Causal Inference by using Invariant Prediction by Peters, Bühlmann and Meinshausen (2016)

Richard Guo

ricguo@stat.washington.edu

Department of Statistics, University of Washington, Seattle

June 14, 2018

Abstract

In this report we study the invariant prediction approach for identifying direct causes of a target from a combination of observational and interventional data [Peters, Bühlmann, and Meinshausen, 2016]. We briefly review the literature in causal discovery. Then we will present the methodology of the paper and its theoretical guarantees. The key technique is to conduct inference by aggregating pieces of information, each of which corresponds to a set of variables that satisfy a conditional invariance condition, in a way that controls error. We provide a software package that implements the algorithm and use it to study the performance of invariant prediction on simulated data and two real-world datasets, including a large gene knockout experiment dataset. Both theory and experiments highlight that invariant prediction guards against false positive discoveries. Finally we discuss its strengths and weaknesses.

1 Introduction

Identifying causes for an outcome is of fundamental importance to many domains of applications. Building such a "causal model" is different from the classical notion of building a regression model, and as addressed by Peters, Bühlmann, and Meinshausen [2016], an essential difference is that the validity of prediction of a causal model should be invariant to changes of the environment. In contrast, regression models tend to only work within the environment where training data was generated. Hence, a regression model can be considered as a first-order approximation to a causal model when the environment is subject to little or no change.

To concretely illustrate such a difference, we consider the following motivating example due to Peters [2015]. Suppose a biologist is interested in identifying the factors that influence a certain phenotype, and she has found that the expression of the phenotype is positively correlated with the expression level of gene A (Fig. 1). Then she wonders whether the expression of the phenotype can be *controlled* by the level of gene A, or simply, if she can suppress the phenotype by removing gene A (i.e., setting the expression level of gene A to zero)? This question cannot be answered based on a regression model; instead, the answer shall depend on the underlying causality. For example, when gene A is a direct cause of the phenotype, removing the gene would likely suppress the phenotype (Fig. 2 left); however, when there is another gene B that is a common cause of both the phenotype and gene A, the phenotype would remain at its typical level even when gene A is removed (Fig. 2 right). From this example, we shall expect that a valid causal model would correctly predict the outcome under an active change of the environment (i.e., removal of gene A), which is called an *intervention* in the context of causal inference.

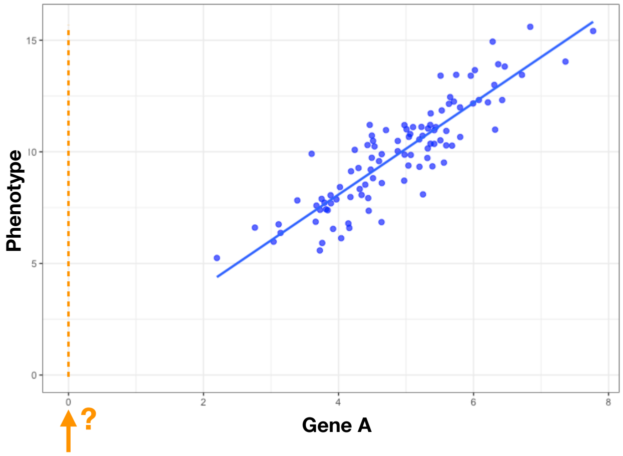

Figure 1: The expression of phenotype is found to be linearly correlated with the expression level of Gene A.

There are mainly two theoretical frameworks to model causality (see Richardson and Robins [2013] for a unified view). One is potential outcomes and counterfactuals, also known as the Neyman-Rubin model [Neyman, 1923, Rubin, 1974, 2005], where causality can be concluded by contrasting the outcomes under two (or more) scenarios. For example, in a clinical trial, we want to contrast the outcome when one has taken the medication ($Y_i(X_i =$

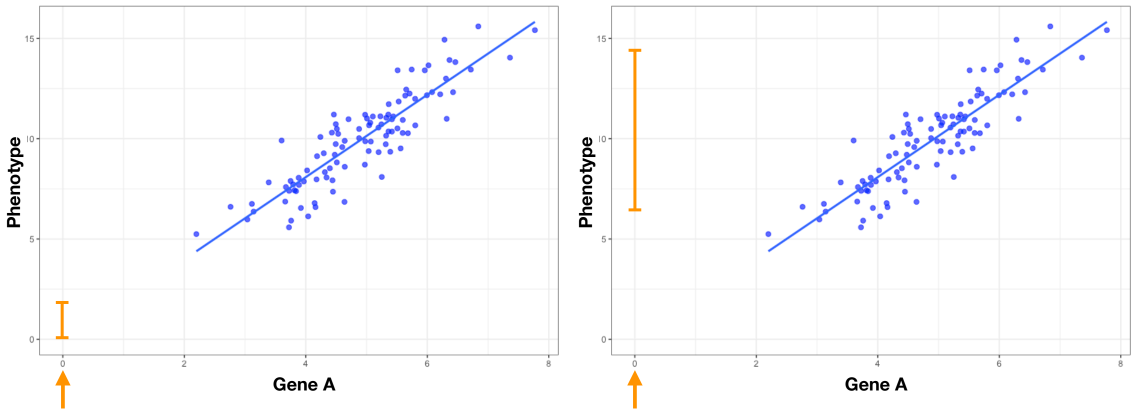

Figure 2: The expression level phenotype when gene A is removed (arrow at 0). Left: when gene A is a direct cause of phenotype; Right: when gene A is not a cause of phenotype.

1)) with the outcome when one has not taken the medication ($Y_i(X_i = 0)$). The fundamental problem is that only one of the two (or more) potential outcomes is observable, and hence the causal effect ($Y_i(X_i = 1) - Y_i(X_i = 0)$) is not identifiable for each individual. Nevertheless, under randomized assignment of X_i the average treatment effect $\mathbb{E}[Y_i(X_i = 1) - Y_i(X_i = 0)]$ is identifiable from data, and works in this framework are mainly concerned with the identification and estimation of the effect under various settings.

The other framework, which the work of Peters, Bühlmann, and Meinshausen [2016] builds upon, is structural equations [Bollen, 1983, Robins et al., 2000, Pearl, 2009] and graphical models [Lauritzen and Spiegelhalter, 1988, Lauritzen, 1996, Spirtes et al., 2000]. A structural equation model (SEM) is specified with a set of structural equations and an associated error distribution. For example, consider the following SEM with $p = 5$ variables X_1, X_2, X_3, X_4, Y:

$$\begin{aligned} &X_4 = f_4(\epsilon_4),\ X_1 = f_1(\epsilon_1),\ X_2 = f_2(X_4, \epsilon_2),\ Y = f_Y(X_1, X_2, \epsilon_Y),\ X_3 = f_3(Y, \epsilon_3), \\ &\epsilon_i \sim G_i \text{ for } i = 1,2,3,4 \text{ and } Y \sim G_Y,\ \text{ all independently}, \end{aligned} \tag{1}$$

where f_i's and f_Y are deterministic functions. Here, $\mathsf{Pa}(\cdot)$ denotes the set of parents of a variable, which appear on the RHS of the corresponding equation and hence are interpreted as *direct causes* of the LHS variable. By drawing a directed edge from each $j \in \mathsf{Pa}(i)$ to i, we have a graphical representation of the SEM (see Fig. 3 left). We assume all the errors are independent, which implies that $\epsilon_i \perp\!\!\!\perp \mathsf{Pa}(i)$ for every i. For the purpose of this report, we assume the graph is a directed acyclic graph (DAG), and therefore we can directly *simulate* $(Y, X_1, \cdots, X_4)$ from the *observational distribution* by following Equation (1) from top to bottom in the graph, i.e., by the topological ordering. In other words, we can first sample the errors, and then, starting from the variables without parents, iteratively substitute into the RHS side of equations by following the arrows, until all the variables are generated. The SEM also describes *interventional distributions* by replacing the equation governing the intervened variables with its interventional condition, while leaving the other equations unchanged. For example, a *do-intervention* [Pearl, 2009] of setting X_2 to zero (Fig. 3 right) amounts to swapping the equation $X_2 = f_2(X_4, \epsilon_2)$ with a new equation $X_2 = 0$ in Eq. (1), while keeping the other equations unchanged. Since X_2 no longer depends on X_4 upon intervention, the edge stemming from its parent is removed. This is illustrated in the right panel of Fig. 3, where we use a hammer to denote do-intervention. We will explain in more details in Section 2.1.

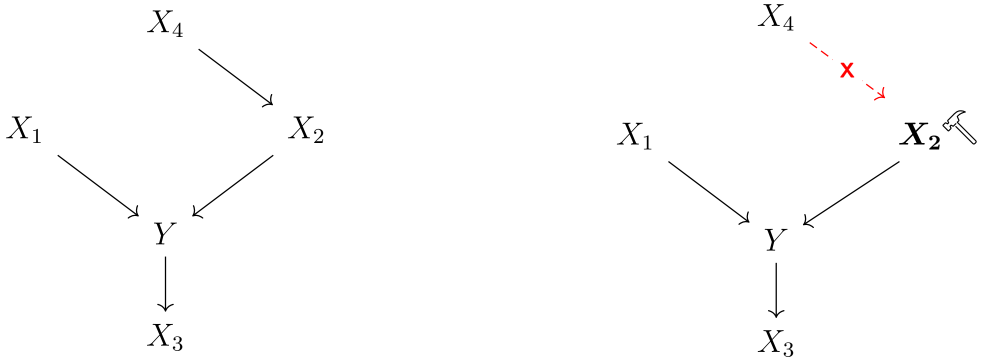

Figure 3: Graphical model of the SEM corresponding to Equation (1) (left) and upon do-intervention on X_2 (right).

