Causal inference using invariant prediction: identification and confidence intervals

Jonas Peters

MPI for Intelligent Systems, Tübingen, Germany
Seminar für Statistik, ETH Zürich, Switzerland

Peter Bühlmann

Seminar für Statistik, ETH Zürich, Switzerland

Nicolai Meinshausen

Seminar für Statistik, ETH Zürich, Switzerland

Summary. What is the difference of a prediction that is made with a causal model and a non-causal model? Suppose we intervene on the predictor variables or change the whole environment. The predictions from a causal model will in general work as well under interventions as for observational data. In contrast, predictions from a non-causal model can potentially be very wrong if we actively intervene on variables. Here, we propose to exploit this invariance of a prediction under a causal model for causal inference: given different experimental settings (for example various interventions) we collect all models that do show invariance in their predictive accuracy across settings and interventions. The causal model will be a member of this set of models with high probability. This approach yields valid confidence intervals for the causal relationships in quite general scenarios. We examine the example of structural equation models in more detail and provide sufficient assumptions under which the set of causal predictors becomes identifiable. We further investigate robustness properties of our approach under model misspecification and discuss possible extensions. The empirical properties are studied for various data sets, including large-scale gene perturbation experiments.

Keywords: causal inference; causal discovery; invariant prediction; confidence intervals

1. Introduction

Inferring cause-effect relationships between variables is a primary goal in many applications. Such causal inference has its roots in different fields and various concepts have contributed to its understanding and quantification. Among them are the framework of potential outcomes and counterfactuals [cf. Dawid, 2000, Rubin, 2005]; or structural equation modelling [cf. Bollen, 1989, Robins et al., 2000, Pearl, 2009] and graphical modeling [cf. Lauritzen and Spiegelhalter, 1988, Greenland et al., 1999, Spirtes et al., 2000], where the book by Pearl [2009] provides a nice overview. Richardson and Robins [2013] make a connection between the frameworks using single-world intervention graphs.

A typical approach for causal discovery, in the context of unknown causal structure, is to characterise the Markov equivalence class of structures (or graphs) [Verma and Pearl, 1991, Andersson et al., 1997, Tian and Pearl, 2001, Hauser and Bühlmann, 2012], estimate

the correct Markov equivalence class based on observational or interventional data [Spirtes et al., 2000, Chickering, 2002, Castelo and Kocka, 2003, Kalisch and Bühlmann, 2007, He and Geng., 2008, Hauser and Bühlmann, 2015, cf.], and finally infer the identifiable causal effects or provide some bounds [Maathuis et al., 2009, VanderWeele and Robins, 2010, cf.]. More recently, within the framework of structural equation models, interesting work has been done for fully identifiable structures exploiting additional restrictions such as non-Gaussianity [Shimizu et al., 2006], nonlinearity [Hoyer et al., 2009, Peters et al., 2014] or equal error variances [Peters and Bühlmann, 2014]. Janzing et al. [2012] exploit an independence between causal mechanisms.

We propose here a new method for causal discovery. The approach of the paper is to note that if we consider all "direct causes" of a target variable of interest, then the conditional distribution of the target given the the direct causes will not change when we interfere experimentally with all other variables in the model except the target itself. This does not necessarily hold, however, if some of the direct causes are ignored in the conditioning.† We exploit, in other words, that the conditional distribution of the target variable of interest (often also termed "response variable"), given the complete set of corresponding direct causal predictors, has to remain identical under interventions on variables other than the target variable. This invariance idea is closely linked to causality and has been discussed, for example, under the term "autonomy" and "modularity" [Haavelmo, 1944, Aldrich, 1989, Hoover, 1990, Pearl, 2009, Schölkopf et al., 2012] or also "stability" [Dawid and Didelez, 2010] [Pearl, 2009, Sec. 1.3.2]. While it is well-known that causal models have an invariance property, we try to exploit this fact for inference. Our proposed procedure gathers all submodels that are statistically invariant across environments in a suitable sense. The causal submodel consisting of the set of variables with a direct causal effect on the target variable will be one of these invariant submodels, with controlled high probability, and this allows to control the probability of making false causal discoveries.

Our method is tailored for (but not restricted to) the setting where we have data from different experimental settings or regimes [Didelez et al., 2006]. For example, two different interventional data samples, or a combination of observational and interventional data [cf. He and Geng., 2008] belong to such a scenario. For known intervention targets, Cooper and Yoo [1999] incorporate the intervention effects as mechanism changes [Tian and Pearl, 2001] into a Bayesian framework and Hauser and Bühlmann [2015] modify the greedy equivalence search [Chickering, 2002] for perfect interventions. Our framework does not require to know the location of interventions. For this setting, Eaton and Murphy [2007] use intervention nodes with unknown children and Tian and Pearl [2001] consider changes in marginal distributions, while Dawid [2012, 2015] make use of different regimes for a decision-theoretic approach. In contrast to these approaches, our framework does not require the fitting of graphical, structural equation or potential outcome models and comes with statistical guarantees. Further advantages are indicated below in Section 1.2.

We primarily consider the situation with no hidden (confounder) variables that influence the target variable. A rigorous treatment with hidden variables would be more involved [see Richardson and Spirtes, 2002, for graphical language] but we provide an example with instrumental variables in Section 5 to illustrate that the method could also work more generally in the context of hidden variables. We do not touch very much on

†We thank a referee for suggesting this succinct description of the main idea.

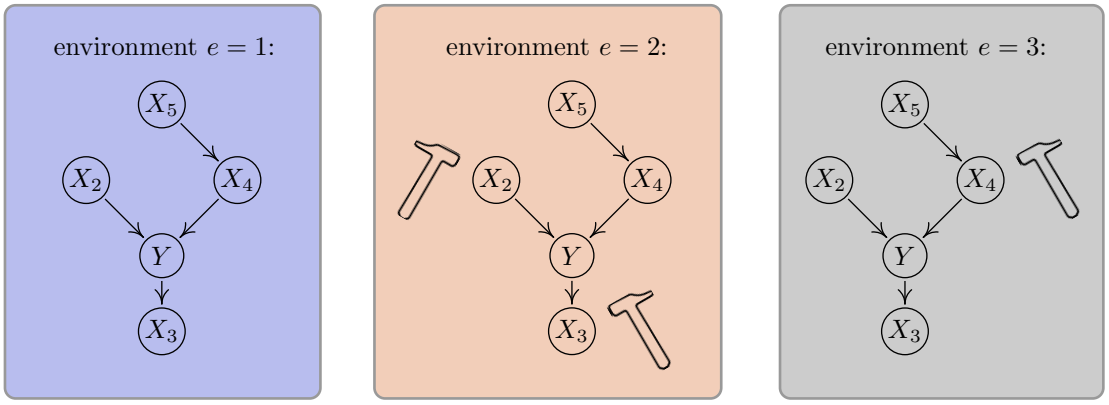

Fig. 1. An example including three environments. The invariance (1) and (2) holds if we consider $S^* = \{X_2, X_4\}$. Considering indirect causes instead of direct ones (e.g. $\{X_2, X_5\}$) or an incomplete set of direct causes (e.g. $\{X_4\}$) may not be sufficient to guarantee invariant prediction.

the framework of feedback models [Lauritzen and Richardson, 2002, Mooij et al., 2011, Hyttinen et al., 2012, cf.], although a constrained form of feedback is allowed. It is an open question whether our approach could be generalised to include general feedback models.

1.1. Data from multiple environments or experimental settings

We consider the setting where we have different experimental conditions $e \in \mathcal{E}$ and have an i.i.d. sample of (X^e, Y^e) in each environment, where $X^e \in \mathbb{R}^p$ is a predictor variable and $Y^e \in \mathbb{R}$ a target variable of interest. While the environments $e \in \mathcal{E}$ can be created by precise experimental design for X^e (for example by randomising some or all elements of X^e), we are more interested in settings where such careful experimentation is not possible and the different distributions of X^e in the environments are generated by unknown and not precisely controlled interventions. If a subset $S^* \subseteq \{1, \ldots, p\}$ is causal for the prediction of a response Y, we assume that

$$\text{for all } e \in \mathcal{E}: \quad X^e \text{ has an arbitrary distribution and} \tag{1}$$

$$Y^e = g(X^e_{S^*}, \varepsilon^e), \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}, \tag{2}$$

where $g : \mathbb{R}^{|S^*|} \times \mathbb{R} \to \mathbb{R}$ is a real-valued function in a suitable function class, $X^e_{S^*}$ is the vector of predictors X^e with indices in a set S^* and both the error distribution $\varepsilon^e \sim F_\varepsilon$ and the function g are assumed to be the same for all the experimental settings. Equations (1) and (2) can also be interpreted as requiring that the conditionals $Y^e \,|\, X^e_{S^*}$ and $Y^f \,|\, X^f_{S^*}$ are identical for all environments $e, f \in \mathcal{E}$ (this equivalence is proved in Section 6.1).

An example of a set of environments can be seen in Figure 1. The invariance (1) and (2) holds if the set S^* consists of all direct causes of the target variable Y and if we do not intervene on Y, see Proposition 1.

Sections 5, 6.2 and 6.3 discuss violations and possible relaxations of this assumption.

1.2. New contribution

The main and novel idea is that we can use the invariance of the causal relationships under different settings $e \in \mathcal{E}$ for statistical estimation, which opens a new road for causal discovery and inference.

For the sake of simplicity, we will mostly focus on a linear model with a target or response variable and various predictor variables, where Equation (1) is unchanged and (2) then reads $Y^e = \mu + X^e\gamma^* + \varepsilon^e$, with μ a constant intercept term. The set S^* of predictors is then given by the support of γ^*, that is $S^* := \{k;\ \gamma_k^* \neq 0\}$. Assumption 1 in Section 2 summarises all requirements. Proposition 1 shows that structural equation models with the traditional notion of interventions [Pearl, 2009] satisfy Assumption 1 if we choose the set S^* to be the parents of Y. Proposition 6 in Appendix D sheds some light on the relationship to potential outcomes.

Obtaining confidence statements for existing causal discovery methods is often difficult as one would need to determine the distribution of causal effects estimators after having searched and estimated a graphical structure of the model. It is unknown how one could do this, except relying on data-splitting strategies which have been found to perform rather poorly in such a setting [Bühlmann et al., 2013]. We propose in Section 3 a new method for the construction of (potentially) conservative confidence statements for causal predictors S^* and of (potentially) conservative intervals for γ_j^* for $j = 1, \ldots, p$ without a-priori knowing or assuming a causal ordering of variables. The method provides confidence intervals without relying on assumptions such as faithfulness or other identifiability assumptions. If a causal effect is not identifiable from the given data, it would automatically detect this fact and not make false causal discoveries.

Another main advantage of our methodology is that we do not need to know how the experimental conditions arise or which type of interventions they induce. We only assume that the intervention does not change the conditional distribution of the target given the causal predictors (no intervention on the target or a hidden confounder): it is simply a device exploiting the grouping of data into blocks, where every block corresponds to an experimental condition $e \in \mathcal{E}$. We will show in Section 3.2 that such grouping can be misspecified and the coverage statements are still correct. This is again a major bonus in practice as it is often difficult to specify what an intervention or change of environment actually means. In contrast, for a so-called do-intervention for structural equation models [Pearl, 2009] it needs to be specified on which variables it acts. Interesting areas of applications include studies where observational data alone are not sufficient to infer causal effects but randomised studies are infeasible to conduct.

We believe that the method's underlying invariance principle is rather general. However, for simplicity, we present our main results for linear Gaussian models, including some settings with instrumental variables and hidden variables.

1.3. Organization

The invariance assumption is formulated and discussed in Section 2. Using this invariance assumption, a general way to construct confidence statements for causal predictors and associated coefficients is derived in Section 3. Two specific methods are shown, using regression effects for various sets of predictors as the main ingredient. Identifiability results for structural equation models are given in Section 4. The relation to instrumental variables and the behaviour in presence of hidden variables is discussed in Section 5. We will discuss extensions to the nonlinear model (2) in Section 6.1 and extenstions to intervened targets in Section 6.2. Some robustness property against model misspecifications is discussed in

Section 6.3.

Simulations and applications to a biological gene perturbation data set and an educational study related to instrumental variables are presented in Section 7. We discuss the results and provide an outlook in Section 8.

1.4. Software

The methods are available in the package `InvariantCausalPrediction` for the `R`-language [R Core Team, 2014].

2. Assumed invariance of causal prediction

We formulate here the invariance assumption and discuss the notion of identifiable causal predictors. Let $\mathcal{E}$ denote again the index set of $|\mathcal{E}|$ possible interventional or experimental settings. As stated above, we have variables (X^e, Y^e) with a joint distribution that will in general depend on the environment $e \in \mathcal{E}$. In the simplest case, $|\mathcal{E}| = 2$, and we have for example in the first setting observational data and interventions of some (possibly unknown) nature in the second setting.

Our discussion will rest on the following assumption. We assume the existence of a model that is invariant under different experimental or intervention settings. Let for any set $S \subseteq \{1, \ldots, p\}$, X_S be the vector containing all variables $X_k, k \in S$.

Assumption 1 (Invariant prediction). *There exists a vector of coefficients* $\gamma^* = (\gamma_1^*, \ldots, \gamma_p^*)^t$ *with support* $S^* := \{k : \gamma_k^* \neq 0\} \subseteq \{1, \ldots, p\}$ *that satisfies*

$$\begin{aligned} \text{for all } e \in \mathcal{E}: \quad & X^e \text{ has an arbitrary distribution and} \\ & Y^e = \mu + X^e\gamma^* + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}, \end{aligned} \tag{3}$$

where $\mu \in \mathbb{R}$ *is an intercept term,* ε^e *is random noise with mean zero, finite variance and the same distribution* F_ε *across all* $e \in \mathcal{E}$.

The distribution F_ε is not assumed to be known in general. If not mentioned otherwise, we will always assume that an intercept μ is added to the model (3). To simplify notation, we will from now on refrain from writing the intercept down explicitly. We discuss the invariance assumption with the help of some examples in Figure 1 and 2; see also Appendix A for another artificial example.

We observe each unit i in only one experimental setting. The distribution of the error ε^e is assumed to stay identical across all environments (though see Sections 6.2 and 6.3 for approaches when this assumption is violated). It is in general not possible to estimate the correlation between the noise variables ε_i^e and ε_i^f for a single unit i in different hypothetical environments e and f, as the outcome is observed for only one environment [Dawid, 2006, 2012]. Knowledge of the correlation would be necessary to answer counterfactual questions about the outcome. Knowledge of the correlation is not necessary for our method.

We deliberately avoid the term "causality" in Assumption 1 in order to keep it purely mathematical. Proposition 1 establishes a link to causality by showing that the parents of Y in a structural equation model (SEM) satisfy Assumption 1. In other words, the variables that have a direct causal effect on Y in a SEM form a set S^* for which Assumption 1 is satisfied. This must not necessarily be true for the variables that have an

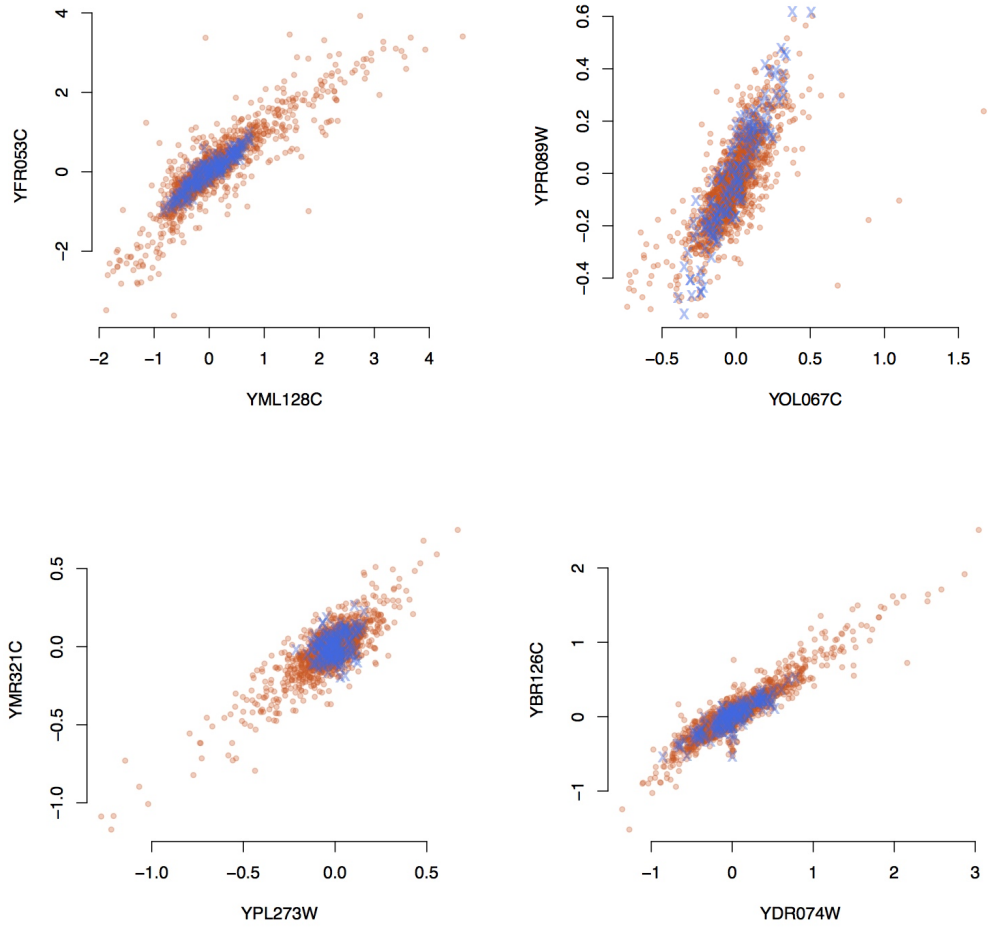

Fig. 2. Some examples from the gene-knockout experiments in Kemmeren et al. [2014], which will be discussed in more detail in Section 7.2. Each panel shows the distribution of a target gene activity Y (on the respective y-axis), conditional on a predictor gene activity X (shown on respective x-axis). Blue crosses show observational data and red dots show interventional data. The interventions do not occur on any of the shown genes. The conditional distribution of Y, given X, is not invariant for the examples in the first row, while invariance cannot be rejected for the two examples in the bottom row. Take the example of the bottom left panel. The variance of the activity of gene $YMR321C$ is clearly higher for interventional than observational data, so we can reject that the invariance assumption holds for the empty set $S = \emptyset$. However, if conditioning on the activity X of gene $YPL273W$, the conditional distribution of the activity Y of gene $YMR321C$ is not significantly different between interventional and observational data, so that the set $S = \{YPL273W\}$ fulfils the invariance assumption (3), at least approximately.

(in)direct effect on Y, i.e., the ancestors of Y. However, the set S^* is not necessarily unique. For a given set of experimental conditions $\mathcal{E}$, there can be multiple vectors γ^* that satisfy (3). For example, if only observational data are available, i.e. all environments are identical, it is apparent that for any model (3) the distribution F_ε of the residuals ε^e does not depend on e. If additionally (X, Y) have a joint Gaussian distribution and X and Y are not independent, for example, then one can find a solution γ^* to (3) for every subset $S^* \subseteq \{1, \ldots, p\}$. The inference we propose works for any possible choice among the set of solutions. We can at most identify the subset of S^* that is common among all possible solutions of (3), see Section 4 for settings with complete identifiability.

