ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2016)
78, *Part* 5, *pp.* 947–1012

Causal inference by using invariant prediction: identification and confidence intervals

Jonas Peters

Max Planck Institute for Intelligent Systems, Tübingen, Germany, and Eidgenössiche Technische Hochschule Zürich, Switzerland

and Peter Bühlmann and Nicolai Meinshausen

Eidgenössiche Technische Hochschule Zürich, Switzerland

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, May 11th, 2016,* Professor C. Leng *in the Chair*]

Summary. What is the difference between a prediction that is made with a causal model and that with a non-causal model? Suppose that 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 (e.g. 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 discovery; Causal inference; Confidence intervals; Invariant prediction

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 (see Dawid (2000) and Rubin (2005)), or structural equation modelling (see Bollen (1989), Robins *et al*. (2000) and Pearl (2009)) and graphical modelling (see Lauritzen and Spiegelhalter (1988), Greenland *et al*. (1999) and Spirtes *et al*. (2000)), where Pearl (2009) provides a nice overview. Richardson and Robins (2013) made a connection between the frameworks by using single-world intervention graphs.

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

Address for correspondence: Jonas Peters, Seminar for Statistics, Department of Mathematics, Eidgenössiche Technische Hochschule Zürich, Rämistrasse 101, 8092 Zürich, Switzerland.
E-mail: peters@stat.math.ethz.ch

 1369–7412/16/78947

correct Markov equivalence class on the basis of 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), and finally to infer the identifiable causal effects or to provide some bounds (Maathuis *et al*., 2009; VanderWeele and Robins, 2010). More recently, within the framework of structural equation models (SEMs), interesting work has been done for fully identifiable structures exploiting additional restrictions such as non-Gaussianity (Shimizu *et al*., 2006), non-linearity (Hoyer *et al*., 2009; Peters *et al*., 2014) or equal error variances (Peters and Bühlmann, 2014). Janzing *et al*. (2012) exploited 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 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 thank a referee for suggesting this succinct description of the main idea.) We exploit, in other words, that the conditional distribution of the target variable of interest (which is often also termed the 'response variable'), given the complete set of corresponding direct causal predictors, must 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) and Pearl (2009), section 1.3.2). Whereas 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 us 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 (see He and Geng (2008)) belong to such a scenario. For known intervention targets, Cooper and Yoo (1999) incorporated the intervention effects as mechanism changes (Tian and Pearl, 2001) into a Bayesian framework and Hauser and Bühlmann (2015) modified the greedy equivalence search (Chickering, 2002) for perfect interventions. Our framework does not require knowledge of the location of interventions. For this setting, Eaton and Murphy (2007) used intervention nodes with unknown children and Tian and Pearl (2001) considered changes in marginal distributions, whereas Dawid (2012, 2015) made use of different regimes for a decision theoretic approach. In contrast with 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 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 the framework of feedback models (Lauritzen and Richardson, 2002; Mooij *et al*., 2011; Hyttinen *et al*., 2012), although a constrained form of feedback is allowed. It is an open question whether our approach could be generalized to include general feedback models.

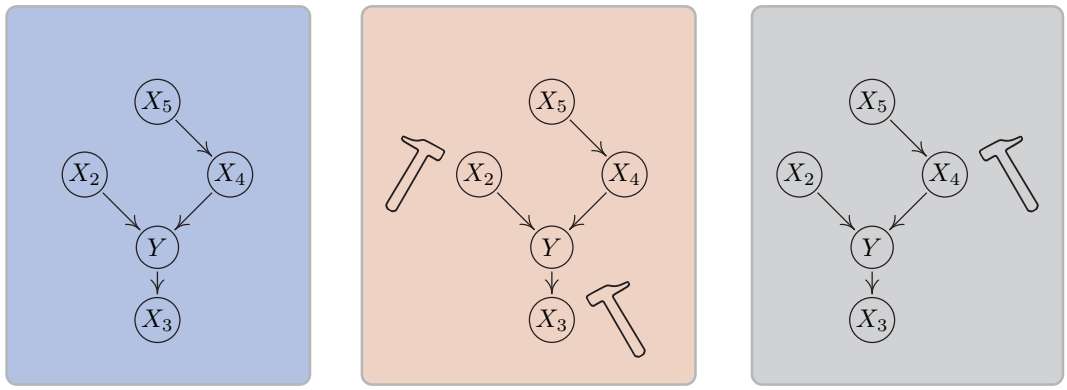

Fig. 1. Example including three environments (the invariance (1) and (2) holds if we consider $S^* = \{X_2, X_4\}$; considering indirect causes instead of direct causes (e.g. $\{X_2, X_5\}$) or an incomplete set of direct causes (e.g. $\{X_4\}$) may not be sufficient to guarantee invariant prediction): (a) environment $e = 1$; (b) environment $e = 2$; (c) environment $e = 3$

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 independent and identically distributed 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. Although the environments $e \in \mathcal{E}$ can be created by precise experimental design for X^e (e.g. by randomizing 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,

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

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

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. Expressions (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 Fig. 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 simplicity, we shall mostly focus on a linear model with a target or response variable and various predictor variables, where expression (1) is unchanged and expression (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^*, i.e. $S^* := \{k;\ \gamma_k^* \neq 0\}$. Assumption 1 in Section 2 summarizes all requirements. Proposition 1 shows that SEMs 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 assume only 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 shall 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 SEMs (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 randomized studies are infeasible to conduct.

We believe that the method's underlying invariance principle is quite 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 SEMs are given in Section 4. The relationship to instrumental variables and the behaviour in the presence of hidden variables are discussed in Section 5. We shall discuss extensions to the non-linear model (2) in Section 6.1 and extensions 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 is a vector of coefficients $\gamma^* = (\gamma_1^*, \ldots, \gamma_p^*)^{\mathrm{T}}$ with support $S^* = \{k : \gamma_k^* \neq 0\} \subseteq \{1, \ldots, p\}$ that satisfies,

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

where $\mu \in \mathbb{R}$ is an intercept term, ε^e is random noise with mean 0, 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 shall always assume that an intercept μ is added to model (3). To simplify the notation, we shall from now on refrain from writing the intercept down explicitly. We discuss the invariance assumption with the help of some examples in Figs 1 and 2; see also Appendix A for another artificial example. (Each panel of Fig. 2 shows the distribution of a target gene activity Y (on the y-axis), conditional on a predictor gene activity X (shown on the x-axis). Blue crosses show observational data and red dots show interventional data. The interventions do not occur on any of the genes shown. The conditional distribution of Y, given X, is not invariant for the examples in Figs 2(a) and 2(b), whereas invariance cannot be rejected for the two examples in Figs 2(c) and 2(d). Take the example of Fig. 2(c). 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 the set $S = \{\text{YPL273W}\}$ fulfils the invariance assumption (3), at least approximately.)

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, 2007, 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 to keep it purely mathematical. Proposition 1 establishes a link to causality by showing that the parents of Y in an SEM satisfy assumption 1. In other words, the variables that have a direct causal effect on Y in an SEM form a set S^* for which assumption 1 is satisfied. This must not necessarily be true for the variables that have an (in)direct effect on Y, i.e. the ancestors of Y. However, the set S^* is not necessarily unique (see the discussion). For a given set of experimental conditions $\mathcal{E}$, there can be multiple vectors γ^* that satisfy condition (3). For example, if only observational data are available, i.e.

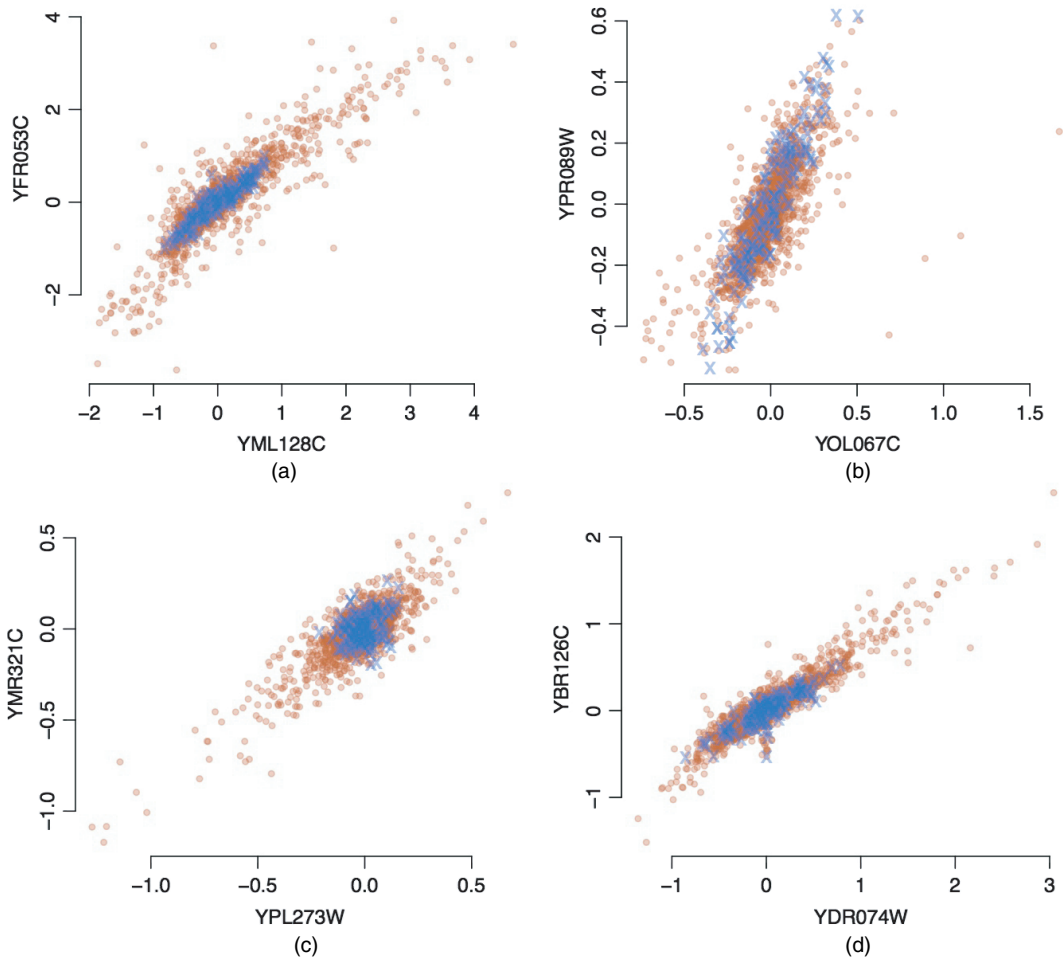

Fig. 2. Some examples from the gene knockout experiments in Kemmeren *et al.* (2014), which are discussed in more detail in Section 7.2

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 condition (3) for every subset $S^* \subseteq \{1, \ldots, p\}$. The inference that 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 condition (3); see Section 4 for settings with complete identifiability.

It is perhaps easiest to think about the example of a linear SEM, as defined in Section 4.1; see also Fig. 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 condition (3).

Proposition 1. Consider a linear SEM, 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 $\varepsilon_1^e \perp\!\!\!\perp \{\varepsilon_j^e; j \in \mathrm{AN}(1)\}$ for all environments $e \in \mathcal{E}$, where AN(1) are the ancestors of Y. Then assumption 1 holds for the parents of Y, namely $S^* = \mathrm{PA}(1)$, 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, also), 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 $Y^e = \sum_{j \in \mathrm{PA}(1)} \beta_{1,j} X_j^e + \varepsilon_Y^e$ for all $e \in \mathcal{E}$, where $\varepsilon_Y^e = \varepsilon_1^e$ is independent of $X_{\mathrm{PA}}(1)$ and has the same distribution for all $e \in \mathcal{E}$. Thus, assumption 1 holds.

We remark that proposition 1 can be generalized 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 because of an intervention, no other structural equations 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 in 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 shall leave it as an open question to what extent the approach can be generalized 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 condition (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 model (3). We therefore define for $\gamma \in \mathbb{R}^p$ and $S \subseteq \{1, \ldots, p\}$ the null hypothesis $H_{0,\gamma,S}(\mathcal{E})$ as

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

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).

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

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

(b) 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 expression (8) (the second equation in expression (6) can be ignored for now) and we define the intersection over an empty index set as the empty set; see the discussion. Under assumption 1, $H_{0,\gamma^*,S^*}(\mathcal{E})$ is true and therefore S^* are plausible causal predictors, i.e. $H_{0,S^*}(\mathcal{E})$ is correct, also. 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 that we define below. Furthermore, the set of identifiable causal predictors under interventions $\mathcal{E}$ is growing monotonically if we enlarge the set $\mathcal{E}$,

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

In particular, if $|\mathcal{E}| = 1$ (for example, there are 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 about which variables are causal.

In Section 4, we examine conditions for SEMs (see proposition 1) under which $S(\mathcal{E})$ is identical to the parents of Y and 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 an 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 fulfils 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 fulfil the null becomes smaller. We call vectors that fulfil 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,

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

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 expression (5) can be simplified. Writing

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

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 $S \subseteq \{1, \ldots, p\}$,

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

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 shall 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 in the following *generic method for invariant prediction*.

Step 1: for each set $S \subseteq \{1, \ldots, p\}$, test whether $H_{0,S}(\mathcal{E})$ holds at level α (we shall discuss concrete examples later).
Step 2: set $\hat{S}(\mathcal{E})$ as

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

Step 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 Fig. 2. Taking the example in Fig. 2(c), we cannot reject $H_{0,S}(\mathcal{E})$ for $S = \{\text{YPL273W}\}$. Hence we can see already from this plot that $\hat{S}(\mathcal{E})$ is either empty or that $\hat{S}(\mathcal{E}) = \{\text{YPL273W}\}$. 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 expression (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 that 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 (see the discussion for a more general version of theorem 1).

Theorem 1. Assume that the estimator $\hat{S}(\mathcal{E})$ is constructed according to expression (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, $\sup_{P:H_{0,S}(\mathcal{E}) \text{ true}} P\{H_{0,S}(\mathcal{E}) \text{ rejected}\} \leqslant \alpha$. 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

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

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

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

Proof. The first property follows immediately since

$$P\{\hat{S}(\mathcal{E}) \subseteq S^*\} = P\left(\bigcap_{S:H_{0,S}(\mathcal{E}) \text{ not rejected}} S \subseteq S^* \right) \geqslant P\{H_{0,S^*}(\mathcal{E}) \text{ not rejected}\} \geqslant 1 - \alpha,$$

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

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

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^* and 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 SEM, i.e. for $\gamma^* = \beta_{1,\cdot}$ and $S^* = \mathrm{PA}(1)$ in the notation of proposition 1.

Remark 1.

(a) We obtain the following empirical version of expression (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.

(b) In expression (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 $\hat{S}(\mathcal{E})$ by using the first equality in expression (15), instead of expression (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 expressions (12) and (13) is a test for $H_{0,S}(\mathcal{E})$ that is valid at level α for any given set of variables $S \subseteq \{1, \ldots, p\}$ and thus implies that

$$P\{H_{0,S^*}(\mathcal{E}) \text{ rejected}\} \leqslant \alpha.$$

To specify a concrete procedure and to 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, $\Sigma_{e\in\mathcal{E}} n_e = n$.

We now propose a way to construct such a test, but we 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 expression (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 circumflex. As mentioned above, under assumption 1, for $S = S^*$, the regression effects are identical to the causal coefficients: for all $e \in E$,

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

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

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

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

3.1.1. Method I: invariant prediction using test on regression coefficients

Step 1: for each $S \subseteq \{1, \ldots, p\}$ and $e \in \mathcal{E}$ do as follows.

(a) 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 ordinary least squares 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.

(b) Under Gaussian errors, if expression (16) is true for a set S, then (Chow, 1960)

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

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

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

letting $\mathbf{1}_n$ 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 expression (17) is below $\alpha/|\mathcal{E}|$ for any $e \in \mathcal{E}$.

Step 2: this step is the same as in the generic algorithm, using expression (12).
Step 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 $\beta^{\mathrm{pred}}(S)$ when using all data simultaneously. For simplicity, we shall use a rectangular confidence region where the constraint for $\beta^{\mathrm{pred}}(S)_k$ is identically 0 if $k \notin S$ and for coefficients in S given by $(\hat{\beta}^{\mathrm{pred}}(S))_S \pm t_{1-\alpha/(2|S|),n-|S|-1} \hat{\sigma}\, \mathrm{diag}\{(\mathbf{X}_S^{\mathrm{T}} \mathbf{X}_S)^{-1}\}$, 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. This procedure has some shortcomings. For example, the inversion of the covariance matrix in expression (17) might be too slow if we have to search many sets and the sample size is large. We 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 non-parametric procedures. Lastly, it is not straightforward to extend this approach to classification and non-linear 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.

3.1.2. *Method II: invariant prediction using fast(er) approximate test on residuals*

Step 1: for each $S \subseteq \{1, \ldots, p\}$ and $e \in \mathcal{E}$ do as follows.

(a) Fit a linear regression model on all data to obtain an estimate $\hat{\beta}^{\mathrm{pred}}(S)$ of the optimal coefficients using set S of variables for linear prediction in regression. Let $R = Y - X \hat{\beta}^{\mathrm{pred}}(S)$.

(b) 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 combining 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.

Step 2: this step is the same as in the generic algorithm, using expression (12).
Step 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 shall use rectangular confidence regions, exactly as in step 3 of method I.

Besides a computational advantage, this method can also easily be extended to non-linear 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

(a) observational data,
(b) a known intervention of a certain type at a known variable,
(c) a random intervention at an unknown and random location or
(d) observational data in a changed environment.

We have used data pooling in methods I and II to obtain 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 so for the procedures that were given above), we can pool data from various $e \in \mathcal{E}$.

To make it more precise, assume that 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 $\mathcal{J}$, i.e.

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

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, i.e. $S(\mathcal{E})$ 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 an 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, i.e. there is no directed path from Y to U, for example as it precedes Y chronologically. (This is similar to 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,

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

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 an SEM (Pearl (2009), theorem 1.4.1). The confidence intervals remain valid but the implication on (partial) identifiability of the causal predictors remains 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), non-linearities (Hoyer *et al.*, 2009; Peters *et al.*, 2014; Bühlmann *et al.*, 2014), equal error variances (Peters and Bühlmann, 2014) or shared information between the 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 non-linearity.

3.4. Computational requirements

The construction of the confidence regions for the set of plausible causal coefficients and the identifiable causal predictors requires us 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 superexponentially 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 obtain a non-empty confidence set) we shall 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 that we are willing to make the assumption that the set of causal variables is at most $s < p$. Then we must just 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) is an adaptation of the confidence interval that was 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), the 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, $P(S^* \subseteq \hat{B}) \geqslant 1 - \alpha$, 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 have been 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 simplicity, we shall 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 structural equation models

We consider linear Gaussian SEMs (e.g. Wright (1921) and Duncan (1975)). We assume that each element $e \in \mathcal{E}$ represents a different interventional set-up. 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 $\varepsilon_j^1 \sim^{\mathrm{IID}} \mathcal{N}(0, \sigma_j^2)$, $j = 1, \ldots, p+1$. The corresponding directed graph is obtained by drawing arrows from variables X_k^1 on the right-hand side of equation (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 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

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

Here, we adapt the usual notation of graphical models (e.g. Lauritzen (1996)). For example, we write $\mathrm{PA}(j)$, $\mathrm{DE}(j)$, $\mathrm{AN}(j)$ and $\mathrm{ND}(j)$ 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, also:

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

Assumption 1 states that the influence of the causal predictors remains the same under interventions, i.e. $Y^e = X^e\gamma^* + \varepsilon_1^e$ for $\gamma^* = (\beta_{1,2}^1, \ldots, \beta_{1,p+1}^1)^{\mathrm{T}}$ and $\varepsilon_1^e =^{\mathrm{d}} \varepsilon_1^1$ 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 those in the observational setting (19). Within this setting, we now define various sorts of intervention.

4.2. Interventions

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

4.2.1. Do interventions

Do types of interventions correspond to the classical do operation from Pearl (2009), for example. In the eth 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 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

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

The do interventions correspond to fixing the intervened variable at a specific value. The following two types of intervention 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 model (20), for $e \neq 1$, as follows:

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

and

$$\varepsilon_j^e \stackrel{\mathrm{d}}{=} \begin{cases} \varepsilon_j^1 & \text{if } j \notin \mathcal{A}^e, \\ A_j^e \varepsilon_j^1 & \text{if } j \in \mathcal{A}^e, \end{cases}$$

or

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

The factors A_j^e and the shifts C_j^e are considered as random but may be constant with probability 1. 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 $k \in \mathrm{DE}(j)$.

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, i.e. $|\mathcal{E}| = 2$ and, for all $j \in \{2, \ldots, p+1\}$,

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

and

$$\varepsilon_j^{e=2} \stackrel{\mathrm{d}}{=} A_j \varepsilon_j^{e=1}$$

or

$$\varepsilon_j^{e=2} \stackrel{\mathrm{d}}{=} \varepsilon_j^{e=1} + C_j.$$

The random variables $A_j \geqslant 0$ are assumed to have a distribution that is absolutely continuous with respect to Lebesgue measure with $E(A_j^2) < \infty$ 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 knockout experiments with off-target effects in biology (e.g. Jackson *et al*. (2003) and 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 expressions (19) and (20) with interventions. Then, with $S(\mathcal{E})$ as in expression (6), all causal predictors are identifiable, i.e.

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

if one of the following three assumptions is satisfied.

(a) 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, i.e. for each $j \in \{2, \ldots, p+1\}$ there is an experiment e with $\mathcal{A}^e = \{j\}$.
(b) 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 $E(C_j^e) \neq 0$ or $0 < \mathrm{var}(C_j^e) < \infty$.
(c) 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 equation (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 1. In general, the conditions that were given above are not necessary. The following remarks, however, provide two specific counterexamples that show the necessity of some conditions.

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

$$Y^1 = X_2^1 + X_3^1 + \varepsilon_Y, \qquad X_2^1 = \varepsilon_2, \qquad X_3^1 = -X_2^1 + \varepsilon_3;$$

$e=2$,

$$Y^2 = X_2^2 + X_3^2 + \varepsilon_Y, \qquad X_2^2 = 0, \qquad X_3^2 = -X_2^2 + \varepsilon_3;$$

$e=3$,

$$Y^3 = X_2^3 + X_3^3 + \varepsilon_Y, \qquad X_2^3 = \varepsilon_2, \qquad X_3^3 = 0,$$

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).

(b) If we check only 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 $e=2$: $e=1$,

$$X^1 = \varepsilon_X, \qquad Y^1 = X^1 + \varepsilon_Y;$$

$e=2$,

$$X^2 = A\varepsilon_X, \qquad Y^2 = X^2 + \varepsilon_Y,$$

with $E(A)^2 = 1$ and $\varepsilon_X, \varepsilon_Y \sim^{\mathrm{IID}} \mathcal{N}(0,1)$. Then we cannot identify the correct set of parents $\mathrm{PA}(Y) = \{X\}$ 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, part (c).