Causal Discovery In the context of unknown structural equations and graph structures, we would want to identify from data, either fully or partially, the set of parents of a target variable Y. This is known as the task of *causal discovery*. A common approach is based on characterizing and estimating the Markov equivalence class of the graph [Spirtes et al., 2000, Chickering, 2002, Kalisch and Bühlmann, 2007], and inferring the causal variables when identifiable [Maathuis et al., 2009, VanderWeele and Robins, 2010]. These methods typically operate as a graph search algorithm and often adopt some greedy or approximation strategy due to the size of the space. Among them, some are able to incorporate both observational and interventional data, or datasets from different interventions [Hauser and Bühlmann, 2012, He and Geng, 2008], but they usually require specifying what variables are intervened.

In the following, we will firstly describe the invariant prediction method and its theoretical guarantees in Section 2. In Section 3, we will examine the performance of the proposed method relative to other methods on simulated datasets and two real-world datasets, including a large gene knockout experiment dataset. Finally in Section 4, we will discuss its strengths, weaknesses and potential future work.

2 Methods

2.1 Formulation and Assumptions

Suppose we want to perform causal discovery for a target variable Y of interest, which is governed by the following equation in an SEM

$$Y = f_Y(\{X_i : i \in \mathsf{Pa}(Y)\}, \epsilon_Y), \quad \epsilon_Y \sim G_Y, \quad \epsilon_Y \perp\!\!\!\perp \mathsf{Pa}(Y). \tag{2}$$

Suppose $Y \in \mathbb{R}$ (generalizable to discrete outcomes) and $f_Y : \mathbb{R}^{|\mathsf{Pa}(Y)|+1} \to \mathbb{R}$ is a deterministic function. $\mathsf{Pa}(Y)$ is the set of direct causes of Y that we want to identify, either fully or partially depending on the amount of information we have. In the following, we also use the notation $X_S := \{X_i : i \in S\}$ for a set of indices S.

f_Y, G_Y and $\mathsf{Pa}(Y)$ in Eq. (2) are unknown; we also do not know the equations and the set of parents for other variables in the SEM. We would like to perform causal inference based on the datasets

$$(\boldsymbol{X}, Y)^{(e=1)}, (\boldsymbol{X}, Y)^{(e=2)}, \cdots, (\boldsymbol{X}, Y)^{(e=|\mathcal{E}|)}$$

that come from a finite set of environments $\mathcal{E}$. Suppose $\boldsymbol{X} := (X_1, \cdots, X_p)^T \in \mathbb{R}^p$ denotes all the covariates that are measured in the dataset, and we assume that *all the direct causes* of Y are included in $\boldsymbol{X}$. In other words, we make the following *causal sufficiency* assumption.

Assumption 1 (causal sufficiency)**.** *There is no unobserved direct cause of the target, namely* $\mathsf{Pa}(Y) \subseteq \{1, \cdots, p\}$.

Note that this assumption is not always testable and we discuss this issue in Section 4. We will use superscript e to denote the environment that $(\boldsymbol{X}, Y)$ come from. For example, $e = 1$ corresponds to observational data (e.g., Fig. 3 left), and $e = 2, \cdots, |\mathcal{E}|$ each corresponds data recorded under a different intervention (e.g., Fig. 3 right corresponds to $e = 2$). In general $(\boldsymbol{X}, Y)^{(e=i)}$ and $(\boldsymbol{X}, Y)^{(e=j)}$ are defined on different probability spaces for $i \neq j$. Note that sample sizes can differ across environments. We can stack all the datasets together as a table

$$\begin{array}{cccccc} X_1 & X_2 & \cdots & X_p & Y & e \\ \hline \vdots & \vdots & & \vdots & \vdots & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots & |\mathcal{E}| \end{array},$$

where e groups the rows that come from the same data generating mechanism.

Here we define an environment as a "version" (or a "modified version") of the SEM, assuming that the equation governing the target stays the same across environments. In other words, target Y is not intervened in all environments.

Assumption 2 (target is not intervened)**.** *For every environment $e \in \mathcal{E}$ it holds that*

$$Y^e = f_Y(X^e_{\mathsf{Pa}(Y)}, \epsilon^e_Y), \quad \epsilon^e_Y \sim G_Y \textit{ and } \epsilon^e_Y \perp\!\!\!\perp X^e_{\mathsf{Pa}(Y)}, \tag{3}$$

where f_Y, G_Y and the set of indices $\mathsf{Pa}(Y)$ do not depend on e.

Under such assumption, an environment e may correspond to one of the following.

1. **Observational** (the original SEM).

2. **Do-intervention** on one or several variables other than Y. When we perform a do-intervention on variables $D^e \subset \{1, \cdots, p\}$ in environment e, we replace the equation of $j \in D^e$ to

$$X^e_j = d^e_j, \tag{4}$$

where d^e_j is the value imposed on X_j. Upon do-intervention, X^e_j no longer depends on $X^e_{\mathsf{Pa}(X_j)}$ and hence the edges from its parents are removed (see Fig. 3 right).

3. **Noise (soft) intervention** on one or several variables other than Y. For an intervened variable $j \in B^e$ in environment e, we keep the function f_j but replace the noise distribution with $G_j^e \neq G_j$. The new equation becomes

$$X_j^e = f_j(X_{\mathsf{Pa}(X_j)}^e, \epsilon_j^e), \quad \epsilon_j^e \sim G_j^e \text{ and } \epsilon_j^e \perp\!\!\!\perp X_{\mathsf{Pa}(X_j)}^e. \tag{5}$$

   For example, G_j^e can be specified as a scaling or a shift of G_j, namely $\epsilon_j^e =_d s_j^e \epsilon_j$ or $\epsilon_j^e =_d \epsilon_j + \delta_j^e$. Contrary to do-interventions, here the edges from its parents are kept.

4. **Observational data under "changed environment"**. This subsumes other arbitrary changes made to the equations other than Equation (3).

Conceptually, Assumption 1 and Assumption 2 are the two main conditions; later we will introduce another assumption on the linearity of f_Y due to technical difficulty. We also remark that, compared to the causal sufficiency assumption, one usually have a better clue on whether the target variable has been intervened or not.

2.2 Inference with Conditional Invariance

2.2.1 Conditional Invariance

For a set of covariates $S \subseteq \{1, \cdots, p\}$, we say that S satisfies *conditional invariance* if for every $e \neq f \in \mathcal{E}$, the conditional probability measure satisfies

$$P(Y^e | X_S^e = \boldsymbol{x}_S) =_d P(Y^f | X_S^f = \boldsymbol{x}_S) \quad \text{for all } \boldsymbol{x}_S. \tag{6}$$

It is immediate from Assumption 2 that $S = \mathsf{Pa}(Y)$ satisfies conditional invariance. To be more explicit, for every $e \in \mathcal{E}$, and for any measurable $A \subseteq \mathbb{R}$ and any $\boldsymbol{x} \in \mathbb{R}^{|\mathsf{Pa}(Y)|}$, by Eq. (3) we have

$$P^e(Y^e \in A | X_{\mathsf{Pa}(Y)}^e = \boldsymbol{x}) = G_Y(\{\epsilon : f_Y(\boldsymbol{x}, \epsilon) \in A\}), \tag{7}$$

which does not depend on e. However, is $S = \mathsf{Pa}(Y)$ the only set that satisfies conditional invariance? Let us consider the following example.

Example 1. Consider the structural equation model under three environments as shown in Fig. 4. We make the following observations on the invariance of $P(Y^e | X_S^e)$ for different S.

1. $S \in \{\emptyset, \{1\}, \{2\}\}$ does not satisfy conditional invariance, because both X_1 and X_2 are intervened in $e = 2$.

2. $S = \{1, 2\} = \mathsf{Pa}(Y)$ satisfies conditional invariance.

3. $S = \{1, 2, 4\}$ also satisfies conditional invariance because $P(Y^e | X_1^e, X_2^e, X_4^e) = P(Y^e | X_1^e, X_2^e)$ by conditional independence. Then the invariance follows from the invariance of $P(Y^e | X_1^e, X_2^e)$.

4. $S = \{1, 2, 3\}$ does not satisfy conditional invariance. In $e \in \{1, 2\}$, Y is not conditionally independent of X_3 given X_1 and X_2; but in $e = 3$, Y is conditionally independent of X_3 given X_1 and X_2.

5. $S = \{1, 4\}$ does *not* satisfy conditional invariance for $e \in \{1, 2, 3\}$, but *does* satisfy for $e \in \{1, 3\}$.

