International Journal of Data Science and Analytics manuscript No.
(will be inserted by the editor)

Introduction to the Foundations of Causal Discovery

Frederick Eberhardt

Received: date / Accepted: date

Abstract This article presents an overview of several known approaches to causal discovery. It is organized by relating the different fundamental assumptions that the methods depend on. The goal is to indicate that for a large variety of different settings the assumptions necessary and sufficient for causal discovery are now well understood.

Keywords causality · graphical models · causal discovery

1 Introduction

Like many scientific concepts, causal relations are not features that can be directly read off from the data, but have to be inferred. The field of causal discovery is concerned with this inference and the assumptions that support it. We might have measures of different quantities obtained from, say, a cross-sectional study, on the amount of wine consumption (for some unit of time) and the prevalence of cardio-vascular disease, and be interested in whether wine consumption is a cause of cardio-vascular disease (positivey or negatively), and not just whether it is correlated with it. That is, we would like to know whether the observed dependence between wine consumption and cardio-vascular disease (suppose there is one) persists even if we change, say, in an experiment, the amount of wine that is consumed (see Fig. 1). The observed dependence between wine consumption and cardio-vascular disease may, after all, be due to a common cause, such as socio-economic-status (SES), where those people with a higher SES consume more wine and are able to afford better health care, whereas those with a lower SES do not consume as much wine and have poorer healthcare[1]. The example illustrates the common mantra that "correlation does not imply causation" and suggests that causal relations can be identified in an experimental setting, such as a randomized controlled trial where each individual in the experiment is randomly assigned to either the treatment or control group (in this case, to different levels of wine consumption) and the effect on cardio-vascular disease is measured. The randomized assignment makes the wine consumption independent of its normal causes (at least in the large sample limit) and thereby destroys the "confounding" effect of SES. Naturally, there are many concerns about such an analysis, starting from the ethical concerns of such a study, the compliance with treatment, the precise treatment levels, the representativeness of the experimental population with respect to the larger population etc., but the general methodological reason, explicitly emphasized in R.A. Fisher's well-known work on experimental design [6], of why randomized controlled trials are useful

This work was supported in part by NSF grant #1564330.

Frederick Eberhardt
California Institute of Technology
Tel.: +1-626-395-4163
E-mail: fde@caltech.edu

[1] See a discussion of this example in Scientific American [22].

for causal discovery becomes evident: randomization breaks confounding, whether due to an observed or unobserved common cause.

Causal relations are of interest because only an understanding of the underlying causal relations can support predictions about how a system will behave when it is subject to intervention. If moderate wine consumption in fact causes the reduction in the risk of cardio-vascular disease (this article takes no stand on the truth of this claim), then a health policy that suggests moderate wine consumption can be expected to be effective in reducing cardio-vascular disease (with due note to all the other concerns about implementation). But if the observed dependence is only due to some common cause, such as SES, then a policy that changes wine consumption independently of SES would have no effect on cardio-vascular disease.

A purely probabilistic representation of these relations is ambiguous with respect to the underlying causal relations: That is, if we let wine consumption be X and cardio-vascular disease be Y, then, without further specification, $P(Y|X)$, the conditional probability of cardio-vascular disease given a particular level of wine consumption, is ambiguous with regard to whether it describes the relation in an experimental setting in which the wine consumption was determined by randomization or whether it describes observational relations, such as in the initial example of a cross-sectional study. Judea Pearl introduced the $do(.)$-operator as a notation to distinguish the two cases [31]. Thus, $P(Y|X)$ is the *observational* conditional probability describing how the probability of Y would change if one observed X (e.g. in a cross-sectional study) while $P(Y|do(X))$ is the *interventional* conditional probability, describing the probability of Y when X has been set experimentally. Of course, not all data can be classified cleanly as observational vs. interventional, since there might well be experiments that do not fully determine the value of the intervened variable. But for the sake of this article, the distinction will suffice (see [28] and [5] for further discussion).

