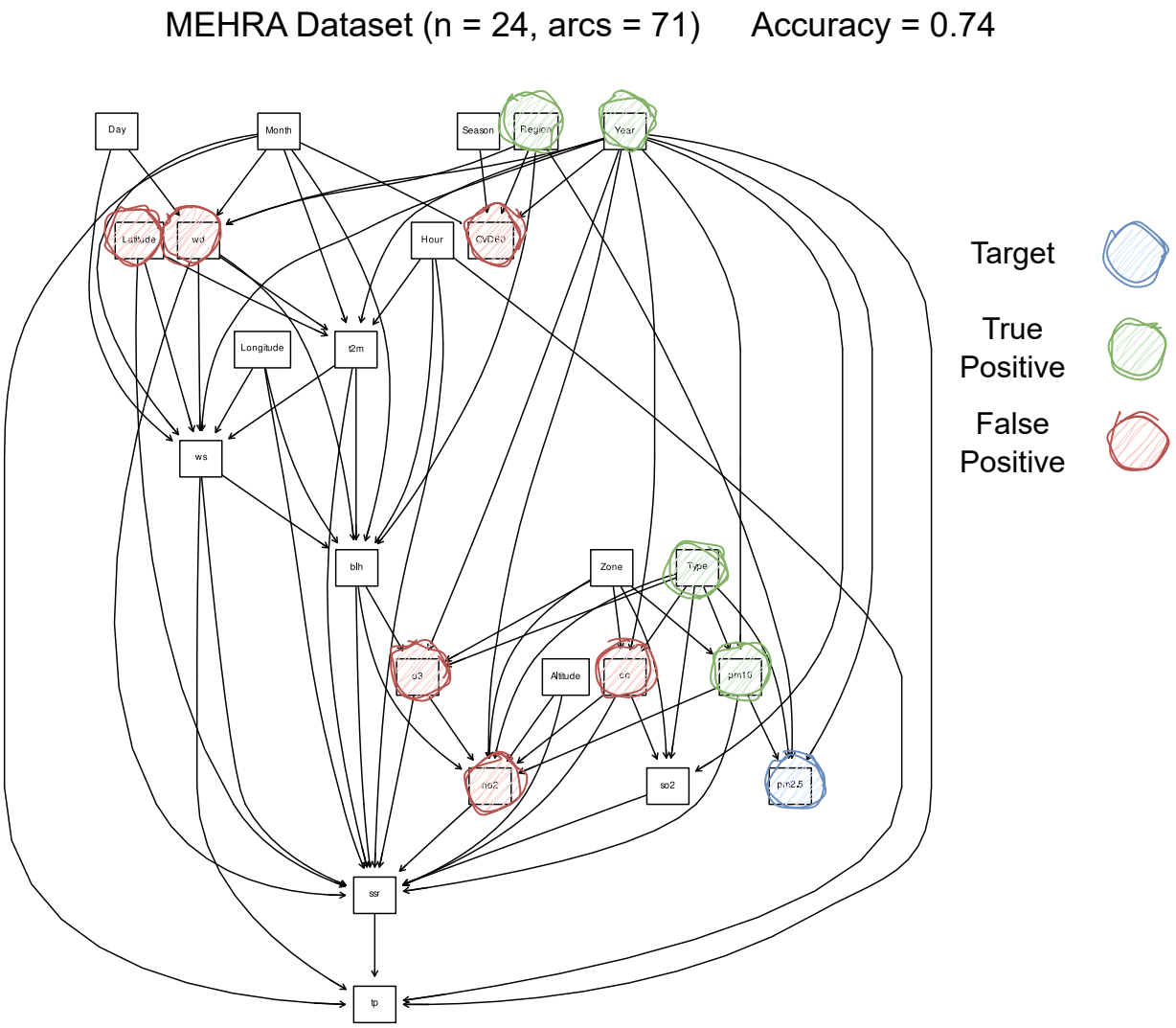


Figure 17: Causal Structure of MEHRA benchmark (Scutari, 2010) and DRCFS's inferred causes. ANDES is a medium size conditional linear Gaussian Bayesian network. DRCFS has good performance with very few false positives and no false negatives.

