Causal Feature Selection via Orthogonal Search

Anonymous authors
Paper under double-blind review

Abstract

The problem of inferring the direct causal parents of a response variable among a large set of explanatory variables is of high practical importance in many disciplines. However, established approaches often scale at least exponentially with the number of explanatory variables, are difficult to extend to nonlinear relationships and are difficult to extend to cyclic data. Inspired by *Debiased* machine learning methods, we study a one-vs.-the-rest feature selection approach to discover the direct causal parent of the response. We propose an algorithm that works for purely observational data while also offering theoretical guarantees, including the case of partially nonlinear relationships possibly under the presence of cycles. As it requires only one estimation for each variable, our approach is applicable even to large graphs. We demonstrate significant improvements compared to established approaches.

1 Introduction

Identifying causal relationships is a profound and hard problem pervading experimental sciences such as biology (Sachs et al., 2005), medicine (Castro et al., 2020), earth system sciences (Runge et al., 2019), or robotics (Ahmed et al., 2020). While randomized controlled interventional studies are considered the gold standard, they are in many cases ruled out by financial or ethical concerns (Pearl, 2009; Spirtes et al., 2000). In order to improve the understanding of a system and help design relevant interventions, the subset of causes that have a direct effect (*direct causes*/*direct causal parents*) often needs to be identified based on observations only. This paper assumes a structural equation model (SEM) comprising (1) a set of d covariates represented by random vector $X \in \mathbb{R}^d$ whose values are determined by a uniquely solvable set of d structural equations, possibly non-linear and possibly including cycles and confounding (2) a response variable $Y \in \mathbb{R}$, who is not a parent of any X and whose value is determined by a linear structural equation of the form,

$$Y := \langle \theta, X \rangle + U \,, \quad \text{with } \theta \in \mathbb{R}^d, \tag{1}$$

where U is an exogenous variable with zero mean, independent from any other exogenous variables of the SEM and $\langle \cdot, \cdot \rangle$ denotes the inner product. Such a SEM is exemplified in Figure 1. Uniquely solvability of SEMs amounts to not having self-cycles in the causal structure, but any other arbitrary non-linear cyclic structure between covariates is allowed (Bongers et al., 2021), possibly including hidden confounders, as long as there is no hidden confounder for the response variable (this would violate the assumption of independence of U). Practically speaking, almost all causal discovery applications lie under the umbrella of simple SCMs (Bollen, 1989; Sanchez-Romero et al., 2019). Besides, the assumption of not having self-cycles is usually assumed not-limiting in the literature (Lacerda et al., 2012; Rothenhäusler et al., 2015; Bongers et al., 2016).

In this paper, we investigate how to find the direct causes of Y among a high-dimensional vector of covariates X. From our formulation, a given entry of θ should be non-zero if and only if the variable corresponding to that particular coefficient is a direct causal parent (Peters et al., 2017), e.g., X_1 and X_2 in Figure 1. We restrict ourselves to the setting of *linear direct causal effects* of Y (LDC, as specified in Equation 1) and *no feature descending from* Y (NFD). LDC is justified as an approximation when the effects of each causal feature are weak such that the possibly non-linear effects can be linearized; NFD is justified in some applications where we can exclude any influence of Y on a covariate. This is, for example, the case when X are genetic factors, and Y is a particular trait/phenotype. Our method, in particular, comes handy in this

case due to the relatively complex non-linear cyclic structure of these genetic factors in high-dimensional regimes (Yao et al., 2015; Meinshausen et al., 2016; Warrell & Gerstein, 2020).

While applicable to full graph discovery rather than the simplified problem of finding causal parents, state-of-the-art methods for causal discovery often rely on strong assumptions or the availability of interventional data or have prohibitive computational costs explained in section 1.1. In addition to and despite their strong assumptions, causal discovery methods may perform worse than simple regression baselines (Heinze-Deml et al., 2018; Janzing, 2019; Zheng et al., 2018). However, estimating θ in high dimensional settings (i.e. # observations $<< d$) using unregularized least squares regression, will lead to identifiability problems since there can be infinitely many possible choices for θ recovered with equivalent prediction accuracy for regressing Y (Bühlmann & Van De Geer, 2011). On the other hand, when using a regularized method such as Lasso, a critical issue is the bias induced by regularization (Javanmard & Montanari, 2018). While regularized and unregularized Least Squares as well as debiased Lasso are applicable to the same problem, they have different statistical characteristics and we will explore their empirical performance in extensive experiments.

Double ML approaches (Chernozhukov et al., 2018a) have shown promising bias compensation results in the context of high dimensional observed confounding of a single variable. In the present paper, we use this approach to find direct causes among a large number of covariates. Our key contributions are:

- We show that under the assumption that no feature of X is a child of Y, the Double ML (Chernozhukov et al., 2018) principle can be applied in an iterative and parallel way to find the subset of direct causes with observational data.
- Our approach has a computational complexity requirement polynomial (fast) time in dimension d.
- Our method provides asymptotic guarantees that the set can be recovered from observational data. Importantly, this result neither requires linear interactions among the covariates, faithfulness, nor acyclic structure.
- Extensive experimental results demonstrate the state-of-the-art performance of our method. Our approach significantly outperforms all other methods (even though underlying data generation conditions favor them), especially in the case of non-linear interactions between covariates, despite relying only on linear projection.

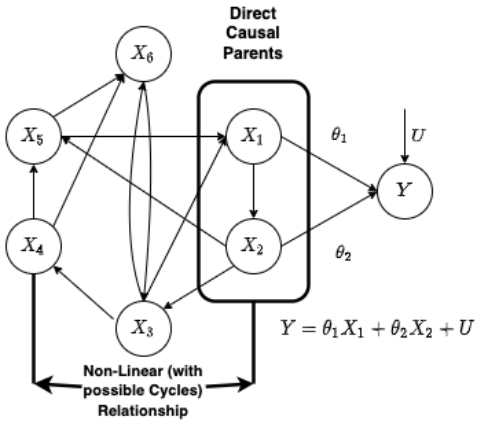


Figure 1: Graphical representation of Causal Feature Selection in our setting, for the case of two direct causal parents of Y, X_1 and X_2, out of variables $\{X_1, \cdots, X_6\}$, such that $Y = \theta_1 X_1 + \theta_2 X_2 + U$, U being an independent zero-mean noise. We propose an approach to find X_1 and X_2 under assumptions discussed in the text. An example of this setup in the real-world is finding genes which directly cause a phenotype.

1.1 Related work

The question of finding direct causal parents is also addressed in the literature as mediation analysis (Baron & Kenny, 1986; Hayes, 2017; Shrout & Bolger, 2002). Several principled approaches have been proposed (relying, for instance, on Instrumental Variables (IVs)) (Angrist & Imbens, 1995; Angrist et al., 1996; Bowden & Turkington, 1990) to test for a single direct effect in the context of specific causal graphs. Extensions of the IV-based approach to generalized IVs-based approaches (Brito & Pearl, 2012; Van der Zander & Liskiewicz, 2016) are the closest known result to discovering direct causal parents. However, no algorithm is provided in Brito & Pearl (2012) to identify the instrumental set. Subsequently, an algorithm is provided in Van der Zander & Liskiewicz (2016) for discovering the instrumental set in the simple setting where all the interactions are linear and the graph is acyclic. In contrast, our method allows non-linear cyclic interaction amongst the variables.