In light of the general underdetermination of causal relations given any probability distribution, it is useful to represent the causal structure explicitly in terms of a directed graph. Unlike other graphical models with directed or undirected edges, which merely represent an independence structure, causal graphical models support a very a strong interpretation: For a given set of variables $\mathbf{V} = \{X_1, \ldots X_n\}$, a causal graph $G = \{\mathbf{V}, \mathbf{E}\}$ represents the causal relations over the set of variables $\mathbf{V}$, in the sense that for any directed edge $e = X_i \rightarrow X_j$ in $\mathbf{E}$, X_i is a direct cause of X_j relative to variables in $\mathbf{V}$. So the claim of an edge in G is that even if you randomize all other variables in $\mathbf{V} \setminus \{X_i, X_j\}$, thereby breaking any causal connection between X_i and X_j through these other variables, X_i still has a causal effect on X_j. Moreover, the causal graph characterizes the effect of an intervention on X_i on the remaining variables precisely in terms of the subgraph that results when all directed edges into X_i are removed from G. Thus, a causal graph not only makes claims about the causal pathways active in an observational setting, but also indicates which causal pathways are active in any experiment on the set of variables in $\mathbf{V}$. Naturally, a direct cause between X_i and X_j may no longer be direct once additional variables are introduced – hence the relativity to the set $\mathbf{V}$.

We use intuitive (and standard) terminology to refer to particular features of the graph: A *path* between two variables X and Y in G is defined as a non-repeating sequence of edges (oriented in either direction) in G where any two adjacent edges in the sequence share a common endpoint and the first edge "starts" with X and the last "ends" with Y. A *directed path* is a path whose edges all point in the same direction. A *descendent* of a vertex Z is a vertex $W \in \mathbf{V}$, such that there is a directed path $Z \rightarrow \ldots \rightarrow W$ in the graph G. Correspondingly, Z is *ancestor* of X. The *parents* of a vertex X are the vertices in $\mathbf{V}$ with a directed edge oriented into X, similarly for the *children* of a vertex.[2] A *collider* on a path p is a vertex on p whose adjacent edges both point into the vertex, i.e. $\rightarrow Z \leftarrow$. A *non-collider* on p is a vertex on p that is not a collider, i.e. it is a mediator ($\rightarrow Z \rightarrow$) or a common cause ($\leftarrow Z \rightarrow$). Note that a vertex can take on different roles with respect to different paths.

2 Basic Assumptions of Causal Discovery

Given the representation of causal relations over a set of variables in terms of causal graphs, causal discovery can be characterized as the problem of identifying as

[2] In a somewhat counter-intuitive usage of terms, a vertex is also its own ancestor and its own descendent, but not its own parent or child.

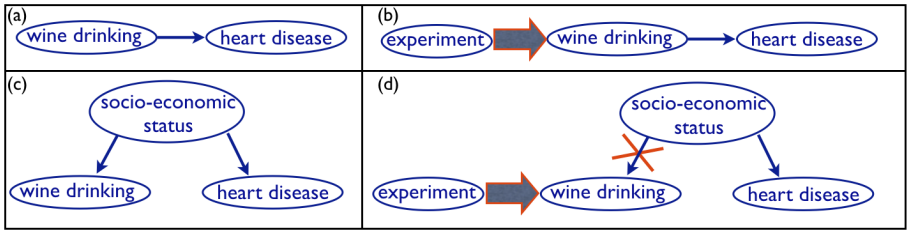

Fig. 1 (a) and (c) are two possible causal models that would explain an observed dependence between wine drinking and heart disease. But only in the case of (a) would that dependence persist if one were to intervene on wine drinking in an experiment. In (b) the intervention would destroy the dependence and make wine drinking independent of heart disease (d).

much as possible about the causal relations of interest (ideally the whole graph G) given a dataset of measurements over the variables $\mathbf{V}$. To separate the causal part from the statistical part of the inference it is – at least for an introduction – useful to think of causal discovery as the inference task from the joint distribution $P(\mathbf{V})$ to the graph G, leaving the task of estimating $P(\mathbf{V})$ from the finite data to the statistician.[3] In principle, there is no *a priori* reason for the joint distribution $P(\mathbf{V})$ to constrain the possible true generating causal structures at all. We noted earlier that correlation does not imply causation (and similarly, the converse is not true either, though that may not be as obvious initially). Yet, we do take both dependencies and independencies as indicators of causal relations (or the lack thereof). For example, it seemed perfectly reasonable above to claim that if a dependence between X and Y was detected in a randomized controlled trial where X was subject to intervention, then X is a cause of Y (again modulo the many other assumptions about successful experiment implementation). Similarly, in the observational case, the dependence between X and Y, if it was not a result of a direct cause, was explained by a common cause. Consequently, there seem to be principles we use – more or less explicitly – that connect probabilistic relations to causal relations.