It is perhaps easiest to think about the example of a linear structural equation model (SEM), as defined in Section 4.1, see also Figure 8 in Appendix A. We show in the following proposition that the set of parents of Y in a linear SEM is a valid set S^* satisfying (3).

PROPOSITION 1. *Consider a linear structural equation model, as formally defined in Section 4.1, for the variables $(X_1 = Y, X_2, \ldots, X_p, X_{p+1})$, with coefficients $(\beta_{jk})_{j,k=1,\ldots,p+1}$, whose structure is given by a directed acyclic graph. The independence assumption on the noise variables in Section 4.1 can here be replaced by the strictly weaker assumption that $\varepsilon_1^e \perp\!\!\!\perp \{\varepsilon_j^e; j \in \mathbf{AN}(1)\}$ for all environments $e \in \mathcal{E}$, where $\mathbf{AN}(1)$ are the ancestors of Y. Then Assumption 1 holds for the parents of Y, namely $S^* = \mathbf{PA}(1)$, and $\gamma^* = \beta_{1,\cdot}$ as defined in Section 4.1, under the following assumption:*

> *for each $e \in \mathcal{E}$: the experimental setting e arises by one or several interventions on variables from $\{X_2, \ldots, X_{p+1}\}$ but interventions on Y are not allowed; here, we allow for do-interventions [Pearl, 2009] (see also Section 4.2.1, and note that the assigned values can be random, too), or soft-interventions [Eberhardt and Scheines, 2007] (see also Sections 4.2.2 and 4.2.3).*

PROOF. It follows by the definition of the interventions in Section 4.2 and because the interventions do not act on the target variable Y, that $Y^e = \sum_{j \in \mathbf{PA}(1)} \beta_{1,j} X_j^e + \varepsilon_Y^e$ for all $e \in \mathcal{E}$, where $\varepsilon_Y^e = \varepsilon_1^e$ is independent of $X_{\mathbf{PA}(1)}$ and has the same distribution for all $e \in \mathcal{E}$. Thus, Assumption 1 holds.

We remark that Proposition 1 can be generalised to include some hidden variables: the exact statement is given in Proposition 4 in Appendix B.

Instead of allowing only do- or soft-interventions in Proposition 1, we can allow for more general interventions which could change the structural equations for $X_2, \ldots, X_{p+1}$ (including for example a change in the graphical structure of the model among the variables $X_2, \ldots, X_{p+1}$), as long as the conditional distribution of Y^e given $X_{S^*}^e$ remains the same. Such a weaker requirement is sometimes referred to as "modularity" [Pearl, 2009] or what is called "autonomy" [Haavelmo, 1944, Aldrich, 1989]; structural equations are autonomous if whenever we replace one of them due to an intervention, all other structural equations do not change, they remain invariant. The remaining part of the condition in Proposition 1 about excluding interventions on the target variable Y is often verifiable in many applications; see Sections 6.2 and 6.3 for violations of this assumption.

Proposition 1 refers to standard linear SEMs that do not allow for feedback cycles. We may, however, include feedback into the SEM and consider equilibrium solutions of the new set of equations. The independence assumption between ε^e and $X_{S^*}^e$ allows for

some feedback cycles in the linear SEM. The independence assumption prohibits, however, cycles that include the target variable Y. We will leave it as an open question to what extent the approach can be generalised to more general forms of feedback models.

It is noteworthy that our inference is valid for *any* set that satisfies Assumption 1 and not only parents in a linear SEM. For the following statements we do not specify whether the set S^* refers to the set of parents in a linear SEM or any other set that satisfies (3), as the confidence guarantees will be valid in either case. Proposition 6 in Appendix D discusses some relationship to the potential outcome framework.

2.1. Plausible causal predictors and identifiable causal predictors

In general, (γ^*, S^*) is not the only pair that satisfies the assumption of invariance in (3). We therefore define for $\gamma \in \mathbb{R}^p$ and $S \subseteq \{1, \dots, p\}$ the null hypothesis $H_{0,\gamma,S}(\mathcal{E})$ as

$$H_{0,\gamma,S}(\mathcal{E}): \quad \gamma_k = 0 \text{ if } k \notin S \quad \text{and} \quad \begin{cases} \exists F_\varepsilon \text{ such that for all } e \in \mathcal{E} \\ Y^e = X^e\gamma + \varepsilon^e, \text{ where } \varepsilon^e \perp\!\!\!\perp X_S^e \text{ and } \varepsilon^e \sim F_\varepsilon. \end{cases} \tag{4}$$

As stated above, we have dropped the constant intercept notationally. The variables that appear in *any* set S that satisfies $H_{0,S}(\mathcal{E})$, we call plausible causal predictors.

Definition 1 (Plausible causal predictors and coefficients).

(i) We call the variables $S \subseteq \{1, \dots, p\}$ plausible causal predictors *under $\mathcal{E}$ if the following null hypothesis holds true:*

$$H_{0,S}(\mathcal{E}): \quad \exists \gamma \in \mathbb{R}^p \text{ such that } H_{0,\gamma,S}(\mathcal{E}) \text{ is true.} \tag{5}$$

(ii) The identifiable causal predictors *under interventions $\mathcal{E}$ are defined as the following subset of plausible causal predictors*

$$S(\mathcal{E}) := \bigcap_{S\,:\,H_{0,S}(\mathcal{E}) \text{ is true}} S = \bigcap_{\gamma \in \Gamma(\mathcal{E})} \{k : \gamma_k \neq 0\}. \tag{6}$$

Here, $\Gamma(\mathcal{E})$ is defined in (13) below (the second equation in (6) can be ignored for now). Under Assumption 1, $H_{0,\gamma^*,S^*}(\mathcal{E})$ is true and therefore S^* are plausible causal predictors, that is $H_{0,S^*}(\mathcal{E})$ is correct, too. The identifiable causal predictors are thus a subset of the true causal predictors,

$$S(\mathcal{E}) \subseteq S^*.$$

This fact will guarantee the coverage properties of the estimators we define below. Furthermore, the set of identifiable causal predictors under interventions $\mathcal{E}$ is growing monotonically if we enlarge the set $\mathcal{E}$,

$$S(\mathcal{E}_1) \subseteq S(\mathcal{E}_2) \quad \text{for two sets of environments } \mathcal{E}_1, \mathcal{E}_2 \text{ with } \quad \mathcal{E}_1 \subseteq \mathcal{E}_2.$$

In particular, if $|\mathcal{E}| = 1$ (for example, there is only observational data), then $S(\mathcal{E}) = \emptyset$ because $H_{0,\emptyset}(\mathcal{E})$ will be true. The set of identifiable causal predictors under a single environment is thus empty and we make no statement as to which variables are causal.

In Section 4, we examine conditions for structural equation models (see Proposition 1) under which $S(\mathcal{E})$ is identical to the parents of Y we thus have complete identifiability of the causal coefficients. In practice, the set $\mathcal{E}$ of experimental settings might often be such that $S(\mathcal{E})$ identifies some but not all parents of Y in a SEM.

2.2. Plausible causal coefficients

We have seen that the null hypothesis (4) $H_{0,\gamma,S}(\mathcal{E})$ is in general not only fulfilled for γ^* and its support S^* but also potentially for other vectors $\gamma \in \mathbb{R}^p$. This is true especially if the experimental settings $\mathcal{E}$ are very similar to each other. If we consider again the extreme example of just a single environment, $|\mathcal{E}| = 1$, and a multivariate Gaussian distribution for (X, Y), we can find for any set $S \subseteq \{1, \ldots, p\}$ a vector γ with support S that fulfills the null hypothesis $H_{0,\gamma,S}(\mathcal{E})$, namely by using the regression coefficient when regressing Y on X_S. If the interventions that produce the environments $\mathcal{E}$ are stronger and we have more of those environments, the set of vectors that fulfill the null becomes smaller. We call vectors that fulfill the null hypothesis plausible causal coefficients.

Definition 2 (Plausible causal coefficients). *We define the set* $\Gamma_S(\mathcal{E})$ *of* plausible causal coefficients for the set $S \subseteq \{1, \ldots, p\}$ *and the global set* $\Gamma(\mathcal{E})$ *of* plausible causal coefficients *under* $\mathcal{E}$ *as*

$$\Gamma_S(\mathcal{E}) := \{\gamma \in \mathbb{R}^p : \ H_{0,\gamma,S}(\mathcal{E}) \text{ is true}\}, \tag{7}$$

$$\Gamma(\mathcal{E}) := \bigcup_{S \subseteq \{1,\ldots,p\}} \Gamma_S(\mathcal{E}). \tag{8}$$

Thus,

$$\Gamma(\mathcal{E}_1) \supseteq \Gamma(\mathcal{E}_2) \quad \text{for two sets of environments } \mathcal{E}_1, \mathcal{E}_2 \text{ with } \quad \mathcal{E}_1 \subseteq \mathcal{E}_2.$$

The global set of plausible causal coefficients $\Gamma(\mathcal{E})$ is, in other words, shrinking as we enlarge the set $\mathcal{E}$ of possible experimental settings.

The null hypothesis $H_{0,S}(\mathcal{E})$ in (5) can be simplified. Writing

$$\beta^{\text{pred},e}(S) := \text{argmin}_{\beta \in \mathbb{R}^p : \beta_k = 0 \text{ if } k \notin S} E(Y^e - X^e\beta)^2 \tag{9}$$

for the least-squares population regression coefficients when regressing the target of interest onto the variables in S in experimental setting $e \in \mathcal{E}$, we obtain the equivalent formulation of the null hypothesis for set $S \subseteq \{1, \ldots, p\}$,

$$H_{0,S}(\mathcal{E}) : \quad \begin{cases} \exists \beta \in \mathbb{R}^p \text{ and } \exists F_\varepsilon \text{ such that for all } e \in \mathcal{E} \text{ we have} \\ \beta^{\text{pred},e}(S) \equiv \beta \text{ and } Y^e = X^e\beta + \varepsilon^e, \text{ where } \varepsilon^e \perp\!\!\!\perp X_S^e \text{ and } \varepsilon^e \sim F_\varepsilon. \end{cases} \tag{10}$$

We conclude that

$$\Gamma_S(\mathcal{E}) = \begin{cases} \emptyset & \text{if } H_{0,S}(\mathcal{E}) \text{ is false} \\ \beta^{\text{pred},e}(S) & \text{otherwise.} \end{cases} \tag{11}$$

In other words, the set of plausible causal coefficients for a set S is either empty or contains only the population regression vector. We will make use of this fact further below in Section 3 when computing empirical estimators.

3. Estimation of identifiable causal predictors

We would like to estimate the set $S(\mathcal{E})$ of identifiable causal predictors (6) when observing the distribution of (X^e, Y^e) under different experimental conditions $e \in \mathcal{E}$. At the same time, we might be interested in obtaining confidence intervals for the linear causal coefficients.

Recall again the definition (5) of the null hypothesis $H_{0,S}(\mathcal{E})$. Suppose for the moment that a statistical test for $H_{0,S}(\mathcal{E})$ with size smaller than a significance level α is available. Then the construction of an estimator $\hat{S}(\mathcal{E})$ and confidence sets $\hat{\Gamma}(\mathcal{E})$ for the causal coefficients can work as follows.

Generic method for invariant prediction

1) For each set $S \subseteq \{1, \ldots, p\}$, test whether $H_{0,S}(\mathcal{E})$ holds at level α (we will discuss later concrete examples).

2) Set $\hat{S}(\mathcal{E})$ as

$$\hat{S}(\mathcal{E}) := \bigcap_{S: H_{0,S}(\mathcal{E}) \text{ not rejected}} S. \tag{12}$$

3) For the confidence sets, define

$$\hat{\Gamma}(\mathcal{E}) := \bigcup_{S \subseteq \{1,\ldots,p\}} \hat{\Gamma}_S(\mathcal{E}), \tag{13}$$

where

$$\hat{\Gamma}_S(\mathcal{E}) := \begin{cases} \emptyset & H_{0,S}(\mathcal{E}) \text{ can be rejected at level } \alpha \\ \hat{C}(S) & \text{otherwise.} \end{cases} \tag{14}$$

Here, $\hat{C}(S)$ is a $(1-\alpha)$-confidence set for the regression vector $\beta^{\text{pred}}(S)$ that is obtained by pooling the data.

As an example, consider again Figure 2. Taking the example in the bottom left panel, we cannot reject $H_{0,S}(\mathcal{E})$ for $S = \{YPL273W\}$. Hence we can see already from this plot that $\hat{S}(\mathcal{E})$ is either empty or that $\hat{S}(\mathcal{E}) = \{YPL273W\}$. The latter case happens if no further set of variables is accepted that does not include the activity of gene *YPL273W* as predictor.

A justification for pooling the data in (14) is given in Section 3.2. (The construction is also valid if the confidence set is based only on data from a single environment, but a confidence set for the pooled data will be smaller in general.) This defines a whole family of estimators and confidence sets as we have flexibility in the test we are using for the null hypothesis (5) and how the confidence interval $\hat{C}(S)$ is constructed.

If the test and pooled confidence interval have the claimed size and coverage probability, we can guarantee coverage of the true causal predictors and the true causal coefficient, as shown below in Theorem 1.

THEOREM 1. *Assume that the estimator $\hat{S}(\mathcal{E})$ is constructed according to* (12) *with a valid test for $H_{0,S}(\mathcal{E})$ for all sets $S \subseteq \{1, \ldots, p\}$ at level α in the sense that for all S, $\sup_{P\,:\,H_{0,S}(\mathcal{E})\ true} P[H_{0,S}(\mathcal{E})\ rejected] \leq \alpha$. Consider now a distribution P over (Y, X) and consider any γ^* and S^* such that Assumption 1 holds. Then, $\hat{S}(\mathcal{E})$ satisfies*

$$P\big[\hat{S}(\mathcal{E}) \subseteq S^*\big] \geq 1 - \alpha.$$

If, moreover, for all (γ, S) that satisfy Assumption 1, the confidence set $\hat{C}(S)$ in (14) *satisfies $P[\gamma \in \hat{C}(S)] \geq 1 - \alpha$ then the set $\hat{\Gamma}(\mathcal{E})$* (13) *has coverage at least level $1 - 2\alpha$:*

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

PROOF. The first property follows immediately since

$$P\big[\hat{S}(\mathcal{E}) \subseteq S^*\big] \;=\; P\Big[\bigcap_{S:H_{0,S}(\mathcal{E}) \text{ not rejected}} S \subseteq S^*\Big] \;\geq\; P\big[H_{0,S^*}(\mathcal{E}) \text{ not rejected}\big] \;\geq\; 1-\alpha,$$

where the last inequality follows by the assumption that the test for $H_{0,S}$ is valid at level α for all sets $S \subseteq \{1,\dots,p\}$. The second property follows since

$$P\big[\gamma^* \notin \hat{\Gamma}(\mathcal{E})\big] \;\leq\; P\big[H_{0,S^*}(\mathcal{E}) \text{ rejected or } \gamma^* \notin \hat{C}(S^*)\big] \;\leq\; \alpha + \alpha = 2\alpha.$$

The confidence sets thus have the correct (conservative) coverage. The estimator of the causal predictors will, with probability at least $1-\alpha$, not erroneously include non-causal predictors. Note that the statement is true for any set of experimental or intervention settings. In the worst case, the set $\hat{S}(\mathcal{E})$ might be empty but the error control is valid nonetheless.

Since Theorem 1 holds for any γ^*, S^* which fulfil Assumption 1, and assuming the setting of Proposition 1, we obtain the corresponding confidence statements for the causal coefficients and causal variables in a linear structural equation model, that is for $\gamma^* = \beta_{1,\cdot}$ and $S^* = \mathbf{PA}(1)$ in the notation of Proposition 1.

REMARK 1. *(i) We obtain the following empirical version of* (6)*:*

$$\hat{S}(\mathcal{E}) \;=\; \bigcap_{\gamma \in \hat{\Gamma}(\mathcal{E})} \{k : \gamma_k \neq 0\} \;=\; \bigcap_{S:H_{0,S}(\mathcal{E}) \text{ not rejected at } \alpha} S \tag{15}$$

provided that if $H_{0,S}(\mathcal{E})$ is not rejected, then for all $\gamma \in \hat{\Gamma}_S(\mathcal{E})$ we have $\mathrm{supp}(\gamma) \subseteq S$ *and $H_{0,\mathrm{supp}(\gamma)}(\mathcal{E})$ is not rejected either.*

(ii) In (14)*, we have constructed confidence sets $\hat{\Gamma}_S(\mathcal{E})$ based on a test for $H_{0,S}(\mathcal{E})$. Alternatively, confidence sets $\hat{\Gamma}_S(\mathcal{E})$ may be available that are not based on a test procedure for $H_{0,S}(\mathcal{E})$. In this case, we may take them as a starting point and define $\hat{S}(\mathcal{E})$ using the first equality in* (15)*, instead of* (12)*. Analogously to Theorem 1, the correct coverage property of $\hat{\Gamma}_{S^*}(\mathcal{E})$ then implies confidence statements for $\hat{\Gamma}(\mathcal{E})$ and $\hat{S}(\mathcal{E})$.*

3.1. Two concrete proposals

The missing piece in the generic procedure given by (12) and (13) is a test for $H_{0,S}(\mathcal{E})$ that is valid at level α for any given set of variables $S \subseteq \{1,\dots,p\}$ and thus implies

$$P\big[H_{0,S^*}(\mathcal{E}) \text{ rejected}\big] \leq \alpha.$$

To specify a concrete procedure and derive its statistical properties, we assume throughout the paper that the data consist of n independent observations. Within each experimental setting e, we assume that we receive n_e independent and identically distributed data points from (X^e, Y^e) and thus, $\sum_{e\in\mathcal{E}} n_e = n$.

We now propose a way to construct such a test, but acknowledge that different choices are possible. Our construction will be based on the fact that the causal coefficients are identical to the regression effects in all experimental settings $e \in \mathcal{E}$ if we consider only variables in the set S^* of causal predictors.