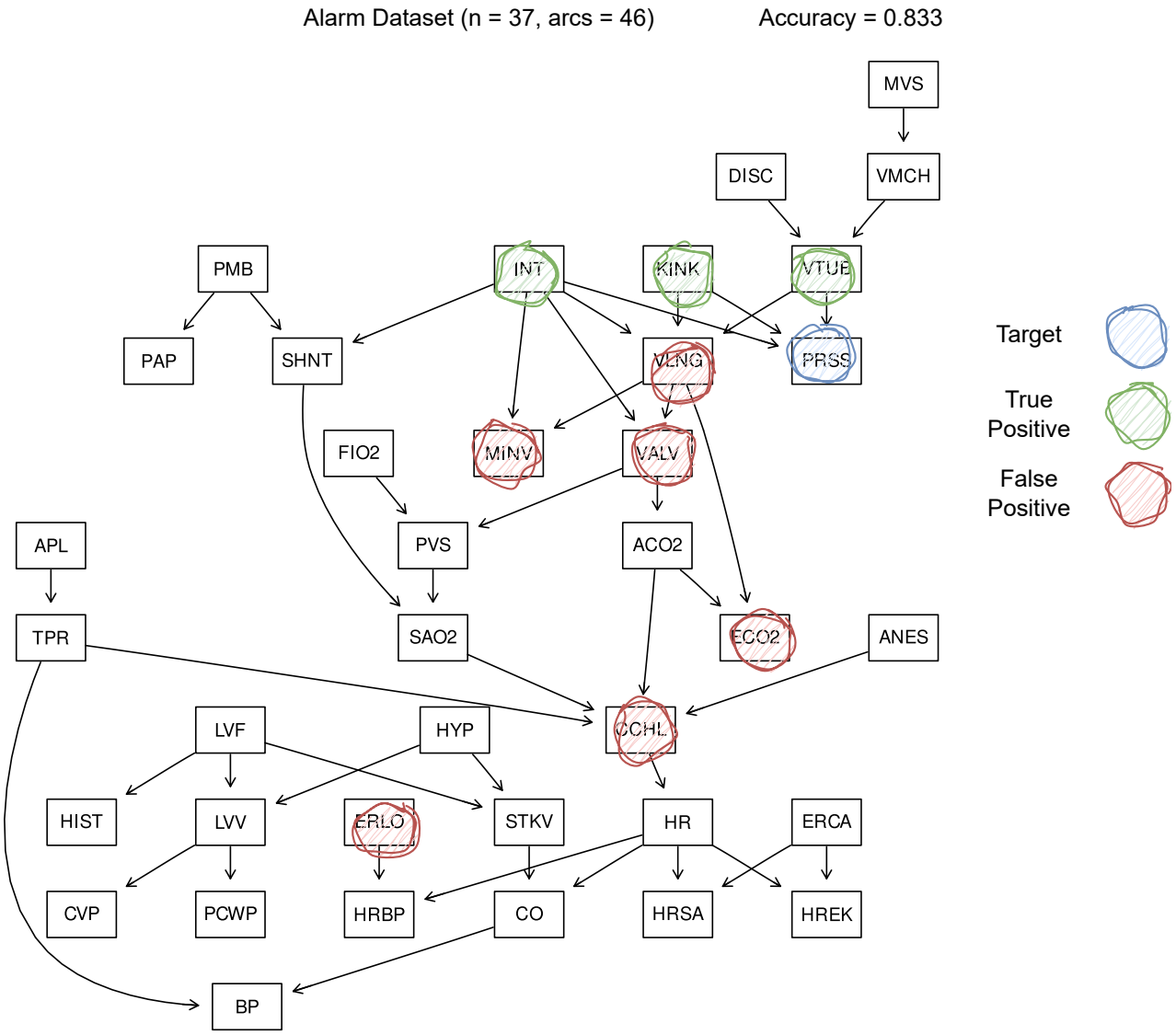


Figure 18: Causal Structure of ALARM benchmark (Scutari, 2010) and DRCFS's inferred causes. ANDES is a medium size discrete Bayesian Network. DRCFS has good performance with very few false positives and no false negatives.

