Theoria, 2022, 37(1), 75-86
https://doi.org/10.1387/theoria.22904

THEORIA

ISSN 0495-4548 – eISSN 2171-679X

Computational causal discovery: Advantages and assumptions

Descubrimiento causal computacional: ventajas y asunciones

Kun Zhang*
Department of Philosophy, Carnegie Mellon University

ABSTRACT: I would like to congratulate James Woodward for another landmark accomplishment, after publishing his *Making Things Happen: A Theory of Causal Explanation* (Woodward, 2003). *Makes Things Happens* gives an elegant interventionist theory for understanding explanation and causation. The new contribution (Woodward, 2022) relies on that theory and further makes a big step towards empirical inference of causal relations from non-experimental data. In this paper, I will focus on some of the emerging computational methods for finding causal relations from non-experimental data and attempt to complement Woodward's contribution with discussions on 1) how these methods are connected to the interventionist theory of causality, 2) how informative the output of the methods is, including whether they output directed causal graphs and how they deal with confounders (unmeasured common causes of two measured variables), and 3) the assumptions underlying the asymptotic correctness of the output of the methods about causal relations. Different causal discovery methods may rely on different aspects of the joint distribution of the data, and this discussion aims to provide a technical account of the assumptions.

KEYWORDS: causal direction; interventionist theory; linear, non-Gaussian causal model; confounders; faithfulness.

RESUMEN: *Quiero dar la enhorabuena a James Woodward por* Flagpoles anyone? *(Woodward, 2022), una contribución que supone un nuevo hito tras la publicación de* Making Things Happen: A Theory of Causal Explanation *(Woodward, 2003).* Making Things Happen *ofrece una elegante teoría intervencionista para entender la explicación y la causación. Esta nueva contribución (Woodward, 2022) se apoya en esa teoría y da grandes pasos hacia la inferencia empírica de relaciones causales a partir de datos no experimentales. En este artículo, me centro en algunos métodos computacionales emergentes para encontrar relaciones causales a partir de evidencia no experimental y trato de complementar la contribución de Woodward discutiendo: 1) cómo estos métodos se conectan con la teoría intervencionista de la causalidad; 2) cómo de informativos son los resultados de estos métodos, incluyendo si producen gráficos causales dirigidos y cómo tratan los confusores (causas no medidas comunes a dos variables medidas); y 3) las asunciones subyacentes a la corrección asintótica de los resultados de estos métodos de descubrimiento causal. Diferentes métodos pueden basarse en aspectos diferentes de a distribución conjunta de los datos. Esta discusión pretende dar una explicación técnica de tales asunciones.*

PALABRAS CLAVE: dirección causal; teoría intervencionista; modelo causal lineal, no-Gaussiano; confusores; fidelidad.

*** *Correspondence to:*** Kun Zhang. Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, USA – kunz1@cmu.edu – https://orcid.org/0000-0002-0738-9958

How to cite: Zhang, Kun (2022). «Computational causal discovery: Advantages and assumptions»; *Theoria. An International Journal for Theory, History and Foundations of Science,* 37(1), 75-86. (https://doi.org/10.1387/theoria.22904).

Received: 2021-06-10; Final version: 2022-01-31.

ISSN 0495-4548 - eISSN 2171-679X / © 2022 UPV/EHU

1. *Introduction: Interventionist Theory of Causality and Discovering Causality*

Given two variables X and Y, we are concerned with the causal direction between them or the direction of explanation. Causal relations are potentially exploitable for the purpose of applying proper manipulations to achieve a certain goal, and it is naturally desirable to provide an interventionist account of causality (Woodward, 2003). Woodward (2022, p. 8) provides a simple version of the interventionist theory:

> (**M**) X causes Z if and only if (i) it is possible to intervene to change the value of X and (ii) under some such intervention on X, the value of Z would change.

An intervention on X is an unconfounded manipulation of X that changes any other variable, if at all, only through the change in X. That is, the intervention on X directly changes X, but does not directly change any other variable in the system. There are "hard" and "soft" interventions. A hard intervention breaks the connection from direct causes of X (except the intervention) to X, and a soft intervention does not break the connection from those direct causes to X but provides X with an exogenous source of variation that is independent from other causes of X (Eberhardt and Scheines, 2007).