For experimental setting $e \in \mathcal{E}$ and a subset S of variables, define the regression coefficients $\beta^{\text{pred},e}(S) \in \mathbb{R}^p$ as above in (9). Define further the population residual standard deviations when regressing Y^e on variables X^e_S as

$$\sigma^e(S) \ := [E(Y^e - X^e \beta^{\text{pred},e}(S))^2]^{1/2}.$$

These definitions are population quantities. The corresponding sample quantities are denoted with a hat. As mentioned above, under Assumption 1, for $S = S^*$, the regression effects are identical to the causal coefficients: for all $e \in E$,

$$\beta^{\text{pred},e}(S^*) \equiv \gamma^* \qquad \text{and} \qquad \sigma^e(S^*) \equiv \text{Var}(F_\varepsilon)^{1/2}.$$

To get a test valid at level α for all subsets S of predictor variables, we first weaken $H_{0,S}(\mathcal{E})$ in (10) to

$$\tilde{H}_{0,S}(\mathcal{E}): \quad \exists (\beta, \sigma) \in \mathbb{R}^p \times \mathbb{R}_+ \text{ such that } \beta^{\text{pred},e}(S) \equiv \beta \text{ and } \sigma^e(S) \equiv \sigma \text{ for all } e \in \mathcal{E}. \quad (16)$$

The null hypothesis $\tilde{H}_{0,S}(\mathcal{E})$ is true whenever the original null hypothesis (10) is true. As in (14), we set

$$\hat{\Gamma}_S(\mathcal{E}) := \begin{cases} \emptyset & \tilde{H}_{0,S}(\mathcal{E}) \text{ can be rejected at level } \alpha \\ \hat{C}(S) & \text{otherwise.} \end{cases}$$

We now give a concrete example which we will use in the numerical examples under the assumption of Gaussian errors and that the design matrix $\mathbf{X}_e$ of all n_e samples in experimental setting $e \in \mathcal{E}$ has full rank. (We write the design matrix in bold letters, as opposed to the random variables X^e.) The whole procedure is then a specific version of the general procedure given further above, where we use a specific test in the first step (the second step is unchanged).

Method I: Invariant prediction using test on regression coefficients

1) For each $S \subseteq \{1, \ldots, p\}$ and $e \in \mathcal{E}$:

   (i) Let I_e with $n_e = |I_e|$ be the set of observations where experimental setting $e \in \mathcal{E}$ was active. Likewise, let $I_{-e} = \{1, \ldots, n\} \setminus I_e$ with $n_{-e} := |I_{-e}|$ be the set of observations when using only observations where experimental setting $e \in \mathcal{E}$ was *not* active. Let $\mathbf{X}_{e,S}$ be the $n_e \times (1 + |S|)$-dimensional matrix when using all samples in I_e and all predictor variables in S, adding an intercept term to the design matrix as mentioned previously. If $S = \emptyset$, the matrix consists only of a single intercept column. Analogously, $\mathbf{X}_{-e,S}$ is defined with the samples in I_{-e}. Let $\hat{Y}_e$ be the predictions for observations in set I_e when using the OLS estimator computed on samples in I_{-e} and let $D := Y_e - \hat{Y}_e$ be the difference between the actual observations Y_e on I_e and the predictions.

   (ii) Under Gaussian errors, if (16) is true for a set S, then [Chow, 1960]

$$\frac{D^t \Sigma_D^{-1} D}{\hat{\sigma}^2 \, n_e} \sim \ F(n_e, n_{-e} - |S| - 1), \quad (17)$$

   where $\hat{\sigma}^2$ is the estimated variance on the set I_{-e} on which the OLS estimator is computed. The covariance matrix Σ_D is given by

$$\Sigma_D = 1_{n_e} + \mathbf{X}_{e,S} (\mathbf{X}_{-e,S}^t \mathbf{X}_{-e,S})^{-1} \mathbf{X}_{e,S}^t,$$

   letting 1_n be the identity matrix in n-dimensions. For any set S, we reject the null hypothesis $\tilde{H}_{0,S}(\mathcal{E})$ if the p-value of (17) is below $\alpha/|\mathcal{E}|$ for any $e \in \mathcal{E}$.

2) As in the generic algorithm, using (12).

3) If we do reject a set S we set $\hat{\Gamma}_S(\mathcal{E}) = \emptyset$. Otherwise, we set $\hat{\Gamma}_S(\mathcal{E})$ to be a $(1-\alpha)$-confidence interval for $\beta^{\text{pred}}(S)$ when using all data simultaneously. For simplicity, we will use a rectangular confidence region where the constraint for $\beta^{\text{pred}}(S)_k$ is identically 0 if $k \notin S$ and for coefficients in S given by $(\hat{\beta}^{\text{pred}}(S))_S \pm t_{1-\alpha/(2|S|),n-|S|-1} \cdot \hat{\sigma}\, \text{diag}((\mathbf{X}_S^t \mathbf{X}_S)^{-1})$, where $\mathbf{X}_S$ is the design matrix of the pooled data when using variables in S, $t_{1-\alpha;q}$ is the $(1-\alpha)$-quantile of a t-distribution with q degrees of freedom, and $\hat{\sigma}^2$ the estimated residual variance.

A justification of the pooling in step 3 is given in Section 3.2. The procedure above has some shortcomings. For example, the inversion of the covariance matrix in (17) might be too slow if we have to search many sets and the sample size is large. One can then just work with a random subsample of the set I_e of size, say, a few hundred, to speed up the computation. It also depends on the assumption of Gaussian errors, although this could be addressed by using rank tests or other nonparametric procedures. Lastly, it is not straightforward to extend this approach to classification and nonlinear models.

We thus provide a second possibility. The fast approximate version below is not fitting a model on each experimental setting separately as in Method I, but is just fitting one global model to all data and comparing the distribution of the residuals in each experimental setting. This is ignoring the sampling variability of the coefficient estimates but leads to a faster procedure.

Method II: Invariant prediction using fast(er) approximate test on residuals

1) For each $S \subseteq \{1, \ldots, p\}$ and $e \in \mathcal{E}$:
   (i) Fit a linear regression model on all data to get an estimate $\hat{\beta}^{\text{pred}}(S)$ of the optimal coefficients using set S of variables for linear prediction in regression. Let $R = Y - X\hat{\beta}^{\text{pred}}(S)$.
   (ii) Test the null hypothesis that the mean of R is identical for each set I_e and $e \in \mathcal{E}$, using a two-sample t-test for residuals in I_e against residuals in I_{-e} and combing via Bonferroni correction across all $e \in \mathcal{E}$. Furthermore, test whether the variances of R are identical in I_e and I_{-e}, using an F-test, and combine again via Bonferroni correction for all $e \in \mathcal{E}$. Combine the two p-values of equal variance and equal mean by taking twice the smaller of the two values. If the p-value for the set S is smaller than α, we reject the set S.
2) As in the generic algorithm, using (12).
3) If we do reject a set S we set $\hat{\Gamma}_S(\mathcal{E}) = \emptyset$. Otherwise, we set $\hat{\Gamma}_S(\mathcal{E})$ to be the conventional $(1-\alpha)$-confidence region for $\beta^{\text{pred}}(S)$ when using all data simultaneously. For simplicity, we will use rectangular confidence regions, exactly as in step 3 of Method I.

Besides a computational advantage, the method can also easily be extended to nonlinear and logistic regression models. For logistic regression, one can test the residuals $R = Y - \hat{f}(X)$ for equal mean across the experimental settings, for example.

3.2. Data pooling

So far, we have assumed that the set $\mathcal{E}$ of experimental settings is given and fixed. An experimental setting $e \in \mathcal{E}$ can for example correspond to

(i) observational data;

(ii) a known intervention of a certain type at a known variable;

(iii) a random intervention at an unknown and random location;

(iv) observational data in a changed environment.

We have used data pooling in Methods I and II to get confidence intervals for the regression coefficients (which is not necessary but increases power in general). A justification of this pooling is in order. The joint distribution of $(X^e_{S^*}, Y^e)$ will vary in general with $e \in \mathcal{E}$. Under Assumption 1, however, the conditional distribution $Y^e \,|\, X^e_{S^*}$ is constant as a function of $e \in \mathcal{E}$, see Section 6.1. As long as our tests and confidence intervals require only an invariant conditional distribution for S^* (which is the case for the procedures given above), we can pool data from various $e \in \mathcal{E}$.

To make it more precise, assume there is a set of countably many experimental settings or interventions $\mathcal{J}$ and (X^j, Y^j) follow a certain distribution F_j for each $j \in \mathcal{J}$. Then each encountered experimental setting e can be considered to be equivalent to a probability mixture distribution over the experimental settings in $\mathcal{J}$, that is

$$F_e = \sum_{j \in \mathcal{J}} w^e_j F_j,$$

where w^e_j corresponds to the probability that an observation under setting e follows the distribution F_j. We can then pool two experimental settings e_1 and e_2, for example, thereby creating a new experimental setting with the averaged weights $(w^{e_1} + w^{e_2})/2$.

Pooling is a trade-off between identifiability and statistical power, assuming that Assumption 1 holds for the settings from $\mathcal{J}$. The richer the set $\mathcal{E}$ of experimental settings, the smaller the set $\Gamma(\mathcal{E})$ of plausible causal coefficients will be and the larger the set of identifiable causal predictors $S(\mathcal{E})$. By pooling data, we make the set of identifiable causal variables smaller, that is $S(\mathcal{E})$ is shrinking as we reduce the number $|\mathcal{E}|$ of different settings. The trade-off can either be settled a-priori (for example if we know that we have "sufficiently" many observations in each known experimental setting, we would typically not pool data) or one can try various pooling procedures and combine all results, after adjusting the level α to account for the increased multiplicity of the associated testing problem. Section 4 discusses conditions on the interventions under which all true causal effects are identifiable.

3.3. Splitting purely observational data

In the case of purely observational data, the null hypothesis (4) is correct for $\gamma = 0$ and $S = \emptyset$. Therefore, $S(\mathcal{E}) = \emptyset$ and $\hat{S}(\mathcal{E}) = \emptyset$ with high probability, i.e., our method stays conservative and does not make any causal claims.

In a reverse operation to data pooling across experiments, the question arises whether we can identify the causal predictors by artificially separating data into several blocks although the data have been generated under only one experimental setting (e.g. the data are purely observational). If the distribution is generated by a SEM (see Section 4.1), we may consider a variable U that is not Y and known to be a non-descendant of the target variable Y, that is, there is no directed path from Y to U, for example as it precedes Y chronologically. (This is similar as in an instrumental variable setting, see Section 5.) We

may now split the data by conditioning on this variable U or any function $h(U)$. Our method then still has the correct coverage for any function $h(U)$ as long as U is a non-descendant of Y, because the conditional distribution of Y given its true causal predictors X_{S^*} does not change and for all z in the image of h,

$$Y \mid X_{S^*} \quad \stackrel{d}{=} \quad Y \mid X_{S^*}, h(U) = z \tag{18}$$

Note that U might or might not be part of the set X_{S^*} but we expect the method to have more power if it is not. Equation (18) is a direct implication of the local Markov property that is satisfied for a SEM [Pearl, 2009, Theorem 1.4.1]. The confidence intervals remain valid but the implication on (partial) identifiability of the causal predictors remains as an open question.

Even without data splitting, there might still be some directional information in the data set that is not exploited by our method; this may either be information in the conditional independence structure [Spirtes et al., 2000, Chickering, 2002], information from non-Gaussianity [Shimizu et al., 2006], nonlinearities [Hoyer et al., 2009, Peters et al., 2014, Bühlmann et al., 2014], equal error variances [Peters and Bühlmann, 2014] or shared information between regression function and target variable [Janzing et al., 2012]. Our method does not exploit these sources of identifiability. We believe, however, that it might be possible to incorporate the identifiability based on non-Gaussianity or nonlinearity.

3.4. Computational requirements

The construction of the confidence regions for the set of plausible causal coefficients and the identifiable causal predictors requires to go through all possible sets of variables in step 1) of the procedures given above. The computational complexity of the brute force scheme seems to grow super-exponentially with the number of variables.

There are several aspects to this issue. Firstly, we often do not have to go through all sets of variables. If we are looking for a non-empty set $\hat{S}(\mathcal{E})$, it is worthwhile in general to start generating the confidence regions $\hat{\Gamma}_S(\mathcal{E})$ for the empty set $S = \emptyset$, then for all singletons and so forth. If the empty set is not rejected, we can stop the search immediately, as then $\hat{S}(\mathcal{E}) = \emptyset$. If the empty set is rejected, we can stop early as soon as we have accepted more than one set S and the sets have an empty overlap (as $\hat{S} = \emptyset$ in this case no matter what other sets are accepted). The method can thus finish quickly if $\hat{S} = \emptyset$. However, in a positive case (where we do hope to get a non-empty confidence set) we will still have to go through all sets of variables eventually. There are two options to address the computational complexity.

The first option is to limit a-priori the size of the set of causal predictors. Say we are willing to make the assumption that the set of causal variables is at most $s < p$. Then we just have to search over all subsets of size at most s and incur a computational complexity that grows like $O(p^s)$ as a function of the number of variables.

A second option (which can be combined with the first one) is an adaptation of the confidence interval defined above, in which the number of variables is first reduced to a subset of small size that contains the causal predictors with high probability. Let $\hat{B} \subseteq \{1, \ldots, p\}$ be, for the pooled data, an estimator of the variables with non-zero regression coefficient when using all variables as predictors. For example, $\hat{B}$ could be the set of

variables with non-zero regression coefficient with square-root Lasso estimation [Belloni et al., 2011], Lasso [Tibshirani, 1996] or boosting [Schapire et al., 1998, Friedman, 2001, Bühlmann and Yu, 2003] with cross-validated penalty parameter. If the initial screening is chosen such that the causal predictors are contained with high probability, $P\big[S^* \subseteq \hat{B}\big] \geq 1 - \alpha$, and we construct the confidence set $\hat{S}(\mathcal{E})$ as above, but just letting S be a subset of $\hat{B}$ instead of $\{1, \ldots, p\}$, it will have coverage at least $1 - 2\alpha$. Sufficient assumptions of such a coverage (or screening) condition are discussed in the literature [e.g. Bühlmann and van de Geer, 2011]. If the second option is combined with the first option, the computational complexity would then scale like $O(q^s)$ instead of $O(p^s)$, where q is the maximal size of the set $\hat{B}$ of selected variables. For the sake of simplicity, we will not develop this argument further here but rather focus on the identifiability results for the low(er)-dimensional case.

4. Identifiability results for structural equation models

The question arises whether the proposed confidence sets for the causal predictors can recover an assumed true set of causal predictors. Such identifiability issues are discussed next. Sections 4.1 and 4.2 describe possible data generating mechanisms and Section 4.3 provides corresponding identifiability results.

4.1. Linear Gaussian SEMs

We consider linear Gaussian structural equation models (SEMs) [e.g. Wright, 1921, Duncan, 1975]. We assume that each element $e \in \mathcal{E}$ represents a different interventional setup. Let the first block of data ($e = 1$) always correspond to an "observational" (linear) Gaussian SEM. Here, a distribution over $(X_1^1, \ldots, X_{p+1}^1)$ is said to be generated from a Gaussian SEM if

$$X_j^1 = \sum_{k \neq j} \beta_{j,k}^1 X_k^1 + \varepsilon_j^1, \qquad j = 1, \ldots, p+1, \tag{19}$$

with $\varepsilon_j^1 \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_j^2)$, $j = 1, \ldots, p+1$. The corresponding directed graph is obtained by drawing arrows from variables X_k^1 on the right-hand side of (19) with $\beta_{jk}^1 \neq 0$ to the variables X_j^1 of the left-hand side. This graph is assumed to be acyclic. Without loss of generality let us assume that $Y^1 := X_1^1$ is the target variable and we write $X := (X_2, \ldots, X_{p+1})$. We further assume that all variables are observed; this assumption can be weakened, see Proposition 4 in Appendix B and Section 5.

The parents of Y are given by

$$\mathbf{PA}(Y) = \mathbf{PA}(1) = \{k \in \{2, \ldots, p+1\} \,:\, \beta_{1,k}^1 \neq 0\}.$$

Here, we adapt the usual notation of graphical models [e.g. Lauritzen, 1996]. For example, we write $\mathbf{PA}(j)$, $\mathbf{DE}(j)$, $\mathbf{AN}(j)$ and $\mathbf{ND}(j)$ for the parents, descendants, ancestors and non-descendants of X_j, respectively.

Let us assume that the other data blocks are generated by a linear SEM, too:

$$X_j^e = \sum_{k \neq j} \beta_{j,k}^e X_k^e + \varepsilon_j^e, \qquad j = 1, \ldots, p+1, \quad e \in \mathcal{E}. \tag{20}$$

Assumption 1 states that the influence of the causal predictors remains the same under interventions, that is $Y^e = X^e\gamma^* + \varepsilon_1^1$ for $\gamma^* = (\beta_{1,2}^1, \ldots, \beta_{1,p+1}^1)^t$ and $\varepsilon_1^e \overset{d}{=} \varepsilon_1^1$ for $e \in \mathcal{E}$. The other coefficients $\beta_{j,k}^e$ and noise variables ε_j^e, $j \neq 1$, however, may be different from the ones in the observational setting (19). Within this setting, we now define various sorts of interventions.

4.2. Interventions

We next discuss three different types of interventions that all lead to identifiability of the causal predictors for the target variable.

4.2.1. Do-interventions

These types of interventions correspond to the classical do-operation from Pearl [2009, e.g.]. In the e-th experiment, we intervene on variables $\mathcal{A}^e \subseteq \{2, \ldots, p+1\}$ and set them to values $a_j^e \in \mathbb{R}$, $j \in \mathcal{A}^e$. For the observational setting $e = 1$, we have $\mathcal{A}^1 = \emptyset$. We specify the model (20), for $e \neq 1$, as follows:

$$\beta_{j,k}^e = \begin{cases} \beta_{j,k}^1 & \text{if } j \notin \mathcal{A}^e \\ 0 & \text{if } j \in \mathcal{A}^e, \end{cases}$$

and

$$\varepsilon_j^e \overset{d}{=} \begin{cases} \varepsilon_j^1 & \text{if } j \notin \mathcal{A}^e \\ a_j^e & \text{if } j \in \mathcal{A}^e. \end{cases}$$

The do-interventions correspond to fixing the intervened variable at a specific value. The following two types of interventions consider "softer" forms of interventions which might be more realistic for certain applications.

4.2.2. Noise interventions

Instead of fixing the intervened variable at a specific value, noise interventions correspond to "disturbing" the variable by changing the distribution of the noise variable. This is an instance of what is sometimes called a "soft intervention" [e.g. Eberhardt and Scheines, 2007]. We now consider a kind of soft intervention, in which we scale the noise distributions of variables $\mathcal{A}^e \subseteq \{2, \ldots, p+1\}$ by a factor A_j^e, $j \in \mathcal{A}^e$. Alternatively, we may also shift the error distribution by a variable C_j^e. More precisely, we specify the model in (20), for $e \neq 1$, as follows:

$$\beta_{j,k}^e = \beta_{j,k}^1 \quad \text{for all } j,$$

and

$$\varepsilon_j^e \overset{d}{=} \begin{cases} \varepsilon_j^1 & \text{if } j \notin \mathcal{A}^e \\ A_j^e \varepsilon_j^1 & \text{if } j \in \mathcal{A}^e, \end{cases} \quad \text{or} \quad \varepsilon_j^e \overset{d}{=} \begin{cases} \varepsilon_j^1 & \text{if } j \notin \mathcal{A}^e \\ \varepsilon_j^1 + C_j^e & \text{if } j \in \mathcal{A}^e. \end{cases}$$

The factors A_j^e and the shifts C_j^e are considered as random but may be constant with probability one. They are assumed to be independent of each other and independent of all other random variables considered in the model except for X_k^e for $k \in \mathbf{DE}(j)$.