Two such principles that have received wide application in the methods of causal discovery are the *causal Markov* and the *causal faithfulness* conditions. The high-level idea is that the causal Markov and faithfulness conditions together imply a correspondence between the (conditional) independences in the probability distribution and the causal connectivity relations within the graph G. Causal connectivity in a graph is defined in terms of *d-separation* and *d-connection* [30]: A path p between X and Y *d-connects* X *and* Y *given a conditioning set* $\mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$ if and only if (i) all colliders on p are in $\mathbf{C}$ or have a descendent in $\mathbf{C}$ and (ii) no non-colliders of p are in $\mathbf{C}$. X and Y are d-separated if and only if there are no d-connecting paths between them. D-separation is often denoted by the single turnstile '$\perp$'.

The causal Markov and the causal faithfulness assumptions (defined and discussed below) together ensure that (conditional) d-separation corresponds to (conditional) probabilistic independence, i.e.

$$X \perp Y \mid \mathbf{C} \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid \mathbf{C} \tag{1}$$

For causal discovery, this type of correspondence is enormously useful as it allows inferences from the (conditional) independence relations testable in data to the underlying causal structure. It can now be seen in what sense the claim that "correlation does not imply causation" still holds true, while a non-zero correlation can still provide an indication about existing causal relations: In particular, for two variables, a non-zero correlation would imply that the variables are d-connected given the empty set, i.e. that one causes the other or vice versa, or that there is a third variable that causes both. So while no specific causal relation can be deter-

[3] In order to separate out limitations and sources of error in the overall inference it can be helpful to make the following three-way distinction: *Statistical inference* concerns the inference from data to the generating distribution or properties of the generating distribution, such as parameter values or (in)dependence relations. *Causal discovery* concerns the inference of identifying as much as possible about the causal structure given the statistical quantities, such as a probability distribution or its features. *Causal inference* concerns the determination of quantitative causal effects given the causal structure and associated statistical quantities. Of course, these three inference steps are not always completely separable and there are plenty of interesting approaches that combine them.

mined, a subset of possible causal relations – an equivalence class of causal structures – can be identified. The correspondence also implies that two independent variables are causally disconnected (d-separated). So in the case of a linear Gaussian model, where no correlation implies independence, it follows that no correlation implies no causation.

Of course, (in)dependence features are only one set of features that a distribution $P(\mathbf{V})$ may exhibit, and to the extent that one is able to characterize other principles that connect other features of the distribution to the underlying causal structure, they can also be exploited for causal discovery – as we shall see below. Causal Markov and causal faithfulness only provide one set of what one might call "bridge principles", and they underlie many methods of so-called "constraint-based causal discovery".

Before proceeding, it is worth making explicit what causal Markov and causal faithfulness claim, and under what circumstances they may be false. The causal Markov condition states that every vertex X in the graph G is probabilistically independent of its non-descendents given its parents, i.e. $X \perp\!\!\!\perp NonDesc(X) \mid Pa(X)$. The causal Markov assumption appears to be a very fundamental assumption of our understanding of causality, since it is quite difficult to come up with situations that we consider to be causal and yet violate causal Markov. There are many ways in which a system may *appear* to violate causal Markov. For example, if one only considers two variables X and Y, but in fact there is an unmeasured common cause L of X and Y, i.e. $X \leftarrow L \rightarrow Y$, then Y is a non-descendent of X but X and Y will be dependent. Naturally, this situation is quickly remedied once L is included in the model and L is conditioned on (as a parent of X). Similar cases of "model-misspecifications" can lead to apparent violations of the Markov conditions when we have mixtures of different populations, there is sample selection bias, misspecified variables or variables that have been excessively coarse-grained (see [13] for more discussion). But in all these cases an appropriate specification of the underlying causal model will provide a causal system that is consistent with the Markov condition. To my knowledge, only in the case of quantum mechanics do we have systems for which we have good reasons to think they are causal and yet there does not appear to be a representation that respects the Markov condition. It is not entirely clear what to make of such cases. As Clark Glymour puts it, "[The Aspect experiments (that test the Einstein-Podolski-Rosen predictions)] create associations that have no causal explanation consistent with the Markov assumption, and the Markov assumption must be applied [...] to obtain that conclusion. You can say that there is no causal explanation of the phenomenon, or that there is a causal explanation but it doesn't satisfy the Markov assumption. I have no trouble with either alternative." [10]

