Huber
Swiss Journal of Economics and Statistics (2024) 160:14
https://doi.org/10.1186/s41937-024-00131-4

Swiss Journal of Economics and Statistics

COMMENT

Open Access

An introduction to causal discovery

Martin Huber[1*]

Abstract

In social sciences and economics, causal inference traditionally focuses on assessing the impact of predefined treatments (or interventions) on predefined outcomes, such as the effect of education programs on earnings. Causal discovery, in contrast, aims to uncover causal relationships among multiple variables in a data-driven manner, by investigating statistical associations rather than relying on predefined causal structures. This approach, more common in computer science, seeks to understand causality in an entire system of variables, which can be visualized by causal graphs. This survey provides an introduction to key concepts, algorithms, and applications of causal discovery from the perspectives of economics and social sciences. It covers fundamental concepts like d-separation, causal faithfulness, and Markov equivalence, sketches various algorithms for causal discovery and discusses the back-door and front-door criteria for identifying causal effects. The survey concludes with more specific examples of causal discovery, e.g., for learning all variables that directly affect an outcome of interest and/or testing identification of causal effects in observational data.

1 Introduction

In socials sciences and economics, causal inference, also known as treatment, program, or impact evaluation, predominantly focuses on evaluating the causal effect of a specific treatment variable, such as an education program, health treatment, or marketing intervention, on an outcome of interest, such as earnings, health, or sales. Comprehensive surveys on causal inference are, for instance, provided in Imbens (2004); Imbens and Wooldridge (2009), and Abadie and Cattaneo (2018), as well as in the textbooks (Angrist and Pischke, 2009; Frölich and Sperlich, 2019; Cunningham, 2021; Huntington-Klein, 2022), and Huber (2023). More recently, causal inference has been combined with machine learning, a subfield of artificial intelligence, which permits considering observed covariates, such as socio-economic characteristics, in a data-adaptive manner, to control for confounding factors that jointly affect the treatment and the outcome (and thus, bias causal estimates) and/or to assess whether effects are heterogeneous across groups with different covariate values. The survey papers by Lieli et al. (2022) and Lechner (2023) as well as the textbooks by Huber (2023) and Chernozhukov et al. (2024) provide an introduction to such causal machine learning methods.

In contrast to causal inference or causal machine learning for assessing the impact of a predefined treatment variable on a predefined outcome, causal discovery, which is more prominent in computer science than in economics and social sciences, aims to learn the causal relationships among several or even many variables in a data-driven manner. In other words, causal discovery seeks to understand or unveil the causal associations within an entire system of variables, which can be depicted by a causal graph. The task of determining which variables influence others, based solely on statistical associations rather than presupposed causal structures (like assuming that a treatment may affect an outcome but not vice versa), poses a significant challenge in empirical settings. A growing number of studies demonstrate the circumstances and assumptions under which this endeavor is at least theoretically feasible, see, for instance, the literature reviews by Kalisch and Bühlmann (2014); Peters et al. (2017), and Glymour et al. (2019).

*Correspondence:
Martin Huber
martin.huber@unifr.ch
[1] Department of Economics, University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland

In economics and social sciences, causal discovery can be useful for identifying key variables that causally affect outcomes such as wages, employment, or education. It might be applied to test theories about which variables influence a particular outcome, explore causality in a data-driven manner when theory is unclear or unavailable, or even refine existing models by discovering previously unrecognized causal relationships. An example would be uncovering social determinants of health or educational attainment, such as determining whether family background, neighborhood characteristics, or peer effects directly contribute to disparities in outcomes. Additionally, specific variants of causal discovery allow for the development of methods to (at least partially) test identifying assumptions underlying causal analysis, such as whether a treatment satisfies multiple assumptions typically imposed for causal evaluation, including exogeneity and instrumental variable assumptions.

This survey provides an introduction to concepts, algorithms, and examples of causal discovery through the lens of economics and social sciences. Section 2 discusses fundamental concepts like d-separation, causal faithfulness, and Markov equivalence. Section 3 sketches various algorithms for causal discovery. Section 4 discusses the back-door and front-door criteria for the identification of causal effects of predefined treatments. Section 5 concludes with examples of causal discovery for learning all variables that directly affect an outcome of interest and/or testing identification of causal effects in observational data. Section 6 concludes.

2 d-separation, causal faithfulness, and Markov equivalence

This section introduces essential concepts in causal discovery: d-separation, causal faithfulness, and Markov equivalence. D-separation, short for "dependency separation", establishes a formal framework to analyze and comprehend the relationships between variables within causal graphs. Causal graphs describe a causal system of variables and consist of nodes, which represent individual variables or sets of variables, and arrows, which indicate causal effects between these variables. For instance, an arrow originating from a node (or variable) D and pointing to node Y signifies that variable D has a causal impact on variable Y, as illustrated in the left graph of Fig. 1. As an example, consider education (D) having a causal effect on earnings (Y). Furthermore, directed acyclic graphs (DAGs) rule out cyclic or simultaneous relations, like arrows going both from D to Y and Y to D. It is therefore ruled out that education (D) simultaneously affects and is affected by earnings (Y) at the point in time when both variables are measured. D-separation enables

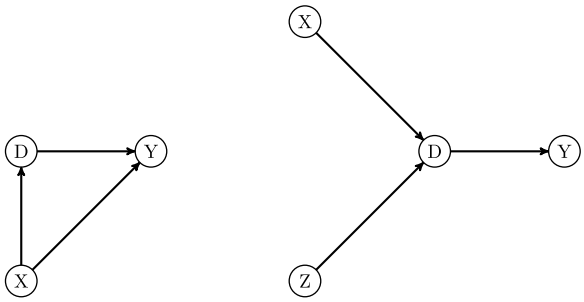

Fig. 1 Selection on observables (left) and Y-learning (right)

the identification of conditional independence relationships among variables based on the structure of a DAG.

The d-separation criterion of Pearl (1988) is founded on the concept of "blocking" or controlling for causal paths in a graph in a manner that establishes statistical (conditional) independence between two variables. This implies that after blocking specific paths connecting the two variables, they are no longer associated with each other. More concisely, a path between two (sets of) variables D and Y is blocked when conditioning on a (set of) control variable(s) C,

1. if the path between D and Y is either a causal chain, implying that $D \rightarrow X \rightarrow Y$ or $D \leftarrow X \leftarrow Y$, or a confounding association, implying that $D \leftarrow X \rightarrow Y$, and variable (set) X is among control variables C (i.e., controlled for),
2. if the path between D and Y contains a collider, implying that $D \rightarrow S \leftarrow Y$, and variable (set) S or any variable (set) causally affected by S is not among control variables C (i.e., not controlled for).

With this definition of blocking in mind, the d-separation criterion states that variables D and Y are d-separated when conditioning on control variable(s) C if and only if C blocks all paths between D and Y. In other words, D and Y are d-separated (i) if we control for all (mediating) variables through which D affects Y or vice versa, as well as all (confounding) variables that jointly affect D and Y, and (ii) if we do not control for any (collider) variables that are jointly affected by D and Y, or any variables influenced by colliders. Since d-separation is a sufficient condition for the (conditional) independence of two variables, it is very useful for causal reasoning in complex causal models. For this reason, d-separation serves as a theoretical basis for causal discovery algorithms designed to learn causal models from data. Reconsidering the left graph of Fig. 1, we observe that D and Y are not d-separated when we control for variable X. Despite the fact that the confounding relationship $D \leftarrow X \rightarrow Y$

is blocked when conditioning on X, the causal effect $D \rightarrow Y$ remains unblocked such that D and Y are dependent conditional on X. As an example, consider the case that education (D) and earnings (Y) are confounded by background characteristics like socio-economic status or innate ability (X). After controlling for such characteristics X to avoid confounding or omitted variable bias, the conditional dependence of Y and D reflects the causal effect of education on earnings under the causal model in the left graph of Fig. 1.