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 counterexamples 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 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 expressions (19) and (20) with all non-zero parameters drawn from a joint density with respect to Lebesgue measure. Let X_{k_0} be a youngest parent of target variable $Y = X_1$, i.e. 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}, i.e. $|\mathcal{E}| = 2$ and $\mathcal{A}^{e=2} = \{k_0\}$ (k_0 does not need to be known).

Then, with probability 1, all causal predictors are identifiable, i.e.

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

if one of the following two assumptions is satisfied.

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

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

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 expect that more identifiability results of similar type can be derived in specific settings. Theorem 3 shows that intervening 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 assumptions made. Intervening on descendants of Y will, in contrast, rule out only 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 no randomized experiments are 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, 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 and g are unknown real-valued functions and η and ε are random-noise variables in p dimensions and one dimension respectively. As is commonly done for SEMs, we require that the noise variables H, η, ε and I are jointly independent. Fig. 3 shows an example of an SEM that satisfies equations (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 model (23) are in order. First, the model requires that I has no direct effect on Y, which is a 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, model (23) allows for feedback between X and Y, i.e. the corresponding graph in an SEM is not required to be acyclic. If feedback exists, the solutions are typically understood to be stable equilibrium solutions of equations (23) but we shall here require only that the solutions satisfy equations (23).

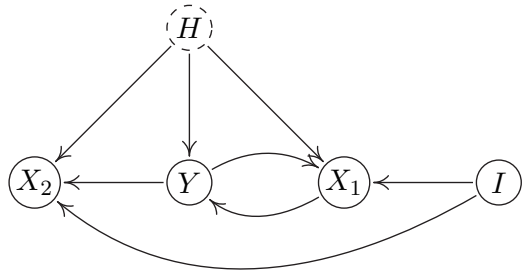

Fig. 3. Graph of a model that satisfies equation (23) with $X := (X_1, X_2)$: variable Y has a direct causal effect on X_2 only, whereas there is a feedback between Y and X_1

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^*, two limitations with this approach are as follows.

(a) The conditional expectation $E(X|I)$ is not allowed to be constant for $I \in \{0, 1\}$.
(b) The predictor X must be univariate for a univariate instrument I, i.e. $p = 1$ 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 that was 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 by 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,

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

with unknown causal coefficients $\gamma^* \in \mathbb{R}^p$ and unknown function $g : \mathbb{R}^q \times \mathbb{R} \to \mathbb{R}$ and the distribution $g(H^e, \varepsilon^e)$ is identical for all e in $\mathcal{E}$.

In the absence of hidden variables and feedback loops, 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 or feedback loops, we cannot require independence of the residuals and the causal predictors X_{S^*} but we can adapt the null hypothesis $H_{0,S}$ in expression (5) to the weaker form

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

Testing the null hypothesis (25) is computationally more challenging than for the corresponding null hypothesis in the absence of hidden confounders (5). In contrast with expression (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 the null hypothesis (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 $H_{0,S,\text{hidden}}(\mathcal{E})$ and define by analogy with expression (12) the estimated set of causal predictors as

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

Then the coverage property follows immediately in the following sense.

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

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

Proof. The hypothesis $H_{0,S,\text{hidden}}(\mathcal{E})$ 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 (a) and (b) for identification under a two-stage instrumental variable approach are violated. Thus, an 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 generalized to include hidden variables that violate assumption (b) (iii) in proposition 4, e.g. by replacing expression (5) with the null hypothesis (25). As a possible implemention of the general approach we must therefore test hypothesis (25) for every set $S \subseteq \{1, \ldots, p\}$. 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 hypothesis (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).

Proposition 2 discussed the coverage of the estimator (26). 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 model (23) a shift in the variance of X under $I = 1$ compared with $I = 0$ observations:

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

where the p-dimensional mean 0 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 $H_{0,S,\text{hidden}}(\mathcal{E})$ has power 1 against any alternative, then

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

A proof is given in Appendix E. Note that the causal variables and coefficients can be identified for model (27), even though the model violates the above-mentioned assumptions (a) and (b) 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 $E(X|I)$ between $I = 1$ and $I = 0$. Although the identifiability $P\{\hat{S}(\mathcal{E}) = S^*\} \geqslant 1 - \alpha$ depends in this specific model (27) on the full rank assumption of Z and this assumption is difficult to verify in practice, we stress again that the coverage property $P\{\hat{S}(\mathcal{E}) \subseteq S^*\} \geqslant 1 - \alpha$ is guaranteed for the general case (23).

6. Further extensions and model misspecification

6.1. Non-linear 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 non-linear setting (2). The equivalent null hypothesis to expression (5) is then

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

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

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

The 'only if' part is immediate and for the 'if' part we can use a similar idea to that in Peters *et al.* (2014), proposition 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 cumulative distribution function of $Y^e | X_S^e = a$.

As in the linear case, we can consider an 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 non-linear SEMs

$$Y_j = f_j(X_{\text{PA}(j)}, \varepsilon_j), \qquad j = 1, \ldots, p+1,$$

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

Proof. Again, the proof is immediate. Only the case with hidden variables requires an argument. From the SEM, we are given $Y^e = f(X^e_{S^0}, X^e_{S^0}, \tilde{\varepsilon}^e)$ 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 $Y^e = g(X^H_{S^0}, \varepsilon^e)$ 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 that we have a test for the null hypothesis $H_{0,S,\text{nonlin}}(\mathcal{E})$. Then, testing all possible sets $S \subseteq \{1, \ldots, p\}$, we can obtain a confidence set for S^* in a similar way to that in the linear setting (15) by

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

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

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

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

Then, we can construct tests for the null hypothesis (28) that are similar to that in the linear case. Analogously to method I in Section 3.1, we can perform non-linear 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 hypothesis (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 that were 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, also. 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 non-linear case can then be weakened to

$$H'_{0,S,\text{nonlin}}(\mathcal{E}): \exists \mathcal{E}' \subseteq \mathcal{E} \text{ with } |\mathcal{E}'| \geqslant |\mathcal{E}| - V \text{ such that } \forall\, e, f \in \mathcal{E}' \text{ the conditional distribution } Y^e | X_S^e = x \text{ and } Y^f | X_S^f = x \text{ are identical } \forall\, x \text{ such that both conditional distributions are well defined.} \tag{32}$$

The null hypothesis $H'_{0,S^*,\text{nonlin}}$ 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 hypothesis (29) can be extended as a test for the weaker null hypothesis (32) by testing all subsets $\mathcal{E}'$ with $|\mathcal{E}'| \geqslant |\mathcal{E}| - V$ at level α, e.g. using a test for hypothesis (28), and rejecting hypothesis (32) only if we can reject all such subsets. We can then treat $H_{0,S,\text{nonlin}}(\mathcal{E})$ 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, non-linear 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 expression (15) as

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

Our approach still satisfies the coverage property $P\{\hat{S}(\mathcal{E}) \subseteq S^*\} \geqslant 1 - \alpha$ 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 an 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 $P\{\hat{S}(\mathcal{E}) \subseteq S_c^*\} \geqslant 1 - \alpha$ is still valid) or H_{0,S_c^*} is no 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 power equal to 1 to detect the alternative hypotheses. Thus, the desired coverage property $P\{\hat{S}(\mathcal{E}) \subseteq S^*_c\} \geqslant 1 - \alpha$ is still valid, even though the method will now have no power to detect the causal variables. It could happen, however, that there is 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 is a different set $\tilde{S}$, for which $H_{0,\tilde{S}}$ 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 an instrumental variable type of setting from educational research.

7.1. Simulation experiments

For the simulations, we generate data from randomly chosen linear Gaussian 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 that were 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 positive results) or whether the correct parents were recovered (true positive results).

For the invariant prediction method proposed, 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 with 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* (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 greedy equivalence search searches greedily over equivalence classes of graph structures to maximize a penalized likelihood score. Here, we apply greedy equivalence search 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 GIES considers soft interventions (at node j) where the conditional $p(x_j|x_{\mathrm{PA}(j)})$ 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 pro-

cedure over interventional Markov equivalence classes maximizing a penalized 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) *GIES with unknown intervention targets*: to obtain a more fair comparison with 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 (LINGAMs)* (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 LINGAMs 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

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

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

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

One may want to model the pooled data set as coming from an SEM of the form

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

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, also. We can neglect this violation of LINGAMs and apply the method nevertheless. There is no theoretical result which would justify LINGAMs 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 familywise 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 Fig. 4 for all methods. The probability of success, $P\{\hat{S}(\mathcal{E}) = S^*\}$, is shown in Fig. 5.

The success probabilities show some interesting patterns. First, there is (as expected) not a method that performs uniformly best over all scenarios. However, *regression* and *marginal regression* are dominated across all 100 scenarios by GIES (both with known and unknown interventions), LINGAMs and the proposed *invariant prediction*). Among the 100 settings, there were three where greedy equivalence search performed best on the given criterion, 14 where GIES (with known interventions) performed best, 54 for LINGAMs and 23 where the proposed invariant prediction was optimal for exact recovery. There is no clear pattern about 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

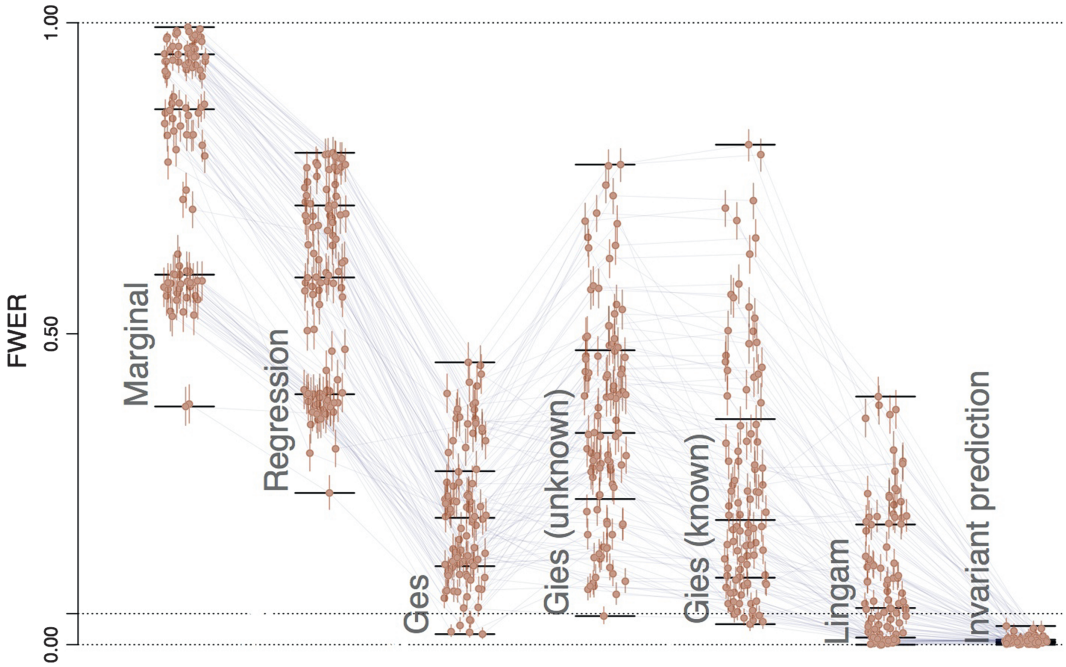

Fig. 4. Probability of erroneous selections $P\{\hat{S}(\mathcal{E}) \not\subseteq S^*\}$ (FWER) for the methods considered, including the proposed invariant prediction to the right: the figure is otherwise generated analogously to Fig. 5; the dotted line indicates the 0.05-level at which the invariant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

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 Fig. 4 is very clear, however. 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 FWERs 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 FWER (and the associated confidence intervals) is the key advantage of the proposed invariant prediction.

7.2. Gene perturbation experiments

7.2.1. Data set

We applied our method to a yeast (*Saccharomyces cerevisiae*) data set (Kemmeren *et al.*, 2014). Genomewide messenger ribonucleic acid expression levels in yeast were measured and we therefore have data for $p = 6170$ genes. There are $n_{\text{obs}} = 160$ 'observational' samples of wild types and $n_{\text{int}} = 1479$ data points for the 'interventional' setting where each of them 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 finding 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 Fig. 6. (We additionally require that the intervention on gene 5954 appears

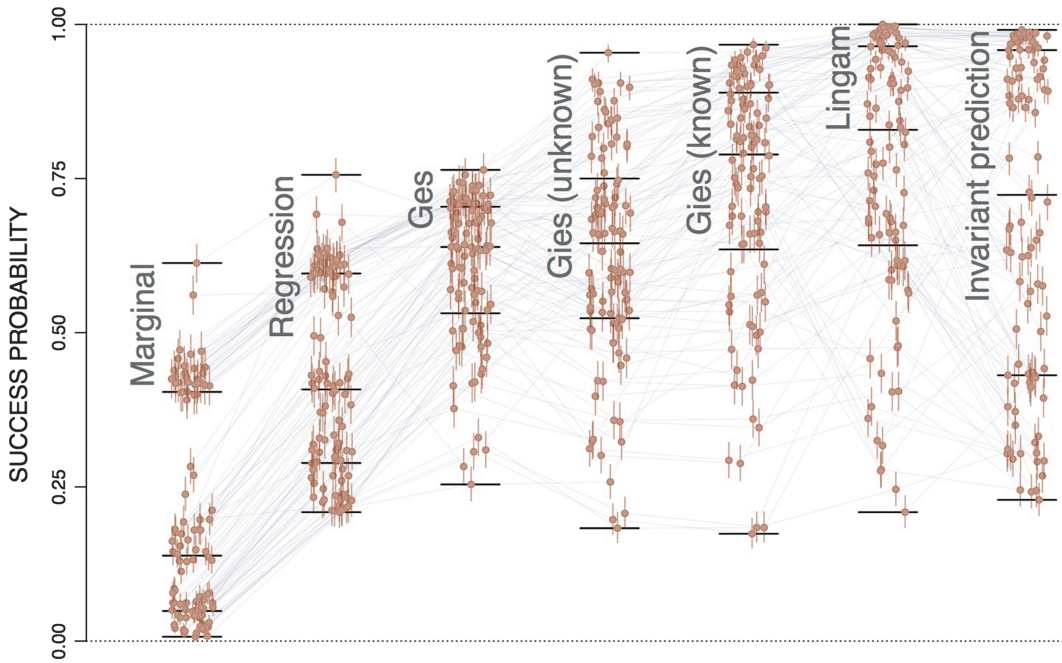

Fig. 5. Probability of success, defined as $P\{\hat{S}(\mathcal{E}) = S^*\}$ for various methods, including our new proposed invariant prediction in the rightmost column: each dot within a column (the x-offset within a column is uniform) corresponds to one of the 100 simulation scenarios; the dot's height shows the empirical probability of success over 1000 simulations and the small bars indicate 95% confidence for the true success probability; identical scenarios are connected by grey lines; for each method, the maximal and minimal values along with the quartiles of each distribution are indicated by horizontal bars

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 so for 38 out of the 1479 interventions.) With this criterion, there are about 9.2% relevant effects, which corresponds to the proportion of true positive findings for a random guessing method.

7.2.2. 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 $n_{\text{int}}\, p$ 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 $3p$ 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.

7.2.3. Goodness of fit and p-values

If we would like to avoid making a single mistake on the data set with high probability $1 - \alpha$,

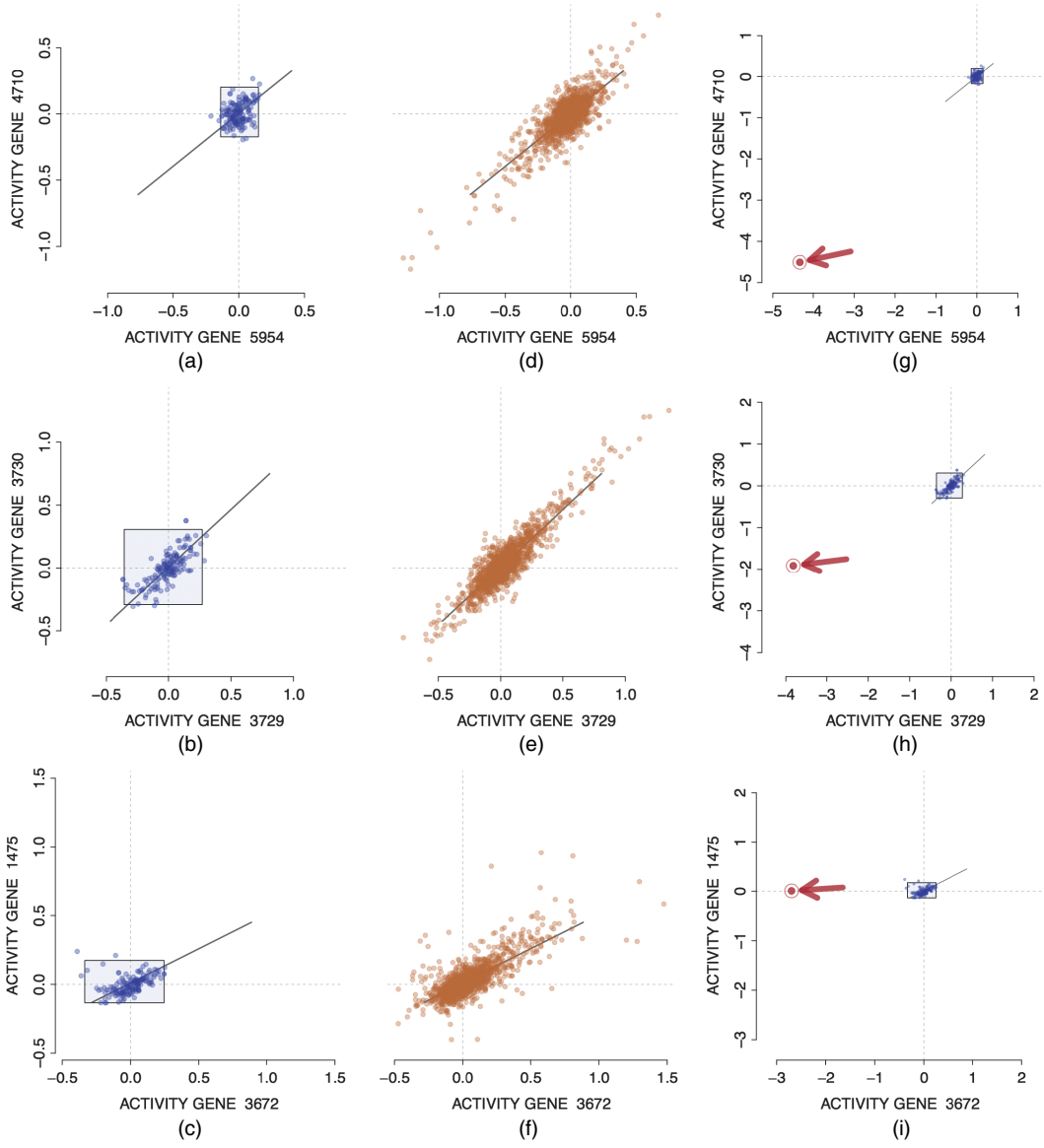

Fig. 6. (a)–(c) Observational data, (d)–(f) interventional data (that are neither using interventions on the target variable itself nor using interventions on the possible causal predictors of the target variable examined) and (g)–(i) test data (with the 1%–99% quantile range of the observational data shown as a shaded box as in (a)–(c)): (d) interventions on genes other than 5954 and 4710; (e) interventions on genes other than 3729 and 3730; (f) interventions on genes other than 3672 and 1475; (g) intervention on gene 5954; (h) intervention on gene 3729; (i) intervention on gene 3672

we can set the level of significance for each gene to α/n_{int}, using a Bonferroni correction to take into account the $p = 6170$ genes. We work with $\alpha = 0.05$ 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 to be violated, and the implications have been partially discussed earlier in Section 6. To guard further against false positive results 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.

7.2.4. 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 10 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 non-negligible probability of losing the few informative data points in the subsampling process. For a computational speed-up we consider only subsets of size 3 or smaller as candidate sets S. Furthermore, we retain only 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.

7.2.5. Comparisons

As alternative methods we consider the 'intervention calculus when the DAG is absent' algorithm 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 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 the LINGAM algorithm (Shimizu *et al.*, 2011) on this data set.

7.2.6. Results

The method proposed (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 shows the three pairs that obtained the highest rank, i.e. smallest p-values. (The two data sets in Figs 6(a)–6(c) and 6(d)–6(f) 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 Figs 6(a)–6(f); we cannot reject the hypothesis that the regression is different for observation and interventional data here. In Figs 6(g)–6(i) 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 positive findings) in the following sense: after successfully intervening on the predicted cause, the target gene shows reduced activity; the third suggested pair is unsuccessful (a false positive finding) since the intervention reduces the activity of the cause but the target gene remains as active as in the observational data.) The rows in Fig. 6 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 them were effectively used for training because of the cross-validation scheme discussed. Many outlying points in the interventional training data of the false positive finding

(Fig. 6(f)) 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 $B = 3$.

To compare with other methods (none of which provide a measure of significance), we always consider the eight highest ranked pairs. Table 1 summarizes the results. In this data set, the alternative methods could exceed random guessing.

To test sensitivity of the results to the chosen implementation details of the method, the variable preselection and the goodness-of-fit cut-off have also all been varied (e.g. by using the lasso instead of boosting as preselection and using a cut-off of 0.1 instead of 0.01). For method II, variable selection with the lasso instead of boosting leads to a true positive rate of 0.63 (five out of eight). Choosing the goodness-of-fit cut-off at 0.01 rather than 0.1 leads to true positive rates of 0.43 (nine out of 21) for boosting and 0.47 (eight out of 17) for the lasso. Method I without forcing eight decisions leads to a true positive rate of 0.75 (three out of four) for boosting and 1.00 (one out of one) for the lasso. Choosing the goodness-of-fit cut-off at 0.01 rather than 0.1 leads to true positive rates of 0.86 (six out of seven) for boosting and 0.75 (three out of four) for the 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 positive findings (e.g. two out of eight for method II in Table 1, there are at least the following options:

(a) noise fluctuations,
(b) non-linearities,
(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 result but is—by chance—classified as false positive by our criterion (see Section 7.2.1 above).

Missing causal variables in the prescreening by boosting or the lasso falls under category (c). We control error (a) and have provided arguments why errors (b) and (c) will lead to rejection of the whole model rather than lead to false positive results. Lowering the goodness-of-fit threshold seemed indeed to lead to more spurious results, as expected from the discussion in Section 6.3 earlier. Validating a potential issue with the experiment as in reason (d) is beyond our possibilities. We could address error (e) if we had access to multiple repetitions of the intervention

Table 1. Number of true effects among the strongest eight effects that have been found in the interventional test data†

	Numbers for the following methods:				*Marginal correlation*		*Random guessing*
	Method I	*Method II*	*GIES*	*IDA*	*Observed*	*Pooled*	
Number of true positive findings (out of 8)	6	6	2	2	1	2	2 (95% quantile) 3 (99% quantile) 4 (99.9% quantile)

†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.