As noted by Woodward (2003, 2022), the notion of intervention is itself a causal notion and as such has a notion of causal direction built into it. However, it provides a way to verify causal claims, if one is able to actually apply changes to the system that are confirmed to be interventions. For instance, that might be possible if time order information is available such that one can check whether the change to X would directly change any other variable. Even without time order information, sometimes we can make sure that the applied changes are valid interventions, thanks to the partial knowledge of the process; for instance, gene knockout is an intervention on a gene that can be exploited for inferring regulatory networks (Pinna *et al.,* 2010). This also suggests that applying proper interventions is usually too expensive, too time-consuming, or even impossible.

On the other hand, if one thinks of the observed data (which clearly have multiple values) as produced under unknown or natural "interventions", then causal direction may be revealed by finding whether certain types of "interventions" actually exist in the data. This idea makes it possible to find causal direction by analyzing observed data, as argued by Woodward (2022) and demonstrated by the algorithms that were recently proposed for distinguishing cause from effects (Shimizu *et al.,* 2006; Zhang and Chan, 2006; Hoyer *et al.,* 2009; Zhang and Hyvärinen, 2009; Janzing *et al.,* 2012; Huang *et al.*).

In fact, the past decades witnessed much progress in discovering causal information from non-experimental data, known as causal discovery, and its successful applications. Since the 1990s, conditional independence relationships in the data have been exploited to recover the underlying causal structure to a certain extent (Spirtes *et al.,* 1993; Chickering, 2002). Recently it has been shown that algorithms based on properly defined functional causal models (FCMs) are able to find causal direction between two variables and hence estimate the underlying causal Directed Acyclic Graph (DAG) uniquely. Can we trust the "causal" information produced by such computational causal discovery methods? How are the methods related to the interventionist theory of causality? Under what assumptions can those method produce causal information? Below we will focus on those questions.

2. Relating Interventionist Theory to Causal Direction Determination between Two Variables

Woodward (2022) suggests that "when one infers causal direction on the basis of non-experimental information what one is in effect doing is inferring what would happen if various interventions were to be performed without actually doing the interventions, relying instead on other features present in such situations—the independence/invariance features." Indeed, we can think of the goal of causal discovery from non-experimental data as finding the footprint of unknown interventions that were applied (by nature, for instance) on the data.

Woodward formulates the Causal to Statistical Independence (**CSI**) assumption, which says that variables that are causally independent are statistically independent. This assumption can be seen as a weaker version of the Causal Markov Condition (CMC) (Kiiveri *et al.,* 1984; Glymour *et al.,* 1987), and it is generally plausible. **CSI** motivated the following principle for inferring causal direction (Woodward, 2022, p. 26):

> (**P**) Suppose there are 3 variables, H, A and S such that either (i) H and A cause S or (ii) A and S cause H. (This assumption, in some way, implies that there are no omitted common causes etc.) Suppose the patterns of dependence among these three variables are as follows: $H \perp\!\!\!\perp A$, $H \not\perp\!\!\!\perp S$, $A \not\perp\!\!\!\perp S$, where $\perp\!\!\!\perp$ means statistical independence and $\not\perp\!\!\!\perp$ means statistical dependence. Then (i) is the correct causal order.

Woodward (2022, p. 28) then connects principle **P** to the interventionist theory **M** in the following way—this connection is desirable since one aims to use principle **P** to find causality which can be understood in terms of interventions:

> According to the interventionist framework, the claim that H causes S and S does not cause H corresponds to the condition that there are interventions on H that will change S but no interventions on S that will change H. Assuming that these are the only two possibilities (i) and (ii) and that there is no common causes, as stated in principle **P**, the pattern of (in)dependencies $H \perp\!\!\!\perp A$, $H \not\perp\!\!\!\perp S$, and $A \not\perp\!\!\!\perp S$ suggests that A functions as a soft intervention variable on S, since it is exogenous and independent of the only other possible cause of S, namely H. [Let us denote this statement by $\mathbf{J}_1$.] Observation shows that changes in this intervention variable A for S are not associated with changes in H, suggesting that S does not cause H. Moreover, if we assume that S causes H, then, under this assumption, there will not be, among the variables in the system, any intervention variable for H that is independent of S, since the only remaining variable, A, is correlated with S. [Denote by $\mathbf{J}_2$ this statement.] Hence, the dependence pattern suggests there is a route to changing S that is independent of H (which is what we expect if H causes S)—namely the route involving A—but no route to changing H that is independent of S, which is what we expect if S causes H.