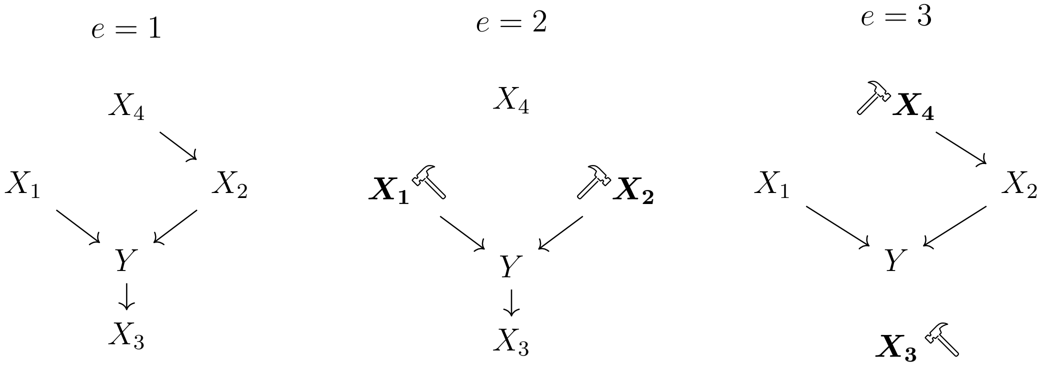

Figure 4: Example of an SEM under $|\mathcal{E}| = 3$ environments. The hammer symbol denotes a do-intervention.

To summarize what we have learned so far:

1. $S = \mathsf{Pa}(Y)$ always satisfies conditional invariance.

2. There might exist other sets $S \neq \mathsf{Pa}(Y)$ that also satisfy conditional invariance.

3. In general, whether or not S satisfies conditional invariance depends on both S and the environments $\mathcal{E}$ considered.

4. It is *not true* that any $S \supset \mathsf{Pa}(Y)$ satisfies conditional invariance.

It is worth mentioning that the conditional invariance of $\mathsf{Pa}(Y)$ has been known as a direct consequence of the local Markov property of SEMs [Pearl, 2009]; however, it has not been previously exploited as a major tool for inference. Supposing we collect *all* the sets that satisfy conditional invariance, then $\mathsf{Pa}(Y)$ must be *one* of these sets although we cannot tell which one it is. Essentially, the idea of Peters, Bühlmann, and Meinshausen [2016] is to perform estimation by aggregation.

2.2.2 Estimation by Aggregation

For every subset of covariates $S \subseteq \{1, \cdots, p\}$, consider testing

$$H_{0,S}(\mathcal{E}) : P(Y^e|X_S^e) \text{ is invariant } \forall e \in \mathcal{E}, \text{ vs. } H_{1,S} : P(Y^e|X_S^e) \text{ is not invariant.} \tag{8}$$

Suppose we can test $H_{0,S}(\mathcal{E})$ versus $H_{1,S}(\mathcal{E})$ from datasets $\{(\boldsymbol{X}, Y)^e : e \in \mathcal{E}\}$ at level α *uniformly* for all $S \subseteq \{1, \cdots, p\}$, namely

$$P_{0,S}(H_{0,S}(\mathcal{E}) \text{ rejected}) \leq \alpha \quad \text{for all } S \subseteq \{1, \cdots, p\}. \tag{9}$$

And suppose we collect all the sets S such that $H_{0,S}(\mathcal{E})$ is not rejected. Can we aggregate them usefully to partially (or fully) identify $\mathsf{Pa}(Y)$, while controlling the error of making a false discovery? It turns out “intersection” is such an aggregation operator.

We describe the following template of the invariant prediction algorithm.

Algorithm 1 Invariant Prediction at level α

```
for $S \subseteq \{1, \cdots, p\}$ do
    Test $H_{0,S} : P(Y^e|X_S^e)$ invariant for all $e \in \mathcal{E}$ at level $\alpha$
    if $H_{0,S}$ not rejected then
        Construct (1-$\alpha$)-level confidence set $\hat{\Gamma}_S$ from data pooled from all the environments
    end if
end for
return the set of estimated direct causes $\hat{S}(\mathcal{E})$ and the confidence set for direct causal effects $\hat{\Gamma}(\mathcal{E})$, as defined in Eq. (10) and Eq. (11) respectively.
```

$$\hat{S}(\mathcal{E}) = \bigcap \{S : H_{0,S}(\mathcal{E}) \text{ not rejected}\}, \tag{10}$$

$$\hat{\Gamma}(\mathcal{E}) = \bigcup \{\hat{\Gamma}_S : H_{0,S}(\mathcal{E}) \text{ not rejected}\}. \tag{11}$$

We will define direct causal effects in the next section, where we assume linearity on f_Y; and we will also leave the description of confidence sets to Section 2.2.4. For now, we focus on the proposed estimator $\hat{S}(\mathcal{E})$. It is guaranteed with high probability to be (i) either an empty set (no discovery) or (ii) a non-empty subset of $\mathsf{Pa}(Y)$. We have the following theorem.

Theorem 1 (family-wise error rate)**.** *Suppose that $H_{0,S}(\mathcal{E})$ is tested at level α uniformly for all set of variables $S \subseteq \{2, \cdots, p+1\}$ and set of environments $\mathcal{E}$, under Assumption 2 we have*

$$P(\hat{S}(\mathcal{E}) \subseteq S^*) \geq 1 - \alpha. \tag{12}$$

Proof. By the definition of $\hat{S}(\mathcal{E})$ in Eq. (10), we know $\{H_{0,\mathsf{Pa}(Y)}$ not rejected$\}$ implies $\hat{S}(\mathcal{E}) \subseteq \mathsf{Pa}(Y)$. Hence, we have

$$P(\hat{S}(\mathcal{E}) \subseteq S^*) \geq P(H_{0,\mathsf{Pa}(Y)}(\mathcal{E}) \text{ not rejected}) \geq 1 - \alpha,$$

where the last inequality comes from the fact that $H_{0,S}$ is tested uniformly at level α. □

Theorem 1 basically states that the family-wise error rate (FWER), namely the probability of making *any* false discovery, is bounded by α. We also have the following natural result that as we collect data from more environments, we make no fewer discoveries.

Proposition 1 (monotonicity)**.** *Suppose $\{H_{0,S}(\mathcal{E}')$ not rejected$\}$ always implies $\{H_{0,S}(\mathcal{E})$ not rejected$\}$ for $\mathcal{E} \subseteq \mathcal{E}'$. Then for $\mathcal{E} \subseteq \mathcal{E}'$ we have $\hat{S}(\mathcal{E}) \subseteq \hat{S}(\mathcal{E}')$ and $\hat{\Gamma}(\mathcal{E}) \supseteq \hat{\Gamma}(\mathcal{E}')$.*

Proof. For $\mathcal{E} \subseteq \mathcal{E}'$, by the supposition we have

$$\{S : H_{0,S}(\mathcal{E}) \text{ not rejected}\} \subseteq \{S : H_{0,S}(\mathcal{E}') \text{ not rejected}\}.$$

Then the proposition follows from the definitions in Eqs. (10) and (11) in terms of intersection and union. □

From this proposition, it is natural to postulate that $\mathsf{Pa}(Y)$ can be eventually fully discovered, i.e., $\hat{S}(\mathcal{E}) = \mathsf{Pa}(Y)$, when the set of environments $\mathcal{E}$ becomes "sufficiently rich". However, we do not have a complete characterization of conditions for full discovery; Peters, Bühlmann, and Meinshausen [2016] propose two sufficient conditions for full discovery under Gaussian linear SEM. We list these results as Theorem 2 for completeness. Since these conditions are hardly met in practice and irrelevant to the experiments performed, we will omit the details and just focus on the case of $S \subsetneq \mathsf{Pa}(Y)$ for the rest of the report.

Theorem 2 (full discovery)**.** *For a linear Gaussian SEM and a set of environments $\mathcal{E}$. We have $P(\hat{S}(\mathcal{E}) = \mathsf{Pa}(Y)) \to 1$ as $\min_{e \in \mathcal{E}} n_e \to \infty$ if any of the following is satisfied:*

1. *For each $j \in \{1, \cdots, p\}$, there exists $e \in \mathcal{E}$ such that $D^e = \{j\}$ and X_j is do-intervened to $d_j^e \neq \mathbb{E}\, X_j$ (see Eq. (4)),*
2. *For each $j \in \{2, \cdots, p+1\}$, there exists $e \in \mathcal{E}$ such that $B^e = \{j\}$ and X_j is noise-intervened with either scaling $1 \neq (s_j^e)^2 < \infty$ or shift $\delta_j^e \neq 0$ (see Eq. (5)).*

2.2.3 Testing Conditional Invariance

From the previous section, the problem is reduced to testing conditional invariance uniformly at level α. However, it is in general difficult to test whether $(Y^e | X_S^e = \boldsymbol{x}_S) =_d (Y^f | X_S^f = \boldsymbol{x}_S)$ holds for every $\boldsymbol{x}_S \in \mathbb{R}^{|S|}$ when the underlying f_Y is nonlinear. To proceed, we circumvent this technical difficulty by assuming Y as a linear combination of $X_{\mathsf{Pa}(Y)}$ and the error term ϵ_Y. We strengthen Assumption 2 to the following linear form.

Assumption 3 (target is not intervened; linearity)**.** *For every environment $e \in \mathcal{E}$ it holds that*

$$Y^e = \sum_{i=1}^{p} \gamma_i^* X_i^e + \epsilon_Y^e \text{ with } \gamma_i^* = 0 \text{ for } i \notin \mathsf{Pa}(Y),\ \epsilon_Y^e \sim G_Y,\ \mathbb{E}\, \epsilon_Y^e = 0 \text{ and } \epsilon_Y^e \perp\!\!\!\perp X_{\mathsf{Pa}(Y)}^e, \quad (13)$$

where $\{\gamma_i^* : i = 1, \cdots, p\}$, G_Y and the set of indices $\mathsf{Pa}(Y)$ do not depend on e.