To illustrate the usefulness of d-separation for causal discovery, consider the causal model depicted in the right graph of Fig. 1, which pertains to Y-learning, as, e.g., discussed in Mani et al. (2012). Variables X and Z are independent of each other when not controlling for D, while they are dependent conditional on D. According to the d-separation criterion, this necessarily implies that D is a collider ($X \rightarrow D \leftarrow Z$). Furthermore, both X and Z are associated with Y, as they both affect Y via D. However, when controlling for D, Y is conditionally independent of both X and Z. According to the d-separation criterion, this necessarily implies that D functions as a mediator, transmitting the causal effects of X and Z to Y ($X \rightarrow D \rightarrow Y, Z \rightarrow D \rightarrow Y$). Put differently, X and Z are so-called instruments for D, as they both affect D but do not directly affect Y (other than through D), thus satisfying an exclusion restriction.

As an example, consider the case that D is participation in a training, Y reflects earnings after training participation, and X and Z are randomly assigned e-mail- and text message-based invitations to participate in the training. In the absence of confounding and if the invitations affect the earnings outcome only through training participation, then the invitations are independent of earnings conditional on training participation. Furthermore, both types of invitations are dependent of each other conditional on training if both types affect training participation. Our Y-learning example demonstrates that in specific causal models, d-separation may enable the comprehensive learning of the entire causal model, but this is not the case in general and unlikely to apply to most causal problems.

While d-separation is sufficient for the conditional independence of two variables, it is important to note that in special cases, two variables might be independent even if d-separation fails. To illustrate this, let us revisit the left graph in Fig. 1. If the confounding path of D and Y due to X ($D \leftarrow X \rightarrow Y$) perfectly offsets the causal effect of D on Y ($D \rightarrow Y$), then D and Y are statistically independent even though d-separation fails (due to the presence of both confounding and the causal effect of D on Y). As an example, consider the case that socio-economic status (X) affects both participation in a training program (D) and earnings (Y) in a way that offsets the effect of the training on earnings, such that training and earnings are statistically independent despite the existence of an earnings effect of training.

Causal discovery methods typically rely on the absence of different paths between variables that exactly cancel each other out. For this reason, we may need to explicitly rule out offsetting paths through the causal faithfulness (or stability) assumption, see, for instance, the discussions in Pearl (2000) and Spirtes et al. (2000). Causal faithfulness imposes that only variables that are d-separated in a DAG are conditionally independent, while any variables that are not d-separated are dependent. Consequently, the assumption mandates that d-separation is not only a sufficient, but also a necessary condition for conditional independence. In other words, two variables are statistically independent if and only if d-separation is met.

However, even with the application of causal faithfulness, d-separation may not (and in most observational data will not) lead to a unique determination of the causal model that underlies the observed associations of variables in the data. This implies that the conditional dependence and independence relationships detected through the d-separation criterion might be compatible with multiple causal models, resulting in ambiguity regarding the true causal structure. This issue motivates the concept of Markov equivalence classes for characterizing causal models that entail identical conditional dependencies and independencies. Distinct causal models belong to the same Markov equivalence class if they exhibit the exact same patterns of conditional dependence and independence in the data, such that the causal models are indistinguishable based on observational data alone.

To provide an example, let us consider two distinct causal models. In the first model, D (e.g., education) affects Y (e.g., earnings) exclusively through a mediator M (e.g., human capital acquisition): $D \rightarrow M \rightarrow Y$. In the second model, Y affects D exclusively through M: $D \leftarrow M \leftarrow Y$. In the absence of any further variables, the two models generate precisely the same patterns of conditional dependence and independence: D and Y are conditionally independent given M, while D and M are dependent given Y and Y and M are dependent given D, and any of D, M, Y are mutually dependent in the absence of conditioning on any variable. As a result, both models belong to the same Markov equivalence class, because we are unable to distinguish between these models based solely on the statistical associations in the data. Causal discovery becomes even more involved when allowing for unobserved variables in a causal model, e.g., an unobserved confounder U jointly affecting D and Y ($D \leftarrow U \rightarrow Y$), such as unobserved personality traits (U)

affecting both education (D) and earnings (Y). Since the d-separation criterion cannot be applied to unobserved variables, their presence tends to exacerbate the uncertainty regarding the true causal model. Consequently, this can increase the number of causal models that align with the same Markov equivalence class.

In scenarios where the correct causal model cannot be learned solely from observational data, the challenge of model ambiguity can potentially be addressed through "external" sources of information concerning the causal structure. Such sources may include domain expertise, theoretical insights, past empirical findings, or knowledge about the temporal sequence of events. These external insights can provide valuable context that helps narrow down potential causal models. For example, if variable D (e.g., training participation) is measured at an earlier time point than variable Y (e.g., earnings after training participation), it becomes evident that a causal path from Y to D (such as $D \leftarrow Y$) can be ruled out. This is because future events cannot influence past events, a principle discussed in works like Reichenbach (1991) concerning the connection between causality and time, and also acknowledged in dynamic (treatment or outcome) models in social sciences, see, for instance, Robins (1986), Abbring and Van den Berg (2003), and Lechner (2009). Moreover, in order to investigate whether an observed association between training D and earnings Y is solely due to confounding ($D \leftarrow U \rightarrow Y$), or if it also involves a nonzero treatment effect ($D \rightarrow Y$), an experiment could be conducted in which the treatment is randomly assigned, as random assignment avoids confounding. Consequently, external sources of information might be helpful for clarifying some of the ambiguous causal associations in a DAG in order to reduce the number of plausible causal models within in a Markov equivalence class.

3 A sketch of algorithms for causal discovery

As discussed in the previous section, the objective of causal discovery is to recognize Markov equivalence classes encompassing all causal models that are indistinguishable (or statistically equivalent) in observed data because they exhibit the same d-separation patterns (i.e., patterns of conditional independence). For the practical application of the d-separation criterion in causal discovery, Verma and Pearl (1990) introduced the IC (or I-equivalence Class) algorithm for observational data. The IC algorithm operates under the assumptions that causal faithfulness holds and that there are no unobserved confounders that jointly affect any pair of observed variables for which the causal associations are to be estimated. The IC algorithm consists of the following steps:

1. For all pairs of variables A and B in the data, search for a set of variables X such that A and B are conditionally independent when controlling for X. Link A and B by an edge, which represents an undirected association in a causal graph, if and only if no X exists that satisfies conditional independence.
2. For all pairs of variables A and B that do not share an edge with each other, but share both an edge with a variable S, verify whether S is in set X of step 1. If this is not the case, S is a collider such that the causal association is $A \rightarrow S \leftarrow B$.
3. In the resulting graph with partially determined causal associations (directed causal arrows) and partially undetermined associations (edges), orient the direction of as many edges as possible subject to two conditions: (i) Any alternative orientation would yield a new collider structure. (ii) Any alternative orientation would yield a directed cycle (i.e., a circular causal relation between variables).

Bluntly speaking, step 1 of the algorithm finds those pairs of variables that are dependent conditional on any feasible set of control variables. Variables in such a pair are then connected by an undirected edge since the specific causal path remains unknown at this stage. Step 2 pinpoints collider paths which unveil the causal directions (arrows) between variables. Step 3 finds further causal associations that adhere to the constraints of not creating circular causal relations (which are prohibited in DAGs) or unwarranted further collider paths (as correct collider paths were already detected in step 2). As highlighted by Pearl (2000), step 3 can be systematized in several ways, e.g., by applying the rules provided in Verma and Pearl (1992) for orienting edges into causal arrows. These rules are sufficient to identify the maximum potential number of causal arrows in a causal graph based on observed data.