Woodward (2022) [Section 9] then goes one step further, to consider the problem of inferring causal direction when only two variables, X and Y, are giving and justify a set of methods to solve this problem. Assume the causal influence from the cause variable and the unmeasured factors (noise) to the effect variable follows some constrained Functional Causal Model (FCM) class, such as the Linear, Non-Gaussian, Acyclic Model (LiNGAM) (Shimizu *et al.,* 2006), Post-NonLinear (PNL) causal model (Zhang and Chan, 2006; Zhang and Hyvärinen, 2009), and the Additive Noise Model (ANM) (Hoyer *et al.,* 2009). First of all, Woodward noted that given only two measured variables X and Y, the error

term is unobserved and must be inferred, in order to apply principle **P**. When we find an error U which is independent of X but no error $U\,l$ which is independent of Y, we infer that U and X are causes of Y. Such asymmetry between two variables X and Y that are assumed to be directly causally related (the estimated error term is independent from the hypothetical cause in only one direction), under the assumption of a properly constrained FCM class, has inspired several approaches to causal discovery that are able to recover the underlying causal DAG uniquely.

Why does the asymmetry in the (in)dependence between the error term and the hypothetical cause imply causal direction between two variables, denoted by H and S? This can be examined from an interventionist perspective, as an application of the arguments in the above quote to justify principle **P** for finding causal directions among 3 variables from the interventionist perspective. We notice that principle **P** makes use of *three variables* H, A, and S, while in the two-variable case, U is not observed but constructed from variables X and Y, in light of the constrained FCM class. In order to apply this principle to find causal direction when *only two variables*, H and S, are given, or more specifically, in order for statements **J**1 and **J**2 to hold true when only variables H and S are measured, the following conditions are expected to hold:

C1) Variable A is a variable that actually exists in the system under consideration. (As a consequence, $\mathbf{J}_1$ is true.)
C2) There does not exist another variable in the system, A', such that $S \perp\!\!\!\perp A'$ and $A' \not\perp\!\!\!\perp H$ (otherwise H and S will be symmetric). (As a consequence, $\mathbf{J}_2$ is true.)

As stated in the condition of principle **P**, it is assumed that H and A are not confounded in the first place. Can we make an alternative set of assumptions that are technically testable or appear weak in terms of the underlying causal model, while guaranteeing that the estimated causal direction from the two given variables is asymptotically correct? We will discuss the required assumptions in Section 4. Specifically, we will see some technical assumptions to imply condition C2 and assumptions on hidden variables in the system under which one can find causal direction even without the unconfounding assumption. Before that, let us review the assumptions that are required for conditional independence-based methods for causal discovery.

3. *Assumptions for Conditional Independence-Based Approaches*

Woodward (2022) focuses on finding the causal direction between two variables which are believed to be directly causally related. The issue of discovering causal information based on conditional independence relations among the measured variables seems somehow irrelevant to it. However, here for completeness of the discussion and for comparative purposes, let us briefly review such methods and discuss their assumptions.

Widely-used conditional independence-based approaches to causal discovery include the PC algorithm and FCI (Spirtes *et al.,* 2001). The PC algorithm returns a Markov Equivalence Class (MEC) of DAGs, and all DAGs in the MEC share the same adjacency and conditional independence relations. FCI allows confounders in the system and returns a Partial Ancestral Graph (PAG). Although the Greedy Equivalence Search (GES) (Chickering, 2002) is a score-based method for causal discovery that assumes no confounding, it

usually assumes a linear-Gaussian model or multinomial data and as a consequence, it also makes use of conditional independence relations among the measured variables, together with a penalty determined by the complexity of the whole causal model, to find the MEC.