Note that without loss of generality, we assume Y is centered and intercept is suppressed. Also, by Eq. (13) we only assume that, in the SEM, the equation governing Y is linear; we *do not* make linear assumptions on the *other* equations. In Equation (13), the coefficients γ_i^* can be interpreted as the *direct causal effect* of X_i on Y.

Non-Gaussian Noise Under Assumption 3, testing $H_{0,S}(\mathcal{E})$ becomes testing

$$\begin{aligned} H_{0,S}(\mathcal{E}) : \exists \gamma_S \in \mathbb{R}^{|S|} \text{ such that } Y^e - \gamma_S^\top X_S^e \perp\!\!\!\perp X_S^e \text{ for every } e \in \mathcal{E}, \\ \text{and } Y^e - \gamma_S^\top X_S^e =_d Y^f - \gamma_S^\top X_S^f \text{ for every } e \neq f \in \mathcal{E}. \end{aligned} \tag{14}$$

Moreover, we can test Eq. (14) by (i) obtaining least square estimate $\hat{\gamma}_S$ by regression Y on X_S with data pooled from all the environments and (ii) testing equality of distribution of residuals $(Y^e - \hat{\gamma}_S^\top X^e)$ across environments $e \in \mathcal{E}$. Such a procedure is called "Method II" in Peters, Bühlmann, and Meinshausen [2016]. Note that in step (ii), when the error is non-Gaussian, we need to use nonparametric tests, such as Kolmogorov-Smirnov test or rank test, for testing identical distribution of residuals. Also, in step (i) we rely on the consistency of $\hat{\gamma}_S$ and hence the resulting test has level α *asymptotically*.

Gaussian Noise When $G_Y = \mathcal{N}(0, \sigma^2)$ in Eq. (13), we can perform *exact* level-α test of $H_{0,S}$ for a pair of environments with the Chow test. The Chow test [Chow, 1960] is invented in econometrics, and is commonly used to test the equality of regression coefficients under Gaussian error with equal variances. Since we might have more than two environments, for each e we can perform a Chow test between data from e and data from all the other environments, and combine the p-values with Bonferroni correction [Bonferroni, 1936].

Algorithm 2 Combined Chow test for $H_{0,S}(\mathcal{E})$ at level α

```
for $e \in \mathcal{E}$ do
    Let $(\boldsymbol{X}_S^e, \boldsymbol{y}^e)$ be the dataset for environment $e$ with sample size $n_e$.
    Let $(\boldsymbol{X}_S^{-e}, \boldsymbol{y}^{-e})$ be the dataset combined from $\mathcal{E} \setminus \{e\}$, with sample size $n_{-e}$.
    Let $\hat{\boldsymbol{\gamma}}_S^{-e}$ be the least-square estimator and $\hat{\sigma}^2$ be the estimated error variance, both from $(\boldsymbol{X}_S^{-e}, \boldsymbol{y}^{-e})$.
    Compute $p$-value from the following $F$-distribution
        $p_e = 1 - F\left(\frac{\boldsymbol{D}_e^T \boldsymbol{\Sigma}_D^{-1} \boldsymbol{D}_e}{\hat{\sigma}^2 n_e}; n_e, n_{-e} - |S| - 1\right), \quad (15)$
    where $\boldsymbol{D}_e = \boldsymbol{y}^e - \boldsymbol{X}_S^e \hat{\boldsymbol{\gamma}}_S^{-e}$ is the prediction residual on $e$, and $\boldsymbol{\Sigma}_D$ is the covariance matrix given by
        $\boldsymbol{\Sigma}_D = \boldsymbol{I}_{n_e} + \boldsymbol{X}_S^e (\boldsymbol{X}_S^{-e\top} \boldsymbol{X}_S^{-e})^{-1} \boldsymbol{X}_S^{e\top}. \quad (16)$
    Reject $H_{0,S}(\mathcal{E})$ if $p_e < \alpha/|\mathcal{E}|$.
end for
```

The combined Chow test retains level α since (i) under $H_{0,S}(\mathcal{E})$, the regression coefficients and error variances are the same across environments and (ii) Bonferroni correction controls Type-I error via the union bound $P_{0,S}(\bigcup_{e \in \mathcal{E}} \{p_e < \alpha/|\mathcal{E}|\}) \leq \sum_{e \in \mathcal{E}} P(p_e < \alpha/|\mathcal{E}|) = \alpha$.

2.2.4 Confidence Intervals

Algorithm 1 aggregates confidence sets by taking union of $\hat{\Gamma}_S$ such that $H_{0,S}(\mathcal{E})$ is not rejected (see Eq. (11)). For simplicity, we choose $\hat{\Gamma}_S$ to be a *rectangular* confidence set for γ^* at level $(1-\alpha)$: (simultaneous) level-$(1-\alpha)$ confidence intervals for $\gamma_i^* : i \in S$, and $\{0\}$ for $\gamma_i^* : i \notin S$. It holds that the aggregated confidence intervals (set) achieves coverage $(1 - 2\alpha)$.

Theorem 3 (confidence intervals)**.** *Suppose that $H_{0,S}(\mathcal{E})$ is tested at level α uniformly for all set of variables $S \subseteq \{2, \cdots, p+1\}$ and set of environments $\mathcal{E}$. And suppose $\hat{\Gamma}_S$ has coverage $(1 - \alpha)$ uniformly for all (γ_S, S) that satisfy Eq.* (14)*. Under Assumption 3 we have*

$$P(\gamma^* \in \hat{\Gamma}(\mathcal{E})) \geq 1 - 2\alpha. \tag{17}$$

Proof. We have

$$P(\gamma^* \notin \hat{\Gamma}(\mathcal{E})) = P\left(\{\gamma^* \notin \hat{\Gamma}(\mathcal{E})\} \bigcap \{H_{0,\mathsf{Pa}(Y)} \text{ rejected}\}\right) \\ + P\left(\{\gamma^* \notin \hat{\Gamma}(\mathcal{E})\} \bigcap \{H_{0,\mathsf{Pa}(Y)} \text{ not rejected}\}\right),$$

where

$$P\left(\{\gamma^* \notin \hat{\Gamma}(\mathcal{E})\} \bigcap \{H_{0,\mathsf{Pa}(Y)} \text{ rejected}\}\right) \leq P(H_{0,\mathsf{Pa}(Y)} \text{ rejected}) \leq \alpha$$

because $H_{0,S}$ is tested at level α uniformly. Also, by the definition in Eq. (11),

$$P\left(\{\gamma^* \notin \hat{\Gamma}(\mathcal{E})\} \bigcap \{H_{0,\mathsf{Pa}(Y)} \text{ not rejected}\}\right) \leq P(\gamma^* \notin \hat{\Gamma}_{\mathsf{Pa}(Y)}) \leq \alpha,$$

where the last inequality follows from the uniform $(1-\alpha)$-coverage of $\hat{\Gamma}_S$. Combining the two, we have $P(\gamma^* \in \hat{\Gamma}(\mathcal{E})) \geq 1 - 2\alpha$. □

2.2.5 Computational Complexity and Screening

To ensure that the truth $\mathsf{Pa}(Y)$ is aggregated in construction of $\hat{S}(\mathcal{E})$ and $\hat{\Gamma}(\mathcal{E})$, Algorithm 1 iterates over 2^p subsets. Suppose we use the Chow test. Each call of Algorithm 2 costs $O(p^2 n_{-e})$ for least square and about $O(n_e^3)$ for matrix inversion in Eq. (15). Assuming n_e is the same for all e and $n_e \gg p$, the worst case complexity is $O(2^p|\mathcal{E}|n_e^3)$. In practice, one can test smaller subsets first in case of early termination when the running intersection becomes empty. Besides, when it is not necessary to report confidence intervals, we can skip testing supersets of the running intersection. These time-saving strategies are implemented in our Julia package `InvariantCausal` (see Section 2.3).

Nevertheless, the computational complexity quickly becomes formidable as p grows. In practice, we have to reduce p to around 10 for a reasonable running time. Peters, Bühlmann, and Meinshausen [2016] suggest restricting to a small set of variables preselected by a screening algorithm, such as lasso [Tibshirani, 1996] and square-root lasso [Belloni et al., 2011]. We also make use of high-dimensional ordinary least square projections [Wang and Leng, 2016] for screening when $p \gg n$. However, in general it is not guaranteed that all variables in $\mathsf{Pa}(Y)$ are kept after screening; and when the screening algorithm fails to cover $\mathsf{Pa}(Y)$, *guarantees of $\hat{S}(\mathcal{E})$ and $\hat{\Gamma}(\mathcal{E})$ are lost.* We illustrate such a phenomenon in Section 3.1.

2.3 Software Package

Accompanying this report, we provide a software package `InvariantCausal`[1] implemented in Julia [Bezanson et al., 2017]. Compared to the original R package `InvariantCausalPrediction`[2] from the authors, our package offers the following features.

- Generally faster computation. Avoids testing supersets of the running intersection when only variable selection is required (option `selection_only=true`).
- More robust high dimensional screening with HOLP algorithm [Wang and Leng, 2016].
- Conditional invariance test for logistic regression with higher power [Perng and Littell, 1976]. See also Section 3.3.