Spirtes and Glymour (1991) introduced a refined method known as the PC algorithm. It reduces computation time by limiting step 1 of searching for sets X that entail the conditional independence of A and B to variables that share edges with (i.e., are adjacent to) either A or B. Glymour et al. (2019) offer a comprehensive discussion and illustration of the PC algorithm based on the causal scenario of Y-learning considered in Sect. 2. Numerous further enhancements have been proposed in causal discovery algorithms, e.g., for integrating external information regarding causal relationships between specific variables. One important contribution is the fast causal inference (FCI) algorithm introduced by Spirtes

et al. (2000), which allows for and in some causal models even detects the presence of unobserved variables. Glymour et al. (2019) discuss an example for the application of the FCI algorithm in the presence of unobserved confounders.

An important question in causal discovery is how to perform statistical inference, e.g., to obtain p-values or confidence intervals for the estimated causal effects between variables. To test the various conditional independence assumptions between pairs of variables in the data, algorithms typically adopt either pairwise test statistics like t-tests or global goodness-of-fit statistics that are computed collectively for all variables (rather than pairs of variables), such as the Bayes information criterion (BIC). In this context, it is important to acknowledge that verifying conditional independencies among multiple variables amounts to jointly testing several hypotheses and therefore introduces multiple hypothesis testing issues. Consequently, the critical values of pairwise tests must be adjusted to account for the number of tested causal associations (which grow exponentially in the number of variables in the causal model), e.g., by a Bonferroni-type adjustment, see, e.g., Holm (1979). Otherwise, there is an increased risk that testing might (by random chance) erroneously indicate the presence of specific causal associations that are actually absent, implying an increased type I error rate in testing. However, on the flip side, such corrections for multiple hypothesis testing may decrease the power of the tests, implying a reduced chance (or probability) to detect nonzero causal effects, implying an increased type II error rate in testing. This issue becomes particularly pronounced when there are many variables in the causal model, which implies more rigorous corrections.

The algorithms discussed so far explore conditional independencies and dependencies to estimate the set of causal models that fit within the Markov equivalent class of the true (albeit unknown) causal model. However, there are alternative avenues to causal discovery that rely on specific assumptions about the functional nature of causal relationships between variables. Under these assumptions, it becomes possible to uniquely determine the true causal model without relying on d-separation or causal faithfulness. As discussed in Hoyer et al. (2008); Zhang and Hyvärinen (2009a), and Peters et al. (2014), causal relationships can be learned from the data under the following conditions:

1. The causal effect of any variable (say, D) on another variable (say, Y) is defined by a nonlinear function.
2. The effects of unobserved variables on Y can be expressed by a single error term that is independent of D and does not interact with D (such that the effects of unobserved variables on Y are homogeneous across values of D).

In other words, if causal associations between observed variables are nonlinear and causal effects of unobserved variables can be represented by additively separable (i.e., non-interacting) and independent error terms, then we can deduce the direction of causality between observed variables from the data. Formally, this requires that Y is characterized by the following model:

$$Y = \mu(D) + \varepsilon, \tag{1}$$

where μ is a nonlinear function of D and ε is an additive error term that is independent of D.

The nonlinearity of μ carries an important implication. When erroneously assuming that Y affects D and estimating a reverse (or "anticausal") association of D as a function of Y, the resulting error term will not be independent of Y. This holds true except for highly specific cases, as elaborated in Peters et al. (2017). Notably, in scenarios beyond these special cases, this phenomenon enables the identification of the correct causal model, wherein D affects Y. In this correct model, the error term is independent of D. In practice, testing the causal direction may be based on running nonlinear regressions of both Y on D and D on Y and verifying in which of these two cases the estimated errors (or residuals) are independent of the regressors. To this end, we can apply a statistical test for the independence of the estimated errors and regressors in either regression and choose that causal model as the supposedly correct one for which the test yields a higher (and statistically insignificant) p-value. To avoid overfitting bias when computing the p-values, sample splitting is advisable, which implies estimating the regression models and conducting the independence tests in distinct subsamples obtained from randomly splitting the full sample. Sample splitting ensures that the regression and testing stages are not statistically associated with each other.

The model in Eq. (1) can be extended or modified in various ways, all while preserving the ability to identify causality. For instance, Breunig and Burauel (2021) consider testing the direction of causality between two variables D and Y when controlling for observed covariates, denoted by X, while allowing for heteroskedasticity of ε in X. As the independence between ε and D is only required to hold conditional on X, this implies a type of selection-on-observables assumption. Formally, the setup is defined by the following nonlinear model:

$$Y = \mu(D, X) + \varepsilon, \tag{2}$$

where the variance of error term ε may differ across values of X.

Furthermore, Zhang and Chan (2006) and Zhang and Hyvärinen (2009b) consider a generalization of Eq. (1), known as post-nonlinear (PNL) model, which consists of two nested nonlinear functions:

$$Y = q(\mu(D) + \varepsilon), \tag{3}$$

where both q and μ are nonlinear functions and q is assumed to be invertible. It is evident that Eq. (1) is a special case of Eq. (3) when q is defined as the identity function. For this reason, Eq. (3) is more general, because it permits Y to be a complex (rather than additive) function of the impact of the treatment, characterized by $\mu(D)$, and of unobserved variables, characterized by ε. Another special case of both Eqs. (1) and (3) is a model in which the causal associations between observed variables D and Y are linear rather than nonlinear:

$$Y = \alpha + \beta D + \varepsilon, \tag{4}$$

where α is the constant term and β is the causal effect of D on Y. Shimizu et al. (2006) show that the correct causal model is identified if the additively separable error term ε is non-normally distributed. The latter assumption is crucial, as unique identification in linear models is generally not feasible if errors are normally distributed, unless additional assumptions are imposed. For instance, Peters and Buhlmann (2013) show that in linear models with normally distributed errors, causality can still be learned under the (strong) assumption that the error terms in the various equations of a causal model all have the same variance.

We subsequently illustrate the implementation of causal discovery algorithms with a simple empirical example using the statistical software R. To this end, the following discussion will present some syntax in R, which can be skipped by readers who are not interested in this detail. We first install the R packages *bnlearn* and *causalweight* created by Scutari (2010) and Bodory and Huber (2018), respectively, using the command *install.packages("bnlearn", "causalweight")*. We then load both packages using the *library* command. Next, we utilize the *data* command to load the *coffeeleaflet* data set, which stems from an experimental study aimed at evaluating the impact of a leaflet discussing coffee production implications on the environmental awareness of students in Bulgaria.

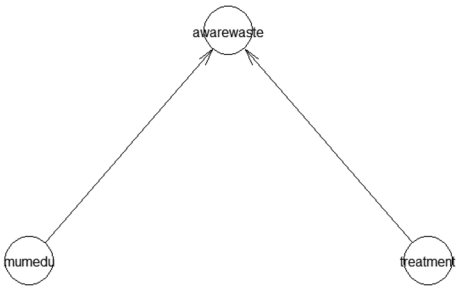

Fig. 2 Collider structure

We create a new frame set named *data*, which only contains the pretreatment characteristic "mother's education" (the variable *mumedu* in the 8th column of the *coffeeleaflet* data), the randomly assigned leaflet treatment (the variable *treatment* in the 32nd column), and the outcome "awareness of waste production due to coffee production" measured on a five-point scale (the variable *awarewaste* in the 38th column): *data=coffeeleaflet[,c(8,32,38)]*. As the variable *treatment* is randomly assigned, it should be independent of *mumedu*, while both variables may affect the outcome *awarewaste*. In this case, we have the causal relation *treatment*→*awarewaste*←*mumedu*, such that *awarewaste* is a collider. We feed our data set into the *pc.stable* function to conduct causal discovery based on the PC algorithm and store the results in an R object named *output*. Finally, we employ the *plot* command to visualize the causal graph derived from the PC algorithm. The box below provides the R code for each of the steps.

```
install.packages("bnlearn", "causalweight") # install packages
library(bnlearn)                            # load bnlearn package
library(causalweight)                       # load causalweight package
data(coffeeleaflet)                         # load coffeeleaflet data
data=coffeeleaflet[,c(8,32,38)]             # select variables nr. 8, 32, and 38
output=pc.stable(data)                      # run the PC algorithm
plot(output)                                # plot the learned causal model
```

Executing the code produces the DAG depicted in Fig. 2. We see that the algorithm detects the previously mentioned collider structure, implying that mother's education and treatment assignment are independent (as expected in a well conducted experiment), while they both affect the awareness outcome. As mentioned before, such algorithms might fail to uniquely determine the direction of the causal arrows in more complicated models with more variables and more involved dependence structures than in our toy model containing a single collider. And even in our toy model, we acknowledge that the causal arrow from mother's education to the awareness outcome might be confounded, as mother's education is not randomly assigned.