4.2.3. Simultaneous noise interventions

The noise interventions above operate on clearly defined variables $\mathcal{A}^e$ which can vary between different experimental settings $e \in \mathcal{E}$. In some applications, it might be difficult to change or influence the noise distribution at a single variable but instead one could imagine interventions that change the noise distributions at many variables simultaneously. As a third example, we thus consider a special case of the preceding Section 4.2.2, in which we pool all interventional experiments into a single data set. That is, $|\mathcal{E}| = 2$ and, for all $j \in \{2, \dots, p+1\}$,

$$\beta_{j,k}^{e=2} = \beta_{j,k}^{e=1} \tag{21}$$

and

$$\varepsilon_j^{e=2} \stackrel{d}{=} A_j \varepsilon_j^{e=1} \quad \text{or} \quad \varepsilon_j^{e=2} \stackrel{d}{=} \varepsilon_j^{e=1} + C_j.$$

The random variables $A_j \geq 0$ are assumed to have a distribution that is absolutely continuous w.r.t. Lebesgue measure with $EA_j^2 < \infty$ and to be independent of all other variables and among themselves. The pooling can either happen explicitly or, as stated above, as we cannot control the target of the interventions precisely and a given change in environment might lead to changes in the error distributions in many variables simultaneously. As an example we mention gene knock-out experiments with off-target effects in biology [e.g. Jackson et al., 2003, Kulkarni et al., 2006].

4.3. Identifiability results

The following Theorem 2 gives sufficient conditions for identifiability of the causal predictors. We then discuss some conditions under which the assumptions can or cannot be relaxed further below. Proofs can be found in Appendix F.

THEOREM 2. *Consider a (linear) Gaussian SEM as in* (19) *and* (20) *with interventions. Then, with* $S(\mathcal{E})$ *as in* (6)*, all causal predictors are identifiable, that is*

$$S(\mathcal{E}) = \mathbf{PA}(Y) = \mathbf{PA}(1) \tag{22}$$

if one of the following three assumptions is satisfied:

i) The interventions are **do-interventions** *(Section 4.2.1) with* $a_j^e \neq E(X_j^1)$ *and there is at least one single intervention on each variable other than* Y*, that is for each* $j \in \{2, \dots, p+1\}$ *there is an experiment* e *with* $\mathcal{A}^e = \{j\}$*.*

ii) The interventions are **noise interventions** *(Section 4.2.2) with* $1 \neq E(A_j^e)^2 < \infty$*, and again, there is at least one single intervention on each variable other than* Y*. If the interventions act additively rather than multiplicatively, we require* $EC_j^e \neq 0$ *or* $0 < \operatorname{Var} C_j^e < \infty$*.*

iii) The interventions are **simultaneous noise interventions** *(Section 4.2.3). This result still holds if we allow changing linear coefficients* $\beta_{j,k}^{e=2} \neq \beta_{j,k}^{e=1}$ *in* (21) *with (possibly random) coefficients* $\beta_{j,k}^{e=2}$*.*

The statements remain correct if we replace the null hypothesis (10) *with its weaker version* (16)*.*

These are examples for sufficient conditions for identifiability but there may be many more. For example, one may also consider random coefficients or changing graph structures (only the parents of Y must remain the same).

Remark. In general, the conditions given above are not necessary. The following remarks, however, provide two specific counter examples that show the necessity of some conditions.

i) We cannot remove the condition $a_j^e \neq E(X_j^1)$ from Theorem 2 i): the following SEMs correspond to observational data in experiment $e = 1$, interventional data with $do(X_2 = 0)$ in experiment $e = 2$, and interventional data with $do(X_3 = 0)$ in experiment $e = 3$:

$$\begin{aligned}
e = 1: &\quad Y^1 = X_2^1 + X_3^1 + \varepsilon_Y, &\quad X_2^1 = \varepsilon_2, &\quad X_3^1 = -X_2^1 + \varepsilon_3,\\
e = 2: &\quad Y^2 = X_2^2 + X_3^2 + \varepsilon_Y, &\quad X_2^2 = 0, &\quad X_3^2 = -X_2^2 + \varepsilon_3,\\
e = 3: &\quad Y^3 = X_2^3 + X_3^3 + \varepsilon_Y, &\quad X_2^3 = \varepsilon_2, &\quad X_3^3 = 0,
\end{aligned}$$

with ε_2 and ε_3 having the same distribution. Then, we cannot identify the correct set of parents $S^* = \{1, 2\}$. The reason is that even $S = \emptyset$ leads to a correct null hypothesis (10).

ii) If we only check the null hypothesis (16) instead of the stronger version (10) (namely whether the residuals have the same variance rather than the same distribution), the condition $E(A_j^e)^2 \neq 1$ is essential. Consider a two-dimensional observational distribution from experiment $e = 1$ and an intervention distribution from experiment $e = 2$:

$$\begin{aligned}
e = 1: &\quad X^1 = \varepsilon_X, &\quad Y^1 = X^1 + \varepsilon_Y,\\
e = 2: &\quad X^2 = A \cdot \varepsilon_X, &\quad Y^2 = X^2 + \varepsilon_Y,
\end{aligned}$$

with $E(A)^2 = 1$ and $\varepsilon_X, \varepsilon_Y \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. Then we cannot identify the correct set of parents $\mathbf{PA}(Y) = \{X\}$ because again $S = \emptyset$ leads to the same residual variance and therefore a correct null hypothesis (16). If we use hypothesis (10), however, condition $E(A_j^e)^2 \neq 1$ can be weakened (if densities exist), see the proof of Theorem 2 (iii).

In practice, we expect stronger identifiability results than Theorem 2. Intuitively, intervening on (some of) the ancestors of Y should be sufficient for identifiability in many cases. Note that the two counter-examples above are non-generic in the way that they violate faithfulness [e.g. Spirtes et al., 2000]. The following theorem shows for some graph structures (which need not to be known) that even one interventional setting with an intervention on a single node may be sufficient, as long as the data generating model is chosen "generically" (see Appendix A for an example).

THEOREM 3. *Assume a linear Gaussian SEM as in* (19) *and* (20) *with all non-zero parameters drawn from a joint density w.r.t. Lebesgue measure. Let X_{k_0} be a youngest parent of target variable $Y = X_1$, that is there is no directed path from X_{k_0} to any other parent of Y. Assume further that there is an edge from any other parent of Y to X_{k_0}. Assume that there is only one intervention setting, where the intervention took place on X_{k_0}, that is $|\mathcal{E}| = 2$ and $\mathcal{A}^{e=2} = \{k_0\}$ (k_0 does not need to be known).*

Then, with probability one, all causal predictors are identifiable, that is

$$S(\mathcal{E}) = \mathbf{PA}(Y) = \mathbf{PA}(1)$$

if one of the following two assumptions is satisfied:

i) The intervention is a **do-intervention** *(Section 4.2.1) with* $a_{k_0}^{e=2} \neq EX_{k_0}^1$.

ii) The intervention is a **noise intervention** *(Section 4.2.2) with* $1 \neq E(A_{k_0}^{e=2})^2 < \infty$ *or* $EC_{k_0}^{e=2} \neq 0$*, respectively.*

It is, of course, also sufficient for identifiability if the interventional setting $\mathcal{A}^{e=2} = \{k_0\}$ is just a member of a larger number of interventional settings. We anticipate that more identifiability results of similar type can be derived in specific settings. Theorem 3 shows that interving on the youngest parent can reveal the whole set of parents of the target variable so this intervention is in a sense the most informative intervention under the made assumptions. Intervening on descendants of Y will, in contrast, only rule out these variables as parents of Y. Some interventions are also completely non-informative; intervening on a variable that is independent of all other variables (including the target) will, for example, not help with identification of the set of parents of the target variable.

5. Instrumental and hidden variables with confounding

We now discuss an extension of the invariance idea that is suitable in the presence of hidden variables. Instrumental variables can sometimes be used when the causal relationship of interest is confounded and there are no randomised experiments available [Wright, 1928, Bowden and Turkington, 1990, Angrist et al., 1996, Didelez et al., 2010]. For simplicity, let us assume that I is binary. We assume that the SEM for a p-dimensional predictor X,

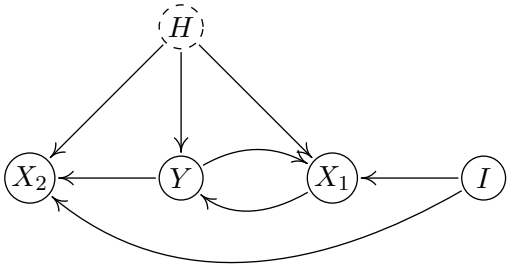

Fig. 3. In this example of a graph of model that satisfies (23), variable Y has a direct causal effect only on X_2, while there is a feedback between Y and X_1.

a univariate target variable Y of interest and a q-dimensional hidden variable H can be written as

$$\begin{aligned} X &= f(I, H, Y, \eta), \\ Y &= X\gamma^* + g(H, \varepsilon), \end{aligned} \tag{23}$$

where γ^* is the unknown vector of causal coefficients, f, g are unknown real-valued functions and η and ε are random noise variables in p dimensions and one dimension respectively. As it is commonly done for SEMs, we require the noise variables H, η, ε, I to be jointly independent. Figure 3 shows an example of a SEM that satisfies (23).

Again, we are interested in the causal coefficient γ^*. Because of the hidden variable H, however, regressing Y on X does not yield a consistent estimator for γ^*.

Two remarks on the model (23) are in place. First, the model requires that I has no direct effect on Y, which is standard assumption for instrumental variable models. For a discussion on why a violation of this assumption usually leads to no false conclusions (only a reduction in power), see Section 6.3. Second, the model (23) allows for feedback between X and Y, that is the corresponding graph in a SEM is not required to be acyclic. If feedback exists, the solutions are typically understood to be stable equilibrium solutions of (23) but we will here only require that the solutions satisfy equations (23).

We can use I as an instrument in a classical sense and estimate γ^* by the following well-known two-stage least squares procedure [Angrist et al., 1996]: first we estimate the influence of I on X and then we regress Y on the predicted values of X given I. For non-linear models one can use two-stage predictor substitution or two-stage residual inclusion; see [Terza et al., 2008] for an overview. If we strive for identification of γ^*, three limitations with this approach are:

(i) The target Y is not allowed to be a parent of any component of X, i.e., $f(I, H, Y, \eta) = f(I, H, \eta)$. This also excludes the possibility of feedback between X and Y.

(ii) The conditional expectation $E(X \mid I)$ is not allowed to be constant for $I \in \{0, 1\}$.

(iii) The predictor X has to be univariate for a univariate instrument I, that is $p = 1$ is required.

What happens if we interpret the two different values of I as two experimental settings? In other words: what happens if I plays the role of the indicator of environment (that we call E at the end of Section 6.1) and we apply the method described above? We can define $\mathcal{E}$ as two distinct environments by collecting all samples with $I = 0$ in the first environment and all samples with $I = 1$ in the second environment. Of course, another split into distinct environments is also possible and allowed as long as the split into distinct environments is not a function of Y, a descendant of Y or the hidden variables H.

We stated in Proposition 1 that SEMs (with interventions) satisfy the assumptions of invariant predictions if there are no hidden variables between the target variable and the causal predictors. Because here there is the hidden variable H we cannot justify our method using Proposition 1 (nor with Proposition 4 in general). However, the invariant prediction procedure (3) can be extended to cover models of the form (23) as these models fulfil

$$\begin{aligned} \text{for all } e \in \mathcal{E}: \quad & X^e \text{ has an arbitrary distribution} \\ & Y^e = X^e \gamma^* + g(H^e, \varepsilon^e), \end{aligned} \tag{24}$$

with unknown causal coefficients $\gamma^* \in \mathbb{R}^p$ and unknown function $g : \mathbb{R}^q \times \mathbb{R} \to \mathbb{R}$.

In the absence of hidden variables, the residuals $Y^e - X^e\gamma^*$ are independent of the causal predictors $X^e_{S^*} = X^e_{\text{supp}(\gamma^*)}$ and have the same distribution across all environments. In the presence of hidden variables, we cannot require independence of the residuals and the causal predictors X_{S*} but can adapt the null hypothesis $H_{0,S}$ in (5) to the weaker form

$$\begin{aligned} H_{0,S,hidden}(\mathcal{E}): \quad & \exists \gamma \in \mathbb{R}^p \text{ such that } \gamma_k = 0 \text{ if } k \notin S \text{ and} \\ & \text{the distribution of } Y^e - X^e\gamma \text{ is identical for all } e \in \mathcal{E}. \end{aligned} \tag{25}$$

Testing the null hypothesis (25) is computationally more challenging than for the corresponding null hypothesis in the absence of hidden confounders (5). In contrast to (5), we

cannot attempt to find for a given set S the vector γ by regressing Y^e on X^e. The reason is that even if (25) holds, it does not require the residuals $Y^e - X^e\gamma$ to be independent of $X^e_{\text{supp}(\gamma)}$.

Suppose nevertheless that we have a test for the null hypothesis $H_{0,S,hidden}(\mathcal{E})$ and define in analogy to (12) the estimated set of causal predictors as

$$\hat{S}(\mathcal{E}) = \bigcap_{S:H_{0,S,hidden}(\mathcal{E}) \text{ not rejected}} S. \tag{26}$$

Then the coverage property follows immediately in the following sense.

PROPOSITION 2. *Consider model* (23) *and let* $S^* = \{k : \gamma^*_k \neq 0\}$. *Suppose the test for* $H_{0,S,hidden}(\mathcal{E})$ *is conducted at level* α *and* $\hat{S}$ *is defined as in* (26). *Then*

$$P[\hat{S}(\mathcal{E}) \subseteq S^*] \;\geq\; 1 - \alpha.$$

PROOF. The hypothesis $H_{0,S,hidden}(\mathcal{E})$ is obviously true for S^* as $Y^e - X^e\gamma^* = g(H^e, \varepsilon^e)$ and the distribution of $g(H^e, \varepsilon^e)$ is invariant across the environments $e \in \mathcal{E}$ (defined by I) as I is independent of H and ε.

The method has thus guaranteed coverage for model (23) even if the necessary assumptions (i)-(iii) for identification under a two-stage instrumental-variable approach are violated. The power of the procedure depends again on the type of interventions, the function class and the chosen test for the null hypothesis. We can ask for specific examples whether $\hat{S}(\mathcal{E}) = S^*$ in the population limit.

PROPOSITION 3. *Assume as a special case of* (23) *a shift in the variance of* X *under* $I = 1$ *compared to* $I = 0$ *observations:*

$$\begin{aligned} X &= f(H, \eta) + Z \cdot 1_{I=1} \\ Y &= X\gamma^* + g(H, \varepsilon), \end{aligned} \tag{27}$$

where the p*-dimensional mean-zero random variable* Z *is independent of* H, ε, η *and* I *and has a full-rank covariance matrix. Then* γ^* *and* S^* *are identifiable in a population sense. Specifically, if the test of* $H_{0,S,hidden}(\mathcal{E})$ *has power 1 against any alternative, then*

$$P[\hat{S}(\mathcal{E}) = S^*] \geq 1 - \alpha.$$

A proof is given in Appendix E. Note that the causal variables and coefficients can be identified for (27), even though the model violates the above-mentioned assumptions (ii) and (iii) for identifiability with a classical two-stage instrumental variable analysis: X can be of arbitrary dimension even though the instrumental variable I is univariate and there is no shift in $E(X \mid I)$ between $I = 1$ and $I = 0$.

A further advantage of the invariance approach might be that no test for a weak influence of I on X is necessary. A weak instrument can lead to amplification of biases in conventional instrumental variable regression [Hernán and Robins, 2006]. With the invariance approach, the confidence intervals for γ^* are naturally wide in case of a weak influence of I on X, leading to small sets $\hat{S}$ of selected causal variables.

Ignoring the computational difficulties, this shows that the approach can be generalised to include hidden variables that violate assumption (ii) c) in Proposition 4, for example by

replacing (5) with the null hypothesis (25). As a possible implementation of the general approach we must therefore test (25) for every set $S \subseteq \{1, \dots, p\}$. We are faced with a formidable computational challenge because the coefficients γ^* cannot be found by simple linear regression anymore. One possibility is to place a stricter constraint on the form of allowed interventions. For shifted soft interventions from Section 4.2.3, for example, such an approach is described in Rothenhäusler et al. [2015]. For general interventions, we can test (25) in a brute-force way by testing the invariance of the distribution over a grid of γ-values. However, the computational complexity of this approach is exponential in the predictor dimension and it would be valuable to identify computationally more efficient ways of testing the null hypothesis (25).

6. Further extensions and model misspecification

6.1. Nonlinear models

We have shown an approach to obtain confidence intervals for the causal coefficients in linear models. We might be interested in identifying the set of causal predictors S^* in the more general nonlinear setting (2). The equivalent null-hypothesis to (5) is then

$$H_{0,S,nonlin}(\mathcal{E}): \quad \begin{array}{l} \text{There exists } g : \mathbb{R}^{|S|} \times \mathbb{R} \to \mathbb{R} \text{ and } \varepsilon^e \text{ such that} \\ Y^e = g(X_S^e, \varepsilon^e), \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X_S^e \text{ for all } e \in \mathcal{E}. \end{array} \tag{28}$$

It is interesting to note that S satisfies (28) if and only if it satisfies

$$H_{0,S,nonlin}(\mathcal{E}): \quad \begin{array}{l} \forall e, f \in \mathcal{E} \text{ the conditional distributions } Y^e \,|\, X_S^e = x \text{ and } Y^f \,|\, X_S^f = x \\ \text{are identical for all } x \text{ such that both cond. distr. are well-defined.} \end{array} \tag{29}$$

The "only if" part is immediate and for the "if" part we can use a similar idea as in [Peters et al., 2014, Prop. 9], for example, and choose a Uniform($[0,1]$)-distributed ε and $g(a,b) = g^e(a,b) := F^{-1}_{Y^e \,|\, X_S^e = a}(b)$, where $F_{Y^e \,|\, X_S^e = a}$ is the cdf of $Y^e \,|\, X_S^e = a$.

As in the linear case, we can consider a SEM with environments corresponding to different interventions and, again, the parents of Y satisfy the null hypothesis. More precisely, we have the following remark.

Remark 2. *Proposition 1 and Proposition 4 still hold if we replace linear SEMs* (19) *with nonlinear SEMs*

$$Y_j = f_j(X_{\mathbf{PA}(j)}, \varepsilon_j), \quad j = 1, \dots, p+1$$

and replace Assumption 1 with the assumption that there exists S^* satisfying (28).

Proof. Again, the proof is immediate. Only the case with hidden variables requires an argument. From the SEM, we are given $Y^e = f(X^e_{S^0}, X^e_{S^0_H}, \tilde{\varepsilon}^e)$ with S^0_H being the hidden parents of Y and $(X^e_{S^0_H}, \tilde{\varepsilon}^e) \perp\!\!\!\perp X^e_{S^0}$. We can then write $Y^e = g(X_{S^0}, \varepsilon^e)$ for a uniformly distributed ε^e that is independent of X_{S^0} and $g(x,n) := F^{-1}_{f(x, X^e_{S^0_H}, \tilde{\varepsilon}^e)}(n)$. The function g does not depend on e because $X^e_{S^0_H}$ and $\tilde{\varepsilon}^e$ have the same distribution for all $e \in \mathcal{E}$.