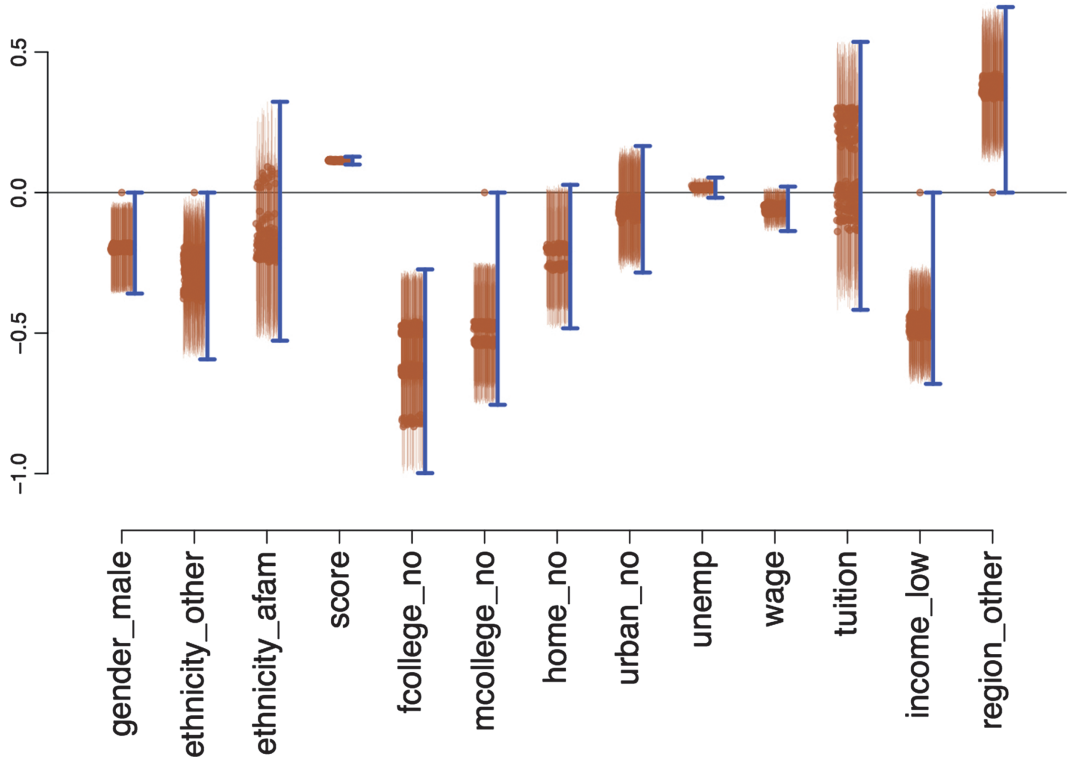

Fig. 7. 90% confidence intervals ( ) for the influence of various variables on the probability of receiving a Bachelor of Arts degree (or higher): of all 8192 possible sets S, we accept 1565 sets (the empty set is not accepted as the probability of receiving a degree is sufficiently different for people within a close distance to a 4-year college and further away)

experiments. We provide code that reproduces the results on our home page. The code may result in minor variations due to updates in the package.

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 that pupils received. We make a binary distinction into whether pupils received a Bachelor of Arts degree or higher (equivalent to at least 16 years of education in the classification that was used in Stock and Watson (2003)) and ask whether we can identify a causal predictive model that allows us to forecast whether students will receive a Bachelor of Arts 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 data set into two parts in the sense of expression (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 must just be independent of the noise in the outcome variable (it must be a non-descendant of the target), which seems satisfied in this data set as the distance to the 4-year college precedes the educational attainment

chronologically. One set of observations is thus all pupils who live closer to a 4-year college than the median distance of 10 miles. The second set is 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 Bachelor of Arts degree or higher. We use the fast approximate method II of Section 3.1, with the suggested extension to logistic regression.

Fig. 7 shows the outcome of the analysis, which is also included as an example in the R package `InvariantCausalPrediction`. (In Fig. 7, 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.) 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 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 Bachelor of Arts 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 groups created, we can potentially have power to detect invariant causal prediction effects.

8. Discussion and future work

An advantage of causal predictors compared with non-causal predictors 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 SEMs 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 non-linear models, feedback models and increased computational efficiency in both the absence and the 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 partially received funding from the People Programme (Marie Curie Actions) of the European Union's seventh framework programme (FP7/2007–2013) under Research Executive Agency grant agreement 326496. The authors thank seven referees for their helpful comments on an earlier version of the manuscript and thank Alain

Hauser, Thomas Richardson, Bernhard Schölkopf and Kun Zhang for helpful discussions, as well as Niels Hansen and Joris Mooij for pointing out two small errors in the original version of the manuscript.

Appendix A: An example

We illustrate here in Fig. 8 the concepts and methodology which have been developed in Sections 2.1, 2.2 and 3. Fig. 8 shows an example of two environments whose data were generated from observational and interventional SEMs. (The first environment corresponds to the graph including the broken edge; the second environment corresponds to an intervention on X_3, the graph excluding the broken edge. Since the structural equation for Y is unchanged, the set $S^* = \{X_2, X_3\} = \mathrm{PA}(1)$ satisfies assumption 1; see proposition 1. We consider the set-up where we know neither S^* nor the SEMs (we do not even require the existence of such an SEM). Instead, we are given two finite samples (one from each environment) and provide an estimator $\hat{S}$ for S^*. In this example, the null hypothesis of invariant prediction is 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 that was described in Section 3.1.)

Appendix B: Hidden variables without confounding

We discuss first a generalization 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 SEM with variables $X_1 = Y, X_2, \ldots, X_p, X_{p+1}, H_1, \ldots, H_q$, where the $H_1, \ldots, H_q$ are unobserved hidden variables with mean 0.

Proposition 4. Consider a linear SEM including variables

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

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

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

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

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

where ε_Y is independent of X_{S^0} and $H_{S^0_H}$.

Then, by choosing $\gamma^* = \{\beta_{Y,j}, j \in S^0\}$ and $S^* = S^0$, assumption 1 holds if one of the following conditions (a) or (b) is satisfied.

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

$$Y^e = \sum_{j \in S^0} \beta_{Y,j} X^e_j + \varepsilon^e_Y \qquad \text{for all } e \in \mathcal{E}, \tag{33}$$

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

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

$$Y^e = \sum_{j \in S^0} \beta_{Y,j} X^e_j + \sum_{k \in S^0_H} \kappa_{Y,k} H^e_k + \varepsilon^e_Y \qquad \text{for all } e \in \mathcal{E}, \tag{34}$$

where $\Sigma_{k \in S^0_H} \kappa_{Y,k} H^e_k + \varepsilon^e_Y$ is independent of $X^e_{S^0}$ and has the same distribution with mean 0 for all $e \in \mathcal{E}$. This holds under the following conditions:

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

(ii) there are no interventions on Y, on nodes in S^0_H or on any ancestor of S^0_H;

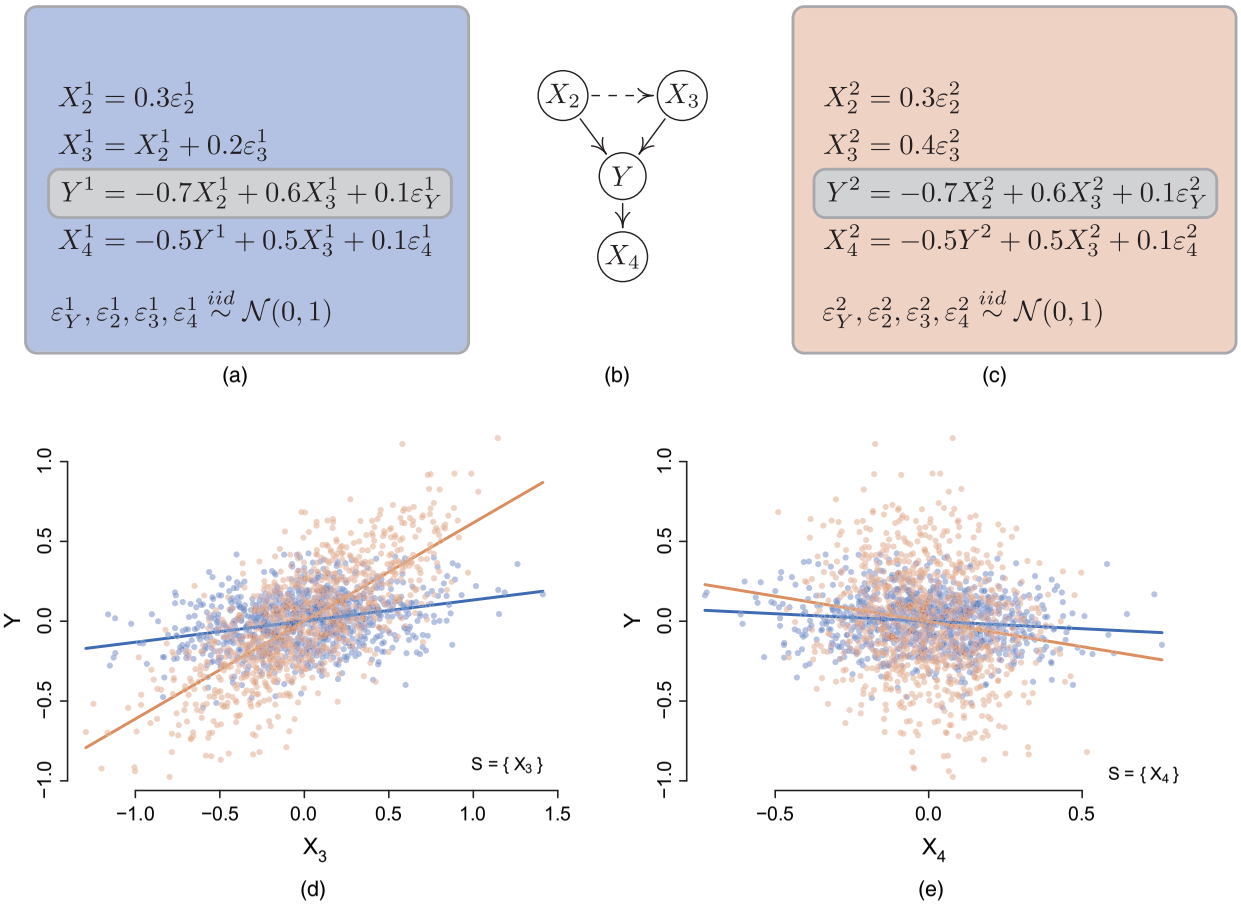

Fig. 8. (a)–(c) Example of two SEMs entailing the two distributions corresponding to two environments (a) $e = 1$ and (c) $e = 2$; (d) for $S = \{X_3\}$, for example, the linear regression coefficients differ in the two environments; for (e) $S = \{X_4\}$, the regression coefficients seem similar but the set is rejected because of varying variances of the residuals (●, data from environment $e = 1$; ●, data from environment $e = 2$); 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 $k_0 = 3$)

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

Proof. Assumption 1 follows immediately from expression (33) or (34). From the definition of the interventions, as described in Section 4.2, the justification for result (33) follows and hence the claim assuming condition (a). When invoking condition (b), we show now that conditions (i)–(iii) imply result (34) and the required conditions. Because of conditions (i) and (ii), 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 condition (iii).

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

Proposition 5. Consider an 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 with respect to the augmented graph. Then

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

In particular, this proposition still holds under model misspecification when for some do interventions, for example, $S^0 = \text{PA}(Y) \cap \{X_2, \ldots, X_{p+1}\}$ does not satisfy $H_{0,S,\text{nonlin}}(\mathcal{E})$ (28); Fig. 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 shall not be able to find any set S that satisfies null hypothesis (28).

Proof. We first note that $H_{0,S,\text{nonlin}}(\mathcal{E})$ (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 $\tilde{S} \subseteq \text{AN}(Y) \cap \{X_2, \ldots, X_{p+1}\}$ 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, i.e. $W \notin \text{AN}(Y)$ 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 $\text{AN}(Y)$, which implies that $S \subseteq \text{AN}(Y)$.) We now prove that for

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

we have that 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 have finished because removing

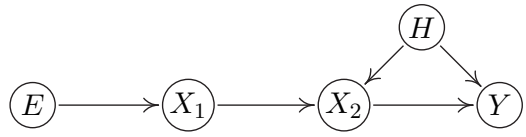

Fig. 9. Graph corresponding to a model misspecification in the sense that the assumptions of proposition 1 and assumption (b), part (iii), of proposition 4 are not satisfied: indeed, we find that $H_{0,S}$ is violated for $S = S^0 := \{X_2\}$, and, since $H_{0,S}$ is satisfied for both $S = \{X_1, X_2\}$ 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 $\text{AN}(Y)$ of Y; see proposition 5

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.

Appendix 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 causal predictors remains invariant. (Here, we discuss the non-linear setting and therefore develop a result that corresponds to remark 2 rather than proposition 1.) Although other formulations may be possible, also, we adopt the counterfactual language that was introduced by Richardson and Robins (2013) who referred to finest fully randomized causally interpretable structured tree graphs (Robins, 1986). We further consider the non-linear version (29) of invariant prediction; see also remark 2.

Similarly to Richardson and Robins (2013), definition 1, we consider random variables $\mathbf{V} := (X_1 = Y, X_2, \ldots, X_p, X_{p+1})$ 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 $j \in \{1, \ldots, p+1\}$. We further assume the following conditions:

(a) 'consistency and recursive substitution' (Richardson and Robins (2013), equation (14)) (condition 1) and
(b) 'finest fully randomized causally interpretable structured tree graphs independence' (Richardson and Robins (2013), equation (17)) (condition 2).

To ease the 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), page 21).

Proposition 6. Consider random variables $\mathbf{V} := (X_1 = Y, X_2, \ldots, X_p, X_{p+1})$ and denote the causes of Y by $\mathbf{P} := \mathrm{PA}(1)$. For each environment $e \in \mathcal{E}$ consider a set $\mathbf{R}^e \subseteq \mathbf{V} \setminus \{Y\}$ of treatment variables and an assignment $\tilde{\mathbf{r}}^e$, i.e. $X_j^e := X_j(\tilde{\mathbf{r}}^e)$. Assuming conditions 1 and 2, i.e. a finest fully randomized causally interpretable structured tree graphs model, we have that

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

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

We have already seen in appendix B that we can allow for some hidden variables, i.e. condition 2 can be relaxed further.

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

$$\begin{aligned}
Y(\tilde{\mathbf{r}}^e) | \mathbf{P}(\tilde{\mathbf{r}}^e) = \mathbf{q} &= Y(\tilde{\mathbf{r}}^e) | (\mathbf{P} \setminus \mathbf{R}^e)(\tilde{\mathbf{r}}^e) = \mathbf{q}_{\mathbf{P} \setminus \mathbf{R}^e}, (\mathbf{P} \cap \mathbf{R}^e)(\tilde{\mathbf{r}}^e) = \tilde{\mathbf{q}}_{\mathbf{P} \cap \mathbf{R}^e} \\
&= Y(\tilde{\mathbf{r}}^e) | (\mathbf{P} \setminus \mathbf{R}^e)(\tilde{\mathbf{r}}^e) = \mathbf{q}_{\mathbf{P} \setminus \mathbf{R}^e} && (36) \\
&= Y(\tilde{\mathbf{r}}^e_{\mathrm{AN}(Y)}) | (\mathbf{P} \setminus \mathbf{R}^e)(\tilde{\mathbf{r}}^e_{\mathrm{AN}(\mathbf{P} \setminus \mathbf{R}^e)}) = \mathbf{q}_{\mathbf{P} \setminus \mathbf{R}^e} && (37) \\
&= Y | (\mathbf{P} \setminus \mathbf{R}^e) = \mathbf{q}_{\mathbf{P} \setminus \mathbf{R}^e}, (\mathbf{P} \cap \mathbf{R}^e) = \tilde{\mathbf{r}}^e_{\mathbf{P} \cap \mathbf{R}^e} && (38) \\
&= Y | (\mathbf{P} \setminus (\mathbf{R}^e \cup \mathbf{R}^f)) = \mathbf{q}_{\mathbf{P} \setminus (\mathbf{R}^e \cup \mathbf{R}^f)}, (\mathbf{P} \cap \mathbf{R}^f) = \mathbf{q}_{\mathbf{P} \cap \mathbf{R}^f}, (\mathbf{P} \cap \mathbf{R}^e) = \tilde{\mathbf{r}}^e_{\mathbf{P} \cap \mathbf{R}^e} \\
&= Y | (\mathbf{P} \setminus (\mathbf{R}^e \cup \mathbf{R}^f)) = \mathbf{q}_{\mathbf{P} \setminus (\mathbf{R}^e \cup \mathbf{R}^f)}, (\mathbf{P} \cap \mathbf{R}^f) = \tilde{\mathbf{r}}^f_{\mathbf{P} \cap \mathbf{R}^f}, (\mathbf{P} \cap \mathbf{R}^e) = \mathbf{q}_{\mathbf{P} \cap \mathbf{R}^e} && (39) \\
&= \ldots \\
&= Y(\tilde{\mathbf{r}}^f) | \mathbf{P}(\tilde{\mathbf{r}}^f) = \mathbf{q},
\end{aligned}$$

where for equation (36) we have used $(\mathbf{P} \cap \mathbf{R}^e)(\tilde{\mathbf{r}}^e) = \tilde{\mathbf{r}}^e_{\mathbf{P} \cap \mathbf{R}^e}$ and $\mathbf{q}_{\mathbf{P} \cap \mathbf{R}^e} = \tilde{\mathbf{r}}^e_{\mathbf{P} \cap \mathbf{R}^e}$ (otherwise equation (35) is not well defined). Equation (37) follows from condition 1 and equation (38) follows from the modularity

property (Richardson and Robins (2013), proposition 16). Equation (39) holds because $\mathbf{q}_{\mathbf{P}\cap\mathbf{R}^f} = \tilde{\mathbf{r}}^f_{\mathbf{P}\cap\mathbf{R}^f}$. All equality signs should be understood as holding in distribution.

Appendix E: Proof of proposition 3

Proof. The residuals $Y - X\gamma$ for $\gamma \in \mathbb{R}^p$ are given by $g(H,\varepsilon) + (\gamma^* - \gamma)f(H,\eta) + Z\mathbf{1}_{I=1}(\gamma^* - \gamma)$. 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, ε, η and 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 $H_{0,S,\text{hidden}}(\mathcal{E})$ will be rejected for $S \neq S^*$, whereas the true null hypothesis $H_{0,S^*,\text{hidden}}(\mathcal{E})$ is accepted with probability at least $1 - \alpha$ by construction of the test and the result follows by the definition of $\hat{S}$ in expression (26).

Appendix F: Proofs of Section 4.3

F.1. Proof of theorem 2, part (a)

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

As in expression (9) we define the regression coefficient

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

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

$$\begin{aligned}\beta^{\text{pred},e_1}(S) &= \beta^{\text{pred},e_2}(S),\\ R^{e_1}(S) &\overset{\text{d}}{=} R^{e_2}(S),\end{aligned}$$

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 $S(\mathcal{E}) \neq \text{PA}(Y)$, then there must be a set $S \not\supseteq \text{PA}(Y)$ 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 $e = 1$:

$$R^1(S) = Y^1 - \sum_{k=2}^{p+1} \beta^{\text{pred},1}(S)_k X^1_k = \sum_{k=2}^{p+1} \alpha_k X^1_k + \varepsilon^1_1,$$

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) $k \in \{2,\dots,p+1\}$.

Among the set of *all* nodes (or variables) X^1_k that have non-zero α_k, we consider a 'youngest' node $X^1_{k_0}$ 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

$$R^1(S) = \alpha_{k_0} X^1_{k_0} + \sum_{k=2,k\neq k_0}^{p+1} \alpha_k X^1_k + \varepsilon^1_1 \tag{40}$$

and

$$R^{e_0}(S) = \alpha_{k_0} a^{e_0}_{k_0} + \sum_{k=2,k\neq k_0}^{p+1} \alpha_k X^1_k + \varepsilon^1_1. \tag{41}$$

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

F.2. Proof of theorem 2, part (b)

Proof. As before we obtain equations (40) and (41) 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 to obtain

$$R^1(S) = \alpha_{k_0}\varepsilon^1_{k_0} + \sum_{k=1,k\neq k_0}^{p+1} \tilde{\alpha}_k \varepsilon^1_k \tag{42}$$

and

$$R^{e_0}(S) = \alpha_{k_0} A^e_{k_0} \varepsilon^1_{k_0} + \sum_{k=1, k\neq k_0}^{p+1} \tilde{\alpha}_k \varepsilon^1_k. \tag{43}$$

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, part (c)

Proof. We start as before and obtain analogously to equations (42) and (43) the equations

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

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

$$\operatorname{var}(A^2_{k_0}) = 0$$

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

F.4. Proof of theorem 3

Proof. The proof of theorem 3 follows directly from lemma 1 (see below) and the fact that faithfulness is satisfied with probability 1 (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 1, we have $\alpha_{k_0} \neq 0$, where α is defined as in equation (44) in lemma 1. (Otherwise, we construct a new SEM by replacing the equation for Y with $Y_{k_0} := \Sigma_{k\in S*\setminus\{k_0\}} \gamma^*_k X_k + \varepsilon_1$ and removing all equations for the descendants of Y. Equation (45) 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 an SEM (19) with all non-zero parameters $\beta_{j,k}$ and σ^2_j drawn from a joint density with respect to Lebesgue measure. Let X_{k_0} denote a youngest parent of target variable $Y = X_1$. Let S be a set with $S^* \setminus S \neq \emptyset$, i.e. some of the true causal parents are missing in the set S. Consider the residuals

$$\begin{aligned} \operatorname{Res}(Y) &= \sum_{k\in S*} \gamma^*_k X_k - \sum_{k\in S} \beta^{\text{pred},1}(S)_k X_k + \varepsilon^1_1 \\ &= \sum_{k\in S*} \alpha_k X_k + \sum_{k\notin S*} \alpha_k \varepsilon^1_k \end{aligned} \tag{44}$$

where the second equation is obtained by iteratively using the structural equations except those for the parents S^* of Y.