The situation is quite different with regard to causal faithfulness. It states the converse of the Markov condition, i.e. if a variable X is independent of Y given a conditioning set $\mathbf{C}$ in the probability distribution $P(\mathbf{V})$, then X is d-separated from Y given $\mathbf{C}$ in the graph G. Faithfulness can be thought of as a simplicity assumption and it is relatively easy to find violations of it – there only have to be causal connections that do not exhibit a dependence. For example, if two causal pathways cancel out each other's effects exactly, then the causally connected variables will remain independent. A practical example is a back-up generator: Normally the machine is powered by electricity from the grid, but when the grid fails, a back-up generator kicks in to supply the energy, thereby making the operation of the machine independent of the grid, even though of course the grid normally causes the machine to work or when it fails it causes the generator to switch on, which causes the machine to work.[4] While such failures of faithfulness require an exact cancellation of the causal pathways, with finite data two variables may often appear independent despite the fact that they are (weakly) causally connected (see [47]).

To keep the present introduction to causal discovery simple initially, we can add additional assumptions about the underlying causal structure. Two commonly used assumptions are that the causal structure is assumed to be *acyclic*, i.e. that there is no directed path from a vertex back to itself in G, and *causal sufficiency*, i.e. that there are no unmeasured common causes of any pair of variables in $\mathbf{V}$. Both of these assumptions are obviously not true in many domains (e.g. biology, social sciences etc.) and below we will see how methods have been developed that do not depend on them. For

[4] This example is taken from [12].

now they help to keep the causal discovery task more tractable and easy to illustrate.[5]

With these conditions in hand (Markov, faithfulness, acyclicity and causal sufficiency), we can now ask what one can learn about the underlying causal relations given the (estimated) joint distribution $P(\mathbf{V})$ over a set of variables $\mathbf{V}$. Can we learn anything about the causal relation at all without performing experiments or having information about the time order of variables?

In fact, substantial information can be learned about the underlying causal structure from an observational probability distribution $P(\mathbf{V})$ given these assumptions alone. In 1990, Verma & Pearl [32] and Frydenberg [7] independently showed that any two acyclic causal structures (without unmeasured variables) that are Markov and faithful to the same distribution $P(\mathbf{V})$ share the same adjacencies (the same undirected graphical skeleton) and the same unshielded colliders. An *unshielded collider* is a collider whose two parents are not adjacent in G. Thus, Markov and faithfulness imply an equivalence structure over directed acyclic graphs, where graphs that are in the same equivalence class have the same (conditional) independence structure, the same adjacencies and the same unshielded colliders. For three variables the Markov equivalence classes are shown in Fig. 2. Note that the graph $X \rightarrow Z \leftarrow Y$ is in its own equivalence class. That means that independence constraints alone are sufficient to uniquely determine the true causal structure G if it is of the form $X \rightarrow Z \leftarrow Y$ (given the conditions stated). This is rather significant, since it implies that sometimes no time order information or experiment is necessary to uniquely determine the causal structure over a set of variables. More generally, knowing the Markov equivalence class of the true causal structure substantively reduces the underdetermination. In general, no closed form is known for how many equivalence classes there are or how many graphs there are per equivalence class, but large scale simulations have been run [9, 11]. It is worth noting that for any number of variables N, there will always be several singleton equivalence classes (e.g. the empty graph, or those containing only unshielded colliders), but that there will also always be at least one equivalence class that contains $N!$ graphs, namely the class containing all the graphs for which each pair of variables is connected by an edge – the set of complete graphs.

Algorithms have been developed that use conditional independence tests to determine the Markov equivalence class of causal structures consistent with a given dataset. For example, the PC-algorithm [41] was developed on the basis of exactly the set of assumptions just discussed (Markov, faithfulness, acyclicity and causal sufficiency) and uses a sequence of carefully selected (conditional) independence tests to both identify as much as possible about the causal structure and to perform as few tests as possible. In a certain sense the PC-algorithm is complete: it extracts all information about the underlying causal structure that is available in the statements of conditional (in)dependence. Or more formally, this bound can be characterized in terms of a limiting result due to Geiger and Pearl [8] and Meek [26]:

Theorem 1 (Markov completeness) *For linear Gaussian and for multinomial causal relations, an algorithm that identifies the Markov equivalence class is complete.*

That is, if the causal relations between the causes and effects in G can be characterized either by a linear Gaussian relation of the form $x_i = \sum_{j \neq i} a_j x_j + \epsilon_i$ with $\epsilon_i \sim N(\mu_i, \sigma_i^2)$ or by conditional distributions $P(X_i \mid pa(X_i))$ that are multinomial, then the PC-algorithm, which in the large sample limit identifies the Markov equivalence class of the true causal model, identifies as much as there is to identify about the underlying causal model.