Assume we have a test for the null hypothesis $H_{0,S,nonlin}(\mathcal{E})$. Then, testing all possible sets $S \subseteq \{1, \dots, p\}$, we can get a confidence set for S^* in a similar way as in the linear

setting (15) by

$$\hat{S}(\mathcal{E}) := \bigcap_{S: H_{0,S,nonlin}(\mathcal{E}) \text{ not rejected}} S. \tag{30}$$

If all tests are conducted individually at level α, we have again the property that for any S^* which fulfills (28) or (29), $P(\hat{S}(\mathcal{E}) \subseteq S^*) \geq 1 - \alpha$ since the null hypothesis for S^* will be accepted with probability at least $1 - \alpha$.

Constructing suitable tests for (29) is easier if we are willing to assume that the function g in (28) is additive in the noise component, that is

$$H_{0,S,additive}(\mathcal{E}): \quad \begin{array}{l} \text{there exists } g : \mathbb{R}^{|S|} \to \mathbb{R} \text{ and } \varepsilon^e \text{ such that} \\ Y^e = g(X_S^e) + \varepsilon^e, \quad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X_S^e \text{ for all } e \in \mathcal{E}. \end{array} \tag{31}$$

Then, we can construct tests for the null hypothesis (28) that are similar as in the linear case. Analogously to Method I in Section 3.1, we can perform nonlinear regression in each environment and test whether the regression functions are identical [e.g. Durot et al., 2013, for isotonic regression functions]. As an alternative, we can also fit a regression model on the pooled data set and test whether the residuals have the same distribution in each environment, see Method II in Section 3.1.

We may also test (29) without assuming additivity of the noise component. This could be addressed by introducing an environment variable E and then performing a conditional independence test for $Y \perp\!\!\!\perp E \,|\, X_S$, see also Appendix C. The details of these approaches lie beyond the scope of this paper.

6.2. Interventions on the target variable and its causal mechanism

So far, we have assumed that the error distribution of the target variable is unchanged across all environments $e \in \mathcal{E}$, see Assumption 1 for linear models. This precludes interventions on Y and precludes a change of the causal mechanism for the target variable. For the gene-knockout experiments mentioned in Section 2 and treated in detail in Section 7.2, we would for example know whether we have intervened on the target gene or not. In other situations, we might not be sure whether an intervention on the target variables occurred or not.

If interventions are sparse, other approaches are possible, too. For any given target variable Y, we might not be sure whether an intervention on Y occurred or not, but we can assume that an intervention on Y happened in at most $V \ll |\mathcal{E}|$ different environments, even if we do not know in which of the environments it occurred, see Kang et al. [2015] for a related setting in instrumental variable regression. The null hypothesis (29) in the general nonlinear case can then be weakened to

$$H'_{0,S,nonlin}(\mathcal{E}): \quad \begin{array}{l} \exists \mathcal{E}' \subseteq \mathcal{E} \text{ with } |\mathcal{E}'| \geq |\mathcal{E}| - V \text{ s.t. } \forall e, f \in \mathcal{E}' \text{ the cond. distr. } Y^e \,|\, X_S^e = x \\ \text{and } Y^f \,|\, X_S^f = x \text{ are identical } \forall x \text{ s.t. both cond. distr. are well-defined.} \end{array} \tag{32}$$

The null hypothesis $H'_{0,S^*,nonlin}$ is then still true even when interventions happen on Y in some environments, where S^* is the causal set of variables that satisfies the invariance assumption in the absence of interventions on Y. Any test for (29) can be extended as a test for the weaker null hypothesis (32) by testing all subsets $\mathcal{E}'$ with $|\mathcal{E}'| \geq |\mathcal{E}| - V$ at level α, e.g. using a test for (28), and rejecting (32) only if we can reject all such

subsets. We can then treat $H_{0,S,nonlin}(\mathcal{E})$ as being "accepted" if we find one subset $\mathcal{E}'$ whose corresponding null hypothesis cannot be rejected.

6.3. Model misspecification

We have shown how the approach can be extended to cover hidden variables, nonlinear models and interventions on the target variable. The question arises how the original approach behaves if these model assumptions are violated but we use the original approach instead of the proposed extensions. We again write $\hat{S}(\mathcal{E})$ as in (15) as

$$\hat{S}(\mathcal{E}) := \bigcap_{S:H_{0,S} \text{ not rejected}} S.$$

Our approach still satisfies the coverage property $P(\hat{S}(\mathcal{E}) \subseteq S^*) \geq 1 - \alpha$ for any set S^* that satisfies Assumption 1. Let S_c^* be a set that is considered to be causal, for example, because it is the set of observed parents of Y in a SEM. Under no model misspecification, Proposition 1 shows that this set will satisfy Assumption 1 or, in the general case Equation (29). If the model assumptions are violated, however, then either H_{0,S_c^*} is still true (in which case the desired confidence statements $P(\hat{S}(\mathcal{E}) \subseteq S_c^*) \geq 1 - \alpha$ is still valid) or H_{0,S_c^*} is not longer true. The latter case thus warrants our attention. There are two possibilities. If $H_{0,S}$ is also false for all other sets $S \subseteq \{1, \ldots, p\}$, then $\hat{S}(\mathcal{E}) = \emptyset$ for a test that has maximal power to reject false hypotheses. Thus, the desired coverage property $P(\hat{S}(\mathcal{E}) \subseteq S_c^*) \geq 1 - \alpha$ is still valid, even though the method will now have no power to detect the causal variables. It could happen, on the other hand, that there exists some set $S' \subseteq \{1, \ldots, p\}$ with $S' \setminus S_c^* \neq \emptyset$ for which $H_{0,S'}$ is true. Proposition 5 in Appendix C shows that under some assumptions even in this case, the mistake is not too severe: then there exists a different set $\tilde{S}$, for which $H_{0,S'}$ is true, and that contains only ancestors of the target Y and no descendants. Then, by construction, the same also holds for $\hat{S}(\mathcal{E})$, with probability greater than $1 - \alpha$.

7. Numerical results

We apply the method to simulated data, gene perturbation experiments from biology with interventional data and and an instrumental variable type setting from educational research.

7.1. Simulation experiments

For the simulations, we generate data from randomly chosen linear Gaussian structural equation models (SEMs) and compare various approaches to recover the causal predictors of a target variable.

The generation of linear Gaussian SEMs is described in Appendix G. We sample 100 different settings and for each of those 100 settings, we generate 1000 data sets. We tried to cover a wide range of scenarios, some (but not all of which) correspond to the theoretical results developed in Section 4.3. After randomly choosing a node as target variable, we can then test how well various methods recover the parents (the causal predictors) of this target. We check whether false variables were selected as parents (false positives) or whether the correct parents were recovered (true positives).

For the proposed invariant prediction method, we divide the data into a block of observational data and a block of data with interventions. Some other existing methods make use of the exact nature of the interventions but for our proposed method this information is discarded or presumed unknown. The estimated causal predictors $\hat{S}(\mathcal{E})$ at confidence 95%, computed as in Method I in Section 3.1, are then compared to the true causal predictors S^* of a target variable in the causal graph (which can sometimes be the empty set). The results of Method II are very similar in the simulations and are not shown separately. We record whether any errors were made ($\hat{S}(\mathcal{E}) \not\subseteq S^*$) and whether the correct set was recovered ($\hat{S}(\mathcal{E}) = S^*$). We compare the proposed confidence intervals with point estimates given by several procedures for linear SEMs:

(a) *Greedy equivalence search (GES)* [Chickering, 2002]. In the case of purely observational data, we can identify the so-called Markov equivalence class of the correct graph from the joint distribution, i.e. we can find its skeleton and orient the v-structures, i.e. some of the edges [Verma and Pearl, 1991]. Although, many directions remain ambiguous in the general case, it might be that we can orient some connections of the target variable $X_j - Y$. If the edge is pointing towards Y, we identify X_j as a direct cause of Y. The GES searches greedily over equivalence classes of graph structures in order to maximise a penalised likelihood score. Here, we apply GES on the pooled data set, pretending that all data are observational.

(b) *Greedy interventional equivalence search (GIES) with known intervention targets* [Hauser and Bühlmann, 2012]. The greedy interventional equivalence search (GIES) considers soft interventions (at node j) where the conditional $p(x_j \,|\, x_{\mathbf{PA}(j)})$ is replaced by a Gaussian density in x_j. One can identify interventional Markov equivalence classes from the available distributions that are usually smaller than the Markov equivalence classes obtained from observational data. GIES is a search procedure over interventional Markov equivalence classes maximising a penalised likelihood score. In comparison, a benefit of our new approach is that we do not need to specify the different experimental conditions. More precisely, we do not need to know which nodes have been intervened on.

(c) *Greedy interventional equivalence search (GIES) with unknown intervention targets.* To obtain a more fair comparison to the other methods, we hide the intervention targets from the GIES algorithm and pretend that every variable has been intervened on.

(d) *Linear non-Gaussian acyclic models (LiNGAM)* [Shimizu et al., 2006]. The assumption of non-Gaussian distributions for the structural equations leads to identifiability. We use an `R`-implementation [R Core Team, 2014] of LiNGAM which is based on independent component analysis, as originally proposed by Shimizu et al. [2006]. In the observational setting, the structural equation of a specific variable X_j reads

$$X_j^1 = \sum_{k \in \mathbf{PA}(j)} \beta_{j,k} X_k^1 + \varepsilon_j^1,$$

whereas in the interventional setting (if the coefficients $\beta_{j,k}$ remain the same), we have

$$X_j^2 = \sum_{k \in \mathbf{PA}(j)} \beta_{j,k} X_k^2 + \varepsilon_j^2.$$

One may want to model the pooled data set as coming from a structural equation model of the form

$$\tilde{X}_j = \sum_{k \in \mathbf{PA}(j)} \beta_{j,k} \tilde{X}_k + \tilde{\varepsilon}_j,$$

where $\tilde{\varepsilon}_j$ follows a distribution of the mixture of ε_j^1 and ε_j^2 and thus has a non-Gaussian distribution (Kun Zhang mentioned this idea to JP in a private discussion). The new noise variables $\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_p$ are not independent of each other: if, for any $j \neq k$, $\tilde{\varepsilon}_j$ comes from the first mixture, then $\tilde{\varepsilon}_k$ does so, too. We can neglect this violation of LiNGAM and apply the method nevertheless. There is no theoretical result which would justify LiNGAM for interventional data.

(e) *Regression.* We pool all data and use a linear least-squares regression and retain all variables which are significant at level α/p, in an attempt to control the family-wise error rate (FWER) of falsely selecting at least a single variable at level α in a regression (not causal) sense. As a regression technique, this method cannot correctly identify causal predictors.

(f) *Marginal regression.* We pool all data and retain all variables that have a correlation with the outcome at significance level α/p. As above, this regression method cannot correctly identify causal predictors.

We show the (empirical) probability of false selections, $P(\hat{S}(\mathcal{E}) \not\subseteq S^*)$, in Figure 5 for all methods. The probability of success, $P(\hat{S}(\mathcal{E}) = S^*)$, is shown in Figure 4.

The success probabilities show some interesting patterns. First, there is (as expected) not a method that performs uniformly best overall scenarios. However, *regression* and *marginal regression* are dominated across all 100 scenarios by *GIES* (both with known and unknown interventions), *LiNGAM* and the proposed *invariant prediction*). Among the 100 settings, there were 3 where *GES* performed best on the given criterion, 14 where *GIES* (with known interventions) performed best, 54 for *LiNGAM* and 23 where the proposed *invariant prediction* were optimal for exact recovery. There is no clear pattern as to which parameter is driving the differences in the performances: Spearman's correlation between the parameter settings and the differences in performances between all pairs of methods was less than 0.3 for all parameters. The interactions between the parameter settings seem responsible for the relative merits of one method over another.

The pattern for false selections in Figure 5 is very clear on the other hand. The proposed invariant prediction method controls the rate at which mistakes are made at the desired 0.05 (and often lower due to a conservativeness of the procedure). All other methods have FWE rates that reach 0.4 and higher. No other method offers a control of FWER and the results show that the probability of erroneous selections can indeed be very high. The control of the FWER (and the associated confidence intervals) is the key advantage of the proposed *invariant prediction.*

7.2. Gene perturbation experiments

Data set. We applied our method to a yeast (Saccharomyces cerevisiae) data set [Kemmeren et al., 2014]. Genome-wide mRNA expression levels in yeast were measured and we therefore have data for $p = 6170$ genes. There are $n_{obs} = 160$ "observational" samples of wild-types and $n_{int} = 1479$ data points for the "interventional" setting where each of them

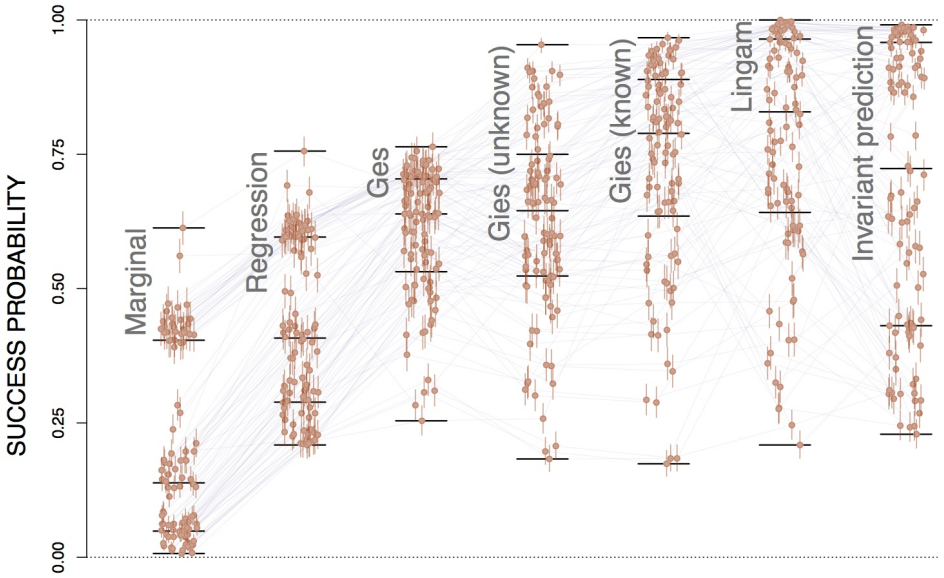

Fig. 4. The probability of success, defined as $P(\hat{S}(\mathcal{E}) = S^*)$ for various methods, including our new proposed invariant prediction in the rightmost column. Each dot within a column (the x-offset within a column is uniform) corresponds to one of the 100 simulation scenarios. The dot's height shows the empirical probability of success over 1000 simulations and the small bars indicate a 95% confidence for the true success probability. Identical scenarios are connected by grey solid lines. For each method, the maximal and minimal values along with the quartiles of each distribution are indicated by horizontal solid bars.

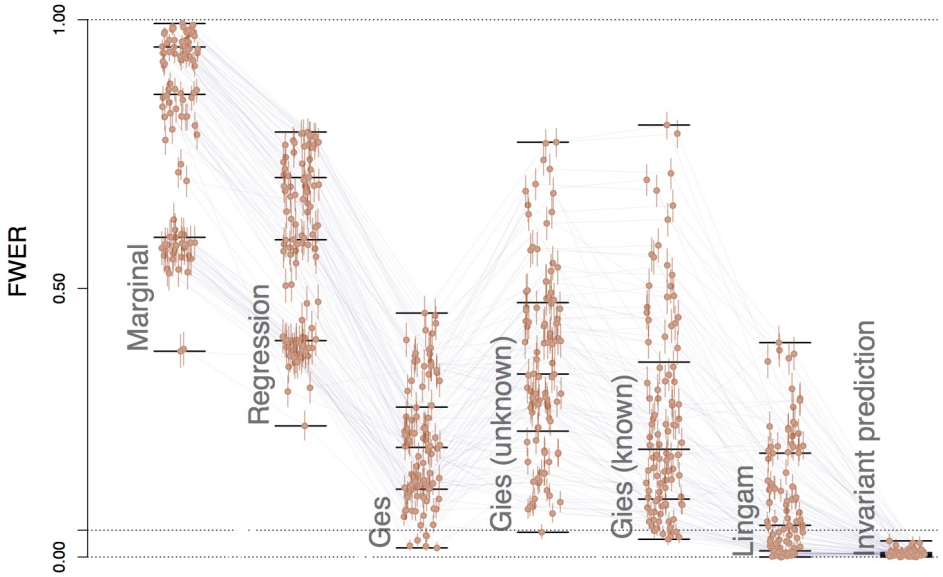

Fig. 5. The probability of erroneous selections $P(\hat{S}(\mathcal{E}) \not\subseteq S^*)$ (FWER) for the considered methods, including the proposed invariant prediction to the right. The figure is otherwise analogously generated as Figure 4. The dotted line indicates the 0.05 level at which the invariant prediction method was (successfully) controlled. All other methods do not offer FWER control.

corresponds to a strain for which a single gene $k \in K := \{k_1, \ldots, k_{1479}\} \subset \{1, \ldots, 6170\}$ has been deleted (meanwhile, there is an updated data set with five more mutants). If the method suggests, for example, gene 5954 as a cause of gene 4710, and there is a deletion strain corresponding to gene 5954, we can use this data point to determine whether gene 5954 indeed has a (possibly indirect) causal influence on 4710. We say that the pair is a true positive if the expression level of gene 4710 after intervening on 5954 lies in the 1% lower or upper tail of the observational distribution of gene 4710, see also Figure 6 below. (We additionally require that the intervention on gene 5954 appears to be "successful" in the sense that the expression level of gene 5954 after intervening on this gene 5954 lies in the 1% lower or upper tail of the observational distribution of gene 5954. This was not the case for 38 out of the 1479 interventions.) With this criterion, there are about 9.2% relevant effects, which corresponds to the proportion of true positives for a random guessing method.

Separation into observational and interventional data. For predicting a causal influence of, say, gene 5954 on another gene we do not want to use interventions on the same gene 5954 (this would use information about the ground truth). We therefore apply the following procedure: for each $k \in K$ we consider the observational data as $e = 1$ and the remaining $1478 = 1479 - 1$ data points corresponding to the deletions of genes in $K \setminus \{k\}$ as the interventional setting $e = 2$. Since this would require $n_{int} \times p$ applications of our method, we instead separate K into $B = 3$ subsets of equal size, consider the

two subsets not containing k as the interventional data, and do not make any use of the subset containing k. This leaves some information in the data unused but yields a huge computational speed-up, since we need to apply our method in total only $3 \times p$ times. Additionally, when looking for potential causes of gene 4710, we do not consider data points corresponding to interventions on this gene (if it exists), see Proposition 1.

Goodness of fit and p-values. If we would like to avoid making a single mistake on the data set with high probability $1 - \alpha$, we can set the significance level to for each gene to α/n_{int}, using a Bonferroni correction in order to take into account the $n_{int} = 1479$ genes that have been intervened on. We work with $\alpha = 0.01$ if not mentioned otherwise. The guarantee requires, however, that the model is correct (for example the linearity assumption is correct and there are no hidden variables with strong effects on both genes of interest). These assumptions are likely violated, and the implications have been partially discussed in the previous Section 6. To further guard against false positives that are due to model misspecification we require that there is at least one model (one subset $S \subseteq \{1, \ldots, p\}$) for which the model fits reasonably well: we define this by requiring a p-value above 0.1 for testing $H_{0,S}(\mathcal{E})$ for the best-fitting set S of variables (the set with the highest p-value), if not mentioned otherwise (but we also vary the threshold to test how sensitive our method is with regard to parameter settings). If no set of variables attains this threshold, we discard the models and make no prediction.