Several other works have also tried to address the problem of discovering causal features. The authors review work on causal feature selection in Guyon & Aliferis (2007). More recent papers on causal feature

selection have appeared since (Cawley, 2008; Paul, 2017; Yu et al., 2018), but none of those claims to recover all the direct causal parents asymptotically or non-asymptotically as we do in our case. There has been another line of works on inferring causal relationships from observational data based on conditional independences, such as the PC-algorithm, which can be used for more general causal inference purposes, at the expense of extra assumptions, such as faithfulness, allowing inference of the Markov equivalence class through testing iterative testing of d-separation statements (Mastakouri et al., 2019; Pearl, 2009; Spirtes et al., 2000). In contrast, under our LDC, NFD and totally independent U assumptions, only a small subset of d-separation relationships are relevant, and further, these assumptions imply that the observed relevant conditional independences and associated d-separation are equivalent, so that a full set of faithfulness assumptions about conditioning sets besides X_{-j} are not necessary.

Another approach is to restrict the class of interactions among the covariates and the functional form of the signal-noise mixing (typically considered additive) or the distribution (e.g., non-Gaussianity) to achieve identifiability (see (Hoyer et al., 2009; Peters et al., 2014)); this includes linear approaches like LiNGAM (Shimizu et al., 2006) and nonlinear generalizations with additive noise (Peters et al., 2011). For a recent review of the empirical performance of structure learning algorithms and a detailed description of causal discovery methods, we refer to (Heinze-Deml et al., 2018). Recently, there have been several attempts at solving the problem of causal inference by exploiting the invariance of a prediction under a causal model given different experimental settings (Ghassami et al., 2017; Peters et al., 2016). The computational cost to run both algorithms is exponential in the number of variables when aiming to discover the full causal graph.

Our method mainly takes inspiration from Debiased/Double ML method (Chernozhukov et al., 2018a) which utilizes the concept of orthogonalization to overcome the bias introduced due to regularization. We will discuss this in detail in the next section. Considering a specific example, the Lasso suffers from the fact that the estimated coefficients are shrunk towards zero, which is undesirable (Tibshirani & Wasserman, 2017). To overcome this limitation, a debiasing approach was proposed for the Lasso in several papers (Javanmard & Montanari, 2014; 2018; Zhang & Zhang, 2014). However, unlike our approach, Debiased Lasso methods do not recover all the non-zero coefficients of the parameter vector θ under the generic assumptions of the present work.

2 Methodology

Before describing the proposed method, we discuss our general strategy as well as Double ML and Neyman orthogonality in the next sections, which will be helpful in building the theoretical framework for our method.

2.1 Reduction to a nonparametric estimation problem

According to Equation (1), determining whether X_j is a parent of Y in our setting amounts to testing whether $\theta_j \neq 0$. Let $X_{-j} = X \setminus X_j$, this can be reduced to testing whether the following estimand vanishes:

$$\chi_j \triangleq \mathbb{E}\left[\left(Y - \mathbb{E}(Y \mid X_{-j})\right)\left(X_j - \mathbb{E}(X_j \mid X_{-j})\right)\right] \tag{2}$$

Indeed, U independent of X entails $Y - \mathbb{E}(Y \mid X_{-j}) = \theta_j \left(X_j - \mathbb{E}(X_j \mid X_{-j})\right) + U$. This leads to

$$\chi_j = \theta_j \mathbb{E}\left[\left(X_j - \mathbb{E}(X_j \mid X_{-j})\right)^2\right] = \theta_j \mathbb{E}\left[X_j \left(X_j - \mathbb{E}(X_j \mid X_{-j})\right)\right] . \tag{3}$$

Under mild assumptions, testing whether $\theta_j \neq 0$ thus reduces to testing whether $\chi_j \neq 0$. Equation (2) shows that χ_j constitutes a *non-parametric estimand*, i.e. a model-free functional of the observed data distribution. Nonparametric estimation results (Robins et al., 2008; Van der Laan et al., 2011; Chernozhukov et al., 2018a) make use of the *efficient influence function* of such estimand (see e.g. Hines et al. (2022)) to derive valid estimates and confidence bounds, while allowing the use of data adaptive estimation strategies, such as machine learning algorithms. The resulting strategies are known as *target learning* and *debiased/double machine learning*, and are suitable in challenging settings such as ours when X is high dimensional with possibly non-linear dependencies among components.

2.2 Double Machine Learning (Double ML)

Double ML constitutes one possible way to derive efficient nonparametric estimates. We introduce it with the partial linear regression setting introduced in Chernozhukov et al. (2018a, Example 1.1). Given a fixed set of policy variables D and control variables X acting as common causes of D and Y, we consider the partial regression model of Equation (4),

$$\begin{aligned} Y &= D\theta_0 + g_0(X) + U, \quad \mathbb{E}\left[U|X, D\right] = 0 \\ D &= m_0(X) + V, \quad \mathbb{E}\left[V|X\right] = 0, \end{aligned} \tag{4}$$

where Y is the outcome variable, U, V are disturbances and $g_0, m_0 : \mathbb{R}^d \to \mathbb{R}$ are (possibly non-linear) measurable functions. An unbiased estimator of the causal effect parameter θ_0 can be obtained via the orthogonalization approach as in Chernozhukov et al. (2018a), which is obtained via the use of the "Neyman Orthogonality Condition" described below.

Neyman Orthogonality Condition: Let W denote the collection of all observed variables. The traditional estimator of θ_0 in Equation (4) can be simply obtained by finding the zero of the empirical average of a score function ϕ such that $\phi(W; \theta, g) = D^\top(Y - D\theta - g(X))$. However, the estimation of θ_0 is sensitive to the bias in the estimation of the function g. Neyman (Neyman, 1979) proposed an orthogonalization approach to get an estimate for θ_0 that is more robust to the bias in the estimation of nuisance parameter (m_0, g_0). Assume for a moment that the true nuisance parameter is η_0 (which represents m_0 and g_0 in Equation (4)) then the orthogonalized "score" function ψ should satisfy the property that the Gateaux derivative operator with respect to η vanishes when evaluated at the true parameter values:

$$\partial_\eta \mathbb{E}\psi(W; \theta_0, \eta_0)[\eta - \eta_0] = 0\,. \tag{5}$$

One way to build such a score, following Chernozhukov et al. (2018a) [eq. (2.7)], is to start from a biased score associated to maximum likelihood-like estimate. Let $\ell(W; (\theta, \boldsymbol{\eta}))$ be the *log likelihood* function or another smooth objective for which the true parameter is the unique maximizer. The true parameter then satisfies $\mathbb{E}\partial_\theta \ell(W; (\theta_0, \boldsymbol{\eta}_0)) = 0$, suggesting to start with $\partial_\theta \ell(W; (\theta_0, \boldsymbol{\eta_0}))$ as a (biased) score. In order to compensate the bias due to the nuisance parameters, we then subtract a linear function of the derivative of the likelihood with respect it, leading to the orthogonalized score

$$\psi(W; \theta, \boldsymbol{\eta}) = \partial_\theta \ell(W; (\theta, \boldsymbol{\eta})) - \boldsymbol{\mu}\partial_\eta \ell(W; (\theta, \boldsymbol{\eta}))\,.$$

where $\boldsymbol{\mu}$ is determined by the constraint of Equation (5) (see proof of Proposition 6 in appendix). The corresponding Orthogonalized or Double/Debiased ML estimator $\check{\theta}_0$ solves a constraint of vanishing empirical average of the orthogonalized score, based on n-iid samples $\{W_i\}_{i=1..n}$ of the observed variables.

$$\frac{1}{n}\sum_{i=1}^{n} \psi(W_i; \check{\theta}_0, \hat{\eta}_0) = 0, \tag{6}$$

where $\hat{\eta}_0$ is the estimator of η_0 and ψ satisfies condition in Equation (5). For the partially linear model discussed in Equation (4), the orthogonalized score function ψ is,

$$\psi(W; \theta, \eta) = (Y - D\theta - g(X))(D - m(X))\,, \tag{7}$$

with $\eta = (m, g)$. This leads to an debiased estimator satisfying

$$\check{\theta}_0 \frac{1}{n}\sum_i D_i(D_i - \check{m}_0(X_i)) = \frac{1}{n}\sum_i (Y_i - \check{g}_0(X_i))(D_i - \check{m}_0(X_i))\,. \tag{8}$$

which relies on the "double" use of machine learning algorithm: once to learn $\check{g}_0(X_i)$ and once to learn $\check{m}_0(X_i)$, hence the name *Double ML* for such estimator. We can further relate this approach to the design an estimator of the non-parametric estimand of previous section.