3 Experimental Results

We study the performance of invariant prediction and compare it to relevant methods with experiments on simulated data and real-world data. We will use our Julia package `InvariantCausal` to run the proposed method.

3.1 Simulated Experiment

3.1.1 Settings

We generate data from random instances of Gaussian linear structural equation models and test the performance of various causal discovery algorithms. In the following, each random instance is called a *setting*, and we study the performance of algorithms on a setting by replicating over 1,000 generated datasets within the setting. We generate 120 random settings to cover a wide variety of scenarios. We firstly describe how a setting is generated.

Each setting consists of two environments: observational and interventional from an SEM. To generate an SEM of p variables, we firstly generate a random acyclic graph by choosing a random ordering and connect two nodes with probability $k/(p-1)$, which gives

[1] Available from `https://github.com/richardkwo/InvariantCausal`

[2] `https://cran.r-project.org/package=InvariantCausalPrediction`

an average degree of k. Given the graph, we sample non-zero coefficients uniformly from an interval $[\mathrm{lb}^{e=1}, \mathrm{ub}^{e=2}]$ bounded away from zero with a random sign. We choose noise variances uniformly between $\sigma^2_{\min}$ and $\sigma^2_{\max}$. Note that the model can be described as

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{\epsilon}, \tag{18}$$

where B_{ij} denotes the coefficient from X_j to X_i, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \mathrm{diag}(\sigma_1^2, \cdots, \sigma_p^2))$. The random vector $\boldsymbol{X}$ can be simply simulated as $\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\epsilon}$. We draw n_{obs} samples as the observational dataset ($e = 1$).

For the interventional dataset, we perform noise interventions and change of coefficients on one or several variables, excluding the target variable Y that is picked uniformly at random. Note that such a modification of the SEM does not change the equation and associated error governing Y, and hence the working condition of invariant prediction is satisfied (see item 4 following Assumption 2). We intervene either a single node or a fraction θ of nodes, and the set of intervened nodes A is randomly picked. For an intervened node $j \in A$, its noise variance is scaled by a factor a_j that is uniformly drawn between $a_{\min}$ and $a_{\min} + \Delta_a$; the interventional coefficients B_{js} are either unchanged (with certain probability), or regenerated as in the previous procedure with bounds $\mathrm{lb}^{e=2}$ and $\mathrm{ub}^{e=2}$. Given the intervened SEM, we draw n_{int} samples as the interventional dataset ($e = 2$).

The parameters for generating a setting are chosen uniformly from the following options (Table 1). Note that given a setting, the dataset $(X^{e=1}, X^{e=2})$ is replicated 1,000 times to obtain the statistics on performance within a setting.

3.1.2 Algorithms in Comparison

We compare the performance of the following algorithms.

Invariant Causal Prediction (`invariant`) When there are more than 10 variables, we use lasso (implemented in `glmnet`) [Friedman et al., 2010] to screen out 10 variables. To speed up the algorithm, we enable option `selection_only=true` in package `InvariantCausal` such that supersets of the running intersection will be skipped.

Greedy Equivalence Search (`GES`) From purely observational data, GES [Chickering, 2002] identifies the Markov equivalence class of the graph, i.e., by identifying the

Table 1: Options for parameters of a randomly generated setting

Setting Parameter	Options
sample size of observational dataset n_{obs}	$100, 200, 300, 400, 500$
sample size of interventional dataset n_{int}	$100, 200, 300, 400, 500$
number of variables p	$5, 6, \cdots, 40$
average degree k of graph	$1, 2, 3, 4$
lower bound for coefficient $\text{lb}^{e=1}$	$0.1, 0.2, \cdots, 2$
$\Delta_b^{e=1} = \text{ub}^{e=1} - \text{lb}^{e=1}$	$0.1, 0.2 \cdots, 1$
bounds for error variances $\sigma_{\min}, \sigma_{\max}$ (subject to $\sigma_{\min} \leq \sigma_{\max}$)	$0.1, 0.2, \cdots, 2$
lower bound for noise multiplier $a_{\min}$	$0.1, 0.2, \cdots, 4$
$\Delta_a = a_{\max} - a_{\min}$	0 (with prob. 1/3), otherwise uniformly from $0.1, 0.2, \cdots, 2$
probability that intervened coefficients unchanged	2/3
bounds for intervened coefficients $\text{lb}^{e=2}, \text{ub}^{e=2}$ ($\text{lb}^{e=2} \leq \text{ub}^{e=2}$)	$0.1, 0.2, \cdots, 2$
probability that only one variable is intervened	1/6
fraction of nodes intervened otherwise	$1/1.1, 1/1.2, \cdots, 1/3$

skeleton and orient v-structures and some edges. However, some edges X_i—Y will still remain ambiguous (meaning that there exist both instances of $X_i \to Y$ and $X_i \leftarrow Y$ in the Markov equivalence class). For this experiment, we treat the pooled dataset as observational data, and treat oriented edges $X_i \to Y$ as stemming from direct causes. We use the implementation from R package `pcalg` [Kalisch et al., 2012].

Greedy Interventional Equivalence Search with known Interventions (`GIES.known`) GIES [Hauser and Bühlmann, 2012] is an extension of GES to a mix of observational and interventional data. GIES require specification of intervened variables; for an intervened variable j, it replaces $p(X_j|X_{\mathtt{Pa}(j)})$ by a Gaussian distribution in X_j. GIES achieves finer identification of Markov equivalence class compared to purely observational data. GIES works by maximizing an ℓ_0-penalized likelihood score and we use the implementation in `pcalg`.

Greedy Interventional Equivalence Search with unknown Interventions (`GIES.unknown`) For fairness of comparison, we hide the information on intervened targets from GIES and treat every variables as intervened in the second environment.

Linear non-Gaussian Acyclic Models (`LiNGAM`) When the errors are non-Gaussian, the graph structure becomes identifiable from observational data [Shimizu et al., 2006]. To apply the method, we pool data from two environments. For example, consider $X_j^{e=1} = \sum_{i \in \mathsf{Pa}(j)} B_{ji} X_i^{e=1} + \epsilon_1$ and $X_j^{e=2} = \sum_{i \in \mathsf{Pa}(j)} B_{ji} X_i^{e=2} + \epsilon_2$ when the coefficients are the same. We can treat the pooled data as $X_j = \sum_{i \in \mathsf{Pa}(j)} B_{ji} X_i + \tilde{\epsilon}$, where $\tilde{\epsilon}$ is distributed as a mixture of two Gaussians and therefore non-Gaussian. However, note that in general the assumptions of LiNGAM are violated since some coefficients are changed in intervention, and the pooled noises are not independent. We ignore such violations for this experiment. LiNGAM is based on independent component analysis and we use its implementation in `pcalg` based on `fastICA` [Marchini et al., 2013].

Regression (`reg`) As a baseline method without guarantee, we run least square regression on pooled data and let $\hat{S}$ be all variables significant at level α/p.

Marginal Correlation (`marginal.corr`) As another baseline method with no guarantee, we let $\hat{S}$ be all variables with a correlation with the target significant at level α/p.

3.1.3 Results

Coverage We show the result on coverage Fig. 5. Here coverage is defined to be one minus the family-wise error rate, namely $P(\hat{S}(\mathcal{E}) \subseteq \mathsf{Pa}(Y))$. The experiment is performed at $\alpha = 0.05$ across 120 settings in Fig. 5 for different methods, where each setting is replicated over 1,000 random datasets. We can observe that `invariant` is the only method that attains nominal coverage of 95% (solid line); however, noticeably there are a few settings that are obviously below the nominal coverage. A closer examination reveals that in these settings the screening algorithm frequently fails to include all the true direct causes of the target variable (the frequency of successful screening is coded by color), and hence the coverage guarantee (see Theorem 1) is lost. When we restrict ourselves to the cases when all the direct causes are retained after/without screening, we can see that the proposed method achieves nominal coverage in all settings (see Fig. 6).