Method. We use L_2-boosting [Friedman, 2001, Bühlmann and Yu, 2003] from the R-package `mboost` [Hothorn et al., 2010] with shrinkage 0.1 as a way to preselect for each response variable ten potentially causal variables, to which we then apply the causal inference methods. We primarily use Method II as Method I requires subsampling for computational reasons. Subsampling can lead to a loss of power as there is a not-negligible probability of loosing the few informative data points in the subsampling process. For a computational speed-up we only consider subsets of size ≤ 3 as candidate sets S. Furthermore, we only retain results where just a single variable has been shown to have a causal influence to avoid testing more difficult scenarios where one would have to intervene on multiple genes simultaneously.

Comparisons. As alternative methods we consider IDA [Maathuis et al., 2009] based on the PC algorithm [Spirtes et al., 2000] and a method that ranks the absolute value of marginal correlation ($j_1 \to j_2$ and $j_2 \to j_1$ obtain the same score and are ranked randomly), both of which make use only of the observational data. We also compare with IDA based on greedy interventional equivalence search (GIES) [Hauser and Bühlmann, 2015] and a correlation-based method that ranks pairs according to correlation on the pooled observational and interventional data. It was not feasible to run LiNGAM [Shimizu et al., 2011] on this data set.

Results. The proposed method (Method II) outputs eight gene pairs that can be checked because the corresponding interventional experiments are available. There are in total eight causal effects that are significant at level 0.01 after a Bonferroni correction. Out of these eight pairs, six are correct (random guessing has a success probability of 9.2%).

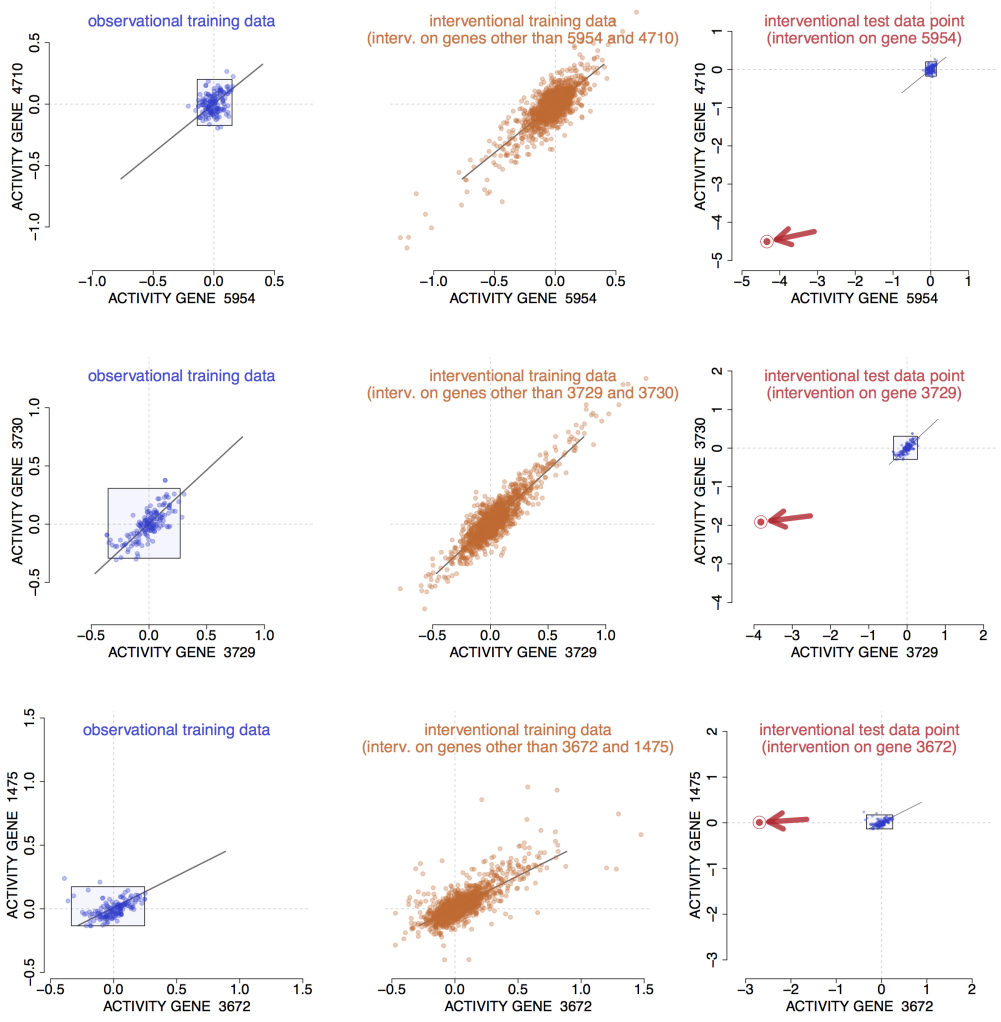

Fig. 6. The three rows correspond to the three most significant effects found by the proposed method (with the most significant effect on top, suggesting a causal effect of gene 5954 on gene 4710). The left column shows the observational data, while the second column shows the interventional data (that are neither using interventions on the target variable itself nor using interventions on the examined possible causal predictors of the target variable); these two data sets are used as two environments for training the invariant prediction model. The regression line for a joint model of observational and interventional data, as proposed in Method II, is shown in both plots; we cannot reject the hypothesis that the regression is different for observational and interventional data here. The third column finally shows the test data (with the 1%-99% quantile-range of the observational data shown as a shaded box as in the first column). There, we use the intervention data point on the chosen gene and look at the effect on the target variable. The first two predicted causal effects can be seen to be correct (true positives) in the following sense: after successfully intervening on the predicted cause, the target gene shows reduced activity; the third suggested pair is unsuccessful (false positive) since the intervention reduces the activity of the cause but the target gene remains as active as in the observational data.

Table 1. The number of true effects among the strongest 8 effects that have been found in the interventional test data (the number 8 has been chosen to correspond to the number of significant effects under the proposed Method II). Method I is based on 1000 samples and required roughly 10 times more computational time than Method II.

method	Method I	Method II	GIES	IDA	marginal corr.		random
					observ.	pooled	guessing
# of true positives (out of 8)	6	6	2	2	1	2	2 (95% quantile) 3 (99% quantile) 4 (99.9% quantile)

Figure 6 shows the three pairs that obtained the highest rank, i.e. smallest p-values. The rows in the figure therefore correspond to the three causal effects in the data set that were regarded as most significant by our method. One note regarding the plot: we plot all available data even though only two-thirds of it was effectively used for training due to the discussed cross-validation scheme. Many outlying points in the interventional training data of the false positive (second column of third row in Figure 6) are in particular not part of the training data and the method might have performed better with a more computationally-intensive validation scheme that would split the data into B blocks with B larger than the currently used $B = 3$.

In order to compare with other methods (none of which provide a measure of significance), we always consider the eight highest-ranked pairs. Table 1 summarises the results. In this data set, the alternative methods were not able to exceed random guessing.

To test sensitivity of the results to the chosen implementation details of the method, the variable pre-selection, the goodness-of-fit cutoff have also all been varied (for example using Lasso instead of boosting as pre-selection and using a cutoff of 0.1 instead of 0.01). For Method II, variable selection with Lasso instead of boosting leads to a true positive rate of 0.63 (5 out of 8). Choosing the goodness-of-fit cutoff at 0.01 rather than 0.1 leads to true positive rates of 0.43 (9 out of 21) for boosting and 0.47 (8 out of 17) for Lasso. Method I without forcing eight decisions leads to a true positive rate of 0.75 (3 out of 4) for boosting and 1.00 (1 out of 1) for Lasso. Choosing the goodness-of-fit cutoff at 0.01 rather than 0.1 leads to true positive rates of 0.86 (6 out of 7) for boosting and 0.75 (3 out of 4) for Lasso. (Using 500 instead of 1000 subsamples for Method I leads to increased speed and worse performance.) We regard it as encouraging that the true positive rate is always larger than random guessing, irrespective of the precise implementation of the method.

Among the reasons for false positives (e.g. 2 out of 8 for Method II in Table 1, there are at least the following options: (a) noise fluctuations, (b) nonlinearities, (c) hidden variables, (d) issues with the experiment (for example the intervention might have changed other parts of the network) and (e) the pair is a true positive but is -by chance- classified as a false positive by our criterion (see "Data set" above). Missing causal variables in the pre-screening by boosting or Lasso falls under category (c). We control (a) and have provided arguments why (b) and (c) will lead to rejection of the whole model rather than lead to false positives. Lowering the goodness-of-fit-threshold seemed indeed to lead to more spurious results, as expected from the discussion in the previous Section 6.3. Validating a potential issue with the experiment as in reason (d) is beyond our possibilities. We could address (e) if we had access to multiple repetitions of the intervention experiments.

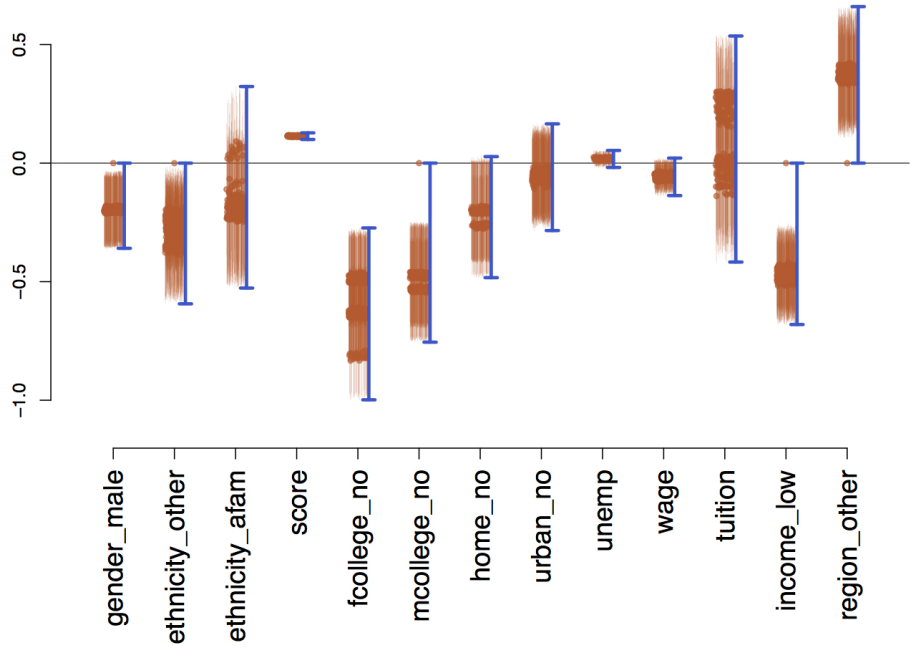

Fig. 7. The 90% confidence intervals for the influence of various variables on the probability of receiving a BA degree (or higher) are shown in blue. Of all 8192 possible sets S, we accept 1565 sets (the empty set is not accepted as the probability of receiving a degree is sufficiently different for people within a close distance to a 4-year college and further away). The point-estimates for the coefficients are shown for these 1565 sets as red dots and the corresponding confidence intervals as vertical red bars. The blue confidence intervals are then the union of all 1565 confidence intervals, as in our proposed procedure. The variables *score* (test score) and *fcollege_no* (active if father did not receive a college degree) show significant effects.

7.3. Educational attainment

We look at a data set about educational attainment of teenagers [Rouse, 1995]. For 4739 pupils from approximately 1100 US high schools, 13 attributes are recorded, including gender, race, scores on relevant achievement tests, whether the parents are college graduates, or family income. Here we work with the data as provided in Stock and Watson [2003], where we can see the length of education pupils received. We make a binary distinction into whether pupils received a BA degree or higher (equivalent to at least 16 years of education in the classification used in Stock and Watson [2003]) and ask whether we can identify a causal predictive model that allows to forecast whether students will receive a BA degree or not and this forms a binary target Y.

The distance to the nearest 4-year college is recorded in the data and we use it to split the dataset into two parts in the sense of (18); we assume that this variable has no *direct* influence on the target variable. As discussed, this variable does not have to satisfy the usual assumptions about instrumental variables for our analysis but just has to be independent of the noise in the outcome variable (it must be a non-descendant of the target), which seems satisfied in this dataset as the distance to the 4-year college precedes

the educational attainment chronologically. One set of observations are thus all pupils who live closer to a 4-year college than the median distance of 10 miles. The second set are all other pupils, who live at least 10 miles from the nearest 4-year college. We ask for a classification that is invariant in both cases in the sense that the conditional distribution of Y, given X, is identical for both groups, where X are the set of collected attributes and Y is the binary outcome of whether they attained a BA degree or higher. We use the fast approximate Method II of Section 3.1, with the suggested extension to logistic regression.

Figure 7 shows the outcome of the analysis, which is also included as an example in the `R`-package `InvariantCausalPrediction`. Factors were split into dummy variables so that "ethnicity_afam" is 1 if the ethinicity is african-american and 0 otherwise, "fcollege_no" is 1 if the father did not receive a college degree and so forth. We provide 90% confidence intervals. All of them include 0 except for the confidence interval for the influence of the test score (positive effect) and the indicator that the father did not receive a college degree (negative effect). A high score on the achievement test thus seems to have a positive causal influence on the probability of obtaining a BA degree, which seems plausible.

As it is difficult to verify the ground truth in this case, we refrain from comparisons with other possible approaches to the same data set and just want to use it as an example of a possible practical application. The example shows that we can use instrumental-variable-type variables to split the data set into different "experimental" groups. If the distributions of the outcome are sufficiently different in the created groups, we can potentially have power to detect invariant causal prediction effects.

8. Discussion and Future Work

An advantage of causal predictors compared to non-causal ones is that their influence on the target variable remains invariant under different changes of the environment (which arise for example through interventions). We have described this invariance and exploit it for the identification of the causal predictors. Confidence sets for the causal predictors and confidence intervals for relevant parameters follow naturally in this framework. In the special case of Gaussian structural equation models with interventions we have proved identifiability guarantees for the set of causal predictors. We discussed some of the questions that require more work: suitable tests for equality of conditional distributions for nonlinear models, feedback models and increased computational efficiency both in the absence and presence of hidden variables.

The approach of invariant prediction provides new concepts and methods for causal inference, and also relates to many known concepts but considers them from a different angle. It constitutes a new understanding of causality that opens the way to a novel class of theory and methodology in causal inference.

Acknowledgements

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/ 2007–2013) under REA grant agreement no 326496. The authors would like to thank seven anonymous referees for their helpful comments on an earlier version of the manuscript and

would like to thank Alain Hauser, Thomas Richardson, Bernhard Schölkopf and Kun Zhang for helpful discussions.

References

J. Aldrich. Autonomy. *Oxford Economic Papers*, 41:15–34, 1989.

S.A. Andersson, D. Madigan, and M.D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics*, 25:505–541, 1997.

J.D. Angrist, G.W. Imbens, and D.B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91:444–455, 1996.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98:791–806, 2011.

K.A. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, New York, USA, 1989.

R.J. Bowden and D.A. Turkington. *Instrumental variables*, volume 8. Cambridge University Press, Cambride, UK, 1990.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, New York, USA, 2011.

P. Bühlmann and B. Yu. Boosting with the L_2-loss: regression and classification. *Journal of the American Statistical Association*, 98:324–339, 2003.

P. Bühlmann, P. Rütimann, and M. Kalisch. Controlling false positive selections in high-dimensional regression and causal inference. *Statistical methods in medical research*, 22: 466–492, 2013.

P. Bühlmann, J. Peters, and J. Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *Annals of Statistics*, 42:2526–2556, 2014.

R. Castelo and T. Kocka. On inclusion-driven learning of Bayesian networks. *Journal of Machine Learning Research*, 4:527–574, 2003.

D.M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

G.C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28:591–605, 1960.

G. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 116–125, 1999.

H. Cramér. Über eine Eigenschaft der normalen Verteilungsfunktion. *Mathemathische Zeitschrift*, 41:405–414, 1936.

A. P. Dawid and V. Didelez. Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, 4:184–231, 2010.

A.P. Dawid. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95:407–424, 2000.

A.P. Dawid. Counterfactuals, hypotheticals and potential responses: a philosophical examination of statistical causality. Technical report, Department of Statistical Science, University College London, Research Report 269, 2006.

A.P. Dawid. *The Decision-Theoretic Approach to Causal Inference*. Wiley Online Library, 2012.

A.P. Dawid. Statistical causality from a decision-theoretic perspective. *Annual Review of Statistics and Its Application*, 2:273–303, 2015.

V. Didelez, A.P. Dawid, and S. Geneletti. Direct and indirect effects of sequential treatments. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artifical Intelligence (UAI)*, pages 138–146, 2006.

V. Didelez, S. Meng, and N.A. Sheehan. Assumptions of IV methods for observational epidemiology. *Statistical Science*, 25:22–40, 2010.

O.D. Duncan. *Introduction to Structural Equation Models*. Academic Press, New York, USA, 1975.

C. Durot, P. Groeneboom, and H. Lopuhaä. Testing equality of functions under monotonicity constraints. *Journal of Nonparametric Statistics*, 25:939–970, 2013.

D. Eaton and K.P. Murphy. Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 107–114, 2007.

F. Eberhardt and R. Scheines. Interventions and causal inference. *Philosophy of Science*, 74:981–995, 2007.

J.H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

S. Greenland, J. Pearl, and J.M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10:37–48, 1999.

T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:S1–S115 (supplement), 1944.

A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464, 2012.

A. Hauser and P. Bühlmann. Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society, Series B*, 77:291–318, 2015.

Y.-B. He and Z. Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9:2523–2547, 2008.

M. Hernán and J. Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006.

K. D. Hoover. The logic of causal inference. *Economics and Philosophy*, 6:207–234, 1990.

T. Hothorn, P. Bühlmann, T. Kneib, M. Schmid, and B. Hofner. Model-based boosting 2.0. *Journal of Machine Learning Research*, 11:2109–2113, 2010.

P.O. Hoyer, D. Janzing, J.M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 689–696, 2009.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13:3387–3439, 2012.

A.L. Jackson, S.R. Bartz, J. Schelter, S.V. Kobayashi, J. Burchard, M. Mao, B. Li, G. Cavet, and P.S. Linsley. Expression profiling reveals off-target gene regulation by RNAi. *Nature Biotechnology*, 21:635–637, 2003.

D. Janzing, J.M. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, 2012.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

H. Kang, A. Zhang, T. Cai, and D.S. Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association (to appear)*, 2015.

P. Kemmeren, K. Sameith, L.A. van de Pasch, J.J. Benschop, T.L. Lenstra, T. Margaritis, E. O'Duibhir, E. Apweiler, S. van Wageningen, C.W. Ko, S. van Heesch, M.M.. Kashani, G. Ampatziadis-Michailidis, M.O.. Brok, N.A. Brabers, A.J. Miles, D. Bouwmeester, S.R. van Hooff, H. van Bakel, E. Sluiters, L.V. Bakker, B. Snel, P. Lijnzaad, D. van Leenen, M.J. Groot Koerkamp, and F.C. Holstege. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157: 740–752, 2014.