Indeed by subtracting $\check{\theta}_0 \frac{1}{n} \sum_i \check{m}_0(X_i)(D_i - \check{m}_0(X_i))$ on both sides of eq. (8), we get

$$\check{\theta}_0 \frac{1}{n} \sum_i (D_i - \check{m}_0(X_i))^2 = \frac{1}{n} \sum_i (Y_i - \check{\theta}_0 \check{m}_0(X_i) - \check{g}_0(X_i))(D_i - \check{m}_0(X_i)) \,. \tag{9}$$

Noticing that $\mathbb{E}[Y|X] = \theta_0 \mathbb{E}[D|X] + g_0(X) = \theta_0 m_0(X) + g_0(X)$, the term $\check{\theta}_0 \check{m}_0(X_i) + \check{g}_0(X_i)$ in eq. (9) appears as an ML estimator of $\mathbb{E}[Y|X]$, such that we recognize on the right hand side of Equation (9) a Double ML estimator of $\mathbb{E}[(Y - \mathbb{E}[Y|X])(D - \mathbb{E}[D|X])]$, which is a special case of the non-parametric estimand χ_j defined in Equation (3), for the setting $X_j = D$ and $X = X_{-j}$. In practice, we directly learn an ML estimator of $\mathbb{E}[Y|X]$ by predicting Y using X, relying on the double robustness of the χ_j estimands (Smucler et al., 2019), as described in section 2.5.

From Double ML to Causal Discovery: The distinction between policy variables and confounding variables is not always known in advance. Fortunately, as described in section 2.1, Double ML relies on estimating a non-parameteric estimand that does only depend on observational data and not on the causal model. This will allow us to exploit the same approach iteratively in the setting of causal discovery. To this end, we consider a set of variables $X = \{X_1, X_2, \cdots X_d\}$ which includes direct causal parents of the outcome variable Y as well as other variables. We also reiterate our assumption that the relationship between the outcome variable and direct causal parents of the outcome variable is linear. The relationship among other variables can be cyclic and nonlinear. We now provide a general approach to scanning putative direct causes scaling "polynomially" in their number (see *Computational Complexity* paragraph in next section), based on the application of a statistical test and Double ML estimators. We describe first the algorithm and then provide theoretical support for its performance.

2.3 Informal Search Algorithm Description

Pseudo-code for our proposed method (CORTH Features) is in Algorithm 1. The idea is to do a one-vs-rest split for each variable in turn and estimate the link between that particular variable and the outcome variable using Double ML. To do so, we decompose Equation (1) to single out a variable $D = X_k$ as policy variable and take the remaining variables $Z = X_{-k} = X \backslash X_k$ as multidimensional control variables, and run Double ML estimation assuming the partial regression model presented in Section 2.2, which now takes the form

$$\begin{aligned} Y &= D\theta_k + g_k(Z) + U, \quad \mathbb{E}\left[U|Z, D\right] = 0 \,, \\ D &= m_k(Z) + V, \quad \mathbb{E}\left[V|Z\right] = 0 \,. \end{aligned} \tag{10}$$

The step-wise description of our estimation algorithm goes as follows:

(a) Select one of the variables X_i to estimate its (hypothetical) linear causal effect θ on Y.
(b) Set all of the other variables X_{-i} as the set of possible confounders.
(c) Use the Double ML approach to estimate the parameter θ i.e. the causal effect of X_i on Y.
(d) If the variable X_i is not a causal parent, the distribution of the conditional covariance χ_i (Proposition 3) is a Gaussian centered around zero. We use a simple normality test for χ_i to select or discard X_i as one of the direct causal parents of Y.

We iteratively repeat the procedure on each of the variables until completion. Pseudo-code for the entire procedure is given below in Algorithm 1. Guaranties for this approach to identify the true parents rely on the assumptions stated in Section 2.5, Equations (13-15). They notably allow for hidden confounders between covariates, as long as those are not direct causes of Y, not descendent of Y. On the contrary, if Y is an ancestor of any covariate, the search algorithm may fail in both directions (false positive and false negative).

Note that Equation (10) is not necessarily a correct structural equation model to describe the true underlying causal structure. In general, for instance, when D actually causes Z, it is non-trivial to show that the Double ML estimation of parameter θ_k will be unbiased (see Section 2.4).

Algorithm 1 Efficient Causal Orthogonal Structure Search (CORTH Features)

Input: response $Y \in \mathbb{R}^N$, covariates $\mathbb{X} \in \mathbb{R}^{N \times d}$, significance level α, number of partitions K.
Split N observations into K random partitions, I_k for $k = 1, \ldots, K$, each having $n = N/K$ samples.
for $i = 1, \ldots, d$ **do**
  for Subsample $k \in [K]$ **do**
    $D_k \leftarrow X_i^{[k]}$ and $Z_k \leftarrow X_{\setminus i}^{[k]}$
    Fit $m_i^{[\setminus k]}(Z_{\setminus k})$ to $D_{\setminus k}$ and fit $g_i^{[\setminus k]}(Z_{\setminus k})$ to $Y^{[\setminus k]}$
    $\hat{V}_{ij}^{[k]} \leftarrow D_{kj} - m_i^{[\setminus k]}(Z_{kj})$, for all $j \in I_k$
    $\check{\theta}_i^{[k]} \leftarrow \left(\frac{1}{n}\sum_{j \in I_k} \hat{V}_{ij}^{[k]} D_{kj}\right)^{-1} \frac{1}{n}\sum_{j \in I_k} \hat{V}_{ij}^{[k]}(Y_{ij}^{[k]} - g_{ij}^{[\setminus k]}(Z_{kj}))$
    $\hat{\chi}_i^{[k]} \leftarrow \frac{1}{n}\sum_{j \in I_k}\left(-Y_j^{[k]} m_{ij}^{[\setminus k]}(Z_{kj}) - D_{kj} g_{ij}^{[\setminus k]}(Z_{kj}) + m_{ij}^{[\setminus k]}(Z_{kj}) g_{ij}^{[\setminus k]}(Z_{kj}) + Y_j^{[k]} D_{kj}\right)$
    $(\hat{\sigma}_i^{[k]})^2 \leftarrow \frac{1}{n}\sum_{j \in I_k}\left(-Y_j^{[k]} m_{ij}^{[\setminus k]}(Z_{kj}) - D_{kj} g_{ij}^{[\setminus k]}(Z_{kj}) + m_{ij}^{[\setminus k]}(Z_{kj}) g_{ij}^{[\setminus k]}(Z_{kj}) + Y_j^{[k]} D_{kj} - \hat{\chi}_i^{[k]}\right)^2$
  end for
  $\hat{\theta}_i \leftarrow \frac{1}{K}\sum_{k \in K} \check{\theta}_i^{[k]}$, $\hat{\chi}_i \leftarrow \frac{1}{K}\sum_{k \in K} \hat{\chi}_i^{[k]}$ and $\hat{\sigma}_i^2 \leftarrow \frac{1}{K}\sum_{k \in K} (\hat{\sigma}_i^{[k]})^2$
end for
for $i \in [d]$ **do**
  Gaussian normality test for $\hat{\chi}_i \approx N\left(0, \frac{\hat{\sigma}_i^2}{N}\right)$ with α significance level and select i^{th} feature if null-hypothesis is rejected.
end for
Return Decision Vector