Generally speaking, conditional independence relations among the measured variables, which reflect only part of the information implied by the joint distribution, do not contain sufficient information for producing a complete picture of the causal relations. First, even under the assumption of no confounding, the methods, such as PC and GES, return a MEC, which may contain multiple DAGs. For instance, if applied to two variables, they cannot determine their causal direction. Second, although there exist algorithms that can produce asymptotically correct results in the presence of confounders, such results are usually not strong enough to determine whether confounders exist. FCI is a remarkable algorithm whose result is asymptotically correct even with confounders. However, in the result by FCI, one usually cannot distinguish between unconfounded pairs of variables with direct causal relations between them and confounded pairs without direct causal relations in between—whenever it is possible to have confounders of the pair of variables, the algorithm will indicate it and, as a consequence, the output usually contains very few variable pairs that are directed causally related without cofounding.

Standard assumptions for the asymptotic correctness of the output of the above algorithms are the CMC and Faithfulness assumption. The Faithfulness assumption (Spirtes *et al.,* 2001; Zhang, 2013; Zhang and Spirtes, 2008) states that there is no 'accidental' conditional independence relation between the variables according to the distribution. More precisely, it says that all conditional independence relations among the variables are implications of the CMC applied to the DAG representing the true causal relations among the variables. Combining the CMC and Faithfulness assumptions, one is then able to recover some information of the underlying DAG from the measured data, given that we have enough data. Faithfulness can be viewed as one version of the "simplicity" assumption of the underlying DAG—if two variables are conditionally independent given any subset of the remaining variables, then they are not adjacent in the causal graph.

4. Independent Noise-Based Approaches: Assumptions on Causal Mechanisms

As Woodward (2022) noted, suitable assumptions on the causal mechanism (which are not made in conditional independence-based methods), such as a LiNGAM model to describe the causal effect, help find causal direction between two variables and hence recover the whole causal DAG, as supported by much empirical evidence. Below we start with an illustration of how and why LiNGAM, which is taken as an example of those properly defined constrained FCMs, helps find causal direction, and then discuss the required assumptions.

4.1. Asymmetry between Two Variables: Illustration

Consider the causal process $X \to Y$ with causal model $Y = dX + U$, where d is the linear coefficient and U is the noise (unmeasured factor) that is independent from X. As a concrete example, one can think of them as atmospheric pressure and the reading of a barometer, respectively. Suppose two linear regressions are done, one predicting X from Y and the other

predicting Y from X. The residual of the regression of X on Y is the difference of the measured variable and its predicted value from the regression. In the anti-causal direction the residual is a random variable, $U' = X - \alpha Z$, that is a function of the predictor variable and the predicted variable, or a function of the underlying noise terms. The coefficient α can be estimated by minimizing the total squares of the residual, which implies that the residual and predictor Z are uncorrelated. Bearing in mind that for simplicity of the presentation, all variables are assumed to be standardized (i.e., they have a zero mean and unit variance), one can see that

$$\alpha = \frac{\mathbb{C}\mathrm{ov}(X,Y)}{\mathbb{V}\mathrm{ar}(Y)} = \mathbb{C}\mathrm{ov}(X,Y) = d$$

where $\mathbb{C}\mathrm{ov}(\cdot)$ and $\mathbb{V}\mathrm{ar}(\cdot)$ denote covariance and variance, respectively. The residual of regressing Y on X (in the causal direction) is U, and the residual for regression in the anti-causal direction is