One can see such a result as a success in that there are methods that reach the limit of what can be discovered about the underlying causal relations, or one can be disappointed about the underdetermination one is left with given that at best this only allows the identification of the Markov equivalence class. Moreover, one might have reason to think that even some of the assumptions required to achieve this limit are unreasonably optimistic about real world causal discovery. Consequently, there are a variety of ways to proceed:

1. One could weaken the assumptions, thereby (in general) increasing the underdetermination of what one will be able to discover about the underlying causal structure. For example, the FCI-algorithm [41] drops

[5] Especially with regard to the assumption of acyclicity it is worth noting that very subtle issues arise both about what exactly we mean when we allow for causal cycles, and how one may infer something about a system in which there are such feedback loops. The interested reader is encouraged to purse the references on cyclic models mentioned below.

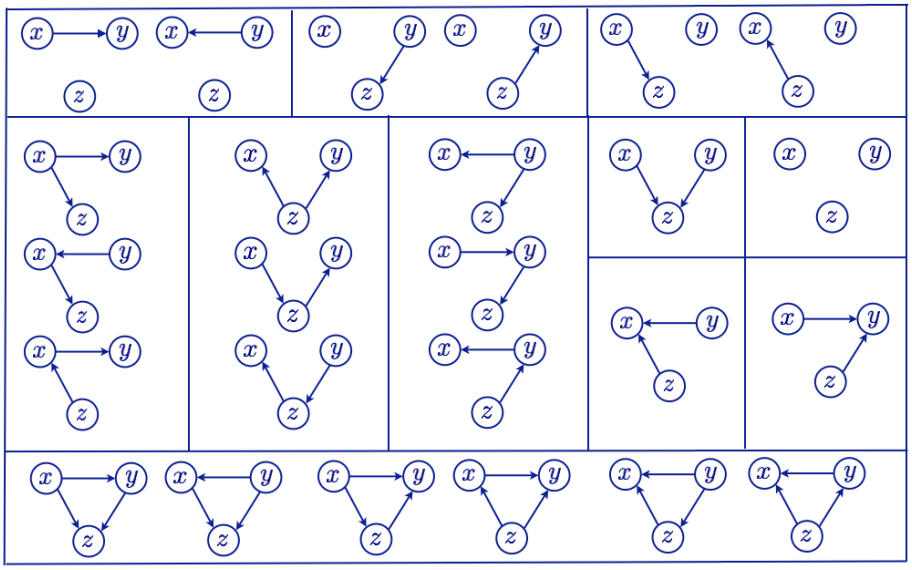

Fig. 2 The Markov equivalence classes for all three variable directed acyclic graphs without latent variables. Graphs in the same equivalence class share the same (conditional) independence structure.

the assumption of causal sufficiency and allows for unmeasured common causes of the observed variables; the CCD-algorithm [36] drops the assumption of acyclicity and allows for feedback, and the SAT-based causal discovery methods discussed below can drop both assumptions. Alternatively, Zhang & Spirtes [49] have worked on weakening the assumption of faithfulness, with corresponding algorithms presented in a paper in this issue. In all cases the aim of these more general approaches is to develop causal discovery methods that identify as much as possible about the underlying causal relations.

2. The limits to causal discovery described in Theorem 1 apply to restricted cases – multinomials and linear Gaussian parameterizations. One can exclude these cases and ask what happens when the distributions are not linear Gaussian or not multinomial. We consider several such approaches below.
3. One could consider more general data collection set-ups to help reduce the underdetermination. For example, one could consider the inclusion of specific experimental data to reduce the underdetermination or use additional "overlapping" datasets that share some but perhaps not all the observed variables (see [44] for an overview).

We will start by pursuing the second option in Sections 3, 4 and 5, and return to consider the first and third option in Section 6.