Remarks on Algorithm 1: $X_i^{[k]}$ is a vector which corresponds to the samples chosen in the kth subsampling procedure, $X_{\setminus i}^{[k]} = (X_1^{[k]}, \ldots, X_{i-1}^{[k]}, X_{i+1}^{[k]}, \ldots, X_d^{[k]})$ for any $i \in [d]$. In general the subscript i represents the estimation for the ith variable and super-script k represents the kth subsampling procedure. K represents the set obtained after sample splitting. $m_i^{[\setminus k]}$ are (possibly nonlinear) parametric functions fitted using $(1^{st}, \ldots, k-1^{th}, k+1^{th}, \ldots, K^{th})$ subsamples.

Computational Complexity: For each subset randomly selected from the data, we fit two lasso estimators. Accelerated coordinate descent (Nesterov, 2012) can be applied to optimize the lasso objective. To achieve ε error, $\mathcal{O}\left(d\sqrt{\kappa_{\max}} \log \frac{1}{\varepsilon}\right)$ number of iterations are required where $\kappa_{\max}$ is the maximum of the two condition number for both the problems and each iteration requires $\mathcal{O}(nd)$ computation. Hence, the computational complexity of running our approach is only polynomial in d.

2.4 Orthogonal Scores

Now we describe the execution of our algorithm for a simple graph with 3 nodes. Let us consider the following linear structural equation model as an example of our general formulation:

$$Y := \theta_1 X_1 + \theta_2 X_2 + \varepsilon_3, \; X_2 := a_{12} X_1 + \varepsilon_2, \text{ and } X_1 := \varepsilon_1. \tag{11}$$

Example 1. *Consider the system of structural equation given in Equation* (10)*. If ε_1, ε_2 and ε_3 are independent uncorrelated noise terms with zero mean, Algorithm 1 will recover the coefficients θ_1 and θ_2.*

A detailed proof is given in Appendix A.1. While the estimation of the parameter θ_1 is in line with the assumed partial regression model of Equation (11), the estimation of θ_2 does not follow the same. However, it can be seen from the proof that θ_2 can also be estimated from the orthogonal score in Equation (7).

We now show that this result holds for a more general graph structure given in Figure 2, allowing for non-linear cyclic interactions among features.

Proposition 2. *Assume the structural causal model of Figure 2, with (possibly non-linear and confounded) assignments between elements of $X = [X_k, X_{-k}^\top]^\top$, with $X_{-k} = [Z_1^\top, Z_2^\top]^\top$, parameterized by $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_{12})$. Assume the unconfounded linear structural assignment $Y := X_k\theta + X_{-k}^\top \boldsymbol{\beta} + U$, with U*

zero mean random variable with finite variance $\sigma_U^2 > 0$, independent of X. Then, the score

$$\psi(W;\theta,\boldsymbol{\beta}) = (Y - X_k\theta - X_{-k}^\top\boldsymbol{\beta})(X_k - r_{XX_{-k}}X_{-k})\,, \tag{12}$$

with $r_{XX_{-k}} = \mathbb{E}[X_k {X_{-k}}^\top]\mathbb{E}[X_{-k}{X_{-k}}^\top]^{-1}$, follows the Neyman orthogonality condition for the estimation of θ with nuisance parameters $\boldsymbol{\eta} = (\beta, \boldsymbol{\gamma})$ which reads

$$\mathbb{E}\left[(Y - X_k\theta - {X_{-k}}^\top\boldsymbol{\beta})(X_k - r_{XX_{-k}}X_{-k})\right] = 0\,.$$

Please refer to Appendix A.2 for the proof. Applying Equation (6), this leads to the debiased estimator

$$\check{\theta} = \frac{\sum_i (Y_i - {X_{-ki}}^\top\check{\boldsymbol{\beta}})(X_{ki} - \check{r}_{XX_{-k}}X_{-k})}{\sum_i X_{ki}(X_{ki} - \check{r}_{XX_{-k}}X_{-ki})}\,.$$

which relies on ML estimates $\check{\boldsymbol{\beta}}$ and $\check{r}_{XX_{-k}}$. Comparing the score in Equation (21) with the score in Equation (7), there are two takeaways from Proposition 2: (i) the orthogonality condition remains invariant irrespective of the causal direction between X_k and Z, and (ii) the second term in Section 2.4 replaces function m by the (unbiased) linear regression estimator for modelling all the relations; given that the relation between Z and Y is linear, even if relationships between Z and X_k are non-linear (See Appendix B for concrete examples). Combining with the Double ML theoretical results (Chernozhukov et al., 2018a), this suggests that regularized predictors based on Lasso or ridge regression are tools of choice for fitting functions (m, g).

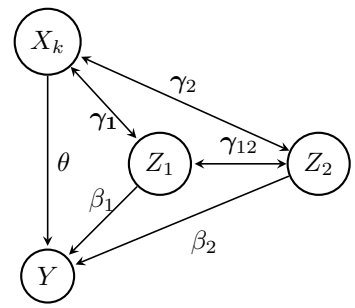


Figure 2: A generic example of identification of a causal effect θ in the presence of causal and anti-causal interactions between the causal predictor and other putative parents, and possibly arbitrary cyclic and nonlinear assignments for all nodes except Y (see Proposition 2). We have $X_{-k} = Z_1 \cup Z_2$.

2.5 Statistical Test

We now provide a theoretically grounded statistical decision criterion for the direct causes after the model has been fitted. Consider (Y, X), $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$, satisfying

$$Y = \langle\theta, X\rangle + U, \tag{13}$$

$$\mathbb{E}(Y^2) < \infty,\ \mathbb{E}(U^2) < \infty,\ \mathbb{E}(U) = 0, \mathbb{E}(U \mid X) = 0, \text{ and } \mathbb{E}(\|X\|_2^2) < \infty, \tag{14}$$

$$\mathbb{E}\left[(X_j - \mathbb{E}(X_j \mid X_{-j}))^2\right] \neq 0, \quad \text{for all } j \in \{1,\dots,d\}\,, \tag{15}$$

where U is an exogenous variable and X_{-j} represents all the variables except X_j. The assumptions made with the above formulation are standard in the orthogonal machine learning literature (Rotnitzky et al., 2019; Smucler et al., 2019; Chernozhukov et al., 2018). They allow identifying causal parents based on estimates of conditional covariances χ_j defined in Equation (3)

Proposition 3. *Let $PA_Y = \{j \in \{1,\dots,d\} : \theta_j \neq 0\}$. Then under the conditions given in Equations* (13) *to* (15)*, for each $j \in \{1,\dots,p\}$*

a) $\chi_j = \theta_j \mathbb{E}\left[(X_j - \mathbb{E}(X_j \mid X_{-j}))^2\right]$ and $j \in PA_Y$ if and only if $\chi_j \neq 0$.

b) We also have (with notations of Prop. 2) $\chi_j = \mathbb{E}\left[(Y - \mathbb{E}(Y \mid X_{-j}))\left(X_j - r_{XX_{-j}}X_{-j})\right)\right]$.