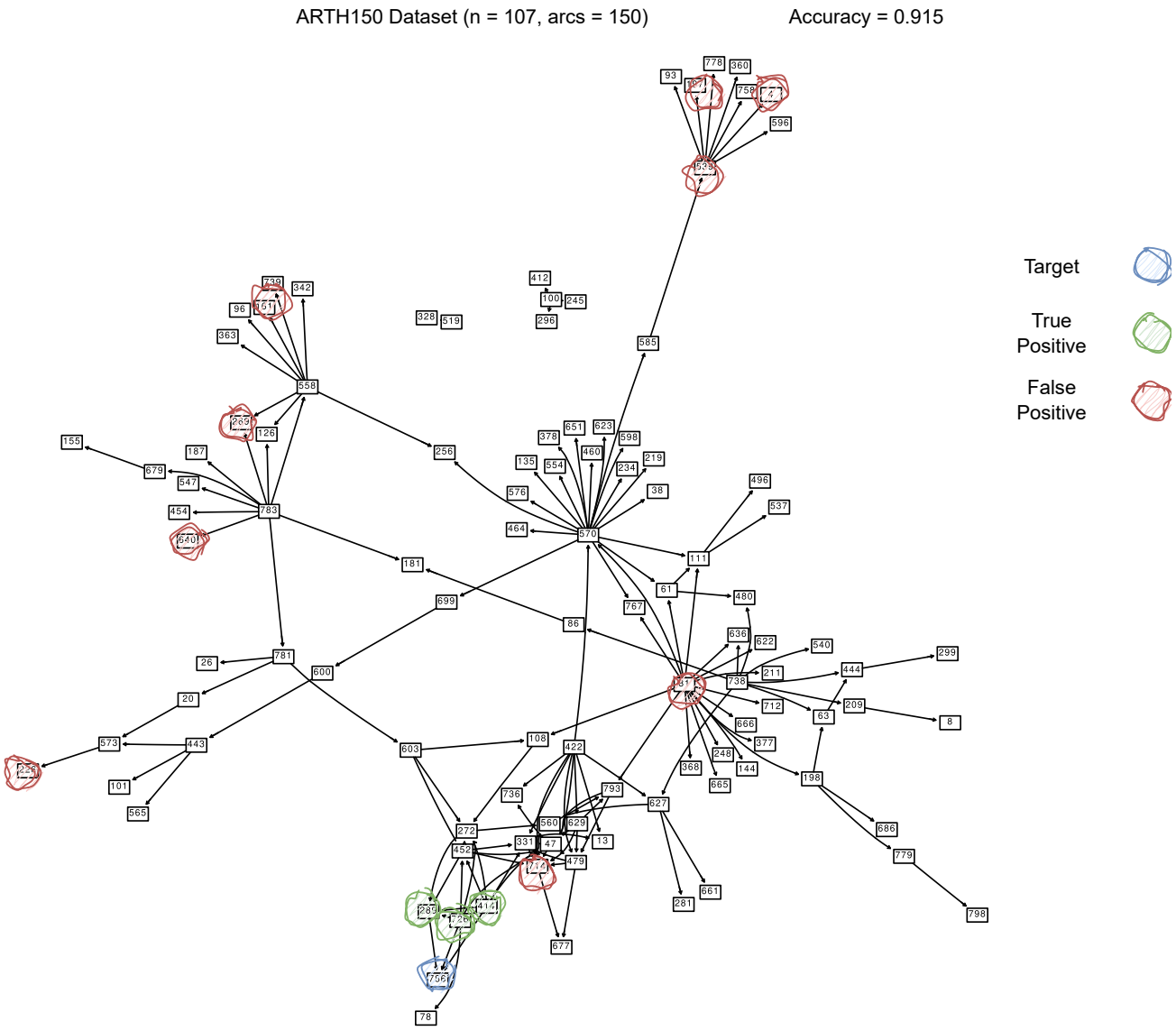


Figure 19: Causal Structure of ARTH150 benchmark (Scutari, 2010) and DRCFS's inferred causes. ANDES is a very large size Gaussian Bayesian network. DRCFS has good performance with very few false positives and no false negatives.