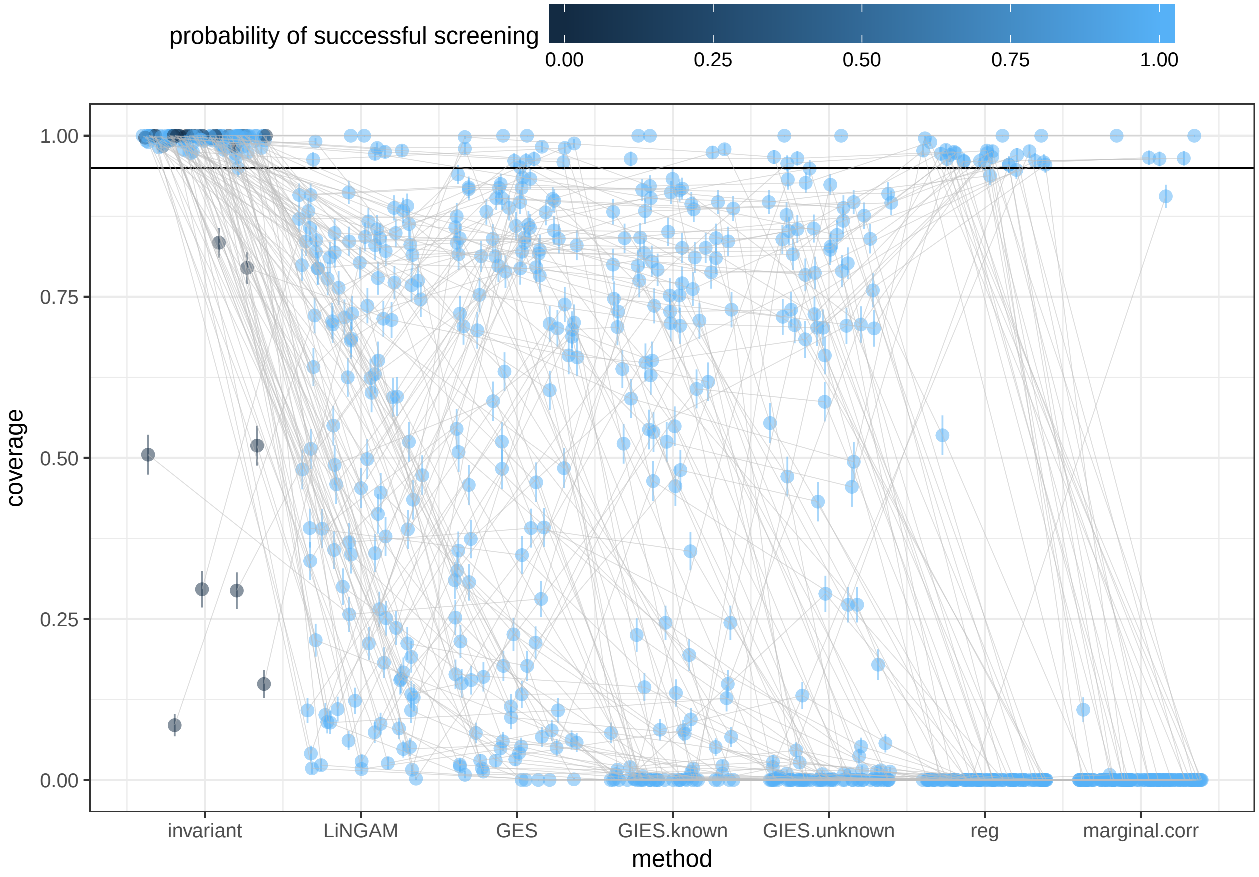

Figure 5: Coverage $P(\hat{S}(\mathcal{E}) \subseteq \mathsf{Pa}(Y))$ of methods in comparison under $\alpha = 0.05$. Each setting corresponds to a dot and the line connects the same settings. Each setting is replicated $1,000$ times to obtain the error bars (95% confidence interval). The solid line marks nominal coverage of 95%. Method `invariant` uses screening algorithm to reduce p to 8, and the color corresponds to the probability that all direct causes are kept after/without screening. Method `invariant` is the only method that achieves nominal coverage except for a few settings where screening algorithm fails to include all the direct causes.

Additionally, we show the results on probability of full discovery $P(\hat{S} = \mathsf{Pa}(Y))$ in Fig. 7. We observe that the `invariant` method is quite conservative and it rarely discovers all the direct causes. In other words, the method has relatively low power. The running time is

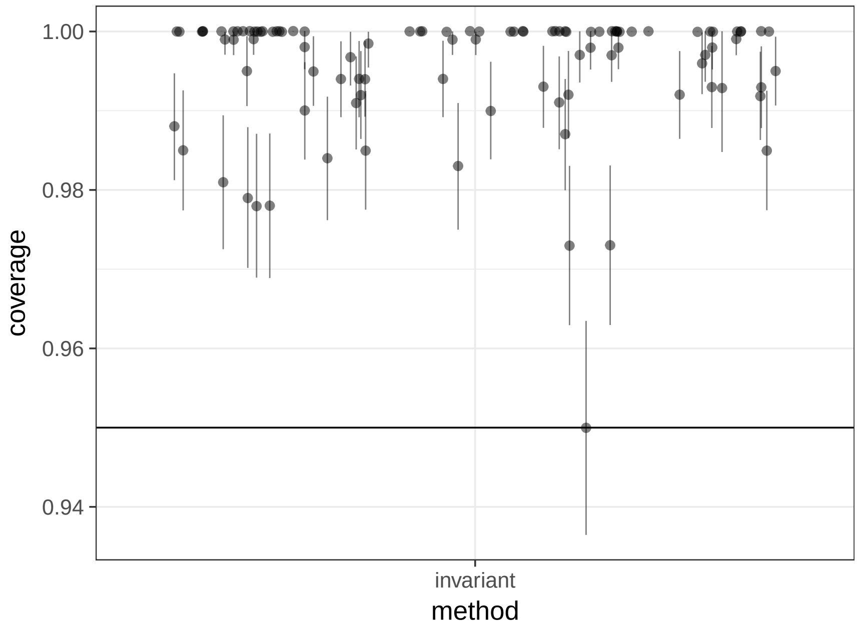

Figure 6: Coverage $P(\hat{S}(\mathcal{E}) \subseteq \mathsf{Pa}(Y))$ of invariant prediction for cases when all true causes are considered by the algorithm (successful screening).

plotted in Fig. 8.

3.2 Gene Knockout Experiments

We apply the invariant prediction method to a large dataset[3] of gene knockout experiments on yeast [Kemmeren et al., 2014]. The dataset comprises of expression levels of $p = 6,170$ genes under different conditions: (i) “wild-type” (observational) expression levels from $n_{\text{obs}} = 160$ measurements and (ii) measurements from $n_{\text{int}} = 1,479$ interventions, where each intervention corresponds to measurements after knocking out a single gene out of $1,479$ genes that are selected by biologists. Note that each intervention is represented by one row of data. We want to perform causal inference to discover direct causal relations among all pairs of genes.

[3] Dataset obtained from `http://deleteome.holstegelab.nl/data/downloads/causal_inference/Kemmeren.hdf5`. See also Meinshausen et al. [2016].

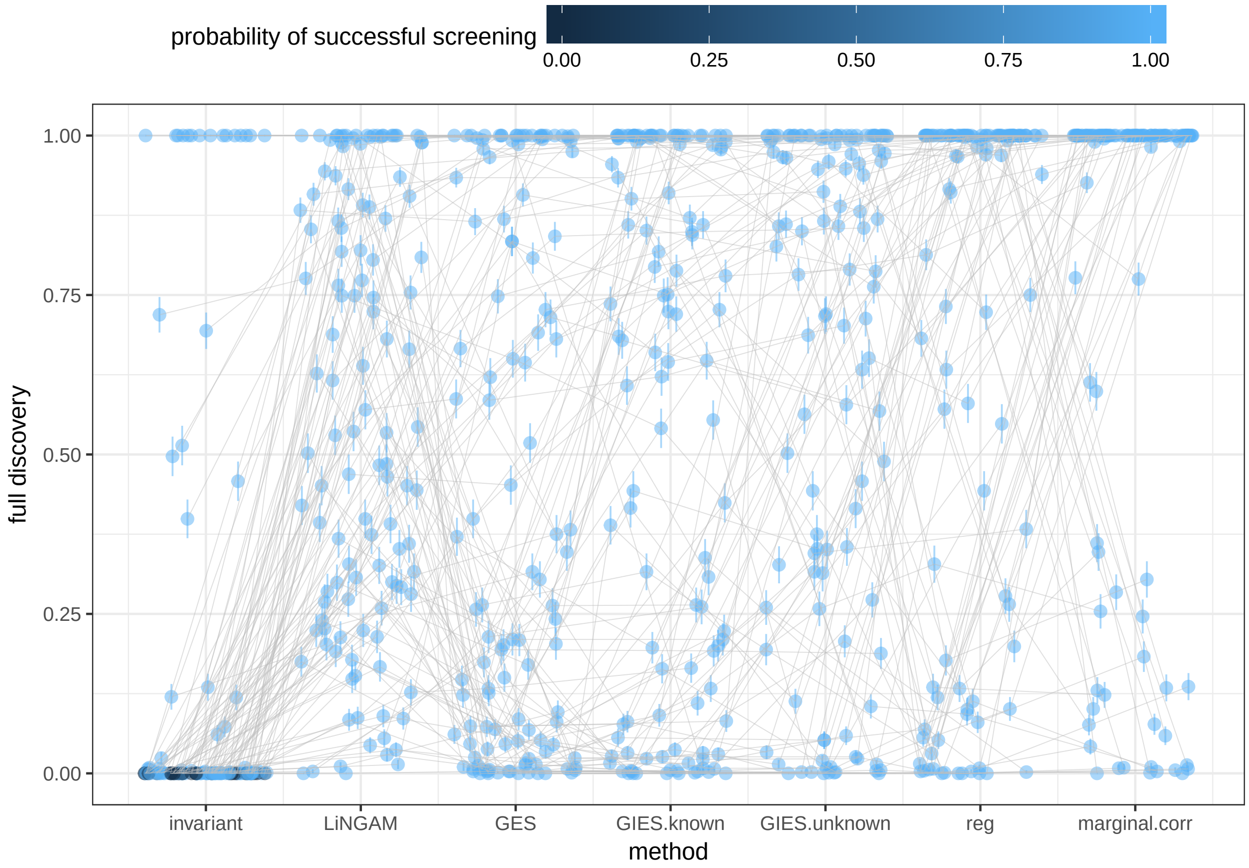

Figure 7: Probability of full discoverey $P(\hat{S} = \mathsf{Pa}(Y))$ of methods under $\alpha = 0.05$. Each setting is replicated $1,000$ times to obtain the error bars (95% confidence interval)

3.2.1 Separation into Environments and Validation Scheme

Since we have data from one observational condition and $1,479$ different interventional conditions, we have a large number of environments. While it is appealing for strengthening power, testing conditional invariance is difficult since each interventional condition only has one measurement. Instead, we define two environments: observational and interventional, the latter of which pools data from different single-gene knockout experiments.