Then, for almost all parameter values, we have $\alpha_{k_0} = 0$ implies that $k_0 \in S$ and

$$X_{k_0} \perp Y_{k_0} | X_{\tilde{S} \setminus \{k_0\}}, \tag{45}$$

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

Proof. With probability 1, 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 shall 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}. Without loss of generality, 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, i.e. 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 $(EX_S^{1\mathrm{T}} X_S^1)$ is the only entry depending (additively) on the parameter σ_s^2; we call this entry d. With

$$(EX_S^{1\mathrm{T}} X_S^1) =: \begin{pmatrix} A & b \\ b^{\mathrm{T}} & d \end{pmatrix}$$

it follows that

$$(EX_S^{1\mathrm{T}} X_S^1)^{-1} = \begin{pmatrix} A^{-1} + \dfrac{A^{-1} b b^{\mathrm{T}} A^{-1}}{d - b^{\mathrm{T}} A^{-1} b} & \dfrac{A^{-1} b}{d - b^{\mathrm{T}} A^{-1} b} \\ \dfrac{b^{\mathrm{T}} A^{-1}}{d - b^{\mathrm{T}} A^{-1} b} & \dfrac{1}{d - b^{\mathrm{T}} A^{-1} b} \end{pmatrix} =: \begin{pmatrix} A^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{d - b^{\mathrm{T}} A^{-1} b} C.$$

Observe that $(EX_S^{1\mathrm{T}} X_S^1)$ is non-singular with probability 1 (if the matrix is non-singular, the full covariance matrix over $(X_2, \ldots, X_{p+1})$ is non-singular, also) and

$$\beta^{\mathrm{pred},1}(S) = (EX_S^{1\mathrm{T}} X_S^1)^{-1} \xi$$

for $\xi := EX_S^{1\mathrm{T}} Y^1 \neq 0$ (otherwise $\beta^{\mathrm{pred},1}(S)$ would be 0 and thus $\alpha_{k_0} = \gamma_{k_0}^* \neq 0$).

According to formula (44) and $\alpha_{k_0} = 0$, computing the linear coefficients $\beta^{\mathrm{pred},1}(S)$, and subsequently using the true structural equations, leads to the following relationship between the true coefficients $\beta_{j,k}$ and γ^*:

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

where η_S depends on the true coefficients $\beta_{j,k}$ and is constructed in the following way: the ith 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 $\eta_{S,i} = \beta_{5,k_0}^1 \beta_{3,5}^1 \beta_{i,3}^1 + \beta_{5,k_0}^1 \beta_{i,5}^1$. All non-descendants of k_0 have a 0-entry in η_S; k_0 itself has the entry 1 in η_S if $k_0 \in S$ (we shall see below that this must be so). But, then, we have

$$\gamma_{k_0}^* = \eta_S^{\mathrm{T}} \beta^{\mathrm{pred},1}(S) = \eta_S^{\mathrm{T}} (EX_S^{1\mathrm{T}} X_S^1)^{-1} \xi = \eta_S^{\mathrm{T}} \begin{pmatrix} A^{-1} & 0 \\ 0 & 0 \end{pmatrix} \xi + \frac{1}{d - b^{\mathrm{T}} A^{-1} b} \eta_S^{\mathrm{T}} C \xi. \tag{46}$$

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 $\eta_S^{\mathrm{T}} C \xi = 0$ since otherwise it follows from equation (46) that

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

which can happen only with probability 0 (it requires a 'fine-tuning' of the parameter σ_s^2; note that d depends on σ_s^2).

But if $\eta_S^{\mathrm{T}} C \xi = 0$ then $\gamma_{k_0}^* = (\eta_1 \ldots \eta_{|S|-1}) A^{-1} (\xi_1, \ldots \xi_{|S|-1}) = \eta_{\tilde{S}_1}^{\mathrm{T}} \beta^{\mathrm{pred},1}(\tilde{S}_1)$ with $\tilde{S}_1 := S \setminus \{s\}$, an equation analogue to the first part of equation (46). 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 $\gamma_{k_0}^* = \eta_{\tilde{S}}^{\mathrm{T}} \beta^{\mathrm{pred},1}(\tilde{S})$, where $\tilde{S} := \tilde{S}_l$ does not contain any descendant of k_0. The only non-zero entry of $\eta_{\tilde{S}}$ is the entry for k_0 (otherwise all remaining $\eta_{\tilde{S}}$-entries would be 0, which implies that $\gamma_{k_0}^* = 0$).

We have thus shown that $k_0 \in S$ and that $\beta^{\mathrm{pred},1}(\tilde{S})_{k_0} = \gamma_{k_0}^*$ with $\tilde{S} := S \cap \mathrm{ND}(k_0)$. We obtain relationship

(45) 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 0 for X_{k_0}.

Appendix 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. 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 $\mathrm{lb}^{e=1}$ and an upper bound $\mathrm{ub}^{e=1} = \mathrm{lb}^{e=1} + \Delta_b^{e=1}$. 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, i.e., 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 to the same procedure with corresponding bounds $\mathrm{lb}^{e=2}$ and $\mathrm{ub}^{e=2}$.

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).

(a) The number n_{obs} of samples in the observational data is chosen uniformly from $\{100, 200, 300, 400, 500\}$.
(b) The number n_{int} of samples in intervention data is chosen uniformly from $\{100, 200, 300, 400, 500\}$.
(c) The number p of nodes in the graph is chosen uniformly from $\{5, 6, 7, \ldots, 40\}$.
(d) The average degree k of the graph is chosen uniformly from $\{1, 2, 3, 4\}$.
(e) The lower bound $\mathrm{lb}^{e=1}$ is chosen uniformly from $\{0.1, 0.2, \ldots, 2\}$.
(f) The maximal difference $\Delta_b^{e=1}$ between largest and smallest coefficients is chosen uniformly from $\{0.1, 0.2, \ldots, 1\}$.
(g) The minimal noise variance $\sigma^2_{\min}$ is chosen uniformly from $\{0.1, 0.2, \ldots, 2\}$.
(h) The maximal noise variance $\sigma^2_{\max}$ is chosen uniformly from $\{0.1, 0.2, \ldots, 2\}$, yet at least equal to $\sigma^2_{\max}$.
(i) The lower bound $a_{j,\min}$ for the noise multiplication is chosen uniformly from $\{0.1, 0.2, \ldots, 4\}$.
(j) The difference Δ_a between upper and lower bound $a_{j,\min}$ for noise multiplication is chosen to be 0 with probability $\frac{1}{3}$ (which results in fixed coefficients) and otherwise uniformly from $\{0.1, 0.2, \ldots, 2\}$.
(k) The interventional coefficients are chosen to be identical ($\beta_{j,s}^{e=2} = \beta_{j,s}^{e=1}$) with probability $\frac{2}{3}$; otherwise they are chosen uniformly between $\mathrm{lb}^{e=2}$ and $\mathrm{ub}^{e=2}$.
(l) The lower bound $\mathrm{lb}^{e=2}$ for new coefficients under interventions is chosen as the smaller value of two uniform values in $\{0.1, 0.2, \ldots, 2\}$.
(m) The upper bound $\mathrm{ub}^{e=2}$ for new coefficients under interventions is chosen as the corresponding larger value.
(n) With probability $\frac{1}{6}$ we intervene only on one (randomly chosen) variable, i.e. $|\mathcal{A}| = 1$.
(o) Otherwise, the inverse fraction $1/\theta$ is chosen uniformly from $\{1.1, 1.2, \ldots, 3\}$, i.e. the fraction of intervened nodes varies between $\theta = \frac{1}{3}$ and $\theta = 1/1.1$.

References

Aldrich, J. (1989) Autonomy. *Oxf. Econ. Pap.*, **41**, 15–34.
Andersson, S. A., Madigan, D. and Perlman, M. D. (1997) A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.*, **25**, 505–541.
Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables. *J. Am. Statist. Ass.*, **91**, 444–455.
Belloni, A., Chernozhukov, V. and Wang, L. (2011) Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, **98**, 791–806.
Bollen, K. A. (1989) *Structural Equations with Latent Variables*. New York: Wiley.
Bowden, R. J. and Turkington, D. A. (1990) *Instrumental Variables*, vol. 8. Cambridge: Cambridge University Press.

Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-dimensional Data: Methods, Theory and Applications*. New York: Springer.
Bühlmann, P., Peters, J. and Ernest, J. (2014) CAM: causal additive models, high-dimensional order search and penalized regression. *Ann. Statist.*, **42**, 2526–2556.
Bühlmann, P., Rütimann, P. and Kalisch, M. (2013) Controlling false positive selections in high-dimensional regression and causal inference. *Statist. Meth. Med. Res.*, **22**, 466–492.
Bühlmann, P. and Yu, B. (2003) Boosting with the *L*2-loss: regression and classification. *J. Am. Statist. Ass.*, **98**, 324–339.
Castelo, R. and Kocka, T. (2003) On inclusion-driven learning of Bayesian networks. *J. Mach. Learn. Res.*, **4**, 527–574.
Chickering, D. M. (2002) Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, **3**, 507–554.
Chow, G. C. (1960) Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, **28**, 591–605.
Cooper, G. and Yoo, C. (1999) Causal discovery from a mixture of experimental and observational data. In *Proc. 15th A. Conf. Uncertainty in Artificial Intelligence*, pp. 116–125. San Francisco: Morgan Kaufmann.
Cramér, H. (1936) Über eine Eigenschaft der normalen Verteilungsfunktion. *Math. Zeits.*, **41**, 405–414.
Dawid, A. P. (2000) Causal inference without counterfactuals. *J. Am. Statist. Ass.*, **95**, 407–424.
Dawid, A. P. (2007) Counterfactuals, hypotheticals and potential responses: a philosophical examination of statistical causality. In *Causality and Probability in the Sciences* (eds F. Russo and J. Williamson), pp. 505–532. London: College Publications.
Dawid, A. P. (2012) The decision-theoretic approach to causal inference. In *Causality: Statistical Perspectives and Applications* (eds C. R. Berzuini, A. P. Dawid and L. Bernardinelli), ch. 4, pp. 25–42. Chichester: Wiley
Dawid, A. P. (2015) Statistical causality from a decision-theoretic perspective. *A. Rev. Statist. Appl.*, **2**, 273–303.
Dawid, A. P. and Didelez, V. (2010) Identifying the consequences of dynamic treatment strategies: a decision-theoretic overview. *Statist. Surv.*, **4**, 184–231.
Didelez, V., Dawid, A. P. and Geneletti, S. (2006) Direct and indirect effects of sequential treatments. In *Proc. 22nd A. Conf. Uncertainty in Artifical Intelligence*, pp. 138–146. Corvallis: Association for Uncertainty in Artificial Intelligence Press.
Didelez, V., Meng, S. and Sheehan, N. A. (2010) Assumptions of IV methods for observational epidemiology. *Statist. Sci.*, **25**, 22–40.
Duncan, O. D. (1975) *Introduction to Structural Equation Models*. New York: Academic Press.
Durot, C., Groeneboom, P. and Lopuhaä, H. (2013) Testing equality of functions under monotonicity constraints. *J. Nonparam. Statist.*, **25**, 939–970.
Eaton, D. and Murphy, K. P. (2007) Exact Bayesian structure learning from uncertain interventions. In *Proc. 11th Int. Conf. Artificial Intelligence and Statistics*, pp. 107–114. JMLR.
Eberhardt, F. and Scheines, R. (2007) Interventions and causal inference. *Philos. Sci.*, **74**, 981–995.
Friedman, J. H. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Statist.*, **29**, 1189–1232.
Greenland, S., Pearl, J. and Robins, J. M. (1999) Causal diagrams for epidemiologic research. *Epidemiology*, **10**, 37–48.
Haavelmo, T. (1944) The probability approach in econometrics. *Econometrica*, **12**, suppl., S1–S115.
Hauser, A. and Bühlmann, P. (2012) Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.*, **13**, 2409–2464.
Hauser, A. and Bühlmann, P. (2015) Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *J. R. Statist. Soc.* B, **77**, 291–318.
He, Y.-B. and Geng, Z. (2008) Active learning of causal networks with intervention experiments and optimal designs. *J. Mach. Learn. Res.*, **9**, 2523–2547.
Hernán, M. and Robins, J. (2006) Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, **17**, 360–372.
Hoover, K. D. (1990) The logic of causal inference. *Econ. Philos.*, **6**, 207–234.
Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. and Hofner, B. (2010) Model-based boosting 2.0. *J. Mach. Learn. Res.*, **11**, 2109–2113.
Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J. and Schölkopf, B. (2009) Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, vol. 21, pp. 689–696. Red Hook: Curran Associates.
Hyttinen, A., Eberhardt, F. and Hoyer, P. O. (2012) Learning linear cyclic causal models with latent variables. *J. Mach. Learn. Res.*, **13**, 3387–3439.
Jackson, A. L., Bartz, S. R., Schelter, J., Kobayashi, S. V., Burchard, J., Mao, M., Li, B., Cavet, G. and Linsley, P. S. (2003) Expression profiling reveals off-target gene regulation by RNAi. *Nat. Biotechnol.*, **21**, 635–637.
Janzing, D., Mooij, J. M., Zhang, K., Lemeire, J., Zscheischler, J., Daniusis, P., Steudel, B. and Schölkopf, B. (2012) Information-geometric approach to inferring causal directions. *Artif. Intell.*, **182–183**, 1–31.
Kalisch, M. and Bühlmann, P. (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.*, **8**, 613–636.

Kang, H., Zhang, A., Cai, T. and Small, D.S. (2015) Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *J. Am. Statist. Ass.*, to be published.
Kemmeren, P., Sameith, K., van de Pasch, L. A., Benschop, J. J., Lenstra, T. L., Margaritis, T., O'Duibhir, E., Apweiler, E., van Wageningen, S., Ko, C. W., van Heesch, S., Kashani, M. M., Ampatziadis-Michailidis, G., Brok, M. O., Brabers, N. A., Miles, A. J., Bouwmeester, D., van Hooff, S. R., van Bakel, H., Sluiters, E., Bakker, L. V., Snel, B., Lijnzaad, P., van Leenen, D., Groot Koerkamp, M. J. and Holstege, F. C. (2014) Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, **157**, 740–752.
Kulkarni, M. M., Booker, M., Silver, S. J., Friedman, A., Hong, P., Perrimon, N. and Mathey-Prevot, B. (2006) Evidence of off-target effects associated with long dsrnas in drosophila melanogaster cell-based assays. *Nat. Meth.*, **3**, 833–838.
Lauritzen, S. L. (1996) *Graphical Models*. New York: Oxford University Press.
Lauritzen, S. L. and Richardson, T. S. (2002) Chain graph models and their causal interpretations. *J. R. Statist. Soc.* B, **64**, 321–348.
Lauritzen, S. L. and Spiegelhalter, D. J. (1988) Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. R. Statist. Soc.* B, **50**, 157–224.
Maathuis, M., Kalisch, M. and Bühlmann, P. (2009) Estimating high-dimensional intervention effects from observational data. *Ann. Statist.*, **37**, 3133–3164.
Mooij, J. M., Janzing, D., Heskes, T. and Schölkopf, B. (2011) On causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems*, vol. 24, pp. 639–647.
Pearl, J. (2009) *Causality: Models, Reasoning, and Inference*, 2nd edn. New York: Cambridge University Press.
Peters, J. and Bühlmann, P. (2014) Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, **101**, 219–228.
Peters, J., Mooij, J. M., Janzing, D. and Schölkopf, B. (2014) Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.*, **15**, 2009–2053.
R Core Team (2014) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Richardson, T. and Robins, J. M. (2013) Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. *Working Paper 128*. Center for the Statistics and the Social Sciences, University of Washington, Seattle.
Richardson, T. and Spirtes, P. (2002) Ancestral graph markov models. *Ann. Statist.*, **30**, 962–1030.
Robins, J. M. (1986) A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Modllng*, **7**, 1393–1512.
Robins, J. M., Hernan, M. A. and Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 550–560.
Rothenhäusler, D., Heinze, C., Peters, J. and Meinshausen, N. (2015) backShift: learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems*, vol. 28. Red Hook: Curran Associates.
Rouse, C. E. (1995) Democratization or diversion?: The effect of community colleges on educational attainment. *J. Bus. Econ. Statist.*, **13**, 217–224.
Rubin, D. B. (2005) Causal inference using potential outcomes. *J. Am. Statist. Ass.*, **100**, 322–331.
Schapire, R. E., Freund, Y., Bartlett, P. and Lee, W. S. (1998) Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Statist.*, **26**, 1651–1686.
Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K. and Mooij, J. (2012) On causal and anticausal learning. In *Proc. 29th Int. Conf. Machine Learning*, pp. 1255–1262.
Shimizu, S., Hoyer, P. O., Hyvärinen, A. and Kerminen, A. J. (2006) A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, **7**, 2003–2030.
Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O. and Bollen, K. (2011) DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model. *J. Mach. Learn. Res.*, **12**, 1225–1248.
Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction, and Search*, 2nd edn. Cambridge: MIT Press.
Stock, J. H. and Watson, M. W. (2003) *Introduction to Econometrics*. Reading: Addison Wesley.
Terza, J., Basu, A. and Rathouz, P. (2008) Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *J. Hlth Econ.*, **27**, 531–543.
Tian, J. and Pearl, J. (2001) Causal discovery from changes. In *Proc. 17th A. Conf. Uncertainty in Artificial Intelligence*, pp. 512–522. San Francisco: Morgan Kaufmann.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
VanderWeele, T. J. and Robins, J. M. (2010) Signed directed acyclic graphs for causal inference. *J. R. Statist. Soc.* B, **72**, 111–127.
Verma, T. and Pearl, J. (1991) Equivalence and synthesis of causal models. In *Proc. 6th A. Conf. Uncertainty in Artificial Intelligence*, pp. 255–270.
Wright, P. G. (1928) *The Tariff on Animal and Vegetable Oils*. New York: Macmillan.
Wright, S. (1921) Correlation and causation. *J. Agric. Res.*, **20**, 557–585.

Discussion on the paper by Peters, Bühlmann and Meinshausen

Peter A. Thwaites (*University of Leeds*)

Peters and his colleagues have produced a stimulating paper, which will be of interest not only to statisticians but also to people working in other communities, such as artificial intelligence. They note that if one can identify all the direct causes or causal predictors of a response variable then the distribution of this variable conditioned on these predictors will be invariant under manipulation of other variables in the system. This could be thought of as a direct consequence of the directed local Markov property that a variable is independent of its non-descendants given its parents (see for example Lauritzen (2001)). They then look for such invariance across different environments to identify these predictors.

The authors have shown that the set of causal predictors is identifiable when manipulations of the system are of certain types (theorem 2), including the rudimentary 'do' interventions of Pearl (2000). However, they also make the assumption (in for example Section 7.1) that the exact nature of the interventions is unknown. If this is indeed so, how probable is it that the interventions are of these types? An urgent task is to demonstrate that the set of predictors is identifiable for a much wider class of interventions—if those listed turn out to be the only ones that allow this set to be identified, then the work in this paper, however interesting, may turn out to be of limited use. I would like to propose investigating the following types of intervention as being among those of interest:

(a) manipulating collections of variables to specific values, where there is not at least one single do intervention on each non-response variable;
(b) *stochastic* manipulations which assign a new probability distribution over the outcomes of manipulated variables;
(c) *functional* manipulations Do $X = g(W)$ for some set of variables W.

We could of course also consider what might be termed *stochastic functional* manipulations.

I shall concentrate here on functional manipulations. So consider the *sprinkler* example from Pearl (2000), a directed acyclic graph for which is given in Fig. 10(a). Here, using the adapted methodology of Section 6.1, we have structural equations models $X_1 = f(\varepsilon_1)$, $X_2 = f_2(X_1, \varepsilon_2)$, $X_3 = f_3(X_1, \varepsilon_3)$, $X_4 = f_4(X_2, X_3, \varepsilon_4)$ and $X_5 = f_5(X_4, \varepsilon_5)$. The do intervention 'Put sprinkler on' removes the edge $X_1 \to X_3$ (as in Fig. 10(b)), and hence X_3 is no longer a function of X_1. But we could consider a manipulation such as 'If it is summer put the sprinkler on; if it is not summer and it is raining put the sprinkler off' (Thwaites, 2013). Here, instead of removing the edge $X_1 \to X_3$, we need to add an edge $X_2 \to X_3$ as in Fig. 10(c), since whether the sprinkler is on depends on both the season and whether it is raining. So a possible structural

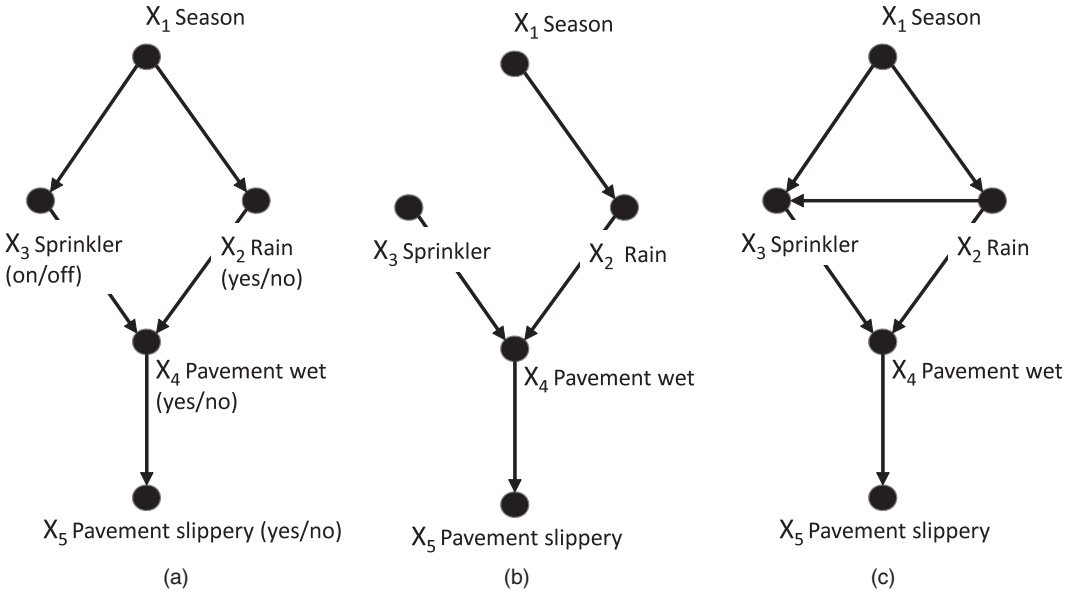

Fig. 10. Example of a *functional* manipulation

equation model for this is $X_3 = f_3'(X_1, X_2)$, implying a deterministic relationship between X_1, X_2 and X_3. But what happens to the sprinkler if it is not summer and not raining?

It is not immediately apparent whether these kinds of scenarios will always satisfy the assumptions that are stated in the paper and, if they do, whether the set of causal predictors will always be identifiable. In this particular example this might not be an issue since the parents of the probable response variables X_4 and X_5 remain unchanged.

The authors have extended their ideas to the non-linear case. The sprinkler example here, which uses discrete variables, suggests to me the further extension to cases where the methodology must necessarily be non-parametric. I would also like to draw attention to the (still relatively small) collection of references on causality which argue that *causes* are more naturally thought of as *events*, rather than random variables (see for example Shafer (1996), Dawid (2000) and Thwaites *et al*. (2010)). Is the analysis in this paper compatible with this interpretation?

As befits a discussion paper, this paper provides plenty of opportunity for debate, argument and further research. It is therefore with great pleasure that I propose a vote of thanks to the authors.

Vanessa Didelez (*Leibniz Institute for Prevention Research and Epidemiology, Bremen, and University of Bremen*)

I thank Peters, Bühlmann and Meinshausen for their stimulating paper, which I believe will have great impact.