$$U' = X - \alpha Y = X - d(dX + U) = (1 - d^2)X - dU.$$

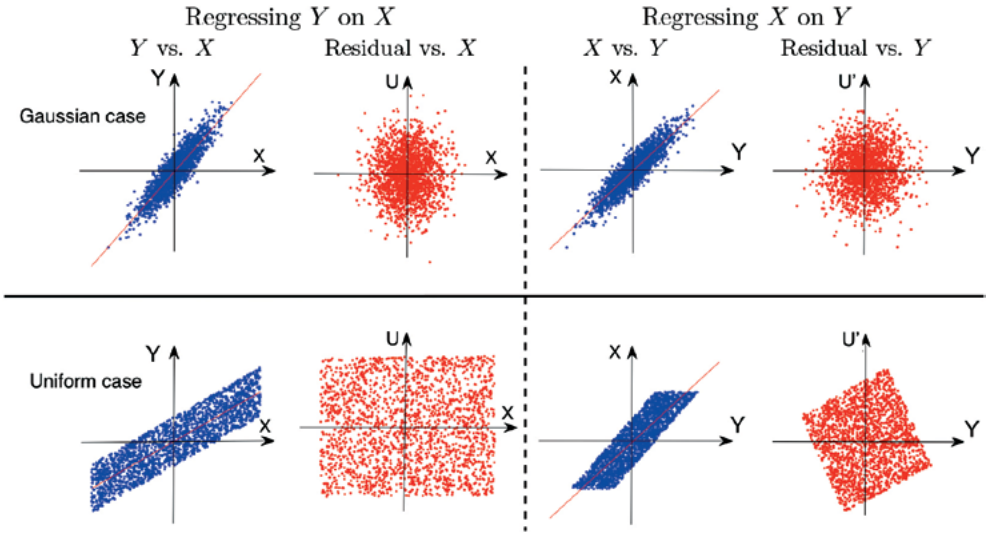


Figure 1

Illustration of the asymmetry between cause and effect in the linear, non-Gaussian case, where X causes Y with $Y = dX + U$. From left to right: the scatter plots for X and Y, for X and the residual of regressing Y on X, for Y and X, and for Y and the residual of regressing X on Y, respectively.
Top row: cause X and noise U are both Gaussian; bottom row: they follow a uniform distribution (a particular type of non-Gaussian distributions).

Can we see the asymmetry between X and Y by looking at the properties of the residuals U and U'?

U is assumed to be independent from X. However, since $Y = dX + U$, U' and Y both involve independent variables X and U, and they CANNOT be independent if at least one of X and U is non-Gaussian, as implied by the Darmois-Skitovich theorem (Kagan *et al.,* 1973) (this is related to condition C2): two linear combinations of a set of independent components cannot be independent from each other if they share any non-Gaussian independent component. So by testing for the independence between residuals and predictors, the direction of the $X - Y$ causal link can be identified. For illustrative purposes, Figure 1 provides scatter plots of variables X and Y and scatter plots of the predictor and regression residual in the causal (left part) and anti-causal (right part) directions, in a joint Gaussian case (top row) and in the case with cause X and the noise term following uniform distributions (bottom row). In the uniform case (as a particular non-Gaussian case), one can see that in the anti-causal direction, residual U' and predictor Y are clearly statistically dependent (because the conditional distribution of one of them given the other taking some specific value is not identical to its marginal distribution).

4.2. What Assumptions on Causal Mechanisms Are Required to Guarantee Asymmetry?

The linear, non-Gaussian model relies on the linearity of the causal mechanism, a particular type of parametric assumption, as well as the non-Gaussianity assumption of the noise terms, making it possible to estimate causal models from non-experimental data. Linear models are thought to be simple in the sense that they involve few parameters (e.g., as linear coefficients).[1] In fact, if there is no proper constraint on the FCM, for any two given random variables, one can always write one of them as a function of the other and some independent error term, as shown by Hyvärinen and Pajunen (1999) and Zhang *et al.* (2015), where the function is related to the conditional distribution of the variables and may be very complex. This is also the case in the example given in Figure 1—although U', the residual of regressing X on Y is not independent from Y in the uniform case, X can still be written as a *rather complex, clearly non-linear* function of Y and some error term that is independent from Y. Assuming linear, non-Gaussian causal models, one would infer the causal direction $X \rightarrow Y$ in this example. The linearity assumption can be checked by inspection on the scatter plots—given the measured data points, one can plot one variable against another variable, and the pattern should be approximately linear. In fact, statistical test of independence between the residual and the predictor (hypothetical cause) can also serve as a test of linearity—if the causal model is linear (resp., non-linear), then the residual produced by linear regression in some direction will be independent from (resp., dependent on) the predictor. If needed, one can also resort to specific tests of linearity, such as the Theil test (Theil, 1950), for this purpose. Test of non-Gaussianity can be performed implicitly or explicitly. If the residual is independent from the predictor only in one direction (i.e., only one DAG gives rise to independent residuals), then the predictor and residual cannot be

[1] Here is a brief explanation of why we prefer linear models from a model selection perspective. If multiple models explain the given data equally well, i.e., with the same likelihood, then suitable model selection approaches, such as Bayesian Information Criterion (BIC) (Schwarz, 1978), would prefer the model with the fewest number of free parameters.

jointly Gaussian. Alternatively, one may directly exploit statistical tests, such as the Shapiro-Wilk test, on the estimated residual or the predictor to check whether it follows the Gaussian distribution.