3 Linear non-Gaussian Models

One way of avoiding the limitation of causal discovery to only identifying the Markov equivalence class of the true causal model is to exclude the restrictions of Theorem 1. We will first consider the case of linear *non*-Gaussian models, that is, we will consider causal models where each variable is determined by a linear function of the values of its parents plus a noise term that has a distribution that is anything (non-degenerate) except Gaussian:

$$x_i = \sum_{j \neq i} a_j x_j + \epsilon_i \quad \text{with} \quad \epsilon_i \sim \text{non-Gaussian} \qquad (2)$$

The remarkable result for causal discovery, shown by Shimizu et al. [39], is that this rather weak assumption about the error distribution is sufficient to uniquely identify the true causal model. Thus,

Theorem 2 (Linear Non-Gaussian) *Under the assumption of causal Markov, acyclicity and a linear non-Gaussian parameterization (Eq. 2), the causal structure can be uniquely determined.*

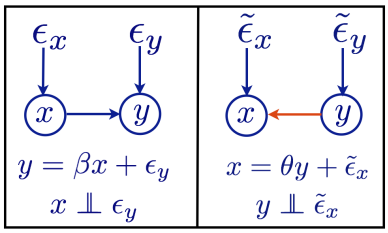

Fig. 3 In the "forwards" model (left) we have $x \perp\!\!\!\perp \epsilon_y$ and $y \not\perp\!\!\!\perp \epsilon_x$, while in the "backwards" model (right) we have $x \not\perp\!\!\!\perp \tilde{\epsilon}_y$ and $y \perp\!\!\!\perp \tilde{\epsilon}_x$. Assuming a linear non-Gaussian parameterization, it is not possible that both the forwards and the backwards model can be fit to the data, hence this assumption can aid the identifiability of causal direction.

Not even faithfulness is required here. Thus, merely the assumption that the causal relations are linear and that the added noise is anything but Gaussian guarantees in the large sample limit that the true causal model can be uniquely identified.

It helps to gain some intuition regarding this result from the two variable case: If we find that x and y are dependent and we assume acyclicity and causal sufficiency, then the Markov equivalence class contains two causal structures, $x \to y$ and $x \leftarrow y$. Consider the "forwards" model in Fig. 3, in which the (unobserved) noise terms are represented in terms of explicit variables:

$$y = \beta x + \epsilon_y \tag{3}$$

D-separation implies that in this model x is independent of the residuals on y, i.e. $x \perp\!\!\!\perp \epsilon_y$. The "backwards" model would take the form:

$$x = \theta y + \tilde{\epsilon}_x \tag{4}$$

We can re-write the equation for the backwards model, and substituting the forwards model for y, we get

$$\tilde{\epsilon}_x = (1 - \theta\beta)x - \theta\epsilon_y \tag{5}$$

Note that Equations 3 and 5 are linear in terms of the random variables x and ϵ_y, which are both non-Gaussian, but – if the forwards model is true – independent of one another. We can now apply the Darmois-Skitovich theorem that states:

Theorem 3 (Darmois-Skitovich) *Let $X_1, \ldots, X_n$ be independent, non-degenerate random variables. If for two linear combinations*

$$l_1 = a_1 X_1 + \ldots + a_n X_n \quad \textit{with} \quad a_i \neq 0$$
$$l_2 = b_1 X_1 + \ldots + b_n X_n \quad \textit{with} \quad b_i \neq 0$$

are independent, then each X_i is normally distributed.

Taking the contrapositive, and substituting the variables of the above example, if x and ϵ_y are independent, non-degenerate random variables that are *not* normally distributed, then the two linear combinations y and $\tilde{\epsilon}_X$ (Equations 3 and 5) are *not* independent. That is, if we mistakenly fit a backwards model to data that in fact came from a forwards model, then we would find that y and the residuals on x would be *dependent*, i.e. $y \not\perp\!\!\!\perp \tilde{\epsilon}_x$, despite the fact that the independence is required by d-separation on the backwards model. In other words, we would notice our mistake and would be able to correctly identify the true (in this case, forwards) model. Of course, this only proves the point for two variables, but the more general proofs can be found in [39] with also some alternative graphical demonstrations that may help the intuition underlying this identifiability result. It should also be noted that the Darmois-Skitovich theorem underlies the method of Independent Component Analysis [20].