They exploit the property of 'invariant prediction' (assumption 1) for causal discovery mostly relying on structural equation models (SEMs). Although it is important to provide results for SEMs as they are extremely popular in numerous fields, they make strong mechanistic assumptions which many researchers find too limiting or unrealistic mostly because of the implied joint distribution of counterfactuals (Dawid, 2000). Careful reading of the properties on which the authors' results rely suggests to me that such assumptions can be much relaxed while still addressing practically relevant questions.

I propose to adopt the decision theoretic framework of Dawid (2002, 2015) and Dawid and Didelez (2010). A brief outline is as follows.

(a) We distinguish between variables that we can (or care) to manipulate and those that we do not. Invariance with respect to the former may not hold but we may be able to characterize sufficient additional information without assuming an underlying SEM. This is similar to randomized controlled trials demonstrating an effect of X on Y: the effect is not necessarily expected to be the same in different populations because of for example different lifestyles habits; even if not of interest in themselves, such habits would need to be taken into account to obtain invariance.

(b) Assumption 1 suggests that the authors' target of inference is the set S^*. However, they define S^* relative to the *actual* experimental settings, i.e. dependent on the available data, and not in a standalone manner; hence it makes no sense to call S^* the 'true causal predictors' (page 954). Further, formulating S^* in the context of available data means that the *prediction* aspect is not obvious. Proposition 1 and other results alternatively suggest that the desired target is $\mathrm{PA}(Y)$ in an SEM; this is shown to be identified under a *particular* set of experiments. Dawid (2002) and Dawid and Didelez (2010) considered causal inference to be about predicting the effects of *future* interventions. A key assumption, then, is invariance across the observed regimes *as well as the potentially different and new one* which we want to predict. Hence, we formulate the target of inference as a (future) decision problem, or ideal experiment. We might ask: given a set of manipulable variables (not assumed complete or direct in any sense) $X = (X_2, \ldots, X_{p+1})$, which subset X_{S*} is the most effective in steering Y? We assume that 'effectiveness' can be decided on the basis of ideal experiments $\mathcal{E}_{\mathrm{ideal}}$, e.g. randomizing all X_i. The available data, however, are gathered under a set of actual experiments $\mathcal{E}_{\mathrm{actual}}$ which may differ from $\mathcal{E}_{\mathrm{ideal}}$, and identification concerns the question whether $\mathcal{E}_{\mathrm{actual}}$ can help to find S^*.

(c) Similarly to Dawid (2002) and Dawid and Didelez (2010), let σ_X be a 'regime indicator' taking values in $\mathcal{E} = \mathcal{E}_{\mathrm{actual}} \cup \mathcal{E}_{\mathrm{ideal}}$. A characterization of valid experimental settings (which is somewhat missing in the authors' approach) can be expressed as assumptions about the set of distributions $P(\cdot; \sigma_X = e), e \in \mathcal{E}$. Then, assumption 1 reads $P(Y|X_{S^*}; \sigma_X = e) = P(Y|X_{S^*}; \sigma_X = e'), e \neq e'$, or

$$Y \perp\!\!\!\perp \sigma_X | X_{S^*}, \tag{47}$$

similarly to using an environment variable E (Section 6.1 and Appendix C). However, σ_X is not a random variable and conditional independence is generalized as in Constantinou and Dawid (2016). Such independences can sometimes be inferred from influence diagrams (Dawid (2002); see the examples in Fig. 11).

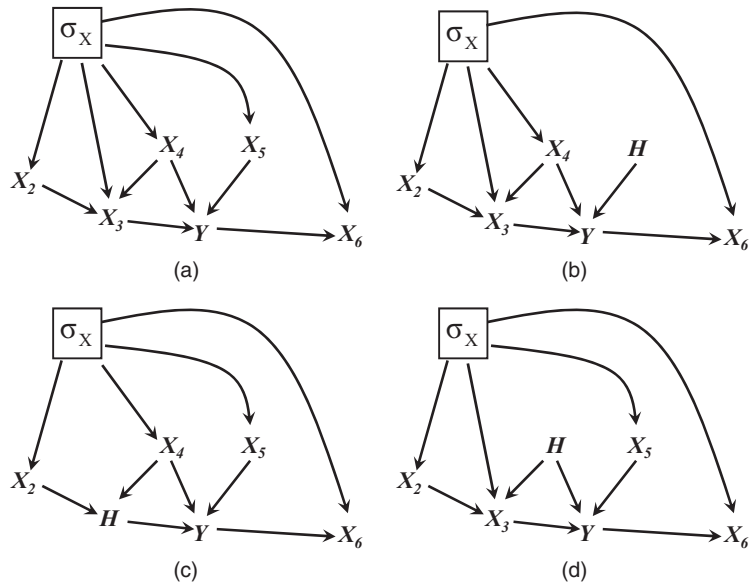

Fig. 11. Examples of influence diagrams

(d) In a more general approach, we could extend assumption (47) to include further variables $H_1, \ldots, H_q$; but here, as above, we do not assume that these are manipulable. It may be more plausible that invariance holds with respect to the enlarged system, analogously to the 'extended stability' of Dawid and Didelez (2010). We can now ask: which of $H_1, \ldots, H_q$ need to be observed in addition to X? Consider the influence diagrams in Fig. 11; for Fig. 11(a) we find $Y \perp\!\!\!\perp \sigma_X | (X_3, X_4, X_5)$ which coincides with PA(Y), but without assuming an SEM; for Fig. 11(b), with an unmanipulable or unobservable H, we find $Y \perp\!\!\!\perp \sigma_X | (X_3, X_4)$, similarly to proposition 4, part (b), of the paper; for Fig. 11(c), we find $Y \perp\!\!\!\perp \sigma_X | (X_2, X_4, X_5)$, so $S^* \neq \mathrm{PA}(Y)$ although still a meaningful quantity; this case is analogous to proposition 5 with X_2 an ancestor of Y, but without considering the existence of H a violation of any assumptions—we are simply not interested in H and find that it can be ignored; finally, in Fig. 11(d), H cannot be ignored as it is required to establish invariance: $Y \perp\!\!\!\perp \sigma_X | (X_3, X_5, H)$. The set $S^* = \{X_3, X_5\}$ is still a meaningful target of inference, and identified from a sufficiently rich $\mathcal{E}_{\mathrm{actual}}$ as long as H is observed and a 'deconfounder' (Dawid, 2002).

The vote of thanks was passed by acclamation.

Ricardo Silva (*University College London*)

I consider that the genuine fundamental problem of causal inference is the need for (partially untestable) invariance assumptions to operationalize interventions, and I thank the authors for emphasizing the role of invariances in a stimulating paper. I shall make some brief comments on how the ideas introduced here can also be helpful in the context of measurement problems.

Much of the contribution involves removing assumptions about the exact target of interventions. This is important: sometimes we may feel uncomfortable to speak of causal effects between some treatment X and outcome Y, not because we cannot think of ways of intervening on X, but because we can think of *too many* ways of intervening. However, perhaps none may plausibly keep the relationship between X and Y invariant. In this case, the methods in Peters and his colleagues cannot be applied.

Many of these problems can be explained as a result of the difficulty of measuring X or Y. Invariance assumptions, fortunately, can be extended to accommodate measurement error. It can also clarify to some extent the nature of unobserved quantities. Consider the classical example of Bollen (1989), of which a simplified version is shown in Fig. 12. Although it may be unrealistic to describe perfect interventions on gross national product that do not directly affect energy consumption, an alternative model postulates an abstract 'industrialization level' index measured indirectly by these two variables. Assumptions of invariance under interventions F on this index could be tested by models that capture different regimes

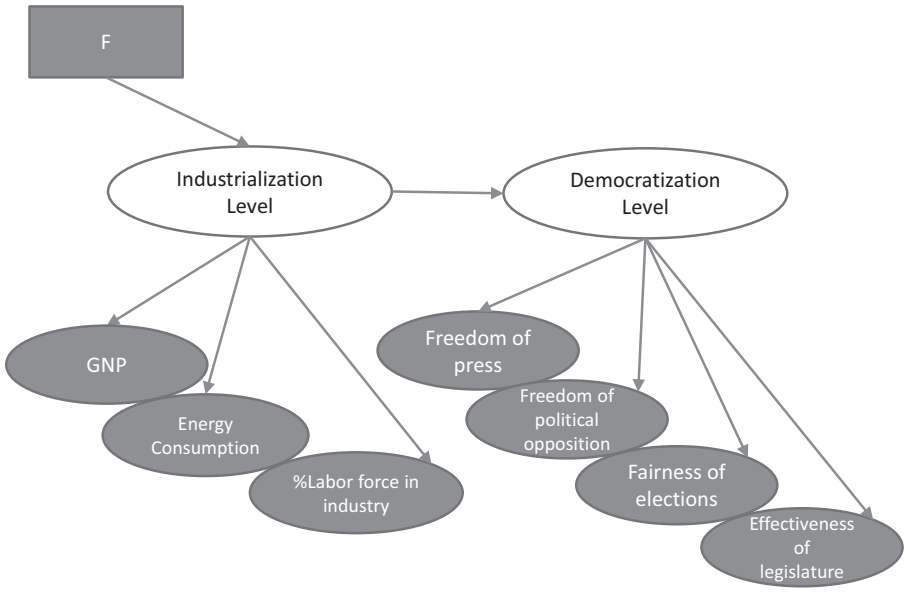

Fig. 12. Dark vertices represent observed vertices, with the square vertex indicating an intervention; latent variables are represented as white vertices (this example is a simplification of that described by Bollen (1989))

among latent variables but share the same measurement model. Identification of measurement models has been studied under the psychometrics (Spearman, 1904), machine learning (Silva *et al.*, 2006) and statistics literatures (Carroll *et al.*, 2006; Allman *et al.*, 2009) and these results can be used to build such a test of invariance.

Moreover, invariance under interventions provides further operational meaning to latent constructs: a quantity that acts as a mediator between an intervention and measurements, as well as other latent variables. Depending on which invariance assumptions are held as primitives (our 'fundamental problem'), violations of measurement invariances may indicate lack of unidimensionality of the latent construct (a different take on the issue of 'versions of a treatment' (VanderWeele and Hernan, 2013), or further unmeasured confounding between measurements and latent variables. In either case, I predict that the valuable ideas introduced by Peters, Bühlmann and Meinshausen will also change the ways we build and interpret latent variable models in the future.

Philip Dawid (*University of Cambridge*)

In the context of the example of Fig. 1, we shall typically not know whether the world is correctly described by the models represented by the graphs shown.

Suppose that we have extensive data—values for all five variables involved—from regime (a), which is purely observational, and from regime (b), where we have intervened on X_2 and X_3. We do not have data from regime (c), where X_4 is intervened on, but we are interested in inferring what would happen there. Adopting the helpful terminology introduced in the discussion contribution of Dr Didelez, regimes (a) and (b) are 'actual' and regime (c) is 'ideal'.

If the actual regimes are indeed correctly represented by Figs 1(a) and 1(b), we shall find some invariances across these two regimes: in both, we have the same joint distribution for (X_4, X_5), the same conditional distribution for Y, given (X_2, X_4), the same conditional distribution for Y, given (X_2, X_5), and indeed the same conditional distribution for Y, given (X_2, X_4, X_5). But this is not enough information to reconstruct the underlying graphs. For example, we shall not know that X_5 precedes X_4, and we shall not be able to deduce whether or not, in regime (c) where X_4 is intervened on, the conditional distribution of Y given (X_2, X_5) will be the same as in (a) and (b)—as it would be if the correct graphical representation of regime (c) had X_4 and X_5 interchanged, but need not be for the situation as pictured.

This example shows that—contrary to an ambiguous impression left by the paper—there is not a unique

minimal S^* determined by assumption 1: when regimes (a) and (b) are the only games in town, we could have $S^* = (X_2, X_4)$, or equally $S^* = (X_2, X_5)$. In this example, the 'correct' choice, which also takes into account the 'ideal' regime represented by (c), is $S^* = (X_2, X_4)$. But we have no way of knowing this. Indeed, we have no warrant to suppose that any invariances discovered from the actual regimes, (a) and (b), will persist into some entirely new regime such as (c). To do so would require making some very strong additional assumptions. But what should these be, and when and how might they be justified?

Adam Foster (*University of Cambridge*)

I thank the authors for a fascinating paper. The aim of this comment is to address the question 'could invariant prediction work for function data?'.

The following motivating example is inspired by Lindquist (2012). Suppose that we wish to study pain processing in the brain. A subject is exposed to a stimulus which can be painful ($P = 1$) or painless ($P = 0$). We measure a subject's brain activity in two regions by using functional magnetic resonance imaging. The resulting functions are $X_1(t)$ and $X_2(t)$. The subject then reports how much pain was experienced as Y.

A possible structural equation model (SEM) for pain processing is shown in Fig. 13.

We have two experimental environments corresponding to $\mathrm{do}(P = 0)$ and $\mathrm{do}(P = 1)$ which we label e and f respectively.

The structural equations implied by this model are functional. For Y we have the structural equation

$$Y = \int X_2(t)\,\gamma_{2Y}(t)\,\mathrm{d}t + \varepsilon_Y.$$

Introducing a suitable inner product, this can be rewritten as

$$Y = \langle X_2, \gamma_{2Y} \rangle + \varepsilon_Y. \tag{48}$$

A common technique for estimating the regression function γ_{2Y} is to expand the data on a finite basis (e.g. a Fourier basis) so the model reduces to a multivariate linear model in the basis coefficients.

This is a causal model, because equation (48) holds for all environments:

$$Y^e = \langle X_2^e, \gamma_{2Y} \rangle + \varepsilon_Y^e,$$
$$Y^f = \langle X_2^f, \gamma_{2Y} \rangle + \varepsilon_Y^f$$

and ε_Y^e and ε_Y^f have the same distribution.

The approximate test, method II in the paper, for invariant prediction can be used in this case, because the residuals are scalar. Suppose now that ε_Y is Gaussian. Then the procedure for estimating the causal predictors is outlined in algorithm 1 (functional invariant prediction for the pain processing example).

Step 1: *for* $S \subseteq \{1, 2\}$ *do* the following steps.
Step 2: fit a functional linear model by regressing $\mathbf{Y}$ on $\mathbf{X}_S$ using data from all environments.
Step 3: test $Y^e - \langle X_S^e, \hat{\gamma} \rangle$ and $Y^f - \langle X_s^f, \hat{\gamma} \rangle$ have the same mean and variance.
Step 4: *end for*.
Step 5: take the intersection of the accepted S.

The approximation being made when using this procedure is the same as for the multivariate case, namely that the estimate $\hat{\gamma}$ is close to the true regression function γ.

As we have seen, the approximate test for invariant prediction can carry over without changes to the functional case when the response is scalar. There are several extensions which could be made in the functional setting which are now discussed.

The motivating example did not have a fully functional flavour, because the residuals—the quantities which are tested in invariant prediction—were scalar.

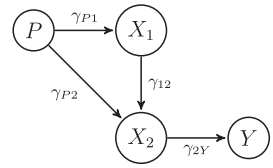

Fig. 13. An SEM for pain processing in the brain

In some cases the response may be functional. In this case we seek a test for the equality of distribution of functions: algorithm 2 (two-samples test for functions).

Step 1: expand the functions F and G on a basis

$$F = \sum C_k \phi_k,$$
$$G = \sum D_k \phi_k.$$

Step 2: *for* $k = 1, \ldots, K$ *do* the following step.
Step 3: test whether C_k and D_k have the same distribution.
Step 4: *end for*.
Step 5: combine the results by using the Bonferroni correction.

In practice, a functional principal component basis is often used. It has been shown (Pomann *et al.*, 2016) that as $K \to \infty$ testing the principal components is equivalent to testing the equality of distribution or the underlying functions.

Steffen Lauritzen (*University of Copenhagen*)
First I congratulate the authors for providing this interesting and stimulating paper. I welcome the fact that stability of causal relationships over varying environments are highlighted, and I wonder whether this usefully could be more formalized. In particular, I should like to draw the attention of the authors to the decision theoretic formalism of *limited memory influence diagrams* (LIMIDs) (Lauritzen and Nilsson, 2001). A LIMID consists of a Bayesian network of *chance nodes* Γ, appropriately augmented with *decision nodes* Δ; for a LIMID there might also be utility nodes which represent consequences of actions.

Every decision node $d \in \Delta$ will have a specified *information set*, represented as the parent set in a directed acyclic graph $\mathcal{D} = (\Gamma \cup \Delta, E)$, for example as illustrated in Fig. 14.

A (randomized) *policy* δ_d for $d \in \Delta$ specifies the distribution of decision d for each configuration $x_{\mathrm{pa}(d)}$ in its information set, and a *strategy* $q = \{\delta_d, d \in \Delta\}$ is a specification of policies for all decisions. Thus a strategy is exactly an environment and the joint distribution of the chance and 'environment' nodes is

$$f(x) = \prod_{\gamma \in \Gamma} p(x_\gamma | x_{\mathrm{pa}(\gamma)}) \prod_{d \in \Delta} \delta_d(x_d | x_{\mathrm{pa}(d)}).$$

Would it be worthwhile to use this notion of environment, for designing causal experiments? Also, it might be fruitful to investigate cases where the environment is partially unknown, corresponding to hidden variables or confounders.

Peng Ding and Avi Feller (*University of California, Berkeley*)
We congratulate the authors on an interesting and compelling contribution to the causal inference literature.

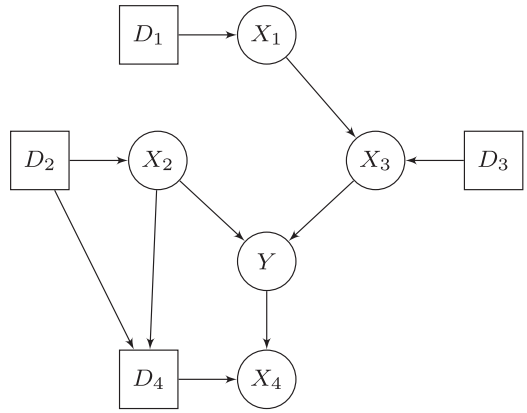

Fig. 14. LIMID describing a set of environments for a causal system involving three potentially explanatory variables and a single response variable

We offer three main comments. First, the authors correctly note that invariance across experimental conditions can be a powerful source for learning about causal relationships. We hope that several examples of this from elsewhere in causal inference will prove useful. For instance, Kling *et al*. (2007) and Reardon and Raudenbush (2013), among others, assumed that key causal relationships are constant across sites in large, multisite experiments, and then leveraged this assumption for inference. Other examples of such *no-interaction* assumptions include identifying causal effects within subgroups defined by non-compliance behaviour (Jo, 2002) and survival status (Ding *et al*., 2011). Finally, this invariance assumption is crucial for inference when *transporting* causal conclusions across populations (Pearl and Bareinboim, 2014).

Second, we note that the paper's fundamental invariance assumption is possibly testable in practice. In recent work, Jiang *et al*. (2016) exploit a similar assumption—which they call *homogeneity*—in the context of evaluating surrogate end points with multiple trials. A key observation from Jiang *et al*. (2016) is that the homogeneity assumption has testable implications when there are a sufficiently large number of trials. We expect that similar results would hold for this paper and we would be curious to see those implemented.

Third, although we found the example of educational attainment interesting, it is not clear how useful the proposed method is in this setting. In particular, we expect that all 13 variables in that example are 'causally' predictive of college graduation; in other words, we imagine the corresponding directed acyclic graph to be very dense *ex ante*, and we gain little in practice from learning that test score and father's education are causally predictive of college graduation. This setting is common in the social sciences, where a truly null relationship between two variables is relatively rare. By contrast, this method seems much more useful in the gene example, in which the directed acyclic graph is presumably much more sparse. We hope that there are ways to extend the proposed method to be more useful in the absence of such sparsity.

Tyler J. VanderWeele (*Harvard University, Boston*)

The causal inference methodology in statistics can be divided into two broad categories. In one category, the causal structure is assumed known and the goal is the identification and estimation of the causal effect of an exposure on an outcome, or of time varying exposures, interactive effects, mediated effects, spillover effects or effects on various latent subpopulations (Pearl, 2009; Morgan and Winship, 2014; Imbens and Rubin, 2015; VanderWeele, 2015; Hernán and Robins, 2016). We might refer to this category as the 'causal effects literature'. In the second broad category, the causal structure is assumed unknown and is to be learned or inferred from the data. We might refer to this category as the 'causal discovery literature'. Prior work on causal discovery has exploited conditional independence relations (e.g. Spirtes *et al*. (2000)), independence relations within reweighted data (e.g. Shpitser *et al*. (2012)), non-Gaussianity (e.g. Shimizu *et al*. (2006)) or non-linearity (e.g. Hoyer *et al*. (2009)) to learn something about the underlying causal structure. Peters and his colleagues propose using the stability of causal coefficients across multiple interventional or experimental settings to infer causal structure, thereby advancing the prior causal discovery literature.

I should like to issue two challenges on causal discovery. First, it would seem that an important next step in advancing the literature would be approaches that integrate what is potentially learned from exploiting several of the various relationships above, rather than just one. Second, I would like to challenge the causal discovery community to find a non-trivial application within the social sciences in which we actually learn something new. In the authors' example, even if we accept the applicability of the methodology, the 'discovery' that test scores have a positive causal influence on the probability of obtaining a college degree is hardly surprising. A difficulty with many of the causal discovery methods in the social science context is that, very often, many underlying structures are compatible with the data (and even more so when unmeasured variables are allowed), and the priors favouring sparsity, that are often used, thus favour structures that, in the social science context, are unreasonable where, typically, everything influences everything else. I have been convinced, through examples, that causal discovery may be of use within gene network contexts, in which the discovery methods can be employed to generate hypotheses about structure that can later be confirmed, or refuted, by experiment. The methods there seem useful for hypothesis generation. However, an application of causal discovery in the social sciences in which the results are neither trivial nor absurd would be of considerable interest.

Federico Crudu (*Pontificia Universidad Catolica de Valparaíso*) **and Freddy López and Emilio Porcu** (*Universidad Federico Santa Maria, Valparaíso*)

We congratulate Peters and his colleagues for their beautiful paper. Our discussion is mainly focused on instrumental variables, which are analysed in Section 5 of the paper. The authors propose a sophisticated version of the model

$$Y = X\gamma + g(H, \varepsilon),$$
$$X = Z\pi + V,$$

where X is the regressor, g the mapping associated with the hidden variable H related to the error term ε and Z is the instrumental variable (denoted I in the paper). Ordinary least squares estimation does not offer consistent estimates whenever X and H are correlated: hence, the need for using the approach above, where the endogenous variable X is related to the instrument Z. In particular, if Z is not related to H, it is a valid instrument and the parameter of interest γ can be consistently estimated with two-stage least squares. One notorious problem for this type of estimation is the *strength* of the instrument, which is governed by the parameter π. If π is large then Z is said to be strong. If π is small Z is said to be weak. A weak instrument tends to lead to awkward inferential results. The authors claim that their method is robust to the presence of a weak instrument, but we do not understand how this can happen. We have some other queries.