In the causal discovery community, this type of asymmetry between X and Y is essential for distinguishing cause from effect: under the assumed (parametric or non-parametric) FCM class, only in the causal direction one can find an independent noise term. This is asymmetry directly implied by the data distribution and the FCM class. In order to give causal claims, e.g., $X \to Y$, Woodward (2022) explicitly assumes that there is no confounder for X and Y. Below we give an alternative formulation of the assumption in the spirit of Faithfulness, to guarantee the connection between asymmetry implied by data and causal direction in the interventionist framework.

4.3. Assumptions on Hidden Variables to Connect Asymmetry to Causal Direction

Under the assumption that there was no confounder for X and Y, the above analysis does not require the traditional Faithfulness assumption for inferring causal directions asymptotically correctly (Shimizu *et al.,* 2006). Note that in reality we usually are not sure whether there are confounders and if yes, how measured variables are confounded in the true processes—can we still trust the result produced by the above LiNGAM analysis (performing linear regression and testing for independence between the residual and hypothetical cause)? Do we need any Faithfulness-like assumptions at all to guarantee the correctness of the statement?

Let us look at an illustrative example.

Example. Let us consider four variables—lifestyle, mortality risk, food consumption, and physical activity—denoted by X, Y, W, and U_X, respectively. Suppose that we can directly measure X and Y but not W and U_X. Measured variables X and Y are generated by the two hidden, independent, non-Gaussian variables, W and U_X, according to the following specific linear model, as shown in Figure 5(a), in which variables in the shaded area are not observed:

$$\begin{aligned} X &= U_X + fW, \\ Y &= fX - fU_X + (1 - f^2)W, \end{aligned}$$

where f is a positive number smaller than 1.

Because of the specification of the coefficients, we can rewrite the above model for Y as

$$Y = fX - fU_X + (1 - f^2)W = f(U_X + fW) - fU_X + (1 - f^2)W = W. \tag{1}$$

That is, mortality risk is determined by food consumption, although physical activity is also its cause. As a consequence, the statistical relationship between X and Y is

$$X = U_X + fW = fY + U_X.$$

Although there is no deterministic relation between measured variables, in this case (X causes Y, together with specific confounders W and U_X), we cannot identify the correct

causal direction with the LiNGAM analysis—in fact, the distribution of X and Y in this case can also be represented by the causal model in Figure 5(b), in which Y causes X without a confounder. What led to the wrong conclusion produced by the LiNGAM analysis? How can we avoid it? We notice here

1. that the confounder U_X (physical activity) actually has a zero effect on Y (mortality risk), although Y is its descendant in the causal graph, and
2. that although X (lifestyle) and Y (mortality risk) are non-deterministically related, they are completely determined by the confounders.

Now suppose neither of the above two properties holds. That is, we make the following assumptions:

AP1. Any hidden variable (which may be a confounder or unobserved noise variable) has a non-zero total causal effect on any of its descendants.
AP2. Each measured variable has non-zero noise relative to its parents (among measured variables and unmeasured confounders).[2]

Under assumptions AP1 and AP2, one can find causal direction in both the unconfounded and confounded cases. In the considered example, let us denote by g_{11}, g_{12}, g_{21}, and g_{22} the direct causal effects (linear coefficients) of U_X on X, U_X on Y, W on X, and W on Y, respectively, and still let f be the causal effect of X on Y. Then X and Y are generated from non-Gaussian independent variables, E_X (unobserved noise in X), E_Y (unobserved noise in Y), U_X, and W, in the following way:

$$X = g_{11}\, U_X + g_{21}\, W + E_X,$$
$$Y = g_{12}\, U_X + fX + g_{22} W + E_Y = (g_{12} + f g_{11}) U_X + (g_{22} + f g_{21}) W + f E_X + E_Y,$$