As a further R example, we consider learning the direction of causality in nonlinear models characterized by Eq. (2) by testing the independence of the estimated nonlinear regression function and the error terms (or residuals). Assuming that all required R packages have been previously installed (using the *install.packages* command), we apply the *library* command to load the *dHSIC* package for independence testing, the *mgcv* package for running nonlinear regressions, and the *datarium* package. The latter contains the *marketing* data set consisting of 200 observations with information on sales and advertising budgets, which we load using the *data* command. We are interested in learning the causal relationship between the variable *newspaper*, which measures the budget of advertisement in newspapers and which we define as treatment *D*, and the variable *sales*, which we define as outcome *Y*.

We set a seed (*set.seed(1)*) for the replicability of our results to follow and use the *gam* function to estimate two different regression models. In the first regression, we estimate the outcome *Y* as a nonlinear function of the treatment *D*: *model1 = gam(Y ∼ s(D))*, where the wrapper *s(D)* implies that the association is measured by series regression. We store the output in an R object named *model1*. In the second (reverse) regression, we estimate the treatment *D* as a nonlinear function of the outcome *Y* and store the output in the object *model2*. We then feed the residuals of the first regression model, *model1$residuals*, and *D* into the *dhsic.test* command to test the independence between both variables and save the results in object *test1*. Using the same command, we test the independence between the residuals of the second regression model and *Y* and save the results in object *test2*. Finally, we investigate the *p*-values of the tests stored in *test1$p.value* and *test2$p.value*.

```
library(dHSIC)                          # load dHSIC package
library(mgcv)                           # load mgcv package
library(datarium)                       # load datarium package
data(marketing)                         # load marketing data
D=marketing$newspaper                   # define treatment (newspaper advertising)
Y=marketing$sales                       # define outcome (sales)
model1=gam(Y ~ s(D))                    # estimate Y as nonlinear function of D
model2=gam(D ~ s(Y))                    # estimate D as nonlinear function of Y
set.seed(1)                             # set seed
test1=dhsic.test(model1$residuals, D)   # independence test for first model
test2=dhsic.test(model2$residuals, Y)   # independence test for second model
test1$p.value; test2$p.value           # show p-values
```

The tests yield p-values of 0.136 (or roughly 14%) and 0.012 (or roughly 1%) for the regressions of Y on D and of D on Y, respectively. Consequently, we find that a causal effect of the supposed treatment D on outcome Y cannot be rejected at the 10% level of statistical significance. In contrast, a reverse causal effect from Y to D is rejected at the 5% level of significance.

4 The back-door and front-door criteria

A concept closely related to d-separation discussed in Sect. 2 is the back-door criterion, see, e.g., Pearl (2000). It provides a causal graph-based framework for determining variables that must be controlled for to prevent confounding bias when measuring the causal effect of a treatment variable D (like education), on an outcome variable Y (like earnings). More concisely, the back-door criterion employs d-separation to detect an appropriate set of control variables such that any confounding association between D and Y (e.g., through characteristics like socio-economic status), also known as back-door path, is eliminated or blocked. More formally, a set of covariates X satisfies the back-door criterion in a DAG where the effect of D on Y is of interest,

1. if X blocks any path between D and Y that contains an arrow (causal effect) into D,
2. if no variable in X is causally affected by D.

To provide an example illustrating the concept of a back-door path and the back-door criterion, let us reconsider the left graph in Fig. 1. In this case, the presence of covariates X (like socio-economic status) that jointly affect D (education) and Y (earnings) forms a back-door path ($D \leftarrow X \rightarrow Y$) which confounds the causal relation of D and Y. Controlling for X satisfies the back-door criterion, as it closes the back-door path

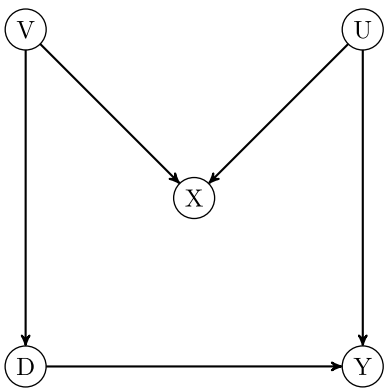

Fig. 3 M-bias

(i.e., controls for confounding), while not including any variable that is causally affected by D. This implies that the so-called selection-on-observables or unconfoundedness assumption, which is frequently invoked in policy or treatment evaluation studies, is satisfied with respect to D, meaning that the treatment is as good as randomly assigned conditional on X. Consequently, the treatment effect of D on Y is identified when controlling for X. For a binary treatment D, the average treatment effect (ATE) of providing everyone vs. no-one with the treatment, henceforth denoted by Δ, is, for instance, identified by the following expression:

$$\Delta = E[E[Y|X, D = 1] - E[Y|X, D = 0]]. \tag{5}$$

See, for instance, Imbens (2004); Imbens and Wooldridge (2009), and Abadie and Cattaneo (2018) for a discussion of the unconfoundedness assumption and evaluation methods for ATE estimation like regression, matching, weighting, or combinations thereof like doubly robust estimation.

In contrast to the previous example, Fig. 3 presents a causal model where controlling for X does not satisfy the back-door criterion. Although the back-door condition that X is not causally affected by D holds, the condition that X blocks any path between D and Y containing an arrow into D is not met. This situation arises because X is affected by both V and U and thus a collider, such that controlling for X introduces dependence between V and U. As V affects D and U affects Y, this collider bias generates a dependence or back-door path between D and Y, leading to the failure of the back-door criterion. This specific form of collider bias is known as M-bias. However, the back-door criterion holds when not controlling for X, thus preventing collider bias, as V (which influences D) is not associated with Y when X is not controlled for. Alternatively, the back-door criterion can be met by controlling for X and either V or U, or both. The rationale here is that controlling for either V or U blocks the confounding bias between D and Y that is introduced by controlling for X.

It is worth noting that a web-based application called "DAGitty" (directed acyclic graph interactive tool) for creating, analyzing, and visualizing causal graphs is available at https://dagitty.net/. It makes use of the back-door criterion to determine if and under which conditions the effect of a treatment D on an outcome Y is identified in a causal model. DAGitty provides a visual representation of causal graphs, allows differentiation between observed and unobserved variables, and highlights all back-door paths between D and Y that need to be controlled for even in complex models with numerous variables. This tool proves useful for researchers and analysts in comprehending the underlying causal structure and making

correct decisions about the feasibility of identifying treatment effects based on observed variables. If identification is possible, DAGitty also indicates the covariates that need to be controlled for. DAGitty is also available as `R` package, which is provided by Textor et al. (2017).

A further criterion for identifying the causal effect of D on Y is the so-called front-door criterion, see Pearl (2000), which does not rely on blocking all back-door paths between D and Y. Yet, it builds on the back-door criterion by applying it sequentially with respect to the effect of D (like education) on a mediating variable M (like human capital acquisition), through which presumably any effect of D on Y (like earnings) operates, and the effect of M on Y. Formally, a set of variables M satisfies the front-door criterion for identifying the causal effect of D on Y,

1. if any effect of D on Y operates via M,
2. if there is no unblocked back-door path from D to M,
3. if all back-door paths from M to Y are blocked when controlling for D.