These powerful identifiability results have been implemented in causal discovery algorithms that go by the acronym of LinGaM, for Linear non-Gaussian Models, and have been generalized (with slight weakenings of the identifiability) to settings where either causal sufficiency [15] or acyclicity [23] is dropped, or where the data generating process satisfies the LinGaM assumptions, but the actual data is the result of an invertible non-linear transformation, resulting in the so-called post-nonlinear model [50,51].

4 Non-linear additive noise models

Alternatively, in the continuous case the restrictions of Theorem 1 can be avoided by considering *non-linear* causal relations, i.e. when each variable x_j is determined by a non-linear function f_j of the values of its parents plus some additive noise

$$x_j = f_j(pa(x_j)) + \epsilon_j \tag{6}$$

We know (from the previous section) that when the f_j are linear, then identifiability requires that the error distributions are non-Gaussian. But one can ask what the

conditions for unique identifiability of the causal structure are when the f_j are non-linear (and there are no restrictions other than non-degeneracy on the error distributions). Identifiability results of this kind are developed in Hoyer et al. [14] and Mooij et al. [27]: The authors characterize a very intricate condition – I will here only refer to it as the *Hoyer condition* – on the relation between the function f, the noise distribution and the parent distribution[6], and provide the following theorem:

Theorem 4 (non-linear additive noise) *Under the assumption of Markov, acyclicity and causal sufficiency and a non-linear additive noise parameterization (Eq. 6),* unless *the Hoyer condition is satisfied, the true causal structure can be uniquely identified.*

In particular, this theorem has the following corrolaries:

- If the (additive) error distributions are all Gaussian, then the *only* functional form that satisfies the Hoyer condition is linearity, otherwise the model is uniquely identifiable.
- If the (additive) error distributions are non-Gaussian, then there exist (rather contrived) functions that satisfy the Hoyer condition, but in general the model is uniquely identifiable.
- If the functions are linear, but the (additive) error distributions are non-Gaussian, then there does not exist a *linear* backwards model (this is the result of the LinGaM approach of the previous section), but there exist cases where one can fit a non-linear backwards model [51].

The basic point of these identifiability results is that – although somewhat more complex than the linear non-Gaussian case – as soon as the functional relation between cause and effect becomes non-linear, and as long as the noise is additive, then (except for the rather special cases that satisfy the Hoyer condition), the true model is uniquely identifiable.

Again, an understanding of these results may be aided with a simple example of two variables (taken from [14]). Fig. 4a-c show first the data from a linear Gaussian model. As the "cuts" through the data indicate, no matter whether one fits the forwards or the backwards model, a Gaussian distribution of the residuals can be found that is independent of the value of the respective cause (x in the forwards, and y in the backwards model). However, panels d-f show that this no longer is true if the true model is in fact a non-linear Gaussian (forwards) model: While the error distribution is independent of the value of the cause in the (correct) forwards model, the error distribution on x is dependent on the value of y if one attempts to construct a backwards model, i.e. we have $y \not\perp\!\!\!\perp \tilde{\epsilon}_x$, when in fact an independence is required for the backwards model to be true.

Causal discovery algorithms have been developed for these settings (see the papers) and the identifiability results have been generalized [35], including to certain types of discrete distributions (see next section). There have – to my knowledge – not been extensions to the causally insufficient or cyclic case.

In light of the identifiability results of this section and the previous one it is ironic that so much of structural equation modeling has historically focused on the linear Gaussian case. The identifiability results mentioned here indicate that this focus on computationally simple models came at the expense of the identifiability of the underlying causal model. So in cases when the true causal model is known, then linear Gaussian parameterizations make the computation of causal effects very easy, but for the identifiability of the model in the first place, the linear Gaussian case is about as bad as it could be.

5 Restrictions on multinomial distributions

Naturally, one can also consider the possibilities of avoiding the limitations placed on causal discovery by Theorem 1 with respect to discrete distributions. This has been a much less explored direction of inquiry, possibly due to the difficulty of estimating specific features of discrete distributions, especially when the state space is finite. Alternatively, the domain of application of discrete distributions may provide only much weaker grounds for the justification of assumptions that pick out specific discrete distributions. The multinomial distribution therefore provides a useful unconstrained model, yet causal identifiability is limited to the Markov equivalence class.

[6] An explicit statement of the condition is omitted here as it requires a fair bit of notation and no further insight is gained by just stating it. The intrigued reader should refer to the original paper, which is a worthwhile read in any case.