The proof is given in appendix A.3. There are two main implications of the results provided in Proposition 3. (i) χ_j is non-zero only for direct causal parents of the outcome variable, and χ_j has double robustness

property as shown in (Rotnitzky et al., 2019; Smucler et al., 2019; Chernozhukov et al., 2018). Having double robustness property means that while computing the empirical version of the χ_j which we denote as $\hat{\chi}_j$, one can use regularized methods like ridge regression or Lasso to estimate the conditional expectation (function m). Afterward, one can perform statistical tests on top of it to decide between zero or non-zero tests. (ii) In line with the above orthogonal score results, we see that this quantity can be estimated using linear (unbiased) regression to fit the function m, although interactions between features may be non-linear. Next, we discuss the variance of our estimator so that a statistical test can be used to identify causal parents. For the sake of convenience, the case of 2 partitions ($K = 2$)[1] is explained here.

Variance of Empirical Estimates of χ_j: Suppose we have n i.i.d. observations indicated by $\mathcal{D}_n = \{(X_i, Y_i), i = 1 \ldots, n\}$. Randomly split the data in two halves, say $\mathcal{D}_{n1}$ and $\mathcal{D}_{n2}$. Take $j \in \{1, \ldots, d\}$. For $k = 1$ let $\overline{k} = 2$, for $k = 2$ let $\overline{k} = 1$. For $k = 1, 2$, compute estimates of $\widehat{\mathbb{E}}^{\overline{k}}(Y \mid X_{-j})$ and $\widehat{\mathbb{E}}^{\overline{k}}(X_j \mid X_{-j})$ using the data in sample $\overline{k}$. Following Smucler et al. (2019), we can use estimates of $\widehat{\mathbb{E}}^{\overline{k}}(Y \mid X_{-j})$ and $\widehat{\mathbb{E}}^{\overline{k}}(X_j \mid X_{-j})$ that are solutions of ℓ_1-regularized regression problems to obtain square root N convergence guaranties. We use Lasso as the estimator for conditional expectation in the experiments. Now, we compute the cross-fitted empirical estimates of χ_j and associated empirical variances

$$\widehat{\chi}_j^k = \mathbb{P}_{nk}\left[-Y\widehat{\mathbb{E}}^{\overline{k}}(X_j \mid X_{-j}) - X_j\widehat{\mathbb{E}}^{\overline{k}}(Y \mid X_{-j}) + \widehat{\mathbb{E}}^{\overline{k}}(Y \mid X_{-j})\,\widehat{\mathbb{E}}^{\overline{k}}(X_j \mid X_{-j}) + YX_j\right] \tag{16}$$

$$\left(\widehat{\sigma}_j^k\right)^2 = \mathbb{P}_{nk}\left[\left(-Y\widehat{\mathbb{E}}^{\overline{k}}(X_j \mid X_{-j}) - X_j\widehat{\mathbb{E}}^{\overline{k}}(Y \mid X_{-j}) + \widehat{\mathbb{E}}^{\overline{k}}(Y \mid X_{-j})\,\widehat{\mathbb{E}}^{\overline{k}}(X_j \mid X_{-j}) + YX_j - \widehat{\chi}_j^k\right)^2\right], \tag{17}$$

where $\mathbb{P}_{nk}$ denotes the empirical average over the k half. Finally, let

$$\widehat{\chi}_j = \frac{1}{2}\left(\widehat{\chi}_j^1 + \widehat{\chi}_j^2\right), \quad \widehat{\sigma}_j^2 = \frac{1}{2}\left(\left(\widehat{\sigma}_j^1\right)^2 + \left(\widehat{\sigma}_j^2\right)^2\right). \tag{18}$$

Consistency of such estimators notably relies on sparsity assumptions for ground truth models in the asymptotic high-dimensional setting where covariate dimension $p(= d - 1)$ and sparsity is allowed to vary with number of samples n. In the notations below, we drop this dependency and specialize to our case.

Definition 4 (Approximate Linear-Sparse class (ALS))**.** *Ground truth predictor $c(X_{-j})$ belongs to the approximately sparse class whenever there exists $\theta^* \in \mathbb{R}^p$ and a function $r(X_{-j})$ satisfying*

$$c(Z) = \langle \theta^*, \phi(Z)\rangle + r(Z)\,, \text{ where } \|\theta^*\|_0 \leq s \text{ and } \mathbb{E}[r(Z)^2] \leq K(s\log(p)/n)\,.$$

Theorem 1 of (Smucler et al., 2019) provides general conditions under which (see also (Chernozhukov et al., 2018)), when the estimators $\widehat{\mathbb{E}}^{\overline{k}}(Y \mid X_{-j})$ and $\widehat{\mathbb{E}}^{\overline{k}}(X_j \mid X_{-j})$ are Lasso-type regularized linear regressions.

Proposition 5. *Given eqs. (13-15) and $j \in \{1, ..., d\}$. Assume: (i) For true parameter in (13), $\|\theta\|_0 \leq s$ with $s\log(p)/n \to 0$; (ii) $r_{XX_{-j}}X_{-j}$ is in the ALS class with $s\log(p)/n \to 0$; (iii) All X_k have support bounded by the same constant; (iv) U is independent of X and has tail decaying at least as fast as an exponential random variable; (v) $\mathbb{E}[X_{-j}X_{-j}^\top]$ has eigenvalues lower and upper bounded; (vi) $\mathbb{E}[Y - \mathbb{E}[Y|X_{-j}]|X_{-j}]$ and $\mathbb{E}[X_j - \mathbb{E}[X_j|X_{-j}]|X_{-j}]$ are bounded with variance lower bounded.*
Then the pair of estimators $(\widehat{\chi}_j, \widehat{\sigma}_j^2)$ of eq. (18), using l_1 regularization coefficient $\lambda \approx \sqrt{\log(p)/n}$ for both Lasso estimates, satisfies $\widehat{\chi}_j \xrightarrow{D} \mathcal{N}\left(\chi_j, \frac{\widehat{\sigma}_j^2}{n}\right)$., as $n \to +\infty$.

All stated bounds are uniform across n and strictly positive. Proof is provided in A.4 with the broader context on the doubly robust estimation framework. The provided conditions correspond to a concise special case of the general statements provided in Smucler et al. (2019). It is possible to relax some of these assumptions, notably to allow some mispecifications of the sparsity assumptions for the Lasso estimates. Overall, the mildness of these assumptions may explain the empirical success of this approach in the following section.

[1]Extension to arbitrary number of data partitions ($K \geq 2$) is straightforward. Check Algorithm 1.