(a) If we understand correctly, the authors point out that one can use only one instrument for one endogenous variable; however, this may not be very convenient as you may wish to use more instruments (when they are available) to increase the precision of your estimates. This is even more relevant in case the function that relates X to Z is unknown and approximable with polynomials or splines. In this case one can automatically generate a potentially large number of instruments.
(b) What happens if the instruments are irrelevant, i.e. $\pi = 0$?
(c) What happens if Z is actually endogenous?

Wenliang Pan and Canhong Wen (*Sun Yat-Sen University, Guangzhou*)
We congratulate Peters and his colleagues for a thought-provoking and fascinating paper on a challenging topic in casual models. Our comments are as follows.

This important work introduces an invariance prediction assumption to construct confidence sets for causal predictors and derives confidence intervals for the associated coefficients. Next, identifiability guarantees for the sets of causal predictors in Gaussian structural equation models are given. Furthermore, extensions to instrumental variable and hidden variables are discussed. The authors also provide discussion on non-linear models, intervened targets and some robustness properties. This work is a substantial contribution to the casual inference problems by providing a solid inference tool.

In particular, we are very interested to know the answers to the following questions.

(a) It is very natural to ask whether similar results can be established in high dimensional and ultrahigh dimensional set-ups. In Section 3.4, the authors address the computational complexity by the shrinkage estimation methods, such as those of Tibshirani (1996) and Fan and Lv (2001). Would it enjoy the oracle property?; what about the sure independence property under the ultrahigh dimensional cases when then dimensionality is much higher than the sample size (Fan and Lv, 2008).

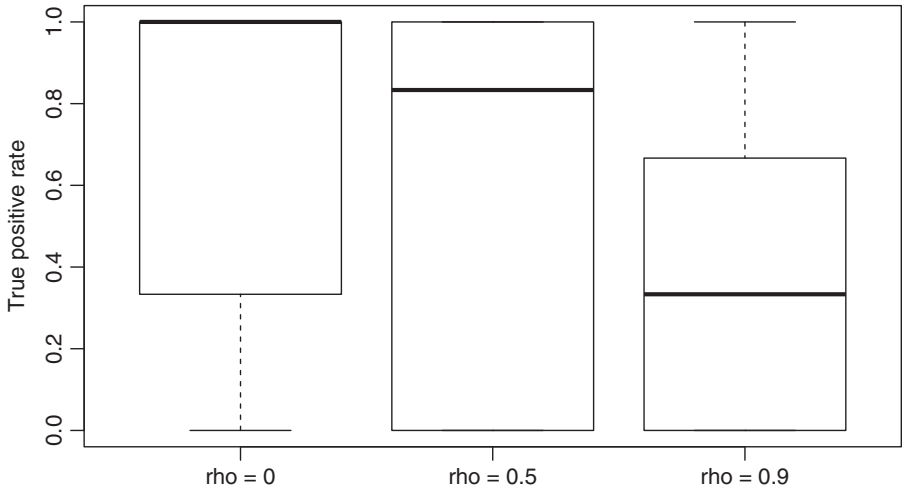

Fig. 15. Simple example: boxplot of the true positive rate in 100 simulated data sets

(b) What are the conditions on the predictors? We have conducted a simple simulation to study the influence of correlation on the performance of invariance prediction. We generated multivariate Gaussian variables X with mean 0 and variance Σ, where Σ has entry $(\sigma_{ij})_{10\times 10} = \rho^{|i-j|}$. The following linear model was considered: $Y = X_1 - X_2 - 1.5X_4 + \epsilon$, where $\epsilon \sim N(0, 1)$. The last variable X_{10} was chosen to be a child of Y, i.e. X_{10} is non-causal, $p = 0, 0.5, 0.9$ for independent, moderately correlated and highly correlated cases respectively. Summary results are given in Fig. 15. It seems that the performance would become worse when the correlation in the predictors becomes higher.

We believe that research along those directions will further enhance the applicability of casual inference by using invariance prediction. Lastly, we conclude this comment by congratulating the authors again for such a wonderful piece of work!

The following contributions were received in writing after the meeting.

Elias Bareinboim (*Purdue University, West Lafayette*)

Causal inference studies the principles and tools necessary for reasoning about cause-and-effect relationships based on heterogeneous data sets and different types of invariance regarding experimental regimes, sampling schemes and environmental conditions (Pearl, 2000; Bareinboim and Pearl, 2016). We commend Peters and his colleagues for leveraging these invariances for structural learning with confidence intervals. Readers may be interested to know that these invariances, which were earlier formulated for understanding, characterizing and testing causal relations (Haavelmo 1995; Aldrich, 1989; Pearl, 2000) have found new applications. In this note, we comment on two different invariances used in the paper.

We start by noting that there is a projection of structural causal models known as causal Bayesian networks (CBNs) (Pearl (2000), pages 23–24), which makes weaker independence assumptions among counterfactuals while still allowing reasoning about interventional distributions. Bareinboim *et al*. (2012) introduced an equivalent formulation of CBNs that makes explicit the modularity condition exploited in the paper, namely

> 'that if we consider all direct causes of a target variable of interest, then the conditional distribution of the target given the direct causes will not change when we interfere experimentally with all other variables in the model except the target itself'.

It was shown explicitly that a CBN encodes a collection of interventional invariances of the form (definition 6, part (iii)) $\forall\ X \subset V, Y \in V, S \subset V, P\{y|\mathrm{do}(x, s, \mathrm{pa}_y)\} = P\{y|\mathrm{do}(s, \mathrm{pa}_y)\}$ whenever there is no arrow from X to Y in G. This property can be seen as the Markov blanket of interventional distributions, which contrasts with its probabilistic counterpart (where X cannot include descendants of Y). It was shown that these invariances can be written in terms of zero direct effects, which Peters and his colleagues could leverage systematically in settings where interventions are not necessarily atomic or precisely identified.

The authors further exploit Y's functional invariance across experimental conditions (sometimes called S-admissibility: definition 8 (Pearl and Bareinboim (2014))), which is a qualitatively different type of assumption. This invariance has also been exploited in the context of transportability theory and is encoded explicitly in the causal diagram through the removal of square nodes. Interestingly, the S-admissibility of Y given its parents is natural in various scenarios, but it is not always necessary for causal inference as discussed in example 9 of Pearl and Bareinboim (2014).

Overall, it is refreshing to see such a refined and systematic use of structural invariances in challenging, real world applications.

Debopam Bhattacharya and Oliver Linton (*University of Cambridge*)

This paper proposes an interesting methodology to detect a relationship between an outcome Y and a set of potential covariates X that is invariant to the 'environment' from which the data were generated. This invariant relationship, to be found by hypothesis tests performed for each selection of covariate sets, is interpreted as a 'causal' model. We shall frame our discussion along the following points.

(a) In many cases of interest, the outcome is non-binary, and the requirement might simply be that $E(Y|X, e)$, rather than the entire distribution of Y given X (as assumed in Section 1.1), to remain invariant as the environment e varies. This would allow for arbitrary heteroscedasticity, and identical forms of heteroscedasticity across environments could represent an 'invariant' feature of the relationship between Y and X. The paper, instead, imposes that the conditional variance of Y in the 'causal' model is independent of the regressors *and* is identical across environments. Furthermore,

Chow-type tests in non-Gaussian heteroscedastic environments are also widely used in practice; see Davidson and MacKinnon (1993), page 377.

(b) The distinction between 'environments' and 'covariates' may not be entirely clear cut in an application. The example in Section 7.3 uses distance to college as a potential example of an environment. This may be problematic, given the long-standing practice in labour economics of using distance to college as an instrument for college attendance, where distance is assumed to affect college attendance without having any direct influence on earnings. This contradicts the paper's assumption that distance to college has no direct 'causal effects' on college attendance. Since the data are cross-sectional at a point in time, it is not sufficient to argue that location decision is chronologically prior to college choice, since one's parents' location choice could depend on how much they value education beyond what is indicated by observed covariates.

(c) It may be a good robustness check to see whether different covariate combinations are selected if one defines environments differently. For example, in Section 7.3, one may use different cut-offs for distance to define far or near. If choice of covariate sets changes with the definition of environment, then how do we interpret 'causality'? Is causality always to be defined in terms of a specific definition of environment? Does it have any other more 'fundamental' definition which does or does not coincide with the invariance-based definition?

(d) How is the concept of 'invariance' useful in regard to formulating public policy or targeting interventions? If a covariate has a statistically insignificant coefficient in every environment, do we include it as part of a causal model?

Andrew Davison (*University of Cambridge*)

Firstly, I congratulate Peters and his colleagues on their interesting and novel approach to causal inference. Although they mainly focus on linear models, by use of the 'conditional formulation' in $H_{0,S,\mathrm{nonlin}}(\mathcal{E})$, it is straightforward to extend the methods to handle generalized linear models, which I shall now detail.

I say that the *invariant prediction assumption* is satisfied if there is a link function g, a column vector $\gamma^* \in \mathbb{R}^p$ with support set S^* and $\eta^* \in \mathbb{R}$ such that, for all $e \in \mathcal{E}$, the $Y^e | X^e_{S*} = x$ belong to an exponential dispersion family with mean parameter μ_x such that $g(\mu_x) = \eta^* + \Sigma_{i \in S^*} x_i \gamma_i^*$ and dispersion parameter ϕ, both independent of e. The analogue of $H_{0,S}(\mathcal{E})$ is now

$$H_{0,S}(\mathcal{E}) : H_{0,\gamma,\eta,S} \text{ is true for some } \gamma \in \mathbb{R}^p \text{ and } \eta \in \mathbb{R} \tag{49}$$

where

$$H_{0,\gamma,\eta,S}(\mathcal{E}) : \begin{cases} \text{there exists } \phi \in (0, \infty) \text{ and a link function } g \text{ such that,} \\ \text{for all } e \in \mathcal{E} \text{ and } x \in \mathbb{R}^{|S|} \text{ when this is defined,} \\ Y^e | X^e_S = x \sim \mathrm{ED}(\mu_x, \phi) \text{ where } g(\mu_x) = \eta + \Sigma_{i \in S} x_i \gamma_i. \end{cases} \tag{50}$$

The population regression coefficients are now defined by

$$(\beta^{\mathrm{pred,e}}(S), \zeta^{\mathrm{pred,e}}(S)) := \underset{\beta \in \mathbb{R}^p : \mathrm{supp}(\beta) \subseteq S;\, \zeta \in \mathbb{R}}{\arg\min} \mathbb{E}[-\log[f\{Y^e; \theta^e(\beta, \zeta), \phi\}]] \tag{51}$$

$$= \underset{\beta \in \mathbb{R}^p : \mathrm{supp}(\beta) \subseteq S;\, \zeta \in \mathbb{R}}{\arg\min} \mathbb{E}[K\{\theta^e(\beta, \zeta)\} - Y^e \theta^e(\beta, \zeta)] \tag{52}$$

where $\theta = \theta(\mu)$ is the natural parameter, $\theta^e(\beta, \zeta) := \theta\{g^{-1}(\zeta + X^e \beta)\}$ and

$$f(y, \theta, \phi) = a(y, \phi) \exp\left[\frac{1}{\phi}\{y\theta - K(\theta)\}\right] \tag{53}$$

is the density of the exponential dispersion family. For Gaussian linear models, this recovers the original definition. I now define the *population residual dispersion parameter* by

$$\phi^{\mathrm{pred},e}(S) := \frac{\mathbb{E}[\{Y^e - \mu^{\mathrm{pred},e}(S)\}^2]}{V\{\mu^{\mathrm{pred},e}(S)\}} \tag{54}$$

where $\mu^{\mathrm{pred},e}(S) := g^{-1}\{\zeta^{\mathrm{pred},e}(S) + X^e \beta^{\mathrm{pred},e}(S)\}$ and $V(\mu) = K''\{\theta(\mu)\}$ is the variance function. Assuming that the X^e are non-degenerate, by Jensen's inequality one can argue that

$$H_{0,S}(\mathcal{E}):\begin{cases}\text{there exists } (\beta,\zeta,\phi)\in\mathbb{R}^p\times\mathbb{R}\times\mathbb{R}_+ \text{ and a link function } g \text{ such that}\\ \text{for all } e\in\mathcal{E} \text{ and } x\in\mathbb{R}^{|S|},\ Y^e|X^e_S=x\sim \mathrm{ED}(\mu_x,\phi) \text{ when this exists,}\\ g(\mu_x)=\zeta+\Sigma_{i\in S}x_i\beta_i, \text{ and } (\beta^{\mathrm{pred},e}(S),\zeta^{\mathrm{pred},e}(S),\phi^{\mathrm{pred},e}(S))\equiv(\beta,\zeta,\phi).\end{cases} \tag{55}$$

I give one possible such test for $H_{0,S}(\mathcal{E})$ at a level α (under sufficient regularity conditions, see for example Jørgensen (1987)), which rejects $H_{0,S}(\mathcal{E})$ if

$$\frac{[1/\{(|S|+1)(|\mathcal{E}|-1)\}](D-\sum_{e\in\mathcal{E}}D^e)}{\hat{\phi}(S)} > F_{(|S|+1)(|\mathcal{E}|-1),n-|\mathcal{E}|(|S|+1)}(\alpha). \tag{56}$$

Here D is the deviance under the pooled model, D^e is the deviance of the model formed using observations in I_e only, $F_{a,b}(\alpha)$ is the upper α-quartile of the $F_{a,b}$-distribution and

$$\hat{\phi}(S):=\frac{1}{n-|\mathcal{E}|(|S|+1)}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

given that

$$\hat{\mu}_i:=g^{-1}\{\hat{\zeta}^{\mathrm{pred},e}(S)+x_i\hat{\beta}^{\mathrm{pred},e}(S)\}, \qquad \text{for } i\in I_e,$$

is a consistent estimator of ϕ. Confidence regions can be obtained as they are for a generalized linear model. The testing procedure in Section 3 can then be used, with the coverage statements in theorem 2 holding asymptotically when the $n_e\to\infty$.

Unfortunately, finding sufficient conditions on $\mathcal{E}$ to ensure that $\hat{S}(\mathcal{E})=S^*$ appears more difficult; for example, the approach of theorem 2 is probably not applicable. As expression (51) seeks to minimize the cross-entropy, I wonder whether an information theoretic approach distinguishing between different (Y^e, X^e) could be fruitful. To end, I again thank the authors, and I hope to see further development of their method.

Jason P. Fine and Michael G. Hudgens (*University of North Carolina, Chapel Hill*)

Peters and his colleagues are to be congratulated on a stimulating and wide reaching presentation, offering a novel approach to causal analysis which has the potential to be widely adopted in real applications.

The theoretical development is quite general, requiring only a valid test of equal conditional distributions across experimental settings. In the linear structural equation model with normal errors, this involves testing the equivalence of the regression and scale parameters, for which tests with good power are already available. Under relaxed model conditions with unspecified error distribution, the construction of omnibus tests is unclear, particularly regarding the infinite dimensional error distribution. We might expect such omnibus tests to be less powerful (potentially much less powerful with small to moderate sample sizes), leading to an increased probability of failing to detect causal predictors. Similar issues would seem to occur whenever semiparametric models are employed and inference about infinite dimensional parameters is needed.

The method proposed seems applicable in controlled experimental settings such as laboratory animal studies, in which case the invariant prediction assumption might be suitable, i.e. it may be reasonable to assume that the conditional distribution of the target given the causal predictors is the same across experiments. The method may be of less utility in other settings such as in human populations where this homogeneity assumption across experiments or environments may be dubious. For example in the educational attainment setting, the likelihood of attaining a Bachelor's degree or higher conditionally on all observed direct causes may differ between subpopulations (environments) because of unmeasured (hidden) causes. If these hidden causes confound the association between the observed direct causes and the target, then the method requires additional information such as an instrumental variable to draw meaningful causal inference.

There are many avenues of possible future research building on the methods proposed. The current approach simultaneously identifies causal predictors with non-zero causal effects and provides inference for those effects; a question arises whether other procedures might be developed under weaker assumptions which identify causal predictors without estimation of the associated effects. The education analysis results in Fig. 7 suggest that one set S was particularly influential, yielding coefficient estimates close to 0 which render the corresponding variables not significant (e.g. income_low); future research might examine the stability of inferences drawn by using the method proposed. The suggested approach adjusts for simul-

taneous inference across multiple experimental settings using the Bonferroni procedure; less conservative inferences might potentially be achieved by using alternative multiplicity adjustments.

Niels Richard Hansen (*University of Copenhagen*)

Peters, Bühlmann and Meinshausen have made a significant contribution to causal inference by systematically exploiting invariance of causal effects. Although they do address the question of model misspecification in Section 6.3 and Appendix C, some issues deserve more attention.

Gene expression is a dynamic process, whose constituents interact according to biochemical processes. The identification of these biochemical mechanisms from experimental data is of great interest; see for example Babtie *et al*. (2014), Finkenstädt *et al*. (2013), Hill *et al*. (2016), Oates *et al*. (2012) and Oates and Mukherjee (2012). For time dynamic processes the causal interpretation of a structural equation model (or a directed acyclic graph) for cross-sectional data may be problematic, and alternatives need to be considered (Aalen *et al*., 2012, 2014; Røysland, 2012; Sokol and Hansen, 2014). At best, a structural equation model is an approximation, and assumption 1 or its non-linear generalization (28) are only approximately true.

When assumption 1 does not hold, can we then still infer an approximately invariant prediction model? If all the hypotheses $H_{0,S}$ are false, the methodology proposed will with high probability find no such model when the tests have sufficient power. Although this prevents false positive results, it does not seem to be an attractive property if there are models that are close to being invariant. This is particularly so because the population value of $(S(\mathcal{E}), \Gamma(\mathcal{E}))$ is highly non-robust to the inclusion of just one environment for which invariant prediction fails. I wonder whether it would be possible to define a method that is more robust to deviations from assumption 1: a method that can produce an approximately invariant prediction model in a well-defined way.

Invariant prediction is likely to be important for disentangling causal gene expression mechanisms, though some work may be needed to transfer the methodology to more realistic models and to make it more robust to model misspecification.

As a final purely technical comment—related to model misspecification—the paper is unclear on how the population quantity $S(\mathcal{E})$ given by expression (6), and its corresponding estimator, given by expression (12), should be computed in the case where all the hypotheses $H_{0,S}$ are false. In Section 6.3 S_{c}^* 'is considered to be causal' but what does that mean if H_{0,S_{c}^*} is false? Also, it is stated that $\hat{S}(\mathcal{E}) = \emptyset$ when all hypotheses are rejected (suggesting that $S(\mathcal{E})$ should be $\emptyset$ when all hypotheses are false), but, in fact, the intersection over the empty index set of true hypotheses is $\{1, \ldots, p\}$ (the nullary intersection). This is a mathematical subtlety, but it also ensures that $S(\mathcal{E})$ indeed is increasing in $\mathcal{E}$ (Section 2.1). Still, when all $H_{0,S}$ are false, $\Gamma(\mathcal{E}) = \emptyset$ as the union of empty sets. This can be contrasted with the case where $H_{0,\emptyset}$ is true, in which case $S(\mathcal{E}) = \emptyset$ and $0 \in \Gamma(\mathcal{E})$.

Kuldeep Kumar (*Bond University, Gold Coast*)

I congratulate the authors for rejuvenating this topic of causal inference by using invariant prediction. However, my trifle disappointment after reading this interesting paper is that there is no mention of Bayesian inference in the context of causal models. Way back in 1978 Rubin had his seminal paper on Bayesian inference on causal effects (Rubin, 1978) and discussed the role of randomization there. Since some prior information about the causal effect may quite often be available can Peters and his colleagues throw some light on the role of Bayesian inference in the context of causal models? It should be mentioned that Bayesian inference has been successfully applied in the context of structural equation modelling. My other concern is related to the results in Section 7.2.6. It seems that the authors have chosen three pairs that obtained the highest rank on the basis of the smallest p-values in spite of the fact that the p-value approach has deep flaws and limitations.

Kuang-Yao Lee, Tianqi Liu and Hongyu Zhao (*Yale University, New Haven*)

Peters and his colleagues are to be congratulated for introducing a new approach to learning causality from both observational and interventional data. There are two major categories of methods in the existing literature for this problem:

(a) two-stage procedures combining the estimations of observational Markov class, and additional structures complemented by the interventional data;
(b) likelihood inferences on integrated observational and interventional data (see, for example, Ellis and Wong (2008) and Luo and Zhao (2011)).

The method proposed by Peters and his colleagues is based on testing the invariant conditional

distributions—a clever idea which we believe will have significant impact on causal inferences and will stimulate further work and extensions under the proposed principle.

We have the following questions inspired from reading the paper.

(i) In regression settings, the accuracy of selecting relevant features often depends on the inner dependence between the predictors, which can be further influenced by the network's complexity. Regarding different types of network, we then wonder how robust the method is. For example, could the random network in the simulation induce simpler structures than a hub network which is more commonly observed in gene networks?

(ii) As mentioned by the authors, the pooling is designed to balance identifiability and statistical efficiency. They have suggested integrating the findings from different poolings. It would be interesting to see whether this procedure can be carried out in a more systematic way. If so, it would also be instructive to know the principles under which it is implemented, e.g. how to make $\mathcal{E}$ 'richer'.

(iii) As suggested by Fig. 4, the method proposed can always pick up a non-trivial set. It is natural to ask whether it is possible to strengthen the result of theorem 1. What we would like to see may be something like

$$P\{\hat{S}(\mathcal{E}) = S^*\}$$

or, at least,

$$P\{\emptyset \neq \hat{S}(\mathcal{E}) \subseteq S^*\}.$$

sufficiently large.

Zudi Lu (*University of Southampton*)

I congratulate Peters and his colleagues warmly for a stimulating contribution of wide application.

Causal inference and invariance principle

The causality concept seems not well defined in general, with inferring the cause described differently by, for example, Encyclopedia Britannica (2014), Shaughnessy *et al.* (2012) and Pearl (2009). Interestingly, the authors propose 'to exploit the invariance of a prediction under a causal model for causal inference', which is made mode based under hypotheses in expressions (4), (25) and (28), and on page 947 put in expressions (1) and (2) as an *assumption* guaranteed by causality.

It will be interesting to see a clearer discussion on 'causal inference' and the assumed 'invariance'. In what sense are inferences on causality and on the invariance equivalent?

Observational and interventional data

Data under interventions are required by using the invariance principle. In most socio-economic or environmental problems, such interventions are unrealistic, with data usually thought of as observational only (see Zhu *et al.* (2004) and Hu *et al.* (2016)). Interestingly, the idea of splitting purely observational data in Section 3.3 can apply with a variable U needed in equation (18) on page 960.

In Section 7.3 on educational data, would different U impact identification of causal predictors? It will also be interesting to have some principle on how to choose a good U for application.