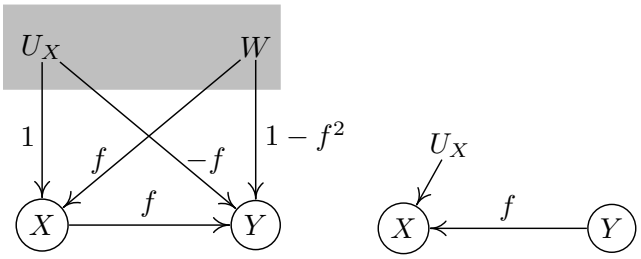
or in matrix form:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \underbrace{\left[\begin{array}{cc:cc} 1 & 0 & g_{11} & g_{21} \\ f & 1 & g_{12}+g_{11} & g_{22}+fg_{21} \end{array}\right]}_{\text{denoted by } \mathbf{A}_1} \cdot \left[\begin{array}{c} E_X \\ E_Y \\ \hdashline U_X \\ W \end{array}\right].$$

The above model, or specifically, the coefficients in matrix **A**1, is identifiable from X and Y with overcomplete ICA (Eriksson and Koivunen, 2004; Hoyer *et al.,* 2008). We can then see that under assumptions AP1 and AP2, we can find the causal relation between X and Y from the estimated matrix $\mathbf{A}_1$. From the second column of $\mathbf{A}_1$, we know that Y does not cause X—otherwise, given that E_Y influences Y, as indicated by the non-zero entry of the

[2] In some work, confounding is modeled by correlated noise; see, e.g. (Mandt *et al.,* 2017). Assumption AP2 is stronger than it. For instance, in Example 1 X and Y are non-deterministically correlated, but Assumption AP2 is violated.

second entry of this column, E_Y must influence X as well, according to Assumption AP1; this means that the first entry of this column cannot be zero, which is not the case. Similarly, we know that it is impossible that X does not influence Y—because if X did not influence Y, its noise term, E_X would influence only X, but not Y, and accordingly, there would be some column of $\mathbf{A}_1$ in which the first entry is non-zero while the second is zero, but there is no such column in $\mathbf{A}_1$. So the causal relation is $X \to Y$.



(a) A causal model with specific confounders, in which variables in the shaded area are not observable.

(b) A causal model with no confounders but a reverse direction between X and Y.

Figure 2

Two causal models used in the Example.
While Causal model (a) is the true one, it and model (b) produce the same joint distribution of X and Y.

Both Assumptions AP1 and AP2 are needed to find the correct causal direction between X and Y. If Assumption AP1 does not hold while AP2 holds, or more specifically, if E_X and E_Y are zero, then one can only recover the last two columns of $\mathbf{A}_1$, which does not imply an acyclic relation between X and Y. If Assumption AP2 does not hold while AP1 holds, e.g., the second entry of the third column of $\mathbf{A}_1$ is zero, i.e., $g_{12} + f g_{11} = 0$, then an alternative causal model consistent with Eq. (2) may be that X and Y do not have a direct causal influence between them (as seen from the second and third columns of $\mathbf{A}_1$) and that there are two confounders, EX and W (each influences both X and Y, as seen from the first and fourth columns). If both AP1 and AP2 are violated, linear, non-Gaussian methods may wrongly infer that Y causes X, as seen at the beginning of the Example.

5. *Conclusion*

Woodward (2022) discusses principles for finding causal direction between two random variables. As philosophical reflections on and justifications of the methods for distinguishing cause from effect recently proposed in machine learning, he connects the computational principles to his interventionist account of causality. In this paper we focus on independent noise-based methods, and attempt to provide alternative formulations of the assumptions that guarantee the correctness of the discover direction. Conditional inde-

pendence-based methods for causal discovery and their assumptions are briefly mentioned for completeness of the comparison.

In order to relate the information that is discovered from empirical data to the underlying causal structure, proper assumptions have to be made. Conditional independence-based methods assume some type of faithfulness: conditional independence in the data is not an accidental statistical property, but a reflection of the underlying causal graph. Independent noise-based methods assume "simplicity" of the causal mechanism, as encoded by the functional class of the functional causal models, to guarantee that cause and effect are asymmetric—only in the correct causal direction, the estimated noise term is independent from the hypothetical cause. Moreover, if one does not directly assume out confounders, some faithfulness-type assumption is needed to guarantee that the asymmetry that is discovered from empirical data actually implies causal direction.