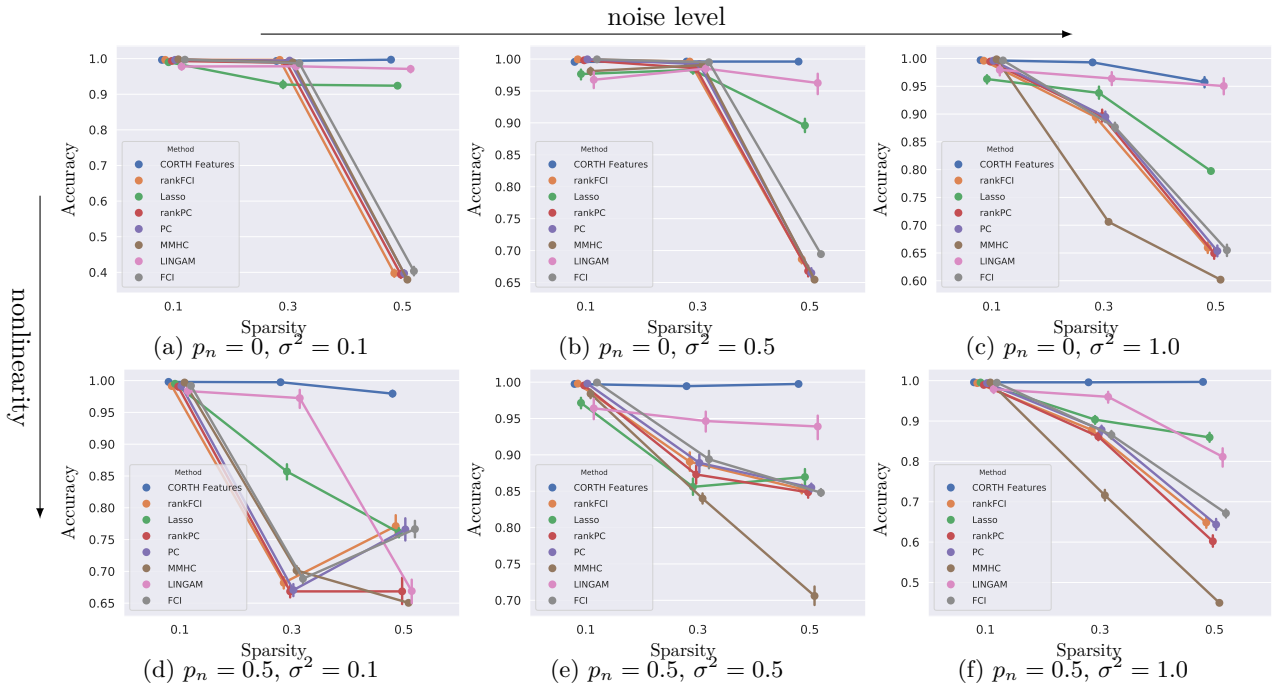
Under conditions of proposition 5, the test that rejects $\chi_j = 0$ when $|\widehat{\chi}_j| \geq 1.96\frac{\widehat{\sigma}_j}{\sqrt{n}}$ will have approximately 95% confidence level. The probability of rejecting the null when it is false is

$$P\left(|\widehat{\chi}_j| \geq 1.96\frac{\widehat{\sigma}_j}{\sqrt{n}}\right) \geq P\left(|\widehat{\chi}_j - \chi_j| \leq |\chi_j| - 1.96\frac{\widehat{\sigma}_j}{\sqrt{n}}\right) \to 1.$$

In order to account for multiple testing, we use Bonferroni correction.

Conditional Independence Tests: Asymptotically, the conditional independence testing between Y and X_j given X_{-j} is also a possible solution for our proposed approach. Indeed, d-separation rules imply that true causes are conditionally dependent according to this test, while non-causes are conditionally independent (because X_{-j} is not a collider under our NFD assumption). However, conditional independence testing is challenging in high-dimensional/non-linear settings. Kernel-based conditional independence testing is computationally expensive (Zhang et al., 2012). We used χ_j in the paper because it was already known from previous works (Smucler et al., 2019; Chernozhukov et al., 2018b) that it has double robustness property, which means one can use regularized methods like Lasso to estimate empirical conditional expectation from a finite number of samples and the empirical estimator is still unbiased with controlled variance. Our work is related to the recent work of (Shah & Peters, 2020), which proposes a conditional independence test whose proofs rely heavily on (Chernozhukov et al., 2018a). In this paper, we use for the first time such double ML-based tests for the search problem.

3 Experiments



(a) $p_n = 0$, $\sigma^2 = 0.1$ (b) $p_n = 0$, $\sigma^2 = 0.5$ (c) $p_n = 0$, $\sigma^2 = 1.0$

(d) $p_n = 0.5$, $\sigma^2 = 0.1$ (e) $p_n = 0.5$, $\sigma^2 = 0.5$ (f) $p_n = 0.5$, $\sigma^2 = 1.0$

Figure 3: Overall performance for a single random DAG with 100 simulations for each setting, having 20 nodes and 500 observations.

3.1 Experimental Setup

To showcase performance of our algorithm, we conducted two sets of experiments: i) Comparison with causal structure learning methods (Casual and Markov Blanket discovery) using data consisted of DAGs with high number of observations-to-number of variables ratio ($n \gg d$) which is applicable to causal structure learning

methods. Markov Blanket discovery methods are included since under NFD, faithfulness, and no-hidden-confounders assumptions, Markov Blanket of the target variable corresponds to the direct parents. Note that, faithfulness and no-hidden-confounders assumptions are not necessary for our method. These experiments are discussed in details in Section 3.1.1 ii) Comparison with inference by regression methods using data consisted of DAGs with high number of observations-to-number of variables ratio ($n \approx d$ and $n \ll d$) to illustrate performance in high-dimensional regimes. This part is explained thoroughly in Section 3.1.2

3.1.1 Causal Structure Learning

For every combination of number of nodes (#nodes), connectivity (p_s), noise level (σ^2), number of observations (n), and non-linear probability (p_n) (see Table C.1), 100 examples (DAGs) are generated and stored as csv files (altogether 72.000 DAGs are simulated, comprising a dataset of overall >10GB). For each DAG, n samples are generated. We provide more details about the parameters (#nodes, p_s, p_n and n) and data generation process in Appendix C.1.1. For future benchmarking, the generated files with the code will be made available later.

The baselines we compare our method against are categorized in two groups which are suitable for observational data: i) Causal Structure Learning methods: LiNGAM (Shimizu et al., 2006), order - independent PC (Colombo & Maathuis, 2014), rankPC, 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) and Lasso[2] (Tibshirani, 1996). ii) Markov Blanket discovery methods: Grow-Shrink (GS (Margaritis & Thrun, 1999)), Incremental Association Markov Blanket (IAMB (Tsamardinos et al., 2003b)), Max-Min Parents & Children (MMPC (Tsamardinos et al., 2003a)), FastIAMB (Yaramakala & Margaritis, 2005). and IAMB with FDR Correction (Pena, 2008). The "CompareCausalNetworks"[3] and "bnlearn: Bayesian Network Structure Learning, Parameter Learning and Inference"[4] R Packages are used to run most of the baselines methods. We use 10-fold cross-validation to choose the parameters of all approaches. As direction of the possible causes in the defined setting is determined, the non-directional edges inferred by some baselines, e.g., PC are evaluated as direct causes of the target variable.