Furthermore, we want to reserve some data from training for *validating* the causal variables found. We partition interventional data into 3 chunks: (i) when the target j is chosen as one of the 1,479 genes, the chunk containing the intervention on the target is reserved for validation; (ii) when the target j is chosen outside the 1,479 genes, we reserve one chunk in a cross-validation fashion. For a causal relation $i \to j$ found, we can only validate when the

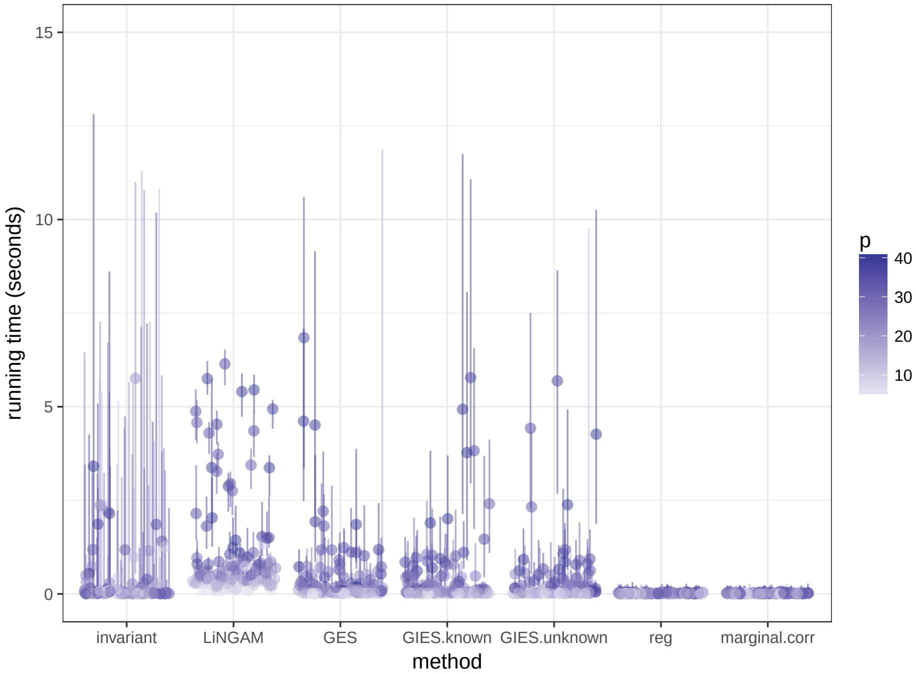

Figure 8: The running time (in seconds) for different methods across settings (error bar denotes maximum and minimum time).

intervention on i happens to be available in the validation set. Since we do not have interventional data available for every gene, we can only validate around 7% of causal relations found. We define the postulated causal relation $i \to j$ as validated if the expression level of gene j after removing gene i is at the 1% upper of lower tail of its observational distribution. The upper or lower tail depends on whether gene i enhances or inhibits gene j.

3.2.2 Methods

We apply invariant prediction method at level $\alpha = 0.01$. Due to $p > n$, we use the high dimensional least square projection [Wang and Leng, 2016] algorithm (`screen = "HOLP"` in the package) for screening out 10 variables. Since the linear model is often misspecified, we want to only keep those where linear models seems a good fit. To this end, we only keep findings where at least one linear model fitted has a p-value larger than 0.1 (see Fig. 9 for why this is a reasonable choice).

For comparison, we use the IDA algorithm [Maathuis et al., 2009], one that is based on parallelized PC algorithm from R package `ParallelPC` (see also Le et al. [2015], and Spirtes et al. [2000]), and ranking based on the magnitude of marginal correlation ($i \to j$ and $j \to i$ are ranked randomly), both of which only use the observational data. We also compare with IDA based on GIES (with known intervention locations, see also Section 3.1.2) and marginal correlation ranking that pools observational and interventional data. Besides, we compare with random guessing that uses no data.

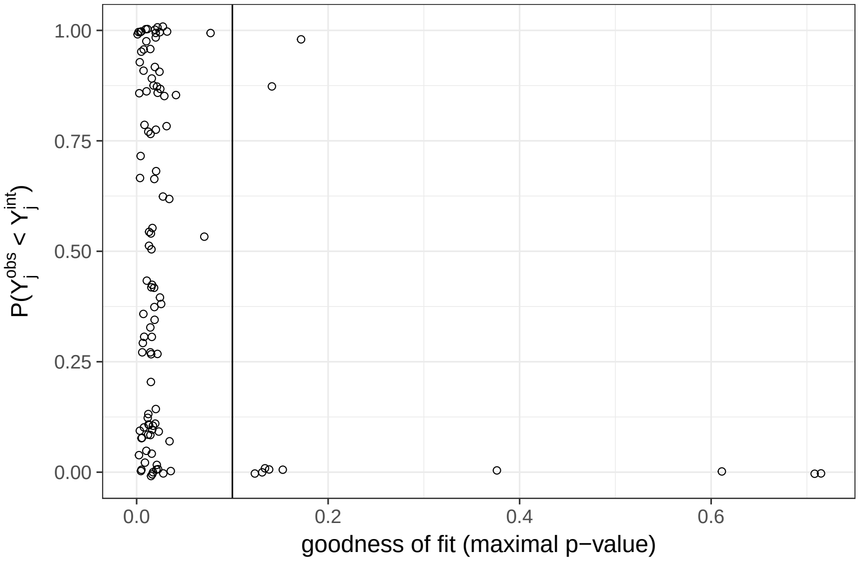

Figure 9: All the causal relations $i \to j$ found by invariant prediction method that can be validated in the yeast gene dataset (x-axis: maximum fitted p-value as a measure of goodness of fit, y-axis: quantile of the expression level of gene j after removing gene i with respect to the observation distribution of gene j's expression level). The vertical line marks 0.1 as a threshold for goodness of fit.

3.2.3 Results

The result is summarized in Table 2. The invariant prediction method found 150 causal relations on 71 target genes, but only 11 of them can be validated. Out of the 11 findings, 9 of them are true positives (82% true discovery rate). Peters, Bühlmann, and Meinshausen [2016] report 8 findings and 6 of them are true positives; we discover more findings since we use our package `InvariantCausal` and can avoid some ad-hoc approximations. The expression levels of two false positives and an example of true positive are plotted in Fig. 10.

For comparison, we look at top-11 ranked findings from other methods that can be validated. Based on the limited validations we have, Table 2 shows that the invariant prediction method produces more true positives compared to other methods.

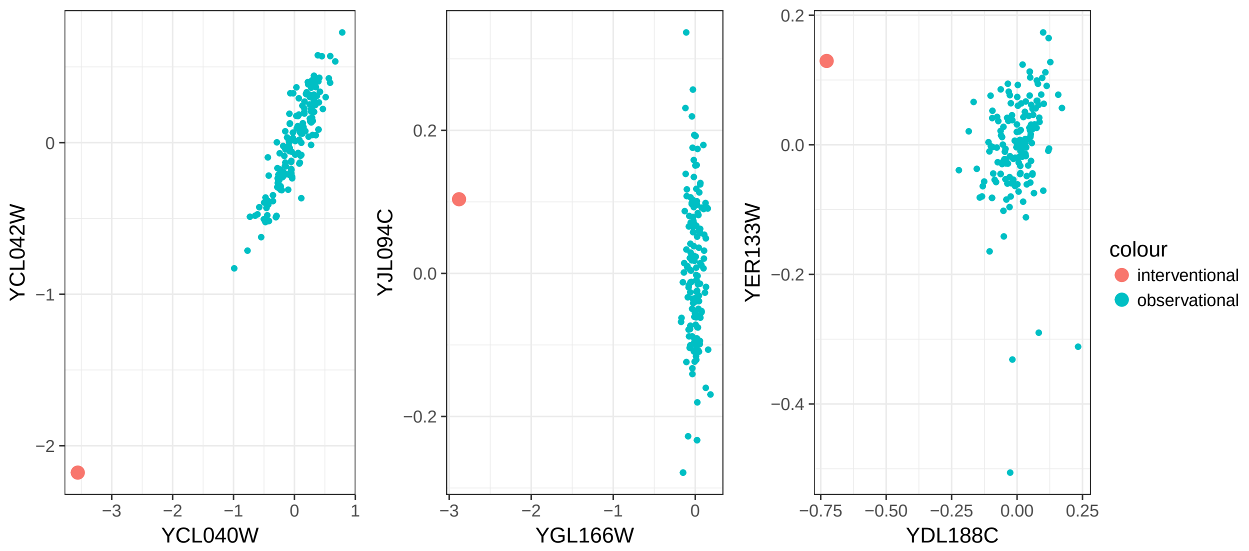

Figure 10: Observational and interventional (knockout of gene i) expression levels of a pair of genes i (x-axis) and j (y-axis), corresponding to verifiable causal relation $i \rightarrow j$ found by invariant prediction (left: a true positive; mid and right: the two false positive findings).