The first assumption states that M fully mediates the effect of D on Y. The second assumption imposes that no variables jointly affect D and M, ensuring that the treatment-mediator relation is unconfounded. The third assumption imposes that conditional on D, no variables jointly affect M and Y, such that the mediator-outcome relation is unconfounded conditional on the treatment. Alongside these assumptions, the identification of causal effects based on the front-door criterion requires a specific common support restriction: For a binary treatment D, it must hold that $0 < \Pr(D = 1|M) < 1$, ensuring that both treated and untreated observations exist for each possible value of M in the population.

The front-door criterion is related to both causal mediation models and instrumental variables, see, for instance, Huber (2021) and Huber and Wüthrich (2019) for literature surveys. The assumptions imply "complete" mediation in the sense that there only exists an indirect effect of D on Y via M but no direct effect, similar to the exclusion restriction in instrumental variable methods. The left graph in Fig. 4 provides an illustration of the front-door criterion. U denotes unobserved confounders of D and Y, where the dashed arrows imply that the causal paths from U cannot be observed. In contrast, confounders of the causal association between treatment D and mediator X (which is on the causal path between D and Y) or between mediator M and outcome Y conditional on D must not exist. This still permits unobservables affecting M (which are omitted in the graph), as long as those unobservables do not affect D or Y, too, and are not associated with U. If the front-door criterion holds, the ATE corresponds to the following expression, see, for instance, Frölich and Sperlich (2019):

$$\Delta = E[\nu(M)|D = 1] - E[\nu(M)|D = 0]. \tag{6}$$

where $\nu(m) = E[E[Y|D, M = m]]$ is a nested mean outcome, in which the conditional mean outcome $E[Y|D, M = m]$ is averaged over values of the treatment D, while keeping the mediator fixed at value $M = m$. For a binary treatment, it holds that $\nu(m) = E[Y|D = 1, M = m] \cdot \Pr(D = 1) + E[Y|D = 0, M = m] \cdot \Pr(D = 0)$, with Pr denoting a probability.

Equation (6) calculates ATE as the average change in outcome Y due to a shift in mediator M induced by switching the treatment from $D = 1$ to $D = 0$, which utilizes the first assumption that any effect of D on Y operates through M. Furthermore, the third assumption ensures that D blocks all back-door (or confounding) paths from M to Y, implying that by controlling for D in the nested mean outcome $\nu(M)$, we can assess the effect of M on Y by varying the values of M in $\nu(M)$. Finally, given the second assumption that there is no back-door path between D and M, the effect of D on M is identified as well. As a result, we can compute the mean potential outcomes and thus, the ATE by simply averaging $\nu(M)$ within the treated and non-treated groups.

The front-door criterion has been rarely applied in empirical research and is virtually absent in social sciences. One potential application discussed in Pearl (2000) is the assessment of the causal effect of smoking (D) on lung cancer (Y) mediated by tar deposits (M). If unobserved genetic factors (U) simultaneously influence smoking behavior and lung cancer, then the back-door criterion fails for D and Y due to confounding. Nevertheless, the effect of smoking on lung cancer can potentially be identified if the front-door criterion is met. This criterion necessitates that there are no unobservable variables affecting both smoking behavior and tar deposits, or tar deposits and lung cancer, and that smoking solely affects lung cancer through tar deposits (exclusion restriction). As noted, e.g., by Imbens (2020), the restrictions on the

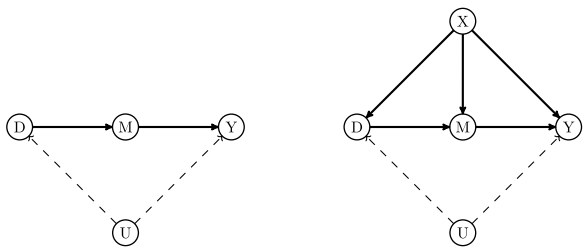

Fig. 4 Front-door criterion without and with controlling for covariates (left and right)

unobservables imposed by the front-door criterion fail in this application if genetic factors influence smoking behavior, the inclination to accumulate tar, and lung cancer altogether, or if unobservables (like a hazardous working environment) jointly affect tar deposits and lung cancer. Additionally, if there exist alternative causal pathways through which smoking behavior affects lung cancer beyond tar deposits, then the exclusion restriction assumption fails. In numerous empirical contexts, meeting the requirements of the front-door criterion appears challenging, potentially explaining its limited prevalence in practical applications.

However, it is worth mentioning that the assumptions underlying the front-door criterion can be relaxed to only hold conditionally on observed variables. Indeed, the assumptions might appear more plausible in empirical applications when allowing covariates X to affect all of D, M, and Y, as illustrated in the right graph of Fig. 4. In this case, the ATE is identified when controlling for X. Consequently, this entails modifying expression (6) to

$$\Delta = E_X[E_M[\nu(M,X)|D=1,X] - E_M[\nu(M,X)|D=0,X]], \tag{7}$$

where $\nu(m,x) = E[E[Y|D,M=m,X=x]|X=x]$ is a nested conditional mean outcome given X. Intuitively, one performs similar calculations as for the previous case without covariates, but this time within groups that share the same values of the covariates X to obtain the conditional ATE (CATE), assuming that the front-door criterion holds conditional on X. Finally, one averages the CATE across all values of X (as indicated by the subscript X in the expectation operator E_X) to obtain the ATE for the entire population.

5 Examples of causal discovery

As discussed in Sect. 3, attempting to uncover all causal relationships between all observed (or even unobserved) variables in a causal model can be very challenging or practically infeasible in many empirical scenarios. This section delves into a few examples where causal discovery is applied to somewhat more modest objectives than learning an entire causal structure. Yet, these objectives may appear interesting from the perspective of economics or social sciences and can be more feasible from a practical standpoint.

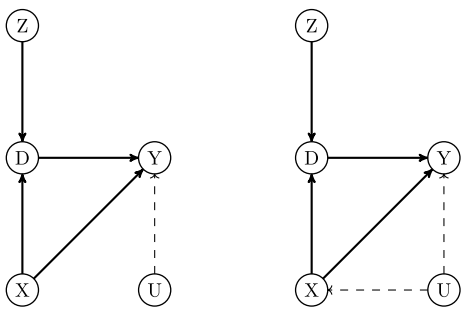

Fig. 5 Learning treatments and/or the identification of treatment effects

For example, researchers or analysts might have an interest in identifying all observed variables that directly influence a specific outcome variable Y. In other words, one aims at detecting all treatments exerting a direct causal impact on Y. For instance, we may want to determine all observed variables that have an effect on earnings (which could include education, profession, work experience, labor market conditions, among others) or affect health (which could include health behavior, genetic factors, living/working environment, among others). Learning treatments requires imposing specific identifying assumptions, such as the condition that any variables jointly affecting any treatment and the outcome are observed in the data. This implies that the selection-on-observables assumption holds for each of the potentially numerous treatments under consideration, which is a strong condition. Furthermore, we need to rule out reverse causal effects of the outcome Y on the treatments. This appears plausible if the observed variables are measured at an earlier point in time than the outcome, as variables measured later cannot impact the values of variables measured earlier.

The graph on the left in Fig. 5 provides a visual illustration, in which the observed variables D and X causally affect the outcome Y. In contrast, Z is not part of the treatments directly affecting Y, as it only exerts an indirect effect on the outcome through D. For both D and X, the selection-on-observables assumption holds in the left graph, as the unobserved term U affecting the outcome (where the non-observability of the effect is indicated by the dashed line) neither affects D nor X. Conversely, the selection-on-observables assumption fails for X in the right graph of Fig. 5, where U jointly affects X and Y.