Predictability and Granger causality

A feature of causal inference in the paper lies in prediction-based invariance across experiments. In econometrics, Granger (1969) causality is predictability based by testing whether one time series is useful in forecasting another, which, it is asserted (see Diebold (2001) and Wikipedia (2016)), finds only 'predictive causality' due to the fallacy of one thing preceding being a proof of causation. In prediction, the invariance causality seems a special case of Granger causality. It will be interesting to see discussions on the relationship between these two causal inferences.

Possible extensions

The results from the paper are inspiring. Some extensions could be made in a natural way:

(a) *quantile causality inference*, with quantile check function $p_\tau(y) := y\{\tau - I(y < 0)\}$, with $0 < \tau < 1$, replacing the least squares in expression (9) on page 955;

(b) *time series causality inference*, with temporal lags of Y added, say $Y_t^e = aY_{t-1}^e + X_t^e\gamma + \mathcal{E}_t^e$, with $|a| < 1$, instead of the linear regression, in equation (4) on page 953;

(c) *spatial causality inference*, with spatial auto-regression replacing structural equation model (19) on page 963.

I hope that further research will allow the authors to extend the applicability of their methods.

Jorge Mateu (*University Jaume I, Castellón*)

Peters and his colleagues are to be congratulated on a valuable contribution and thought-provoking paper in this timely topic of causal inference when the influence of causal predictors on a target variable remains invariant under different changes of the environment. This context can be found in a variety of transdisciplinary problems. As they state, the approach of invariant prediction provides new concepts and methods for causal inference while relating to many known concepts, viewed from a different angle. This is the point I would like to emphasize here in relation to dealing with spatial and spatiotemporal dependences and graph theory.

The previous decade has witnessed an extraordinary increase in interest in the analysis of network-related data within numerous disciplines: interest caused by a strongly expanded availability of network data, and by the fact that underlying relational structures of (process) data have gained severe attention. Thus, a swift move towards network-centric perspectives has taken place. In this context, an alternative graph-based approach of analysing point patterns in space and space–time that occur on network structures introduces several different graph-related intensity measures (Eckardt and Mateu, 2016). These patterns occur on undirected and directional as well as partially directed network structures. These intensity measures can be parametrically formulated when covariate information is available on the network and are considered (causal) predictors for the number of events happening per unit area of the network. Fig. 3 in the paper is one of the multiple cases that arise in this context. We hardly know (perhaps they are hidden or just missing) all potential predictors of the target counting variable, but we know that there are continuous experimental changes and interventions on these predictors. Following the roots of this paper, if we at least control all direct causes on the target variable, then the conditional distribution of the target given the direct causes will not change under experimental changes. This idea of invariance and causality is fundamental in network-driven intensity models. Note that these graph-based models deal with spatial or spatiotemporal dependences, and these dependences pose non-linear complex dependences that the corresponding statistical methodological approach must handle. Invariance and causality bring a new insight into the specific network context problem and open new avenues for sound research in computing and statistics. In addition, this methodology sets the basis for approaching many real problems from a variety of applied scientific fields.

Joris M. Mooij (*University of Amsterdam*)

I congratulate Peters and his colleagues on an original and thought-provoking paper. In my opinion, the main contribution of this work is the innovative conservative way to use statistical tests to arrive at decisions regarding causal relations, while allowing control of the probability of making false causal discoveries. A remarkable aspect of the invariant prediction method proposed by the authors is that it does not require the faithfulness assumption (Spirtes *et al.*, 2000), unlike most other, if not all, constraint-based causal discovery methods.

Let us first consider the causally insufficient setting, as discussed in Appendix C. The main idea of the invariant prediction method in that setting essentially boils down to the local causal discovery method proposed by Cooper (1997). Indeed, treat the environment as a random variable $E \in \mathcal{E}$, as the authors do in proposition 5. Assume that the environment E is not caused by any of the observed variables, and that the target variable Y is dependent on the environment ($Y \not\perp\!\!\!\perp E$). If we find a set of observed variables S such that

$$Y \perp\!\!\!\perp E | S \tag{57}$$

then (as shown by Cooper (1997)) we can conclude that $S \subseteq \mathrm{AN}(Y)$, i.e. the variables in S must be ancestors of Y. This reasoning depends critically on the assumption of faithfulness.

In the causally sufficient setting, a crucial observation made by the authors is that the parent set $\mathrm{PA}(Y)$ actually satisfies $Y \perp\!\!\!\perp E | \mathrm{PA}(Y)$ even when faithfulness is not assumed (the local Markov condition and the fact that E is assumed to be a non-descendant of Y suffice for this to hold), and therefore, when conservatively taking the intersection of all sets S that satisfy condition (57), we must obtain a subset of the parents (direct causes) of Y. Interestingly, this strategy is still valid in the presence of faithfulness violations. Indeed, these can only lead to *more* sets S that satisfy condition (57) and, by taking the intersection of all such sets, the worst thing that can happen is that we obtain a smaller subset of the parents of Y.

Not relying on faithfulness is potentially a huge advantage of the method, as faithfulness is likely to be violated in practice. However, other strong assumptions then must be made: in particular the absence of latent confounders (which is also likely to be violated in practice). An intriguing question is whether this work can be generalized to allow for latent confounders *without* assuming faithfulness.

Chris. J. Oates (*University of Technology Sydney and Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers, Melbourne*), **Jessica Kasza** (*Monash University, Melbourne, and Victorian Centre for Biostatistics, Melbourne*) **and Sach Mukherjee** (*German Center for Neurodegenerative Diseases, Bonn*)

The information theorist is taught to list invariances and then to derive models that exhibit those invariances. We warmly congratulate Peters and his colleagues on their insight and creativity in bringing ideas of invariance to bear on causal inference.

An attractive feature of the method proposed is that it allows integration of multiple data sources, even when the precise nature or target of perturbations is unknown. However, a softer approach that allows for variation in causal structure across data subsets might sometimes be appropriate (see for example Oates *et al*. (2014)).

On invariance, we wonder how far we can push the information theorist: consider estimation of the causal effect θ_{ij} of one variable X_i on another X_j. A correct causal graph G can be interrogated to produce a minimal sufficient set S of variables to adjust for in estimation of θ_{ij} (Pearl, 2009). Call such an estimator $\hat{\theta}_{ij}(S)$. Often, minimal sufficient adjustment sets are not unique, in which case any other such set S' will allow consistent estimation. Then, we would expect, for large sample sizes, $\hat{\theta}_{ij}(S) \approx \hat{\theta}_{ij}(S')$. However, there seems no particular reason to expect that these two estimates would coincide if the graph G were incorrect. This seems to suggest another invariance that could be exploited for causal discovery. Potentially, other invariances could play a role. This may in future lead to having to ask which invariances are most useful in practice.

There has long been (in our view justifiable) empirical scepticism towards *de novo* causal discovery (Freedman and Humphreys, 1999). The issue is that it is difficult empirically to validate causal discovery by using data at hand. This goes further than familiar issues of statistical uncertainty, since the underlying concern is of a potentially profound mismatch between critical assumptions and the real data-generating system. The authors' insightful discussion of model misspecification is therefore welcome and the conservative behaviour of their procedure very appealing. We note also that background scientific knowledge may itself be misspecified but that in some circumstances it may be possible to effect 'repair' on the relevant causal structures (Oates *et al*., 2016). We see it as a positive development that empirical validation of causal discovery is becoming more common (see for example Hill *et al*. (2016)). In the near future, empirical work, not least in biology, ought to give us a better sense of the practical efficacy of causal discovery.

T. S. Richardson (*University of Washington, Seattle*) **and J. M. Robins** (*Harvard School of Public Health, Boston*)

We thank Peters and his colleagues for a thought-provoking and highly innovative paper that links disparate approaches to causality.

In Section 5 they describe two assumptions that they associate with the instrumental variable (IV) method. However, these do not fully represent the IV literature. There are IV papers that assume a parametric structural model with the error term independent of the instrument (e.g. Newey (1990), page 110, and Robins and Tsiatis (1991)) contrary to conditions (a) and (b); these papers also describe semiparametric efficient methods for estimation and testing that avoid search. Identification in such models rests on the parametric assumptions; in their absence non-parametric bounds may be found.

We found Section 7.3 confusing because the authors initially mention IV methods which allow for unmeasured confounding (or feedback). However, algorithm II, applied in Section 7.3, explicitly assumes the absence of hidden variables (even though one might expect them to be present).

In Section 5 the authors present another method based on direct search that allows for unmeasured confounders, but this method is not applied. In Appendix C, the authors discuss algorithm II in the presence of unmeasured confounders. Then, as shown in Fig. 9, the method may fail to find 'direct' causal predictors.

Proposition 5 in Appendix C shows that under the assumptions of faithfulness, and exogeneity of E, the set $\mathcal{S}(\mathcal{E})$ will consist solely of ancestors of Y. It follows from Richardson (1996), lemma 4, and Acid and de Campos (1996) that this conclusion holds, even without faithfulness, provided that E is temporally prior to all other variables and some set d-separates E and Y in the augmented graph. The structural conclusions in proposition 5 are essentially those resulting from the fast causal inference algorithm (Spirtes *et al.*, 2000) when selection bias is assumed absent; otherwise $\mathcal{S}(\mathcal{E})$ need not contain ancestors of Y. See Fig. 16.

The authors assume that the allocation of units to environments is exogenous (so E has no parents, observed or unobserved). Though plausible in an experimental context such as in Section 7.2, this often may not hold when 'environments' are constructed *post hoc* (Section 7.3). A related point: in Section 2 the authors assume that data are independent and identically distributed for each environment $e \in \mathcal{E}$.

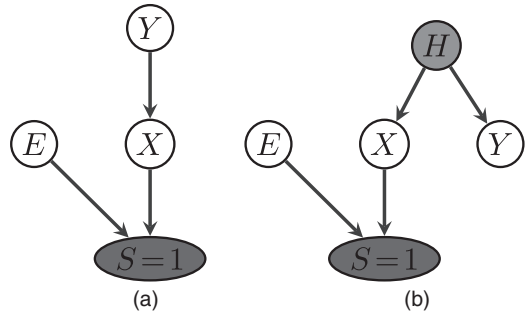

Fig. 16. Two directed acyclic graphs representing selection bias: observations are only recorded for $S = 1$; in (b) H is unobserved; in both, X will be an identified causal predictor of Y, given $P(X, Y|E, S = 1)$

However, in Section 3.2 they pool data from different experimental settings to create a single environment. Such pooled data will no longer be independent and identically distributed in general.

Milan Stehlík (*Johannes Kepler University in Linz and University of Valparaíso*) **and Silvia Stehlíková**
We congratulate Peters and his colleagues on their paper, introducing readers to the challenging world of causal inference.

The assumption of independently distributed errors with finite variance is unrealistic for genetic data examples. This is too oversimplifying an assumption for large-scale genetic perturbation experiments (Kemmeren *et al.* (2014), e.g. pages 740, 743, 745 and 747). Heavy-tailed errors denote the presence of highly connected hubs. We can accept additive errors, but more difficult is the assumption of independent errors but finite variance, which cannot be assumed without proper testing (Stehlík *et al.*, 2014). This invites the question: how convincing is Fig. 2(c) regarding invariance of set $S = \{YPL273W\}$? The basic theory of independent additive errors gives either a normal distribution or heavy-tailed stable laws. Moreover one can make a tree of possibilities of estimators of $\mathrm{corr}(\epsilon_i^e, \epsilon_i^f)$ for a single unit i in different environments e and f. Thus, the counterfactual question can be answered partially.

In genetic data examples, where strong correlations are present (Kemmeren *et al.* (2014), e.g. pages 740, 741, 743, 747 and 750), weighted least squares should be applied instead of ordinary least squares to obtain best linear unbiased estimators. It is disturbing to read on page 949 that '... we are more interested in settings where such careful experimentation is not possible...' and on page 957 to request a full rank linear model (without careful experimental design this is overrealistic).

Finally we would prefer more foundation and justification of the invariance introduced. Invariance should have deep algebraical or logical bases. Assumption 1 in Section 2 relies on 'the existence of a model that is invariant under different experimental or intervention settings'. Especially in genetics such models can be easily empty sets. There is a well-known approach on invariant statistical models based on groups. We mention for example James (1954), Obenchein (1971) and Francis *et al.* (2016) for orthogonal, linear and finite reflection groups. Having an invariant statistic, we can invert it to confidence intervals. This form of invariance relates directly to algebraic–geometric foundations of statistical information. What are these algebraic structures in the case of invariance in the paper? One can think about losing symmetry, or other particular group axioms, but it needs justification. What is its relationship to the fiducial statistics of Hora and Buehler (1967)?

Linbo Wang (*Harvard University, Boston*) **and Shizhe Chen and Ali Shojaie** (*University of Washington, Seattle*)
We congratulate Peters and his colleagues on a thought-provoking proposal to identify (a subset of) causal predictors by using the invariance principle. Here, we make two comments about the approach proposed, concerning the effect of selection variables and the design of perturbation experiments in high dimensions.

First, consider the simple directed acyclic graph (DAG) $X \rightarrow S \leftarrow Y$, where the data are observed given a particular value of the *selection variable* S. Although X does not directly affect Y, intervening on X would, in general, change the distribution of Y in the *observed data*. Consequently, in this case, the method proposed will be anticonservative, as it detects X as a (spurious) causal predictor. This is noteworthy since

the approach is valid (albeit potentially conservative) under many other deviations, such as confounding and model misspecification. Moreover, (under the faithfulness assumption) the 'traditional' structure learning approach can draw valid causal inferences in the presence of selection variables by using maximal ancestral graphs (Richardson and Spirtes, 2002; Colombo *et al*., 2012). We wonder whether the approach proposed can be extended to handle selection variables.

Second, the design of perturbation experiments is of practical interest for applications in high dimensions. Consider, for example, the *do* interventions of Section 4.2.1. Theorem 2, part (a), shows that the causal predictors are identifiable if there is at least one intervention on each of p variables potentially affecting the response Y. However, with large p, intervening on all variables can be too costly. In such cases, learning a superset of the parents of Y, pa_Y, from observational data can help to design more efficient experiments. For this, one could, for example,

(a) learn the (partially directed) *skeleton* of the DAG by using, for example, the PC algorithm (Spirtes *et al.*, 2000; Kalisch and Bühlmann, 2007), or
(b) learn the Markov blanket (Pearl, 2014) of Y by using, for example, lasso regression (Meinshausen and Bühlmann, 2006).

For linear structural equation models with Gaussian noise, both methods consistently select (a superset of) all variables in pa_Y (Kalisch and Bühlmann, 2007; Meinshausen and Bühlmann, 2006). In practice, however, Markov blanket learning using lasso regression (Meinshausen and Bühlmann, 2006) may select more members of pa_Y, as seen in Fig. 17. Moreover, the faithfulness assumption is not necessary when learning the Markov blanket and lasso inference procedures do not require a 'beta-min' condition (see, for example, van de Geer *et al*. (2014)). These preliminary findings suggest that investigating the trade-off between experimental cost and statistical power when learning a superset of pa_Y may be fruitful.

Qingyuan Zhao, Charles Zheng, Trevor Hastie and Robert Tibshirani (*Stanford University*)

We congratulate Peters and his colleagues on this throught-provoking paper. Statistical inference of causality has been thoroughly studied in randomized experiments or observational studies but is seldom considered when data from both *observational* and *interventional* settings are available. Peters and his colleagues

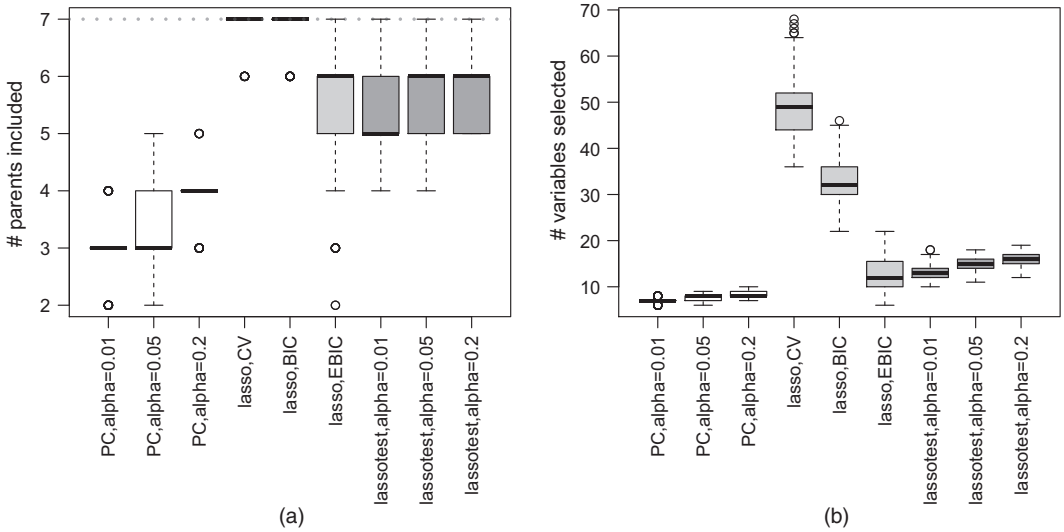

Fig. 17. Results of a simulation study to compare the performance of structure learning procedures for selecting a superset of the parents of variable Y in a DAG (in each of $B = 100$ simulated data sets, $n = 500$ observations are generated from a linear structural equation model based on a randomly generated DAG with $p = 100$ variables; the DAG is the same for all simulated data sets; the results suggest that Markov blanket learning using the lasso may miss fewer parents of Y; however, the performance of lasso-based procedures depends heavily on the choice of tuning parameter): (a) distribution of the number of true parents of node Y selected by using each estimation method (·····, total number of true parents of Y); (b) distribution of the total number of nodes included in the selected sets

Table 2. Robustness properties of the ICP procedure†

	Issues	*ICP's behaviour*
(a)	Intervene on Y (or a missing cause)	$\bigcap_\emptyset$
(b)	Non-linear, non-additive and/or heteroscedastic	$\bigcap_\emptyset$
(c)	Not enough interventions	False causal positive findings
(d)	Small sample size	$\emptyset$
(e)	Left out a confounder	$\bigcap_\emptyset$
(f)	Left out an unconfounding predictor	Okay
(g)	Misspecified model or noise distribution	False positive findings

†Under certain types of model misspecification, ICP will return a 'model reject', denoted by $\bigcap_\emptyset$ (i.e. all subsets including the empty set are not invariant), rather than produce false positive results: (a) when interventions are performed on Y, no predictor set can be invariant; (b) when the homoscedastic linear model is misspecified, the prediction rule will vary depending on the range of the predictors; (c) without enough interventions, the set of causal parents is unidentifiable, and non-causal invariant sets exist; (d) when the sample size is small, the hypothesis test for invariance has insufficient power to reject the invariance null, and hence many sets are accepted as invariant; (e) if a confounder is left out, this is equivalent to intervening on Y; (f) when an uncounfounding predictor is left out, its effect is equivalent to independently and identically distributed noise; (g) under a misspecified noise model, the hypothesis test may not be sensitive to differences in the noise distribution, leading to low power.

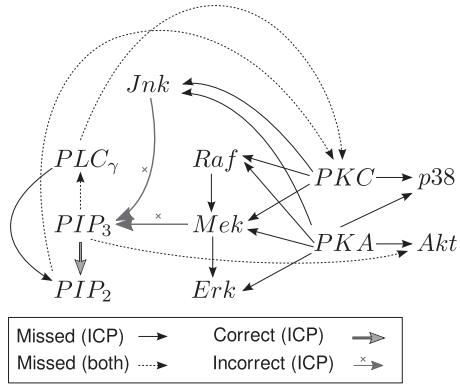

Fig. 18. Application of the ICP procedure to recover the protein signalling network, taking in turn each of the 11 variables as the response of interest and selecting the subset of environments in which the response was not perturbed: the invariant set for each variable can be identified as the parents of that variable in the graph; for nine of the 11 proteins, ICP rejected the model and reported no discoveries; for protein PIP2, ICP correctly identified one parent, PIP3; for protein PIP3, ICP reported Mek and Jnk as part of the invariant set, but these do not match any interactions known in the literature

have made an important contribution by tackling this problem with their notion of invariant causal prediction (ICP).

At first look, ICP is a corollary of structural equation models, but we think that its value might be much more substantial. Dawid (2000) noted that causal researchers are predominantly Laplacian determinists, for whom

'nothing short of a functional model relating outputs to inputs will do as a description of nature'.

Peters and his colleagues provide an alternative approach that defines causality by *predictability* instead of

determinism: two different concepts that are not logically connected (Hoefer, 2016). In light of Breiman's (2001) two cultures of statistics, determinism roughly corresponds to the data modelling culture and predictability is the spirit of Breiman's algorithmic modelling culture.

Bearing this difference in mind, Peters and his colleagues do not take a downright predictability approach in this paper. Rather, they consider two types of assumption: invariant prediction to define causality and deterministic modelling assumptions such as linearity. This hybrid perspective becomes clear when comparing the assumptions in equation (4) with those of expressions (24), (28) or (31). As a consequence, ICP can make causal discovery only when the modelling assumptions are correct. The authors take this as a robustness property, but in our view it also limits the applicability in practice. We did not find in the paper a summary of the robustness of ICP, so we tried to outline in Table 2 the behaviour of linear ICP when some of its assumptions are not met. We would welcome the authors' comments on this summary.

To test the empirical performance of ICP, we use the authors' software on a protein signalling network data set. Sachs *et al*. (2005) collected a combination of observational and nine interventional data sets to infer the causal structure of 11 proteins (Fig. 18). Using their own method, Sachs *et al*. (2005) reportedly recovered 15 of the known directed arcs and discovered two new putative links (which are not shown), and missed three of the interactions which were known in the literature. In contrast, ICP makes only three causal discoveries. Among them, only one belongs to the known arcs. The poor performance of ICP on this data set could be explained by the overly restrictive linear model.

The **authors** replied later, in writing, as follows.

We thank all the contributors to the discussion for many insightful and interesting comments. We shall address some of the points that have been raised but, for brevity, we cannot respond to all the issues mentioned.

Non-uniqueness of S^*

Several comments (including Didelez and Dawid) addressed the point that for a given set of environments $\mathcal{E}$ there might be different sets S^* satisfying assumption 1; see the discussion in the paragraphs following this assumption. We acknowledge that our exposition in the paper is perhaps confusing, mainly because the formulation of assumption 1 depends on the set of environments $\mathcal{E}$. A better way and a stronger result is as follows.

We regard the set $\mathcal{E}$ of the *observed environments* as a subset of any larger set $\mathcal{F} \supseteq \mathcal{E}$ of *possible environments*. Given $\mathcal{F}$, we are interested in a set $S^* = S^*(\mathcal{F})$ for which the invariance assumption 1 holds with respect to $\mathcal{F}$. (If $\mathcal{F}$ does not contain sufficiently many interventions, there could be multiple sets $S^*(\mathcal{F})$ that satisfy the invariance assumption with respect to $\mathcal{F}$ and all these sets will be covered by our confidence statements below.)

We have the following more general version of theorem 1.