Table 2: Number of direct causal relations found and validated from yeast gene dataset. The total number of findings is set to match that of invariant prediction.

	invariant	IDA	IDA	Corr.	Corr.	random
		(PC)	(GIES)	(obs.)	(pooled)	guess
Number of validated true positive findings (out of 11)	9	3	3	2	2	1
99% confidence interval	—	—	—	$[0, 5]$	$[0, 4]$	$[0, 3]$
Total number of findings	150	—	278	—	—	—

3.3 Educational Attainment

We extend the method to logistic regression and apply it to an educational attainment dataset [Rouse, 1995] using instrumental variable. The dataset consists of 4,739 students from 1,100 US high schools and 13 attributes are recorded, including their test scores, family backgrounds (e.g., whether parents attended college and family income), and demographic attributes (e.g., gender and ethnicity). We want to estimate the causal effect of these attributes on the probability of obtaining a Bachelor degree, via a logistic regression model. Additionally, the distance from home to the nearest college is recorded for each student and used as an instrumental variable. The distance to nearest college is postulated to (i) have no direct effect on the outcome and (ii) exogenous (not a descendent of the outcome variable). Therefore, we can use the instrumental variable to divide data into environments. For this case, we use data provided by Stock and Watson [2012] and divide the dataset into two environments by comparing the distances to the population median.

For a level-α test of invariance, we test equal mean and variance of residuals $(Y_i - \hat{p}(Y_i = 1))/\sqrt{\hat{p}(Y_i = 1)(1 - \hat{p}(Y_i = 1))}$ for environments e and its complement $\mathcal{E} \setminus \{e\}$, where the predicted probability comes from fitting logistic regression on data pooled from all environments (see paragraph "non-Gaussian noise" under Section 2.2.3). Again, we use Bonferroni correction to obtain a combined p-value. We use Sukhatme-Fisher test [Sukhatme, 1935, Perng and Littell, 1976] for a two sample test of equal mean and variance, and this is implemented as `method="logistic-SF"` in package `InvariantCausal`. We run the algorithm with level $\alpha = 0.01$. Out of 8,192 subsets of $p = 13$ variables, 1,317 are accepted and only one variable 'score' is found to be a direct cause. In Figure 11, we show aggregated (blue, see Eq. (11)) and individual (orange, obtained with option `iterate_all=true`) confidence intervals of regression coefficients, and only the aggregated confidence interval for variable 'score' does not contain zero.

4 Discussion

To conclude, we discuss the strengths and weaknesses of the invariant prediction method and prospective future work.

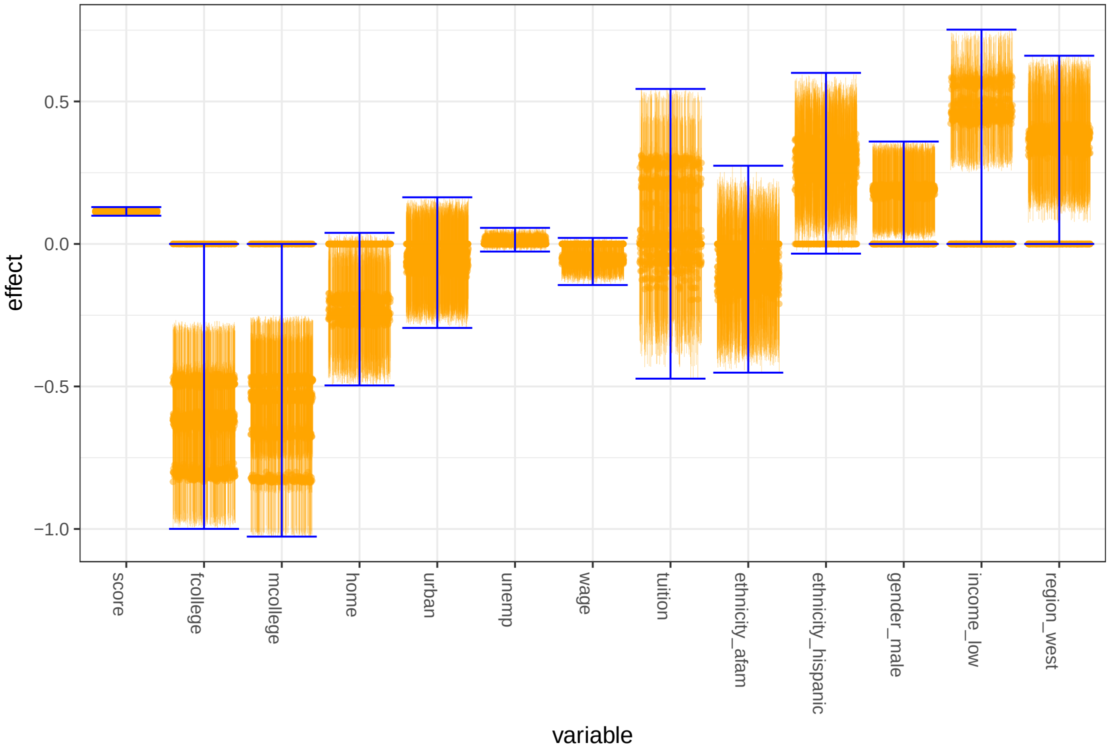

Figure 11: Confidence intervals (blue) aggregated (by Eq. (11)) from 1,317 accepted subsets (orange) (8,192 subsets in total) from logistic regression model on the educational attainment dataset. Distance from home to nearest college is treated as an instrumental variable and is used to divide data into $|\mathcal{E}| = 2$ environments.

Strengths Firstly, invariant prediction is a flexible framework. It uses both observational and interventional data, and contrary to previous methods, it incorporates a wide range of interventions and does not require specifying the location of interventions. Secondly, it is designed to guard against false discoveries and is well-suited for applications where one wants to infer an underlying "sparse" causal network, as commonly seen in biomedical sciences.

Weaknesses The computational complexity scales exponentially with p and hence, under moderate or large p, a screening procedure is nececessary to reduce p to around 10 for a reasonable running time. This is only valid when the underlying causal graph is "sparse"; otherwise the method faces limitation when every variable has a direct causal effect on almost every other variable (e.g., in some datasets from social sciences). Even when the causal graph is sparse, the screening algorithm does not always cover all the direct causes, and all the guarantees will be lost upon such failures (as we have seen in Fig. 5). Essentially, the method provides guarantee only when one is confident that all the direct causes are observed and included in a small set of variables.

Nevertheless, it is worth mentioning that the causal sufficiency assumption (see Assumption 1) is *not entirely untestable* this framework (nor is it entirely testable). When it does not hold, $H_{0,S}$ might be rejected for *every* $S \in \{1, \cdots, p\}$, and then the software will alert the user to *model rejection*. And yet, model rejection can also occur when the linear model in Eq. (13) is misspecified or the target is intervened.

Future work It is interesting to study how to perform causal discovery sequentially. For example, in the gene knockout setting, we would like the algorithm to return "key-clue" genes, by knocking out which we can glean the most information for uncovering the underlying causal graph. In other words, we would want to perform causal discovery and experimental design interactively. However, it might be difficult to develop such a procedure since a *post-selection inference* problem surfaces, as we want to construct environments *post-hoc* based on earlier inference.

References

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

Kenneth Bollen. Structural Equations with Latent Variables. *New York: Wiley*, 1983.

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

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Gregory C Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, pages 591–605, 1960.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13, 2012.

Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(Nov):2523–2547, 2008.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H Maathuis, Peter Bühlmann, et al. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

Patrick Kemmeren, Katrin Sameith, Loes AL van de Pasch, Joris J Benschop, Tineke L Lenstra, Thanasis Margaritis, Eoghan O'Duibhir, Eva Apweiler, Sake van Wageningen, Cheuk W Ko, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157(3):740–752, 2014.

Steffen L Lauritzen. Graphical models, 1996.

Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–224, 1988.

Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, and Shu Hu. Parallelpc: an r package for efficient constraint based causal exploration. *arXiv preprint arXiv:1510.03042*, 2015.

Marloes H Maathuis, Markus Kalisch, Peter Bühlmann, et al. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

JL Marchini, C Heaton, and BD Ripley. fastica: Fastica algorithms to perform ica and projection pursuit. *R package version*, 1(0), 2013.

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.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

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

SK Perng and Ramon C Littell. A test of equality of two normal population means and variances. *Journal of the American Statistical Association*, 71(356):968–971, 1976.

Jonas Peters. Script on Causality. `http://www.math.ku.dk/~peters/jonas_files/scriptChapter1-4.pdf`, 2015.

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.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128, 2013.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

Cecilia Elena Rouse. Democratization or diversion? the effect of community colleges on educational attainment. *Journal of Business & Economic Statistics*, 13(2):217–224, 1995.

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

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

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.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

James H Stock and Mark W Watson. *Introduction to econometrics: Global edition*. Pearson Education Boston, MA, 2012.

PV Sukhatme. A contribution to the problem of two samples. In *Proceedings of the Indian Academy of Sciences-Section A*, volume 2, pages 584–604. Springer, 1935.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Tyler J VanderWeele and James M Robins. Signed directed acyclic graphs for causal inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1): 111–127, 2010.

Xiangyu Wang and Chenlei Leng. High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):589–611, 2016.