Soleymani et al. (2022) provide an algorithm in this context that selects treatment variables in a data-driven way, while also controlling for observed covariates in a data-adaptive manner based on the double machine learning (DML) framework, a causal machine learning approach suggested by Chernozhukov et al. (2018). The algorithm sequentially considers each of the observed variables that may affect Y as treatment D to estimate the direct effect of that candidate treatment on the outcome Y, while considering all remaining observed variables as covariates X to be controlled for by DML. Finally, the algorithm retains only those observed variables as treatment variables that exhibit statistically significant effects on Y (where judging statistical significance should account for issues related to testing multiple

hypotheses based on multiple candidate treatments). Another machine learning-based algorithm proposed by Quinzan et al. (2023) adopts a somewhat distinct testing approach. In a first step, the algorithm estimates the outcome as a function of all observed variables by DML. In a second step, it repeats the estimation of the outcome when excluding one candidate treatment D. If the omission of D statistically significantly alters the outcome estimate compared to using all observed variables (as done in the first step), then D is retained as treatment variable; otherwise, it is dropped. By sequentially applying this process to all candidate treatments, the algorithm provides an estimated set of treatments that have a direct effect on Y.

Another example of applying causal discovery concepts involves learning whether sufficient conditions for identifying the causal effect of a treatment of interest (D) on an outcome of interest (Y) hold in observational data when imposing causal faithfulness (see Sect. 2). In contrast to the previous example, the objective here is not to learn all treatments influencing an outcome, but rather to assess whether the effect of a predefined treatment is identifiable within the available data. As discussed in de Luna and Johansson (2014); Black et al. (2015), and Huber and Kueck (2023), if there exists an instrument Z that affects D and is conditionally independent of Y given X, then this is a sufficient condition for the selection-on-observables assumption. Consequently, the causal effect of D on Y is identified conditional on X. This result follows from an application of the d-separation criterion: If covariates X permit controlling for all back-door paths (or confounders) between Z and Y as well as D and Y and if Z is a valid instrument in the sense that it does not directly affect the outcome Y other than through D, then controlling for D in addition to X leads to the independence of Z and Y. In particular, conditioning on D does not introduce collider bias (see Sect. 2) between Z and Y given X, implying that D is as good as randomly assigned conditional on X, such that the selection-on-observables assumption holds. In both graphs of Fig. 5, the testable implication holds for treatment D, implying that the effect of D on Y is identified conditional on X. Instrument Z affects outcome Y only through treatment D and the unobservable U does not jointly affect Z and Y or D and Y conditional on X. Therefore, Z and Y are conditionally independent given X and D.

As an example, consider the case that conditional on observed socio-economic characteristics X, there are no unobserved factors jointly affecting participation in a training program (D) and earnings (Y), such that selection-on-observables holds. Furthermore, assume there is a random encouragement or invitation to attend the program (Z), which induces some of the (randomly chosen) invitees to attend the training, but has no direct effect on earnings, such that Z is an instrument for D. In this setup satisfying selection-on-observables and instrument validity, it holds that the encouragement is statistically independent of the earnings outcome conditional on training participation and the socio-economic characteristics. It is interesting to note that this approach can also be applied in the context of comparing treatment effects from experimental and observational data, as seen in the study by Morucci et al. (2023). In this context, Z is a binary indicator for experimental or observational data, while D, X, and Y represent the treatment, covariates, and outcome measurements in the respective datasets. The conditional independence of Z and Y given D and X in the joint data again implies the satisfaction of both the selection-on-observables assumption with respect to D and the instrument validity of Z. However, the latter assumption now bears a specific interpretation: Conditional on X, treatment effects are homogeneous across the experimental and observational data; otherwise, Z would have a direct association with Y through effect heterogeneity. Such heterogeneous effects could, for example, stem from differences in the timing or geographic location in which the experimental and observational data were collected, meaning that the treatment (like a training program) is more effective in some regions or periods (e.g., with a particularly high or low unemployment rate) than in others.

The previous insights on the implications of conditional independence can be used in a causal discovery approach to learn sets of instruments Z and covariates X that satisfy the conditional independence of Z and Y given X and D in the data (if such sets exist at all), as outlined in Apfel et al. (2024), rather than predefining Z and X. The method is based on considering all observed variables which are not the treatment D or the outcome Y and assuming one of them to be a candidate instrument Z, while the remaining ones (or subsets thereof) constitute the covariates X. In a first step, the first-stage effect of Z on D given X is estimated by DML to verify whether the conditional association between Z and D is strong enough, which is a precondition for testing. Only if this is the case, the second step consists of testing the conditional independence of Z and Y given X and D by DML. This two-step approach is iteratively repeated to each variable that is neither D nor Y, effectively cycling through all candidates for the role of the instrument Z.

For instance, assume that D is college education and Y is later life earnings, while the remaining variables are socio-economic characteristics like parental education, parental income, and place of residence, among others. The algorithm would initially designate parental education as Z and the remaining variables (or their subsets)

as X, followed by treating place of residence as Z with the remaining variables as X, and so forth. If for one or even several definitions of Z and X the conditional independence assumption is not rejected by the conditional independence tests, this signals the possible fulfillment of the selection-on-observables assumption, particularly in large samples with a reduced uncertainty in testing. Consequently, one can opt for the definition of control variables X (and Z) that yields the highest p-value when running the conditional independence test, thereby implying the lowest probability of violating conditional independence. Relatedly, Hassanpour and Greiner (2019) and Wu et al. (2021) propose so-called deep learning algorithms to simultaneously (rather than iteratively) learn sets of variables which are (1) observed confounders X that jointly affect a predefined treatment D and outcome Y, (2) instruments Z that only affect the treatment D, and (3) outcome predictors that only affect the outcome Y.

Another causal discovery method that combines elements of the previous examples is suggested in Peters et al. (2015). This approach aims at the identification of multiple treatments (like in the first example), which directly affect an outcome Y and satisfy a selection-on-observables assumption, by means of instruments (like in the second example), also referred to as "environments" in the computer science literature. If an instrument Z affects outcome Y only through one or several other observed variables and is independent of Y conditional on a set of candidate treatment variables (rather than a single treatment), then this set will include variables that have a direct causal effect on Y. Another way of putting this is that if, e.g., the conditional mean outcome does not depend on (or is invariant across) variations in instrument Z conditional on a set of candidate treatments $\tilde{D}$, such that $E[Y|\tilde{D}, Z] = E[Y|\tilde{D}]$, then $\tilde{D}$ contains variables affecting Y because this set intercepts any (average) effect of Z on Y.

In light of this, Peters et al. (2015) propose an algorithm that aims to find sets of $\tilde{D}$ for which the estimation of $E[Y|\tilde{D}]$ remains invariant across different values of Z. Verifying the invariance of the influence of $\tilde{D}$ on Y differs importantly from simply estimating the influence of $\tilde{D}$ on Y (without checking invariance across Z), as done in conventional regression approaches, where the predictive power of $\tilde{D}$ may spuriously reflect the causal effects of other variables rather than its own effect. However, the invariance algorithm's capability of detecting treatments directly affecting the outcome hinges on the strength of the association between Z and the treatments. If the first-stage association is weak or absent, the estimation of $E[Y|\tilde{D}]$ will be (almost) invariant across values of Z even if important treatments are missing in set $\tilde{D}$. For this reason, more and stronger instruments increase the chance of finding true treatments that directly affect Y.

A final example concerns testing the identification of the causal effect of D on Y (as in the second example) when substituting instrumental variable assumptions on Z, which rule out a direct effect of Z on Y or an association between Z and unobservables U affecting Y, with a different set of assumptions. Karlsson and Krijthe (2023) present a method under the assumption that Z, representing specific environments like distinct observational datasets, directly influences all of the covariates X, treatment D, outcome Y, and unobservables U (which impact Y). However, the mechanisms through which Z affects each of X, D, U, and Y must be independent of one another. In other words, Z is presumably characterized by a set of variables that are mutually independent, with each variable only influencing one of X, D, U, and Y. When represented as $Z = (Z_X, Z_D, Z_U, Z_Y)$, the assumption of independent causal mechanisms implies, for example, that the effect of Z_D on D must not be associated with the effect of Z_U on U.