Theorem 1. Consider an observed set of environments $\mathcal{E}$. Consider a distribution P over (Y, X) and assume a valid test for $H_{0,S}(\mathcal{E})$, in expression (12) for all sets $S \subseteq \{1, \ldots, p\}$ at level α in the sense that, for all S, $\sup_{P: H_{0,S}(\mathcal{E}) \text{ true}} P\{H_{0,S}(\mathcal{E}) \text{ rejected }\} \leqslant \alpha$. Then, $\hat{S}(\mathcal{E})$ satisfies

$$\text{for every } \mathcal{F} \supseteq \mathcal{E} \text{ and every } S^* \text{ satisfying assumption 1 with respect to } \mathcal{F},$$
$$P\{\hat{S}(\mathcal{E}) \subseteq S^*\} \geqslant 1 - \alpha.$$

If, moreover, $\hat{C}(S)$ in expression (14) is a valid confidence interval satisfying $P\{\beta^{\text{pred}}(S) \in \hat{C}(S)\} \geqslant 1 - \alpha$ for all sets S satisfying $H_{0,S}(\mathcal{E})$, then the set $\hat{\Gamma}(\mathcal{E})$ in expression (13) has coverage at least level $1 - 2\alpha$:

$$\text{for every } \mathcal{F} \supseteq \mathcal{E} \text{ and every } \gamma^* \text{with support } S^* \text{satisfying assumption 1 with respect to } \mathcal{F},$$
$$P\{\gamma^* \in \hat{\Gamma}(\mathcal{E})\} \geqslant 1 - 2\alpha.$$

Important is the fact that the confidence statements hold for *all* (possibly future) environments (or interventions) $\mathcal{F}$ which include the observed data (but potentially many more interventions) and satisfy the invariance assumption 1 with respect to $\mathcal{F}$. This property is interesting for, say, predictive tasks where one wants to be protected against new possibly adversarial environments. Applied to Fig. 1, for possible environments $\mathcal{F} = \{1, 2, 3\}$ and observed environments $\mathcal{E} = \{1, 2\}$, we have a unique target of interest $S^* = \{2, 4\}$ which fulfils assumption 1 with respect to $\mathcal{F}$. If we are interested in other environments $\mathcal{F} \supseteq \mathcal{E}$, then the set S^* of interest can be (depending on $\mathcal{F}$) $\{2, 4\}$, $\{2, 5\}$ or $\{2, 4, 5\}$ and the coverage statement in theorem 1 is true whichever of these sets we are interested in.

In the special case of structural equation models (SEMs), if we care about *any* possible set $\mathcal{F} \supseteq \mathcal{E}$ of interventions (excluding those on the response variable Y), then the set of parents of Y always satisfies assumption 1; see proposition 1. The coverage statement is then true for the parents of Y even if the observed set of environments $\mathcal{E}$ is much smaller than the set $\mathcal{F}$ of future intervention environments of interest.

The assumption of structural equation models

In response to Didelez: assumption 1 and theorem 1 do not use the terminology of SEMs. Instead, we discuss SEMs as an example; see proposition 1. The confidence statements in theorem 1 rely only on assumption 1 and do not require a specification of a type of intervention. SEMs can be used to model observational distributions, interventional distributions and counterfactuals. We stress that, in our work, we use them solely in the first two ways. We therefore never make use of or assume the existence of any joint distribution of counterfactuals

Weak or unspecified interventions

In response to Thwaites, the coverage property is true under all possible interventions, including the examples that were mentioned, as theorem 1 is not making any assumption about the nature of the interventions except for excluding direct interventions on the response variable (although identifiability requires obviously some further work and assumptions about the strength of interventions in each specific scenario).

Zhao, Zheng, Hastie and Tibshirani observed numerically that too weak or too few interventions lead to false positive results. This is guaranteed not to happen *on average* under the assumptions of the simulations since the coverage property does not require strong interventions or specific types of interventions) and we think the observed result is thus most likely to be a non-representative realization (the other model robustness properties are in agreement with our experience and writing).

Instrumental variables

Applied to instrumental variables (IVs) we regard our method as robust with respect to weak instruments in the following sense: if, in the extreme case, the influence of the instrument on the predictor X is zero, then all environments (which correspond to different values of the IV) are identical. Then even the empty set is accepted and the confidence intervals thus include zero. In response to Crudu, López and Porcu, we do not provide any guarantee for situations in which the instrument is endogenous, but an IV can be a non-descendant of Y in the graph (see the next paragraph).

Covariates and environments

The framework of limited memory influence diagrams proposed by Lauritzen is definitely of interest in this context (as is the related decision theoretic framework that was developed by Dawid and Didelez). A decision node can be used to define an environment as long as it is not a descendant of Y. This point of view allows a richer class of possible models than those where the decision node cannot have ancestors in the graph (as would be so for an IV if we view the IV as a decision node).

Given purely observational data, it is in particular possible to use covariates for a '*post hoc*' construction of environments, as long as these covariates can be assumed to be non-descendants of the target Y; see Section 3.3. Note that these covariates are allowed to have parents. Different candidate covariates may lead to different outcomes of the method, since some covariates may lead to a more informative splitting into environments than others.

Model misspecifications

In response to Richardson and Robins, we have formulated proposition 5 without assuming the existence of a set d-separating E and Y (this is violated if E points directly to Y, for example); instead we assume faithfulness but this can be regarded as a matter of taste.

Furthermore, it has been pointed out by Richardson and Robins, and Wang, Chen and Shojaie that, if the model is misspecified because of implicit conditioning ('selection bias'), the method may falsely regard non-ancestors of Y as causal predictors. If the implicit conditioning and the invariance hold in every 'possible' environment (see above) $e \in \mathcal{F} \supseteq \mathcal{E}$, predictions using those causal predictors remain valid in the sense that prediction intervals for Y will have correct coverage and the false selection is in this sense unproblematic. We acknowledge, however, that problems do occur if one tries to model interventions on such predictors if these kinds of interventions have not appeared in $\mathcal{E}$ (there is no set S^* satisfying assumption 1 with respect to a set $\mathcal{F} \supseteq \mathcal{E}$ including these interventions).

Latent variables

Richardson and Robins, and Mooij wondered whether one can allow for latent variables beyond the

discussion in the paper. We believe that this is possible by exploiting a slightly different form of invariance. For some progress in this direction see Rothenhäusler *et al*. (2016).

Data pooling

Richardson and Robins correctly point out that data pooled from different environments are not independently and identically distributed in general. It is important to note that the coverage statement only relies on an independently and identically distributed data assumption of the noise in the response variable Y, not on an independently and identically distributed data assumption for the joint distribution of (X, Y). Phrased differently, for correct coverage, we must make sure only that we accept the invariance of the *conditional* distribution $Y|X_{S^*} = x$ with high probability (conditional on the true causal variables X_{S^*}). The distribution of X_{S^*} can be arbitrary, as we condition on it for the test.

Sachs data and interventions on response

Hansen also asked whether there is a way for causal inference if invariance does not hold for any set, as happens for example under interventions on the target and Zhao wondered about the pooling of data. In the same context, Zhao, Zheng, Hastie and Tibshirani apply our method to the Sachs data. We have done this in other work as well: since interventions may occur on the response in some environments, a direct application of the method to all eight environments will not produce interesting discoveries (as also pointed out by the comment of Zhao and his colleagues). A simple way to avoid environments with interventions on the target (even though the precise location of the targets is unknown) is to consider all pairs of two environments and to combine the discovered causal predictors with a union operation (after adjusting for multiplicity). The reasoning is as follows: if an intervention on the outcome occurs for a given pair of environments, an empty set will in general be returned. If no interventions occur on the target, however, we can and will make causal discoveries. We obtain rather different results which are in reasonable agreement with the speculated 'ground truth' in Sachs (2005); see Meinshausen *et al*. (2016).

Sparseness of graph

It has been correctly noted by VanderWeele, and Ding and Feller that the assumption of a sparse graph can be doubtful in social sciences (and elsewhere). We agree that in these contexts we must treat the estimated set $\hat{S}(\mathcal{E})$ with caution. Even in the absence of interest in a stipulated set S^*, however, we note that the method offers confidence intervals on the causal coefficients. We support the idea of the challenge that was proposed by VanderWeele. The difficult question in the absence of very convincing examples with a ground truth is whether the data have been so far of too poor quality, the questions ill posed or whether there is genuinely nothing interesting to learn (or all of these).

Further extensions

We appreciate the discussion of several other extensions, some of which already contain detailed procedures; these ideas include functional (Foster), dynamical (Hansen) or high dimensional data (Pan and Wen), latent factor models (Silva), non-linear and generalized linear models (Davison and Lu), network models (Mateu), non-Gaussian error distributions (Stehlík and Stehlíková), more details on the relationship to modularity (Bareinboim), the use of different types of invariance (Bhattacharya and Linton, and Oates, Kasza and Mukherjee), formalizations within the framework of limited memory influence diagrams (Lauritzen), applicability under weaker assumptions on homogeneity (Fine and Hudgens), finite sample identifiability statements and how they depend on the dependence structure of the covariates (Zhao), the method's relationship to measurement problems (Silva) and Bayesian reasoning (Kumar).

References in the discussion

Aalen, O., Røysland, K., Gran, J., Kouyos, R. and Lange, T. (2014) Can we believe the DAGs?: a comment on the relationship between causal DAGs and mechanisms. *Statist. Meth. Med. Res.*, to be published.

Aalen, O. O., Røysland, K., Gran, J. M. and Ledergerber, B. (2012) Causality, mediation and time: a dynamic viewpoint. *J. R. Statist. Soc.* A, **175**, 831–861.

Acid, S. and de Campos, L. M. (1996) An algorithm for finding minimum d-seperating sets in belief networks. In *Proc. 12th A. Conf. Uncertainty in Artificial Intelligence* (eds F. V. Jensen and E. Horvitz), pp. 3–10. San Francisco: Morgan Kaufmann.

Aldrich, J. (1989) Autonomy. *Oxf. Econ. Pap.*, **41**, 15–34.

Allman, E., Matias, C. and Rhodes, J. (2009) Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, **6**, 3009–3132.

Babtie, A. C., Kirk, P. and Stumpf, M. P. H. (2014) Topological sensitivity analysis for systems biology. *Proc. Natn. Acad. Sci. USA*, **111**, 18507–18512.

Bareinboim, E., Brito, C. and Pearl, J. (2012) Local characterizations of causal Bayesian networks. In *Graph Structures for Knowledge Representation and Reasoning* (eds M. Croituru, S. Rudolph, N. Wilson, J. Howse and O. Corby), pp. 1–17. Berlin: Springer.
Bareinboim, E. and Pearl, J. (2016) Causal inference and the data-fusion problem. *Proc. Natn. Acad. Sci. USA*, **113**, 7345–7352.
Bollen, K. (1989) *Structural Equations with Latent Variables*. New York: Wiley.
Breiman, L. (2001) Statistical modeling: the two cultures (with comments). *Statist. Sci.*, **16**, 199–231.
Carroll, R., Ruppert, D. and Crainiceanu, C. (2006) *Measurement Error in Nonlinear Models: a Modern Perspective*. Boca Raton: Chapman and Hall.
Colombo, D., Maathuis, M. H., Kalisch, M. and Richardson, T. S. (2012) Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.*, **40**, 294–321.
Constantinou, P. and Dawid, A. P. (2016) Extended conditional independence and applications in causal inference. *Preprint arXiv: 1512.00*
Cooper, G. F. (1997) A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Minng Knowl. Discov.*, **1**, 203–224.
Davidson, R. and MacKinnon, J. G. (1993) *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
Dawid, A. P. (2000) Causal inference without counterfactuals (with discussion). *J. Am. Statist. Ass.*, **95**, 407–448.
Dawid, A. P. (2002) Influence diagrams for causal modelling and inference. *Int. Statist. Rev.*, **70**, 161–189.
Dawid, A. P. (2015) Statistical causality from a decision-theoretic perspective. *A. Rev. Statist. Appl.*, **2**, 273–303.
Dawid, A. P. and Didelez, V. (2010) Identifying the consequences of dynamic treatment strategies: a decision-theoretic overview. *Statist. Surv.*, **4**, 184–231.
Diebold, F. X. (2001) *Elements of Forecasting*, 2nd edn, p. 254. Cincinnati: South Western.
Ding, P., Geng, Z., Yan, W. and Zhou, X. (2011) Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *J. Am. Statist. Ass.*, **106**, 1578–159l.
Eckardt, M. and Mateu, J. (2016) Point patterns occurring on complex structures in space and space-time: an alternative network approach. To be published.
Ellis, B. and Wong, W. H. (2008) Learning causal Bayesian network structures from experimental data. *J. Am. Statist. Ass.*, **103**, 778–789.
Encyclopedia Britannica (2014) Causal inference. In *Encyclopedia Britannica*. Encyclopedia Britannica.
Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.
Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc.* B, **70**, 849–911.
Finkenstädt, B., Woodcock, D. J., Komorowski, M., Harper, C. V., Davis, J. R. E., White, M. R. H. and Rand, D. A. (2013) Quantifying intrinsic and extrinsic noise in gene transcription using the linear noise approximation: an application to single cell data. *Ann. Appl. Statist.*, **7**, 1960–1982.
Francis A., Stehlík, M. and Wynn, H. (2016) "Building" exact confidence nets. *Bernoulli*, to be published.
Freedman, D. and Humphreys, P. (1999) Are there algorithms that discover causal structure? *Synthese*, **121**, 29–54.
van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, **42**, 1166–1202.
Granger, C. W. J. (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **137**, 424–438.
Haavelmo, T. (1995) *The Foundations of Econometric Analysis* (eds D. F. Hendry and M. S. Morgan), pp. 440–453. Cambridge: Cambridge University Press.
Hernán, M. A. and Robins, J. M. (2016) *Causal Inference*. Boca Raton: Chapman and Hall.
Hill, S. M., Heiser, L. M., Cokelaer, T., Unger, M., Nesser, N. K., Carlin, D. E., Zhang, Y., Sokolov, A., Paull, E. O., Wong, C. K., Graim, K., Bivol, A., Wang, H., Zhu, F., Afsari, B., Danilova, L. V., Favorov, A. V., Lee, W. S., Taylor, D., Hu, C. W., Long, B. L., Noren, D. P., Bisberg, A. J., HPN-DREAM Consortium, Mills, G. B., Gray, J. W., Kellen, M., Norman, T., Friend, S., Qutub, A. A., Fertig, E. J., Guan, Y., Song, M., Stuart, J. M., Spellman, P. T., Koeppl, H., Stolovitzky, G., Saez-Rodriguez, J. and Mukherjee, S. (2016) Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Meth.*, **13**, 310–318.
Hoefer, C. (2016) Causal determinism. In *The Stanford Encyclopedia of Philosophy* (ed. E. N. Zalta). Stanford: Center for the Study of Language and Information.
Hora, R. B. and Buehler, R. J. (1967) Fiducial theory and invariant prediction. *Ann. Math. Statist.*, **38**, 795–801.
Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J. and Schölkopf, B. (2009) Nonlinear causal discovery with additive noise models. In *Advances in Neural information Processing Systems*, vol. 21, pp. 689–696. Vancouver: Curran Associates.
Hu, F., Lu, Z., Wong, H. and Yuen, T. P. (2016) Analysis of air quality time series of Hong Kong with graphical modeling. *Environmetrics*, **27**, 169–181.
Imbens, G. and Rubins, D. B. (2015) *Causal Inference for Statistics, Social and Biomedical Sciences*. Cambridge: Cambridge University Press.
James, A. T. (1954) Normal multivariate analysis and the orthogonal group. *Ann. Math. Statist.*, **25**, 40–75.

Jiang, Z., Ding, P. and Geng, Z. (2016) Principal causal effect identification and principal surrogate end point evaluation by multiple trials. *J. R. Statist. Soc.* B, **79**, 829–848.
Jo, B. (2002) Estimation of intervention effects with noncompliance: alternative model specifications. *J. Educ. Behav. Statist.*, **27**, 385–409.
Jørgensen, B. (1987) Exponential dispersion models (with discussion). *J. R. Statist. Soc.* B, **49**, 127–162.
Kalisch, M. and Bühlmann, P. (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.*, **8**, 613–636.
Kemmeren, P., Sameith, K., van de Pasch, L. A., Benschop, J. J., Lenstra, T. L., Margaritis, T., O'Duibhir, E., Apweiler, E., van Wageningen, S., Ko, C. W., van Heesch, S., Kashani, M. M., Ampatziadis-Michailidis, G., Brok, M. O., Brabers, N. A., Miles, A. J., Bouwmeester, D., van Hooff, S. R., van Bakel, H., Sluiters, E., Bakker, L. V., Snel, B., Lijnzaad, P., van Leenen, D., Groot Koerkamp, M. J. and Holstege, F. C. (2014) Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, **157**, 740–752.
Kling, J. R., Liebman, J. B. and Katz, L. F. (2007) Experimental analysis of neighborhood effects. *Econometrica*, **75**, 83–119.
Lauritzen, S. L. (2001) Causal inference from graphical models. In *Complex Stochastic Systems* (eds O. E. Barndorff-Nielsen, D. R. Cox and C. Klüppelberg). Boca Raton: Chapman and Hall–CRC.
Lauritzen, S. L. and Nilsson, D. (2001) Representing and solving decision problems with limited information. *Mangmnt Sci.*, **47**, 1235–1251.
Lindquist, M. A. (2012) Functional causal mediation analysis with an application to brain connectivity. *J. Am. Statist. Ass.*, **107**, 1297–1309.
Luo, R. and Zhao, H. (2011) Bayesian hierarchical modeling for signaling pathway inference from single cell interventional data. *Ann. Appl. Statist.*, **5**, 725–745.
Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34**, 1436–1462.
Meinshausen, N., Hauser, A., Mooij, J. M., Peters, J., Versteeg, P. and Bühlmann, P. (2016) Methods for causal inference from gene perturbation experiments and validation. *Proc. Natn. Acad. Sci. USA*, to be published.
Morgan, S. L. and Winship, C. (2014) *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd edn. Cambridge: Cambridge University Press.
Newey, W. K. (1990) Semiparametric efficiency bounds. *J. Appl. Econmetr.*, **5**, 99–135.
Oates, C. J., Hennessy, B. T., Lu, Y., Mills, G. B. and Mukherjee, S. (2012) Network inference using steady state data and Goldbeter–Koshland kinetics. *Bioinformatics*, **28**, 2342–2348.
Oates, C. J., Kasza, J., Simpson, J. A. and Forbes, A. B. (2016) A pre-processing approach to repair of misspecified causal diagrams. To be published.
Oates, C. J., Korkola, J., Gray, J. W. and Mukherjee, S. (2014) Joint estimation of multiple related biological networks. *Ann. Appl. Statist.*, **8**, 1892–1919.
Oates, C. J. and Mukherjee, S. (2012) Network inference and biological dynamics. *Ann. Appl. Statist.*, **6**, 1209–1235.
Obenchein R. L. (1971) Multivariate procedures invariant under linear transformations. *Ann. Math. Statist.*, **42**, 1569–1578.
Pearl, J. (2000) *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.
Pearl, J. (2009) Causal inference in statistics: an overview. *Statist. Surv.*, **3**, 96–146.
Pearl, J. (2014) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo: Morgan Kaufmann.
Pearl, J. and Bareinboim, E. (2014) External validity: from do-calculus to transportability across populations. *Statist. Sci.*, **29**, 579–595.
Peters, J., Bühlmann, P. and Meinshausen, N. (2015) Causal inference using invariant prediction: identification and confidence intervals. *Preprint arXiv:1501.01332*. Eidgenössiche Technische Hochschule Zürich, Zürich.
Pomann, G.-M., Staicu, A.-M. and Ghosh, S. (2016) A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *Appl. Statist.*, **65**, 395–414.
Reardon, S. F. and Raudenbush, S. W. (2013) Under what assumptions do site-by-treatment instruments identify average causal effects? *Sociol. Meth. Res.*, **42**, 143–163.
Richardson, T. S. (1996) A discovery algorithm for directed cyclic graphs. In *Proc. 12th A. Conf. Uncertainty in Artificial Intelligence* (eds F. V. Jensen and E. Horvitz), pp. 454–461. San Francisco: Morgan Kaufmann.
Richardson, T. and Spirtes, P. (2002) Ancestral graph Markov models. *Ann. Statist.*, **30**, 962–1030.
Robins, J. M. and Tsiatis, A. A. (1991) Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communs Statist. Theor. Meth.*, **20**, 2609–2631.
Rothenhäusler, D., Heinze, C., Peters, J. and Meinshausen, N. (2015) BACK-SHIFT: learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems*, pp. 1513–1521.
Røysland, K. (2012) Counterfactual, analyses with graphical models based on local independence. *Ann. Statist.*, **40**, 2162–2194.
Rubin, D. B. (1978) Bayesian inference for causal effects: the role of randomization. *Ann. Statist.*, **6**, 34–58.
Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. and Nolan, G. P. (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.

Shafer, G. (1996) *The Art of Causal Conjecture*. Cambridge: MIT Press.
Shaughnessy, J., Zechmeister, E. and Zechmeister, J. (2012) *Research Methods in Psychology*, 9th edn, p. 447. New York: McGraw-Hill.
Shimizu, S., Hoyer, P. O., Hyvärinen, A. and Kerminen, A. J. (2006) A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, **7**, 2003–2030.
Shpitser, I., Richardson, T. S., Robins, J. M. and Evans, R. (2012) Parameter and structure learning in nested Markov models. *Preprint arXiv: 1207.5058*. Johns Hopkins University, Baltimore.
Silva, R., Scheines, R., Glymour, C. and Spirtes, P. (2006) Learning the structure of linear latent variable models. *J. Mach. Learn. Res.*, **7**, 191–246.
Sokol, A. and Hansen, N. R. (2011) Causal interpretation of stochastic differential equations. *Electron. J. Probab.*, **19**, 1–24.
Spearman, C. (1904) "General intelligence," objectively determined and measured. *Am. J. Psychol.*, **15**, 210–293.
Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction and Search*, 2nd edn. Cambridge: MIT Press.
Stehlík, M., Thulin, M. and Střelec, L. (2014) On robust testing for normality in chemometrics. *Chemometr. Intell. Lab. Syst.*, **130**, 98–108.
Thwaites, P. A. (2013) Causal identifiability via chain event graphs. *Artif. Intell.*, **195**, 291–315.
Thwaites, P. A., Smith, J. Q. and Riccomagno, E. M. (2010) Causal analysis with chain event graphs. *Artif. Intell.*, **174**, 889–909.
VanderWeele, T. J. (2015) *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York: Oxford University Press.
VanderWeele, T. and Hernan, M. (2013) Causal inference under multiple versions of treatment. *J. Causl Inf.*, **1**, 1–20.
Wikipedia (2016) Granger causality. In *Wikipedia*. (Available from `https://en.wikipedia.org/wiki/Granger_causality`.)
Zhu, H., Lu, Z., Wang, S. and Soofi, A. S. (2004) Causal linkages among Shanghai, Shenzhen, and Hong Kong stock markets. *Int. J. Theoret. Appl. Finan.*, **7**, 135–149.