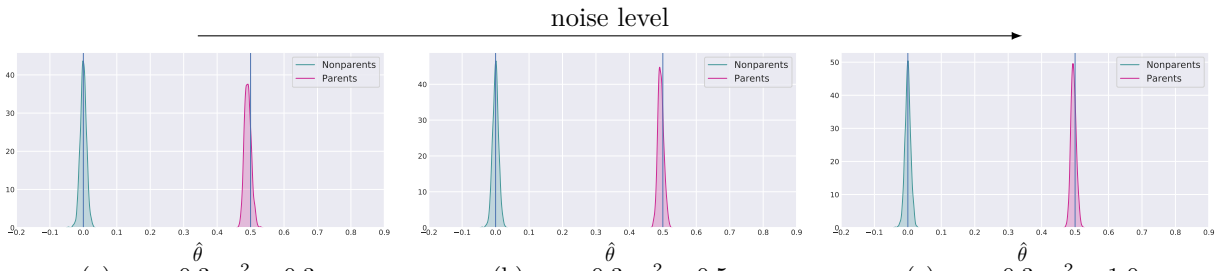
3.1.2 Inference by Regression

Similar to the previous section, for every combination of parameters, 50 examples are generated and stored, which means 15000 DAGs overall. Details are provided in Appendix C.1.2 We compare our algorithm to methods for inference in regression models: Standard 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). "selectiveInference: Tools for Post-Selection Inference" R Package [5] is leveraged to run most of these baselines. We used cross-validation to choose hyperparameters and confidence level for hypothesis testing considered is 90%.

Regression Technique and Hyper-parameters: We use Lasso as the estimator of conditional expectation for our method because the variance bound for χ_j with Lasso type estimator of conditional expectation is provided in equation 17. Further, using more splits than 2 splits in the experiment relatively increases the performance of parameter estimation. See Figure 4 for parameter estimations.

Evaluation: Recall, Fall-out, Critical Success Index, Accuracy, F1 Score, and Matthews correlation coefficient (Matthews, 1975) are considered as metrics for the evaluation. These metrics are described in Appendix C.2.

[2]None-zero coefficients are reported.
[3]https://cran.r-project.org/web/packages/CompareCausalNetworks/index.html
[4]https://cran.r-project.org/web/packages/bnlearn/
[5]https://cran.r-project.org/web/packages/selectiveInference/



(a) $p_n = 0.3$, $\sigma^2 = 0.3$ (b) $p_n = 0.3$, $\sigma^2 = 0.5$ (c) $p_n = 0.3$, $\sigma^2 = 1.0$

Figure 4: Distribution of the estimated θ values for the true and false causal parents in 100 simulations of the graph with 20 nodes, 20000 observations and 0.3 as connectivity. The vertical lines indicate the ground truth values for the linear coefficients corresponding to causal parents.

3.2 Results

3.2.1 Causal Structure Learning

Results aggregated by the number of nodes (corresponding to 18000 simulations per entry in the table), connectivity level (corresponding to 24000 simulations per entry in the table), the number of observations (corresponding to 24000 simulations per entry in the table) are illustrated in Tables 1 to 3 respectively[6]. Our method performs better than the competing baselines in terms of accuracy and F1 score, especially for more connected structures, despite data being generated according to DAG causal structures, which, dissimilar to our method, is an essential condition for them. To provide a visual comparison, we plot the accuracy of all methods w.r.t. the connectivity parameter (p_s) in Figure 3 for different values of p_n and σ^2 on 1800 samples.

It can be observed that the accuracies of the competing baselines significantly drop with increasing noise level and nonlinearity, while our method is more robust to them. We also extensively compare all the metrics (Recall, Fall-out, Critical Success Index, Accuracy, F1 Score, and Matthews correlation coefficient) for all the methods in Appendix C.3.1. According to these metrics, our approach performs better than baselines in most cases regardless of the set of parameters used for generating data. Our method shows in particular stability in performance w.r.t. the number of nodes (Table C.3), partially non-linear relationships (Table C.4), connectivity (Table C.5), number of observations (Table C.7), and noise level (Table C.6). We also show the plot of parameter estimation for direct causal parents vs. non-causal parents in Figure 4. In the plots and tables, we denote our approach as `CORTH Features`.

3.2.2 Inference by Regression

Analogous to previous part, results are aggregated by nonlinear probability (corresponding to 3750 simulations per entry in the table), number of observations (3000 simulations per entry in the table), connectivity (5000 simulations per entry in the table) and beta distribution parameters are provided in Tables C.8 to C.11. Based on these results, our method suggests more robustness w.r.t. the set of parameters used for generating data and relatively better performance compared to other methods.

3.3 Scaling Causal Inference to Large Graphs

Figure 5 shows the runtime of the method in secs as a function of the graph's size. Notice that the runtime of our algorithm in the log-log plot is roughly linear, supporting our above statement about the computational time being polynomial in d. As we used 5000 observations, additional overhead comes from cross-validation.

[6]Please refer to Appendix C.3.1 for thorough tables for all parameters.

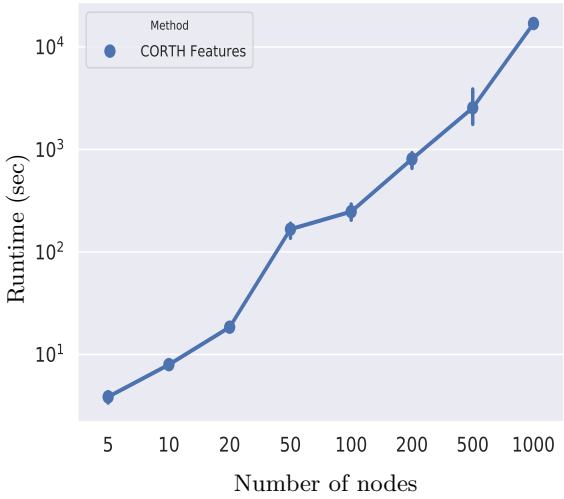


Figure 5: Runtime as a function of the number of variables for 10 simulations per number of nodes. In these simulations connectivity, number of observations, nonlinaer prob., and noise level are set to 0.3, 5000, 0, and 1 respectively.

Method	Number of Observations					
	100		500		1000	
	ACC	F1	ACC	F1	ACC	F1
GES	0.80	0.59	0.81	0.65	0.81	0.67
rankGES	0.79	0.56	0.81	0.64	0.81	0.65
ARGES	0.78	0.49	0.80	0.58	0.80	0.59
rankARGES	0.78	0.47	0.79	0.57	0.80	0.58
FCI+	0.84	0.67	0.86	0.75	0.87	0.78
LINGAM	0.84	0.65	0.91	0.74	0.94	0.88
PC	0.83	0.64	0.86	0.73	0.87	0.75
rankPC	0.82	0.62	0.85	0.71	0.85	0.73
MMPC	0.77	0.37	0.82	0.53	0.83	0.57
MMHC	0.80	0.56	0.82	0.62	0.83	0.64
GS	0.79	0.43	0.84	0.59	0.86	0.62
IAMB	0.74	0.39	0.81	0.57	0.83	0.61
Fast-IAMB	0.80	0.46	0.84	0.59	0.86	0.62
IAMB-FDR	0.78	0.37	0.84	0.58	0.85	0.61
PCI	0.83	0.59	0.91	0.85	0.93	0.89
Lasso	0.87	**0.81**	0.89	0.85	0.89	0.85
`CORTH Features`	**0.88**	0.78	**0.93**	**0.91**	**0.94**	**0.92**

Table 3: Performance across all the settings for different number of observations (100, 500 and 1000). Each single entry in the table is averaged over 24000 simulations. Our method is almost state of the art in every case.