Some of the assumptions, such as Woodward (2022)'s Causal to Statistical Independence (**CSI**) assumption, are widely accepted in the machine learning and philosophy communities. Some of the assumptions, such as linearity of the relations and non-Gaussianity of the noise, are generally testable. Some of them, including the faithfulness-type assumption AP1, are not generally testable.

It is worth noting that causal discovery is typically different from traditional machine learning problems such as regression, classification, and clustering, although both of them learn from data: causal discovery aims to find the underlying truth, while machine learning usually aims at good predictions. Therefore, first, it is essential to connect the principles underlying the computational methods to the interventionist theory of causality, to make sure that the causal discovery result actually has a causal interpretation. Second, researchers and practitioners in causal discovery have to pay close attention to the assumptions to guarantee the correctness (relative to the ground truth) of the result produced by computational methods for causal discovery, as pointed out by Woodward (2022).

Acknowledgements

I would like to acknowledge the support by the United States Air Force under Contract No. FA8650-17-C-7715, by National Institutes of Health under Contract No. R01HL159805, and by a grant from Apple. The United States Air Force or National Institutes of Health is not responsible for the views reported in this article. I am grateful to Clark Glymour for stimulating discussions and all his support.

REFERENCES

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507-554.

Eberhardt, F. & Scheines, R. (2007). Interventions and causal inference. *Philosophy of Science*, 74: 981-995.

Eriksson, J. & Koivunen, V. (2004). Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Processing Letters*, 11(7):601-604.

Glymour, C., Scheines, R., Spirtes, P., & Kelly, K. (1987). *Discovering Causal Strucure*. Academic Press.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., & Palviainen, M. (2008). Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362-378.

Hoyer, P. O., Janzing, D., Mooji, J., Peters, J., & Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, Vancouver, B.C., Canada.

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., & Schölkopf, B. (2020). Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21:1-53.

Hyvärinen, A. & Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429-439.

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B., & Schölkopf, B. (2012). Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1-31.

Kagan, A. M., Linnik, Y. V., & Rao, C. R. (1973). *Characterization Problems in Mathematical Statistics*. Wiley, New York.

Kiiveri, H., Speed, T., & Karlin, J. B. (1984). Recursive causal models. *Journal of theAustralian Mathematical Society (Series A)*, 36:30-52.

Mandt, S., Wenzel, F., Nakajima, S., Cunningham, J., Lippert, C., & Kloft, M. (2017). Sparse probit linear mixed model. *Machine Learning*, 106:1621-1642.

Pinna, A., Soranzo, N., & de la Fuente, A. (2010). From knockouts to networks: Establishing direct cause-effect relationships through graph analysis. *PLoS ONE*, 5.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461-464.

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

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search*. Spring-Verlag Lectures in Statistics.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis i. *Proceedings of the Royal Netherlands Academy of Sciences*, 53:386-392.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press, New York.

Woodward, J. (2022). Flagpoles anyone? Causal and explanatory asymmetries. *THEORIA. An International Journal for Theory, History and Foundations of Science,* 37(1), 7-52 (https://doi.org/10.1387/theoria.21921).

Zhang, J. (2013). A comparison of three occam's razors for markovian causal models. *British Journal of Philosophy of Science*, 64:423-448.

Zhang, J. & Spirtes, P. (2008). Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18:239-271.

Zhang, K. & Chan, L. (2006). Extensions of ICA for causality discovery in the Hong Kong stock market. In *Proc. 13th International Conference on Neural Information Processing (ICONIP 2006)*.

Zhang, K. & Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada.

Zhang, K., Wang, Z., Zhang, J., & Schölkopf, B. (2015). On estimation of functional causal models: General results and application to post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technologies*.

Kun Zhang is an associate professor of philosophy and an affiliate faculty in the machine learning department at Carnegie Mellon University. He has been actively developing methods for automated causal discovery from various kinds of data, investigating machine learning problems including transfer learning and representation learning from a causal perspective, and studying philosophical foundations of causation and various machine learning tasks.

Address: Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
E-mail: kunz1@cmu.edu
ORCID: 0000-0002-0738-9958