M.M. Kulkarni, M. Booker, S.J. Silver, A. Friedman, P. Hong, N. Perrimon, and B. Mathey-Prevot. Evidence of off-target effects associated with long dsrnas in drosophila melanogaster cell-based assays. *Nature Methods*, 3:833–838, 2006.

S.L. Lauritzen. *Graphical Models*. Oxford University Press, New York, USA, 1996.

S.L. Lauritzen and T.S. Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society, Series B*, 64:321–348, 2002.

S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50:157–224, 1988.

M.H. Maathuis, M. Kalisch, and P. Bühlmann. Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37:3133–3164, 2009.

J.M. Mooij, D. Janzing, T. Heskes, and B. Schölkopf. On causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 639–647, 2011.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, USA, 2nd edition, 2009.

J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101:219–228, 2014.

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

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org`.

T. Richardson and J.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.

T. Richardson and P. Spirtes. Ancestral graph markov models. *Annals of Statistics*, 30: 962–1030, 2002.

J.M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393 – 1512, 1986.

J.M. Robins, M.A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560, 2000.

D. Rothenhäusler, C. Heinze, J. Peters, and N. Meinshausen. backShift: Learning causal cyclic graphs from unknown shift interventions. *arXiv preprint arXiv:1506.02494*, 2015.

C.E. Rouse. Democratization or diversion? The effect of community colleges on educational attainment. *Journal of Business & Economic Statistics*, 13:217–224, 1995.

D.B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100:322–331, 2005.

R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26:1651–1686, 1998.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1255–1262, 2012.

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

S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P.O. Hoyer, and K. Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

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

J.H. Stock and M.W. Watson. *Introduction to econometrics*, volume 104. Addison Wesley, Reading, USA, 2003.

J. Terza, A. Basu, and P. Rathouz. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of health economics*, 27(3):531–543, 2008.

J. Tian and J. Pearl. Causal discovery from changes. In *Proceedings of the 17th Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 512–522, 2001.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

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

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 255–270, 1991.

P.G. Wright. *The tariff on animal and vegetable oils*. The Macmillan company New York, 1928.

S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

A. An Example

We illustrate here in Figure 8 the concepts and methodology which have been developed in Sections 2.1, 2.2 and 3. The figure shows an example of two environments whose data were generated from observational and interventional structural equation models.

B. Hidden variables without confounding

We discuss first a generalisation of Proposition 1, allowing for some hidden variables but excluding confounding between the observable causal variables and the target variable. Another setting allowing for such confounding is presented in Section 5. Consider the structural equation model with variables $X_1 = Y, X_2, \ldots, X_p, X_{p+1}, H_1, \ldots, H_q$, where the latter $H_1, \ldots, H_q$ are unobserved, hidden variables with mean zero.

PROPOSITION 4. *Consider a linear structural equation model including variables*

$$(X_1 = Y, X_2, \ldots, X_p, X_{p+1}, H_1, \ldots, H_q),$$

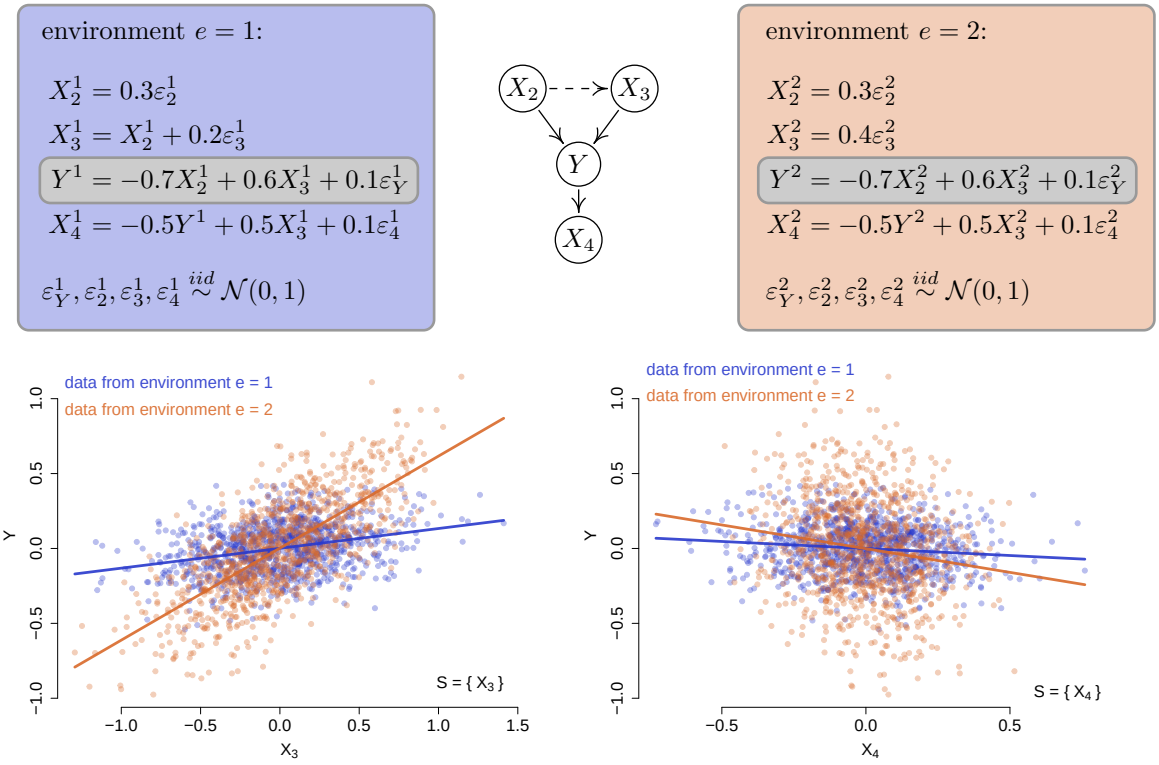

Fig. 8. The top row shows the example of two structural equation models (SEMs) entailing the two distributions corresponding to two environments $e = 1$ and $e = 2$. Here, the first environment corresponds to the graph including the dashed edge, the second environment corresponds to an intervention on X_3, the graph excluding the dashed edge. Since the structural equation for Y is unchanged, the set $S^* = \{X_2, X_3\} = \mathbf{PA}(1)$ satisfies Assumption 1, see Proposition 1. We consider the setup where we know neither S^* nor the SEMs (we do not even require the existence of such a SEM). Instead, we are given two finite samples (one from each environment) and provide an estimator $\hat{S}$ for S^*. In the above example, the null hypothesis of invariant prediction gets rejected for any set S of variables except for $S = \{X_2, X_3\}$ and $S = \{X_2, X_3, X_4\}$ (using the methodology described in Section 3.1). The bottom row shows that for $S = \{X_3\}$, for example, the linear regression coefficients differ in the two environments. For $S = \{X_4\}$, the regression coefficients seem similar but the set is rejected because of varying variances of the residuals. We then propose to consider the intersection of the sets of variables for which the hypothesis of invariance is not rejected; this leads to the (conservative) estimate $\hat{S}$ for the set of identifiable predictors S^*: $\hat{S} = \{X_2, X_3\} \cap \{X_2, X_3, X_4\} = \{X_2, X_3\}$. We thus have for this case $\hat{S} = S^*$, see also Theorem 3 with $k_0 = 3$.

whose structure is given by a directed acyclic graph. Denote by

$$S^0 := \mathbf{PA}(1) \cap \{2, \dots, p+1\}$$

the indices of the observable direct causal variables for Y and by S_H^0 the set of indices having a directed edge from the hidden variables $H_1, \dots, H_q$ to Y, i.e., $S_H^0 = \mathbf{PA}(1) \setminus S^0$. The structural equation for Y is

$$Y = \sum_{j \in S^0} \beta_{Y,j} X_j + \sum_{k \in S_H^0} \kappa_{Y,k} H_k + \varepsilon_Y,$$

where ε_Y is independent of X_{S^0} and $H_{S_H^0}$.
Then, by choosing $\gamma^* = \{\beta_{Y,j},\ j \in S^0\}$ and $S^* = S^0$, Assumption 1 holds if one of the following conditions (i) or (ii) is satisfied.

(i) There are no direct causal effects from the hidden variables $H_1, \dots, H_q$ to the target variable Y, i.e., $S_H^0 = \emptyset$, and it holds that

$$Y^e = \sum_{j \in S^0} \beta_{Y,j} X_j^e + \varepsilon_Y^e \text{ for all } e \in \mathcal{E}, \tag{33}$$

where ε_Y^e is independent of $X_{S^0}^e$ and has the same distribution for all $e \in \mathcal{E}$. In particular, this holds under do- or soft-interventions on the variables $\{X_2, \dots, X_{p+1}\} \cup \{H_1, \dots, H_q\}$ given that $S_H^0 = \emptyset$.

(ii) There are hidden variables which have a direct effect on the target variable Y, i.e., $S_H^0 \neq \emptyset$. It holds that

$$Y^e = \sum_{j \in S^0} \beta_{Y,j} X_j^e + \sum_{k \in S_H^0} \kappa_{Y,k} H_k^e + \varepsilon_Y^e \text{ for all } e \in \mathcal{E}, \tag{34}$$

where $\sum_{k \in S_H^0} \kappa_{Y,k} H_k^e + \varepsilon_Y^e$ is independent of $X_{S^0}^e$ and has the same distribution with mean zero for all $e \in \mathcal{E}$. This holds under the following conditions (a)-(c):

(a) the experiments $e \in \mathcal{E}$ arise as do- or soft-interventions;

(b) there are no interventions on Y, on nodes in S_H^0 or on any ancestor of S_H^0;

(c) there is no d-connecting path between any node in S^0 and S_H^0.

PROOF. Assumption 1 follows immediately from (33) or (34), respectively. From the definition of the interventions, as described in Section 4.2, the justification for (33) follows and hence the claim assuming condition (i). When invoking condition (ii), we show now that (a)-(c) imply (34) and the required conditions. Due to (a) and (b), we have Equation (34) and we know that the distribution of

$$\eta^e := \sum_{k \in S_H^0} \kappa_{Y,k} H_k^e + \varepsilon_Y^e$$

is the same for all $e \in \mathcal{E}$. Furthermore, η^e is independent of $X_{S^0}^e$ because of (c).

C. Model Misspecification

Under model misspecification $S(\mathcal{E})$ may not be a subset of the direct causes of Y anymore. The following proposition shows that in most cases it is still a subset of the ancestors of

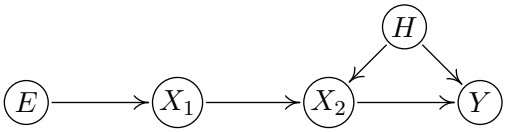

Fig. 9. This graph corresponds to a model misspecification in the sense that the assumptions of Proposition 1 and assumption (ii) c) of Proposition 4 are not satisfied. Indeed, we find that $H_{0,S}$ is violated for $S = S^0 := \{X_2\}$. And since $H_{0,S}$ is satisfied for both $S = \{X_1, X_2\}$ and $S = \{X_1\}$, we obtain $S(\mathcal{E}) = \{X_1\}$. Therefore, $S(\mathcal{E})$ is not a subset of S^0 but it is still a subset of the ancestors $\mathbf{AN}(Y)$ of Y, see Proposition 5.

Y (and is therefore a subset of possibly indirect causes of Y). The proposition is formulated in the general case, see Section 6.1. In order to formulate the required faithfulness assumption, we consider an environment variable E.

PROPOSITION 5. *Consider a SEM over nodes $(Y, X_2, \ldots, X_{p+1}, H_1, \ldots, H_q)$ with hidden variables $H_1, \ldots, H_q$. We now augment the corresponding graph by a discrete environment variable $E \in \mathcal{E}$ [e.g. Pearl, 2009] that satisfies $P(E = e) > 0$ for all $e \in \mathcal{E}$ and has a directed edge to any node that is do- or soft-intervened on. Let us assume that the joint distribution over $(Y, X_2, \ldots, X_{p+1}, H_1, \ldots, H_q, E)$ is faithful w.r.t. the augmented graph. Then*

$$S(\mathcal{E}) := \bigcap_{S\,:\,H_{0,S,nonlin}(\mathcal{E})\ is\ true} S \ \subseteq\ \mathbf{AN}(Y) \cap \{X_2, \ldots, X_{p+1}\}.$$

In particular, this proposition still holds under model misspecification when for some do-interventions, for example, $S^0 = \mathbf{PA}(Y) \cap \{X_2, \ldots, X_{p+1}\}$ does not satisfy $H_{0,S,nonlin}(\mathcal{E})$ (28); Figure 9 shows an example. The following proof also shows that there are model misspecifications where we expect $S(\mathcal{E}) = \emptyset$. If Y is directly intervened on, for example, under the assumption of Proposition 5, we will not be able to find any set S that satisfies (28).

PROOF. We first note that $H_{0,S,nonlin}(\mathcal{E})$ (29) holds if and only if $Y \perp\!\!\!\perp E \,|\, X_S$. Because of faithfulness this is the same as Y and E being d-separated given X_S in the augmented graph. Assume now that the latter holds for some set $S \subseteq \{X_2, \ldots, X_{p+1}\}$. (Such a set S does not exist if Y is directly intervened on.) The proposition follows if we can construct a set $\tilde{S} \subseteq \mathbf{AN}(Y) \cap \{X_2, \ldots, X_{p+1}\}$ that satisfies Y and E being d-separated given $X_{\tilde{S}}$.

Assume that not all nodes in S are ancestors of Y. Define then $W \in S$ to be one "youngest" non-ancestor in S, that is, $W \notin \mathbf{AN}(Y)$ and there is no directed path from W to any other node in S. (Such a node must exist since otherwise all youngest nodes of S are in $\mathbf{AN}(Y)$, which implies $S \subseteq \mathbf{AN}(Y)$.) We now prove that for

$$\tilde{S} := S \setminus \{W\}$$

we have Y and E are d-separated given $X_{\tilde{S}}$. To see this, consider any path from E to Y. If this path does not go through W, the path is blocked by $\tilde{S}$ because it was blocked by $S = \tilde{S} \cup \{W\}$ (removing nodes outside a path can -if anything- only block it). Consider now a path that passes W and the two edges connected to W that are involved in this path. If both edges are into W, we are done because removing W does not open the path. If one of these edges goes out of W, there must be a collider on this path which is a

descendant of W (E does not have incoming edges and W is not an ancestor of Y). But because W is the youngest node in S neither the collider nor any of its descendants is in S. We can therefore remove W and the path is still blocked.

D. Potential Outcomes and Invariant Prediction

We now sketch that the assumption of invariant prediction can also be satisfied in a potential outcome framework [e.g. Rubin, 2005]: as long as we do not intervene on the target variable Y, the conditional distributions of Y given the of causal predictors remains invariant. (Here, we discuss the nonlinear setting and therefore develop a result that corresponds to Remark 2 rather than Proposition 1.) Although other formulations may be possible, too, we adopt the counterfactual language introduced by Richardson and Robins [2013] who refer to finest fully randomised causally interpretable structured tree graphs (FFR-CISTG) [Robins, 1986]. We further consider the nonlinear version (29) of invariant prediction, see also Remark 2.

Similar as in [Richardson and Robins, 2013, Definition 1], we consider random variables $\mathbf{V} := (X_1 = Y, X_2, \dots, X_p, X_{p+1})$ and assume the existence of counterfactual variables $X_j(\tilde{\mathbf{r}})$, for any assignment $\tilde{\mathbf{r}}$ to a subset $\mathbf{R} \subseteq \mathbf{V}$ and for all $j \in \{1, \dots, p+1\}$. We further assume

(C1) "consistency and recursive substitution" [Richardson and Robins, 2013, equation (14)] and

(C2) "FFR-CISTG independence" [Richardson and Robins, 2013, equation (17)] .

To ease notation, we require $X_j(x_j = \tilde{r}) = \tilde{r}$ rather than $X_j(x_j = \tilde{r}) = X_j$ [Richardson and Robins, 2013, p. 21].

PROPOSITION 6. *Consider random variables* $\mathbf{V} := (X_1 = Y, X_2, \dots, X_p, X_{p+1})$ *and denote the causes of* Y *by* $\mathbf{P} := \mathbf{PA}(1)$. *For each environment* $e \in \mathcal{E}$ *consider a set* $\mathbf{R}^e \subseteq \mathbf{V} \backslash \{Y\}$ *of treatment variables and an assignment* $\tilde{\mathbf{r}}^e$, *that is* $X_j^e := X_j(\tilde{\mathbf{r}}^e)$. *Assuming (C1) and (C2), i.e. an FFR-CISTG model, we have that*

$$Y(\tilde{\mathbf{r}}^e) \,|\, \mathbf{P}(\tilde{\mathbf{r}}^e) = \mathbf{q} \quad \stackrel{d}{=} \quad Y(\tilde{\mathbf{r}}^f) \,|\, \mathbf{P}(\tilde{\mathbf{r}}^f) = \mathbf{q} \tag{35}$$

for all $e, f \in \mathcal{E}$ *and for all* $\mathbf{q}$ *such that both sides of* (35) *are well-defined. Therefore, the set* $\mathbf{P}$ *of parents satisfies* (29).

We have already seen in Appendix B, that we can allow for some hidden variables, i.e., the assumption (C2) can be relaxed further.

PROOF. We have for all $e \in \mathcal{E}$

$$\begin{aligned} Y(\tilde{\mathbf{r}}^e) \,\big|\, \mathbf{P}(\tilde{\mathbf{r}}^e) = \mathbf{q} \quad &= \quad Y(\tilde{\mathbf{r}}^e) \,\big|\, (\mathbf{P} \setminus \mathbf{R})(\tilde{\mathbf{r}}^e) = \mathbf{q}_{\mathbf{P}\setminus\mathbf{R}}, (\mathbf{P} \cap \mathbf{R})(\tilde{\mathbf{r}}) = \tilde{\mathbf{r}}_{\mathbf{P}\cap\mathbf{R}} \\ &\stackrel{(*)}{=} \quad Y(\tilde{\mathbf{r}}^e) \,\big|\, (\mathbf{P} \setminus \mathbf{R})(\tilde{\mathbf{r}}^e) = \mathbf{q}_{\mathbf{P}\setminus\mathbf{R}} \\ &\stackrel{(+)}{=} \quad Y \,\big|\, (\mathbf{P} \setminus \mathbf{R}) = \mathbf{q}_{\mathbf{P}\setminus\mathbf{R}}, (\mathbf{P} \cap \mathbf{R}) = \tilde{\mathbf{r}}_{\mathbf{P}\cap\mathbf{R}}, \end{aligned}$$

where we have used $(\mathbf{P} \cap \mathbf{R})(\tilde{\mathbf{r}}) = \tilde{\mathbf{r}}_{\mathbf{P}\cap\mathbf{R}}$ in $(*)$ and both (C1) and the modularity property [Richardson and Robins, 2013, Proposition 16] in $(+)$. This proves the statement because the latter expression is an observational distribution. All equality signs should be understood as holding in distribution.