3.4 Real-World Data

We also apply our algorithm to a recent COVID-19 Dataset (Einstein, 2020) where the task is to predict COVID-19 cases (confirmed using RT-PCR) amongst suspected ones. For an existing and extensive analysis of the dataset with predictive methods, we refer to Schwab et al. (2020). We apply our algorithm to discover the features which directly cause the diagnosed infection. We found that the following were the most common causes across different runs of our approach: Patient age quantile, Arterial Lactic Acid, Promyelocytes, and Base excess venous blood gas analysis. Lacking medical ground truth, we report these not as corroboration of our approach but rather as a potential contribution to causal discovery in this challenging problem. It is encouraging that some of these variables are consistent with other studies Schwab et al. (2020). Details on data preprocessing and more results are available in Appendix D.

4 Discussion

A recent empirical evaluation of different causal discovery methods highlighted the desirability of more efficient search algorithms (Heinze-Deml et al., 2018). In the present work, we provide identifiability results for the set of direct causal parents, including the case of partially nonlinear cyclic models, as well as a highly efficient algorithm that scales well w.r.t. the number of variables and exhibits state-of-the-art performance across extensive experiments. Our approach builds on the Double ML method Chernozhukov et al. (2018a) and properties of conditional covariance estimands, which can leverage Lasso predictors to obtain consistent estimators in high dimensional settings Smucler et al. (2019). Theoretical properties of such approaches (Smucler et al., 2019, Section 2), such as *rate double robustness*, which accommodates non-strict (approximate) sparsity assumptions, as well as model double robustness, which allows some degree of model mispecification, may explain the empirically observed performance of our approach. Whilst not amounting to full causal graph discovery, identification of causal parents is of major interest in real-world applications, e.g., when assaying the causal influence of genes on the phenotype. A natural direction worth exploring is to extend this approach for discovering direct causal parents in the case when nonlinear relationships exist between the output variable and its direct causal parents.

References

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Manuel Wüthrich, and Stefan Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296*, 2020.

Joshua D Angrist and Guido W Imbens. Identification and estimation of local average treatment effects. Technical report, National Bureau of Economic Research, 1995.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6), 1986.

Kenneth A Bollen. *Structural equations with latent variables*, volume 210. John Wiley & Sons, 1989.

Stephan Bongers, Jonas Peters, Bernhard Schölkopf, and Joris M Mooij. Theoretical aspects of cyclic structural causal models. *arXiv preprint arXiv:1611.06221*, 2016.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

Roger J Bowden and Darrell A Turkington. *Instrumental variables*, volume 8. Cambridge university press, 1990.

Carlos Brito and Judea Pearl. Generalized instrumental variables. *arXiv preprint arXiv:1301.0560*, 2012.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020.

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

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018a.

Victor Chernozhukov, Whitney Newey, and James Robins. Double/de-biased machine learning using regularized riesz representers. *arXiv preprint arXiv:1802.08667*, 2018b.

Victor Chernozhukov, Whitney K. Newey, and Rahul Singh. Learning L2 Continuous Regression Functionals via Regularized Riesz Representers. *arXiv preprint arXiv:1809.05224*, Sep 2018.

David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3 (null):507–554, March 2003. ISSN 1532-4435. doi: 10.1162/153244303321897717. URL `https://doi.org/10.1162/153244303321897717`.

Tom Claassen, Joris M. Mooij, and Tom Heskes. Learning sparse causal models is not np-hard. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, pp. 172–181, Arlington, Virginia, USA, 2013. AUAI Press.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(116):3921–3962, 2014. URL `http://jmlr.org/papers/v15/colombo14a.html`.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Hospital Israelita Albert Einstein. Diagnosis of COVID-19 and its clinical spectrum, 2020. https://www.kaggle.com/einsteindata4u/covid19.

Patrick Forré and Joris M Mooij. Markov properties for graphical models with cycles and latent variables. *arXiv preprint arXiv:1710.08775*, 2017.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. In *Advances in Neural Information Processing Systems*, pp. 3011–3021, 2017.

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

Andrew F Hayes. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications, 2017.

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

Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, pp. 1–13, 2022.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pp. 689–696, 2009.

Dominik Janzing. Causal regularization. In *Advances in Neural Information Processing Systems*, 2019.

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

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1), 2014.

Adel Javanmard and Andreas Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A), 2018.

Gustavo Lacerda, Peter L Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. *arXiv preprint arXiv:1206.3273*, 2012.

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

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

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. *Advances in neural information processing systems*, 12, 1999.

Atalanti Mastakouri, Bernhard Schölkopf, and Dominik Janzing. Selecting causal brain features with a single conditional independence test per feature. In *Advances in Neural Information Processing Systems 32*, pp. 12532–12543, 2019.

Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

Nicolai Meinshausen, Alain Hauser, Joris M Mooij, Jonas Peters, Philip Versteeg, and Peter Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016.

Preetam Nandy, Alain Hauser, and Marloes Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *Annals of Statistics*, 46, 03 2016. doi: 10.1214/17-AOS1654.

Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Jerzy Neyman. C (α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A*, 1979.

Michael Paul. Feature selection as causal inference: Experiments with text classification. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 163–172, 2017.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jose M Pena. Learning gaussian graphical models of gene networks with false discovery rate control. In *European conference on evolutionary computation, machine learning and data mining in bioinformatics*, pp. 165–176. Springer, 2008.

Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 589–598, 2011.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

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

James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and statistics: essays in honor of David A. Freedman*, 2:335–421, 2008.

Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. *Advances in Neural Information Processing Systems*, 28, 2015.

Andrea Rotnitzky, Ezequiel Smucler, and James M. Robins. Characterization of parameters with a mixed bias property. *arXiv preprint arXiv:1904.03725*, 2019.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):1–13, 2019.

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

Ruben Sanchez-Romero, Joseph D Ramsey, Kun Zhang, Madelyn RK Glymour, Biwei Huang, and Clark Glymour. Estimating feedforward and feedback effective connections from fmri time series: Assessments of statistical methods. *Network Neuroscience*, 3(2):274–306, 2019.

Patrick Schwab, August DuMont Schütte, Benedikt Dietz, and Stefan Bauer. predcovid-19: A systematic study of clinical predictive models for coronavirus disease 2019. *arXiv preprint arXiv:2005.08302*, 2020.

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

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

Patrick E Shrout and Niall Bolger. Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological methods*, 7(4), 2002.

Ezequiel Smucler, Andrea Rotnitzky, and James M Robins. A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. Technical report, 2019.

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

Robert Tibshirani. Regression shrinkage and others. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Ryan Tibshirani and Larry Wasserman. Sparsity, the lasso, and friends, 2017.

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

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

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

Ioannis Tsamardinos, Laura Brown, and Constantin Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65:31–78, 10 2006. doi: 10.1007/s10994-006-6889-7.

Mark J Van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 10. Springer, 2011.

Benito Van der Zander and Maciej Liskiewicz. On searching for generalized instrumental variables. In *AISTATS*, pp. 1214–1222, 2016.

Jonathan Warrell and Mark Gerstein. Cyclic and multilevel causation in evolutionary processes. *Biology & Philosophy*, 35(5):1–36, 2020.

Shun Yao, Shinjae Yoo, and Dantong Yu. Prior knowledge driven granger causality analysis on gene regulatory network discovery. *BMC bioinformatics*, 16(1):1–18, 2015.

Sandeep Yaramakala and Dimitris Margaritis. Speculative markov blanket discovery for optimal feature selection. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pp. 4–pp. IEEE, 2005.

Kui Yu, Lin Liu, and Jiuyong Li. A unified view of causal and non-causal feature selection. *arXiv preprint arXiv:1802.05844*, 2018.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 2014.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pp. 9472–9483, 2018.