Under these conditions, the selection-on-observables assumption for treatment D can be tested using the following algorithm. First, an observation with index i is randomly selected (e.g., $i = 1$ for the first randomly sampled observation) in each environment $z \in 1, 2, ..., \mathcal{Z}$, where $\mathcal{Z}$ is the number of different environments. Denoting $D_i(z)$ as the treatment state of the observation with index i in environment z, the next step involves aggregating the treatments of index i across all environments into a treatment vector, denoted as $\mathcal{D}_i = (D_i(1), D_i(2), ..., D_i(\mathcal{Z}))$. Proceeding analogously for covariates X and outcome Y, the vectors $\mathcal{X}_i$ and $\mathcal{Y}_i$ are obtained. The final step is to test whether, for any distinct indices $i \neq j$ (e.g., $i = 1$ and $j = 2$), $\mathcal{D}_j$ is conditionally independent of $\mathcal{Y}_i$ when controlling for $\mathcal{D}_i$, $\mathcal{X}_i$, and $\mathcal{X}_j$. A rejection of conditional independence indicates a violation of the selection-on-observables assumption. It is worth noting that the choice of observations assigned indices i and j within an environment is arbitrary if the samples in each environment are drawn randomly; the only requirement is that i and j cannot be assigned to the same observation within an environment.

As an empirical example in R, let us consider testing whether the causal effect of a predefined treatment is identified in observational data based on a predefined instrument, using the method of Huber and Kueck (2023). We load the *causalweight* package and use the *data* command to load the *JC* data from an experimental study of the Job Corps program, a training program for disadvantaged youth in the US, see Schochet et al. (2008). The goal is to test whether the average effect of training

participation in the first year after program assignment (D) on the health state four years after assignment (Y) is identified when controlling for the baseline covariates (X) measured prior to training. To this end, we test whether random assignment to Job Corps (Z), which has a strong first-stage effect on D, is conditionally mean independent of Y given D and X. To perform this test, we define the variables *Z=JC[,1]* (as the first column in the *JC* data contains the binary assignment variable), *D=JC[,37]*, *X=JC[,2:29]* (as columns 2 to 29 contain the pretreatment covariates), and *Y=JC[,46]*. Next, we feed these variables into the *identificationDML* command, a DML-based test that by default applies lasso regression to control for X and D in a data-driven way. We store the results in an `R` object named *output* and inspect the p-value of the test by calling *output$pval*. The box below provides the `R` code for each step.

The test yields a p-value of 0.26 (or 26%), which indicates that the null hypothesis of conditional mean independence between Z and Y given D and X cannot be rejected at conventional levels of statistical significance. For this reason, we do not find compelling statistical evidence for a violation of the selection-on-observables assumption concerning D or the instrument validity of Z when controlling for X. This suggests that the baseline covariates might be sufficiently rich to control for confounders jointly affecting D and Y.

In a next step, we load the *InvariantCausalPrediction* package developed by Meinshausen (2019) and apply the algorithm of Peters et al. (2015) to identify sets of observed variables that entail invariant estimations of the conditional mean outcome when assuming a linear outcome model. We now consider all variables in the *JC*

```
library(causalweight)                          # load causalweight package
data(JC)                                       # load JC data
Z=JC[,1]                                       # define instrument (assignment)
D=JC[,37]                                      # define treatment (training)
X=JC[,2:29]                                    # define covariates
Y=JC[,46]                                      # define outcome (health state)
output=identificationDML(y=Y, d=D, x=X, z=Z)   # run identification test
output$pval                                    # p-value of the test
```

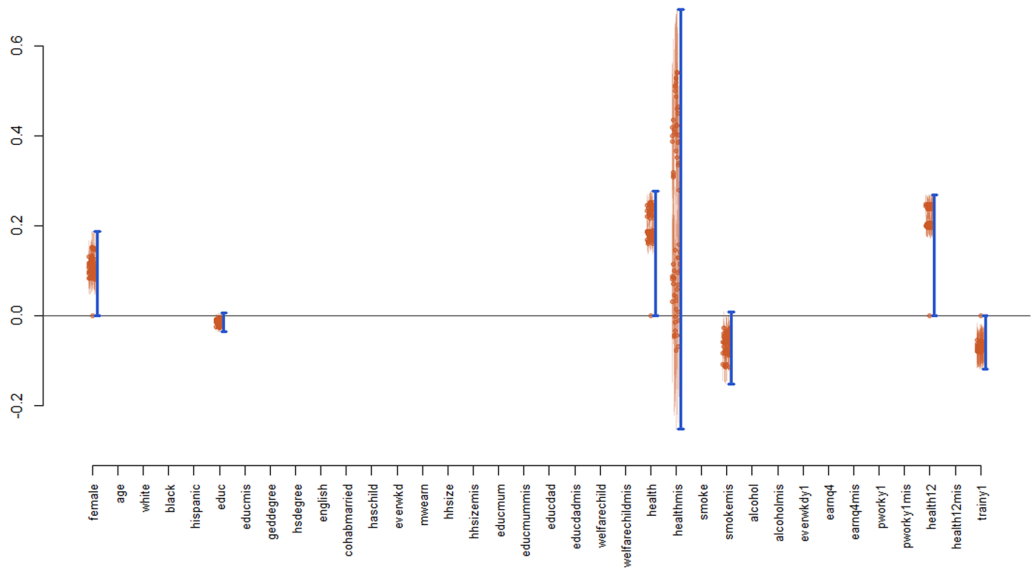

Fig. 6 Invariant causal prediction

data that were measured either before or during the first year after the random assignment to Job Corps as candidate treatment variables, namely, columns 2 to 37 of the data set. To this end, we define *X=as.matrix(JC[,c(2:37)])*, where the *as.matrix* command converts our data frame into a numeric matrix, as required by the algorithm. We note that X includes pretreatment covariates as well as training participation and covariates measured in the first year after program assignment. As before, random assignment to Job Corps is used as the instrument (Z), which presumably does not directly affect the health outcome Y, but only indirectly through elements in X (in particular, training participation). We feed X, Y, Z into the *ICP* command to execute the algorithm and set the argument *alpha=0.05* for testing the impact of candidate treatments at the 5% level of significance (including a correction for multiple hypothesis testing). We save the results in an object named *output* and wrap the latter by the *plot* command to plot the results. See the box below for the R code of the various steps.

```
library(InvariantCausalPrediction)      # load InvariantCausalPrediction package
X=as.matrix(JC[,c(2:37)])               # observables (candidate treatments)
output=ICP(X=X,Y=Y,ExpInd=Z,alpha=0.05) # algorithm invariant causal prediction
plot(output)                            # plot results
```

Running the code generates the graph in Fig. 6, which conveys information regarding the selection of variables in X as regressors in various outcome regression models. Each of these regression models adheres to the criterion that the estimation of the conditional mean outcome ($E[Y|\tilde{D}]$) based on the respective set of selected regressors ($\tilde{D}$) is rather invariant across values of the instrument Z. In this graph, the dots correspond to the estimates of the regression coefficients (y-axis) for different variables coming from X (x-axis) for each of the models that meet the invariance criterion. The slim vertical bars are the 95% confidence intervals of the estimated coefficients. In contrast, the fatter vertical bars correspond to the union (or combination) of the confidence intervals across all models satisfying the invariance condition, i.e., they are derived from the maximum upper and minimum lower bounds of any confidence interval across all relevant models.