E. Proof of Proposition 3

PROOF. The residuals $Y - X\gamma$ for $\gamma \in \mathbb{R}^p$ are given by $g(H, \varepsilon) + (\gamma^* - \gamma)f(H, \eta) + Z1_{I=1}(\gamma^* - \gamma)$. The two environments $\mathcal{E}$ are equivalent to conditioning on $I = 0$ for the first environment and $I = 1$ for the second environment. Since $I, H, \varepsilon, \eta, Z$ are independent and Z has a full-rank covariance matrix, the distribution of the residuals can only be invariant between the two environments if $\gamma - \gamma^* \equiv 0$. Hence the test of $H_{0,S,hidden}(\mathcal{E})$ will be rejected for $S \neq S^*$, whereas the true null $H_{0,S^*,hidden}(\mathcal{E})$ is accepted with probability at least $1 - \alpha$ by construction of the test and the result follows by the definition of $\hat{S}$ in (26).

F. Proofs of Section 4.3

F.1. Proof of Theorem 2 (i)

PROOF. As shown in Proposition 1 we have $S(\mathcal{E}) \subseteq \mathbf{PA}(Y)$ because the null hypothesis (5) is correct for $S^* = \mathbf{PA}(Y)$. We assume that $S(\mathcal{E}) \neq \mathbf{PA}(Y)$ and deduce a contradiction.

As in (9) we define the regression coefficient

$$\beta^{\text{pred},e}(S) := \text{argmin}_{\beta \in \mathbb{R}^p : \beta_k = 0 \text{ if } k \notin S} E(Y^e - X^e\beta)^2.$$

We then look for sets $S \subseteq \{1, \ldots, p\}$ such that for all $e_1, e_2 \in \mathcal{E}$

$$\beta^{\text{pred},e_1}(S) = \beta^{\text{pred},e_2}(S) \quad \text{and} \quad R^{e_1}(S) \stackrel{d}{=} R^{e_2}(S),$$

with $R^{e_1}(S) := Y^{e_1} - X^{e_1}\beta^{\text{pred},e_1}(S)$ and $R^{e_2}(S) := Y^{e_2} - X^{e_2}\beta^{\text{pred},e_2}(S)$ ("constant beta" and "same error distribution"). If $S(\mathcal{E}) \neq \mathbf{PA}(Y)$, then there must be a set $S \not\supseteq \mathbf{PA}(Y)$ whose null hypothesis is correct and that satisfies $\beta^{\text{pred},e}(S) \neq \beta^{\text{pred},e}(S^*) = \gamma^*$. This set S leads to the following residuals for $e = 1$:

$$R^1(S) = Y^1 - \sum_{k=2}^{p+1} \beta^{\text{pred},1}(S)_k X_k^1 = \sum_{k=2}^{p+1} \alpha_k X_k^1 + \varepsilon_1^1,$$

with $\alpha_k := \gamma_k^* - \beta^{\text{pred},1}(S)_k = \gamma_k^* - \beta^{\text{pred},e}(S)_k$ for any $e \in \mathcal{E}$ and $\alpha_k \neq 0$ for some (possibly more than one) $k \in \{2, \ldots, p+1\}$.

Among the set of *all* nodes (or variables) X_k^1 that have non-zero α_k, we consider a "youngest" node $X_{k_0}^1$ with the property that there is no directed path from this node to any other node with non-zero α_k. We further consider experiment e_0 with $\mathcal{A}^{e_0} = \{k_0\}$. This yields

$$R^1(S) = \alpha_{k_0} X_{k_0}^1 + \sum_{k=2, k\neq k_0}^{p+1} \alpha_k X_k^1 + \varepsilon_1^1 \quad \text{and} \tag{36}$$

$$R^{e_0}(S) = \alpha_{k_0} a_{k_0}^{e_0} + \sum_{k=2, k\neq k_0}^{p+1} \alpha_k X_k^1 + \varepsilon_1^1, \tag{37}$$

Since $E(X_{k_0}^1) \neq a_{k_0}^{e_0}$, $R^{e_0}(S)$ and $R^1(S)$ cannot have the same distribution. This yields a contradiction.

F.2. Proof of Theorem 2 (ii)

PROOF. As before we obtain equations (36) and (37) for a "youngest" node $X^1_{k_0}$ among all nodes with non-zero α_{k_0} and an experiment e_0 with $\mathcal{A}^{e_0} = \{k_0\}$. We now iteratively use the structural equations in order to obtain

$$R^1(S) = \alpha_{k_0}\varepsilon^1_{k_0} + \sum_{k=1,k\neq k_0}^{p+1} \tilde{\alpha}_k \varepsilon^1_k \quad \text{and} \tag{38}$$

$$R^{e_0}(S) = \alpha_{k_0} A^e_{k_0}\varepsilon^1_{k_0} + \sum_{k=1,k\neq k_0}^{p+1} \tilde{\alpha}_k \varepsilon^1_k. \tag{39}$$

Since all ε^e_k are jointly independent and $E(A^{e_0}_{k_0})^2 \neq 1$, $R^1(S)$ and $R^{e_0}(S)$ cannot have the same distribution. This contradicts the fact that the null hypothesis (5) is correct for S. The proof works analogously for the shifted noise distributions.

F.3. Proof of Theorem 2 (iii)

PROOF. We start as before and obtain analogously to equations (38) and (39) the equations

$$R^1(S) = \alpha_{k_0}\varepsilon^1_{k_0} + \sum_{k=1,k\neq k_0}^{p+1} \tilde{\alpha}_k \varepsilon^1_k \quad \text{and}$$

$$R^2(S) = \alpha_{k_0} A_{k_0}\varepsilon^1_{k_0} + \sum_{k=1,k\neq k_0}^{p+1} \tilde{D}_k \varepsilon^1_k,$$

where the $\tilde{D}_k$ are continuous functions of the random variables $A_s, s \in \{2,\ldots,p+1\}\setminus\{k_0\}$ and $\beta^{e=2}_{j,s}, j, s \in \{2,\ldots,p+1\}$ (and therefore random variables themselves). $R^1(S)$ and $R^2(S)$ are supposed to have the same distribution. It follows from Cramér's theorem [Cramér, 1936] that $A_{k_0}\varepsilon^1_{k_0}$ must be normally distributed. But then it follows that

$$\begin{aligned} E[(A_{k_0})^4]\, E[(\varepsilon^1_{k_0})^4] &= E[(A_{k_0}\varepsilon^1_{k_0})^4] = 3E[(A_{k_0}\varepsilon^1_{k_0})^2]^2 \\ &= 3E[(A_{k_0})^2]^2\, E[(\varepsilon^1_{k_0})^2]^2 = E[(A_{k_0})^2]^2\, E[(\varepsilon^1_{k_0})^4] \end{aligned}$$

and therefore

$$\mathrm{Var}(A^2_{k_0}) = 0$$

which means $P[A_{k_0} \in \{-c, c\}] = 1$ for some constant $c \geq 0$. This contradicts the assumption that A_{k_0} has a density.

F.4. Proof of Theorem 3

PROOF. The proof follows directly from Lemma 1 (see below) and the fact that faithfulness is satisfied with probability one [Spirtes et al., 2000, Theorem 3.2]. Assume that the null hypothesis (10) is accepted for S with $S^* \setminus S \neq \emptyset$. Lemma 1 implies that with probability one, we have $\alpha_{k_0} \neq 0$, where α is defined as in (40). (Otherwise, we construct a new SEM by replacing the equation for Y with $Y_{k_0} := \sum_{k\in S^*\setminus\{k_0\}} \gamma^*_k X_k + \varepsilon_1$ and removing all equations for the descendants of Y. Equation (41) then reads a violation of faithfulness since there is a path between k_0 and Y_{k_0} via nodes in $S^* \setminus S$ that is unblocked

given $S \setminus \{k_0\}$.) But if $\alpha_{k_0} \neq 0$, we can use exactly the same arguments as in the proof of Theorem 2.

LEMMA 1. *Assume that the joint distribution of $(X_1, \ldots, X_{p+1})$ is generated by a structural equation model* (19) *with all non-zero parameters $\beta_{j,k}$ and σ_j^2 drawn from a joint density w.r.t. Lebesgue measure. Let X_{k_0} denote a youngest parent of target variable $Y = X_1$. Let S be a set with $S^* \setminus S \neq \emptyset$, that is, some of the true causal parents are missing in the set S. Consider the residuals*

$$\begin{aligned} Res(Y) &= \sum_{k \in S^*} \gamma_k^* X_k - \sum_{k \in S} \beta^{\text{pred},1}(S)_k X_k + \varepsilon_1^1 \\ &= \sum_{k \in S^*} \alpha_k X_k + \sum_{k \notin S^*} \alpha_k \varepsilon_k^1 \end{aligned} \tag{40}$$

where the second equation is obtained by iteratively using the structural equations except the ones for the parents S^* of Y.
Then for almost all parameter values, we have: $\alpha_{k_0} = 0$ implies $k_0 \in S$ and

$$X_{k_0} \perp Y_{k_0} \,|\, X_{\tilde{S} \setminus \{k_0\}}, \tag{41}$$

where $Y_{k_0} := \sum_{k \in S^* \setminus \{k_0\}} \gamma_k^* X_k + \varepsilon_1$ and $\tilde{S} := S \cap \mathbf{ND}(k_0)$ with $\mathbf{ND}(k_0)$ being the non-descendants of k_0.

PROOF. With probability one, we have $\gamma_{k_0}^* \neq 0$. Hence, $\alpha_{k_0} = 0$ can happen only if $k_0 \in S$ or S contains a descendant of X_{k_0} (otherwise $\alpha_{k_0} = \gamma_{k_0}^* \neq 0$). We will now show that in fact $k_0 \in S$ must be true. Let the random vector X_S contain all variables X_k with $k \in S$ and let it be topologically ordered such that if X_{k_2} is a descendant of X_{k_1}, it appears after X_{k_1} in the vector X_S. Assume now that S contains a descendant of X_{k_0}. W.l.o.g., we can assume that the $|S|$-entry of X_S (i.e. its last component) is a "youngest" descendant X_s of X_{k_0} in S, that is, there is no directed path from X_s to any other descendant of X_{k_0} in S. The entry $(|S|, |S|)$ of the matrix $\left(E X_S^{1\,t} X_S^1\right)$ is the only entry depending (additively) on the parameter σ_s^2, we call this entry d. With

$$\left(E X_S^{1\,t} X_S^1\right) =: \begin{pmatrix} A & b \\ b^T & d \end{pmatrix}$$

it follows

$$\left(E X_S^{1\,t} X_S^1\right)^{-1} = \begin{pmatrix} A^{-1} + \frac{A^{-1} b b^T A^{-1}}{d - b^T A^{-1} b} & \frac{A^{-1} b}{d - b^T A^{-1} b} \\ \frac{b^T A^{-1}}{d - b^T A^{-1} b} & \frac{1}{d - b^T A^{-1} b} \end{pmatrix} =: \begin{pmatrix} A^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{d - b^T A^{-1} b}\, C$$

Observe that $\left(E X_S^{1\,t} X_S^1\right)$ is non-singular with probability one (if the matrix is non-singular, the full covariance matrix over $(X_2, \ldots, X_{p+1})$ is non-singular, too) and

$$\beta^{\text{pred},1}(S) = \left(E X_S^{1\,t} X_S^1\right)^{-1} \xi$$

for $\xi := E X_S^{1\,t} Y^1 \neq 0$ (otherwise $\beta^{\text{pred},1}(S)$ would be zero and thus $\alpha_{k_0} = \gamma_{k_0}^* \neq 0$).

According to formula (40) and $\alpha_{k_0} = 0$, computing the linear coefficients $\beta^{\text{pred},1}(S)$ and subsequently using the true structural equations, leads to the following relationship between the true coefficients $\beta_{j,k}$ and γ^*:

$$\gamma_{k_0}^* = \eta_S^t \, \beta^{\text{pred},1}(S),$$

where η_S depends on the true coefficients $\beta_{j,k}$ and is constructed in the following way: the i-th component of η_S is obtained by multiplying the path coefficients between X_{k_0} and X_i. For example, the two directed paths $X_{k_0} \to X_5 \to X_3 \to X_i$ and $X_{k_0} \to X_5 \to X_i$, lead to the corresponding ith entry $\eta_{S,i} = \beta^1_{5,k_0}\beta^1_{3,5}\beta^1_{i,3} + \beta^1_{5,k_0}\beta^1_{i,5}$. All non-descendants of k_0 have a zero entry in η_S, k_0 itself has the entry one in η_S if $k_0 \in S$ (we will see below that this must be the case). But then, we have:

$$\gamma^*_{k_0} = \eta_S^t \; \beta^{\text{pred},1}(S) = \eta_S^t \; \left(E {X_S^1}^t X_S^1\right)^{-1} \xi = \eta_S^t \; \begin{pmatrix} A^{-1} & 0 \\ 0 & 0 \end{pmatrix} \; \xi + \frac{1}{d - b^T A^{-1} b} \; \eta_S^t \; C \; \xi. \tag{42}$$

If $X_s \neq X_{k_0}$ then ξ does not depend on σ_s^2 (it does if $X_s = X_{k_0}$). We must then have that $\eta_S^t C \xi = 0$ since otherwise it follows from (42) that

$$d = b^T A^{-1} b \; + \; \frac{\eta_S^t \; C \; \xi}{\gamma^*_{k_0} - \eta_S^t \begin{pmatrix} A^{-1} & 0 \\ 0 & 0 \end{pmatrix} \xi},$$

which can happen only with probability zero (it requires a "fine-tuning" of the parameter σ_s^2; note that d is depending on σ_s^2).

But if $\eta_S^t C \xi = 0$ then $\gamma^*_{k_0} = (\eta_1 \cdots \eta_{|S|-1}) A^{-1} (\xi_1, \cdots \xi_{|S|-1}) = \eta^t_{\tilde{S}_1} \; \beta^{\text{pred},1}(\tilde{S}_1)$ with $\tilde{S}_1 := S \setminus \{s\}$, an equation analogue to the first part of (42). We can now repeat the same argument for $\tilde{S}_1$ (assume that $\tilde{S}_1$ contains a descendant of k_0, then consider the youngest descendant of k_0 in $\tilde{S}_1 \ldots$) and obtain $\tilde{S}_2$. After ℓ iterations, we obtain $\gamma^*_{k_0} = \eta^t_{\tilde{S}} \; \beta^{\text{pred},1}(\tilde{S})$, where $\tilde{S} := \tilde{S}_\ell$ does not contain any descendant of k_0. The only non-zero entry of $\eta_{\tilde{S}}$ is the one for k_0 (otherwise all remaining $\eta_{\tilde{S}}$ entries would be zero which implies $\gamma^*_{k_0} = 0$).

We have thus shown that $k_0 \in S$ and that $\beta^{\text{pred},1}(\tilde{S})_{k_0} = \gamma^*_{k_0}$ with $\tilde{S} := S \cap \mathbf{ND}(k_0)$. We obtain (41) with the following argument: regressing Y on $\tilde{S}$ yields a regression coefficient $\gamma^*_{k_0}$ for X_{k_0}; thus, regressing $Y_{k_0} = Y - \gamma^*_{k_0} X_{k_0}$ on $\tilde{S}$ yields a regression coefficient zero for X_{k_0}.

G. Experimental settings for numerical studies

We sample n_{obs} data points from an observational and n_{int} data points from an interventional setting ($|\mathcal{E}| = 2$). We first sample a directed acyclic graph with p nodes that is common to both scenarios. In order to do so, we choose a random topological order and then connect two nodes with a probability of $k/(p-1)$. This leads to an average degree of k. Given the graph structure, we then sample non-zero linear coefficients with a random sign and a random absolute value between a lower bound $lb^{e=1}$ and an upper bound $ub^{e=1} = lb^{e=1} + \Delta_b^{e=1}$. We consider normally distributed noise variables with a random variance between $\sigma^2_{\min}$ and $\sigma^2_{\max}$. We can then sample the observational data set ($e = 1$).

For the interventional setting ($e = 2$), we choose simultaneous noise interventions (Section 4.2.2) with the extension of changing linear coefficients, that is for $j \in \mathcal{A}$ (where even $\mathcal{A}$ is random and can include the later target of interest Y), we have $\varepsilon_j^{e=2} = A_j \varepsilon_j^{e=1}$ and (possibly) $\beta_{j,s}^{e=2} \neq \beta_{j,s}^{e=1}$. The set $\mathcal{A}$ of intervened nodes contains either a single node or a fraction θ of nodes. We chose A_j to be uniformly distributed random variables that take values between a_{min} and $a_{min} + \Delta_a$. The linear coefficients $\beta_{j,s}^{e=2}$ are chosen either equal to $\beta_{j,s}^{e=1}$ or according the same procedure with corresponding bounds $lb^{e=2}$ and $ub^{e=2}$.

All parameters were sampled independently for each of the scenarios, uniformly in a given range that is shown below in brackets (or with given probability for discrete parameters). (1) The number n_{obs} of samples in the observational data is chosen uniformly from $\{100, 200, 300, 400, 500\}$. (2) The number n_{int} of samples in intervention data is chosen uniformly from $\{100, 200, 300, 400, 500\}$. (3) The number p of nodes in the graph is chosen uniformly from $\{5, 6, 7, \ldots, 40\}$. (4) The average degree k of the graph is chosen uniformly from $\{1, 2, 3, 4\}$. (5) The lower bound $lb^{e=1}$ is chosen uniformly from $\{0.1, 0.2, \ldots, 2\}$. (6) The maximal difference $\Delta_b^{e=1}$ between largest and smallest coefficients is chosen uniformly from $\{0.1, 0.2, \ldots, 1\}$. (7) The minimal noise variance $\sigma^2_{\min}$ is chosen uniformly from $\{0.1, 0.2, \ldots, 2\}$ and (8) the maximal noise variance $\sigma^2_{\max}$ uniformly from $\{0.1, 0.2, \ldots, 2\}$, yet at least equal to $\sigma^2_{\max}$. (9) The lower bound $a_{j,min}$ for the noise multiplication is chosen uniformly from $\{0.1, 0.2, \ldots, 4\}$. (10) The difference Δ_a between upper and lower bound $a_{j,min}$ for noise multiplication is chosen to be zero with probability $1/3$ (which results in fixed coefficients) and otherwise uniformly from $\{0.1, 0.2, \ldots, 2\}$. (11) The interventional coefficients are chosen to be identical ($\beta_{j,s}^{e=2} = \beta_{j,s}^{e=1}$) with probability $2/3$, otherwise they are chosen uniformly between $lb^{e=2}$ and $ub^{e=2}$. (12) The lower bound $lb^{e=2}$ for new coefficients under interventions is chosen as the smaller value of two uniform values in $\{0.1, 0.2, \ldots, 2\}$ and (13) the upper bound $ub^{e=2}$ for new coefficients under interventions as the corresponding larger value. (14) With probability $1/6$ we intervene only on one (randomly chosen) variable, that is $|\mathcal{A}| = 1$. (15) Otherwise, the inverse fraction $1/\theta$ is chosen uniformly from $\{1.1, 1.2, \ldots, 3\}$, that is the fraction of intervened nodes varies between $\theta = 1/3$ and $\theta = 1/1.1$.