Variables without any coefficient estimates and confidence intervals are never selected to be in the set $\tilde{D}$. Among the variables having nonzero coefficients (and thus, nonzero effects on the health outcome) in models satisfying the invariance criterion are gender, education, health at baseline and after the first year, missing dummies for smoking behavior and health at baseline, and training participation in the first year. However, for any of those variables, the union of the confidence intervals includes a zero effect. Nevertheless, we can observe, for instance, that the coefficients on training participation (*trainy1*) are often negative and never positive, suggesting a health-increasing effect (due to the inverse coding of the outcome). As already mentioned, more instruments (particularly those with large first-stage effects on elements in X) may be helpful in reducing estimation uncertainty and thus the length of the confidence intervals. Put differently, having more and stronger instruments reduces the ambiguity about which and how candidate treatments affect the outcome of interest.

6 Conclusion

The application of causal discovery, aimed at learning causal relationships among multiple variables in a data-driven manner, remains so far relatively uncommon in economics and social sciences. One likely reason is that assumptions and causal setups that permit unambiguous learning of causality in entire systems of variables appear typically too restrictive to match real-world complexity. Another reason is that rather than understanding causality among many variables, researchers frequently focus on studying the effects of specific treatment variables based on more narrowly defined (but depending on the empirical context still challenging) assumptions. Even if learning causal structures in complex systems might be a too ambitious goal, the concepts of causal discovery can nevertheless be useful for somewhat more modest, but from an economic or social science perspective yet interesting causal problems. For instance, they can help determine the set of treatments directly impacting a specific outcome or for deriving and verifying conditions that (when tested and not rejected in large samples) imply the identification of treatment effects in observational data. For this

reason, this study provided an introduction to fundamental concepts of causal discovery such as d-separation, causal faithfulness, Markov equivalence, or the back-door and front-door criteria, along with several empirical illustrations using the statistical software (R).

Acknowledgements
I am grateful to the editor and an anonymous referee for their helpful comments and suggestions.

Author contributions
Martin Huber is responsible for all contributions of this article.

Funding
This research was not financed by any funding agency.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Competing interests
The author declares no competing interests.

Received: 25 June 2024 Accepted: 27 September 2024
Published online: 29 October 2024

References

Abadie, A., & Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics, 10*, 465–503.
Abbring, J. H., & Van den Berg, G. J. (2003). The nonparametric identification of treatment effects in duration models. *Econometrica, 71*, 1491–1517.
Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.
Apfel, N., Hatamyar, J., Huber, M., & Kueck, J. (2024). Learning control variables and instruments for causal analysis in observational data. Preprint retrieved from arXiv:2407.04448.
Black, D. A., Joo, J., LaLonde, R. J., Smith, J. A., & Taylor, E. J. (2015). *Simple tests for selection bias: Learning more from instrumental variables*. In IZA Discussion Paper No 9346.
Bodory, H., & Huber, M. (2018). *The causalweight package for causal inference in R*. In SES Working Paper 493, University of Fribourg.
Breunig, C., & Burauel, P. (2021). Testability of reverse causality without exogeneous variation. Preprint retrieved from arXiv:2107.05936.
Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal, 21*, C1–C68.
Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. (2024). *Applied Causal Inference Powered by ML and AI*.
Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale University Press.
de Luna, X., & Johansson, P. (2014). Testing for the unconfoundedness assumption using an instrumental assumption. *Journal of Causal Inference, 2*, 187–199.
Frölich, M., & Sperlich, S. (2019). *Impact evaluation: Treatment effects and causal analysis*. Cambridge University Press.
Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics, 10*, 1–15.
Hassanpour, N., & Greiner, R. (2019). Learning disentangled representations for counterfactual regression.
Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65–70.
Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems, 21*, 689.
Huber, M. (2021). *Mediation analysis*. Springer.
Huber, M. (2023). *Causal analysis: Impact evaluation and causal machine learning with applications in R*. MIT Press.
Huber, M., & Kueck, J. (2023). Testing the identification of causal effects in observational data. Preprint retrieved from arXiv:2203.15890.
Huber, M., & Wüthrich, K. (2019). Local average and quantile treatment effects under endogeneity: A review. *Journal of Econometric Methods, 8*, 1–28.
Huntington-Klein, N. (2022). *The effect: An introduction to research design and causality*. Chapman and Hall/CRC.
Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics, 86*, 4–29.
Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature, 58*, 1129–1179.
Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*, 5–86.
Kalisch, M., & Bühlmann, P. (2014). Causal structure learning and inference: A selective review. *Quality Technology & Quantitative Management, 11*, 3–21.
Karlsson, R. K. A., & Krijthe, J. H. (2023). Detecting hidden confounding in observational data using multiple environments. Preprint retrieved from arXiv:2205.13935.
Lechner, M. (2009). Sequential causal models for the evaluation of labor market programs. *Journal of Business and Economic Statistics, 27*, 71–83.
Lechner, M. (2023). Causal machine learning and its use for public policy. *Swiss Journal of Economics and Statistics, 159*, 8.
Lieli, R. P., Hsu, Y.-C., & Reguly, Á. (2022). *The use of machine learning in treatment effect estimation*. In working paper, Central European University.
Mani, S., Spirtes, P. L., & Cooper, G. F. (2012). A theoretical study of y structures for causal discovery. Preprint retrieved from arXiv:1206.6853.
Meinshausen, N. (2019). *Invariantcausalprediction: Invariant causal prediction*. R package.
Morucci, M., Orlandi, V., Parikh, H., Roy, S., Rudin, C., Volfovsky, A. (2023). A double machine learning approach to combining experimental and observational data. Preprint retrieved from arXiv:2307.01449.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.
Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.
Peters, J., & Bühlmann, P. (2013). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika, 101*, 219–228.
Peters, J., Bühlmann, P., & Meinshausen, N. (2015). Causal inference using invariant prediction: Identification and confidence intervals. Preprint retrieved from arXiv:1501.01332.
Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. The MIT Press.
Peters, J., Mooij, J. M., Janzing, D., & Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research, 15*, 2009–2053.
Quinzan, F., Soleymani, A., Jaillet, P., Rojas, C. R., & Bauer, S. (2023). Drcfs: Doubly robust causal feature selection. In *International Conference on Machine Learning*, pp. 28468–28491. PMLR.
Reichenbach, H. (1991). *The direction of time* (Vol. 65). University of California Press.
Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Mathematical Modelling, 7*, 1393–1512.
Schochet, P. Z., Burghardt, J., & McConnell, S. (2008). Does job corps work? Impact findings from the national job corps study. *The American Economic Review, 98*, 1864–1886.
Scutari, M. (2010). Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software, 035*(i03), 1.
Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., & Jordan, M. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research, 7*, 2003.
Soleymani, A., Raj, A., Bauer, S., Schökopf, B., & Besserve, M. (2022). Causal feature selection via orthogonal search. *Transactions on Machine Learning Research*.
Spirtes, P., & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review, 9*, 62–72.
Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

Textor, J., van der Zander, B., Gilthorpe, M. S., Liskiewicz, M., & Ellison, G. T. (2017). Robust causal inference using directed acyclic graphs: The R package 'dagitty'. *International Journal of Epidemiology, 45*, 1887–1894.

Verma, T., & Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 255–270.

Verma, T., & Pearl, J. (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. *Uncertainty in artificial intelligence* (pp. 323–330). Elsevier.

Wu, A., Kuang, K., Yuan, J., Li, B., Wu, R., Zhu, Q., Zhuang, Y., & Wu, F. (2021). Learning decomposed representation for counterfactual inference. Preprint retrieved from arXiv:2006.07040.

Zhang, K., & Chan, L.-W. (2006). Extensions of ICA for causality discovery in the hong kong stock market. In *International Conference on Neural Information Processing*, pp. 400–409. Springer.

Zhang, K., Hyvärinen, A. (2009a). Causality discovery with additive disturbances: An information-theoretical perspective. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II 20*, pp. 570–585. Springer.

Zhang, K., & Hyvärinen, A. (2009b). On the identifiability of the post-nonlinear causal model. In *25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pp. 647–655. AUAI Press.

Publisher's Note