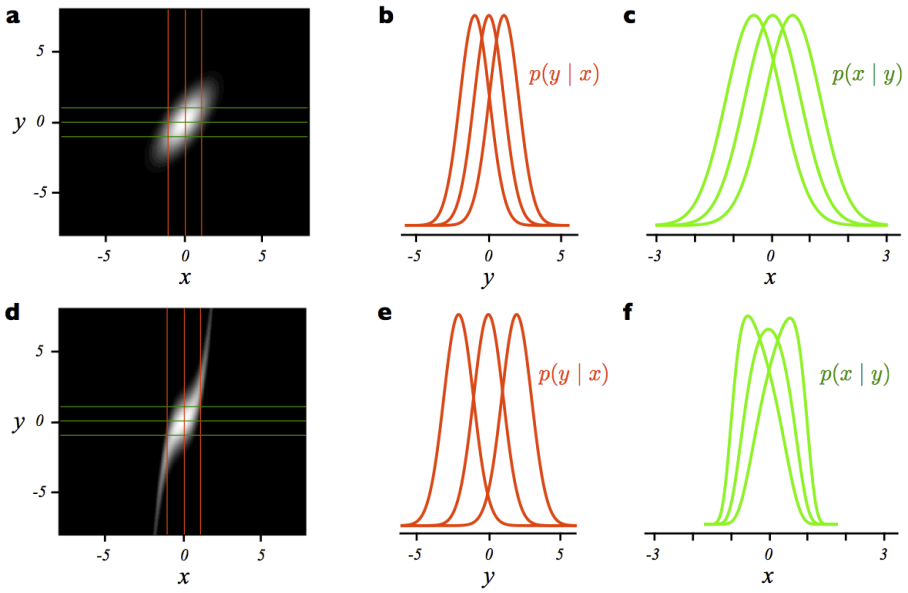

Fig. 4 (a) Linear Gaussian model with $x = \epsilon_x$ and $y = x + \epsilon_y$ with ϵ_x, ϵ_y distributed according to independent Gaussians. Both a "forwards" model ($x \to y$) and a "backwards" model ($x \leftarrow y$) can be fit to the data (panels b & c). However, in the case of a non-linear Gaussian model as in (d), where $x = \epsilon_x$, but $y = x + x^3 + \epsilon_y$ with ϵ_x, ϵ_y distributed according to independent Gaussians, we see that when fitting the "backwards" model (f), the distribution of the residuals on x are dependent on the value of y, while the residuals on y are independent of x when fitting the (correct) "forwards" model (e). (Graphics taken from [14].)

However, in a couple of papers by Peters et al. [33, 34], the authors extend the additive noise approach discussed in the previous section to the discrete case. While the variables take on discrete values, the causal relations follow the formal restrictions of the continuous case:

$$Y = f(X) + N \tag{7}$$

where the noise term N and the variable X are probabilistic and the addition now is in the space of integers $\mathbb{Z}$ or some "cyclic" space of values $\mathbb{Z}/m\mathbb{Z}$ for some integer m. The associated identifiability results under the assumption of causal sufficiency and acyclicity of the causal structure show that only for very specific choices of functions f and distributions over N is it possible to fit both a forwards model $X \to Y$ and backwards model $X \leftarrow Y$ to the data. In the generic case the causal direction is identified.

Instead of considering additive noise models, Park & Raskutti [29] consider discrete variables with Poisson distributions. Again, the causal structure can be identified as long as the variables have non-zero variances in specific settings (see their Theorem 3.1 for the precise condition). The key idea that drives the identifiability result in this case is *overdispersion*. For a variable X that is marginally Poisson distributed, we have $E(X) = Var(X)$, but for a variable $Y \mid X$ that is conditionally Poisson distributed, we have $Var(Y) > E(Y)$. The argument is nicely illustrated with the simple bivariate example on p.3 in [29].

To my knowledge there is very little work (other than some subcases of the additive noise models referred to above) that has developed general restrictions to enable identifiability of the causal structure for discrete models with finite state spaces, even though it is known that the assumption of a so-called "noisy-OR" parameterization enables in some cases identifiability beyond that of Markov equivalence.

6 Experiments and background knowledge

The previous several sections have considered the challenge of causal discovery in terms of finding weak generic assumptions about the nature of the underlying causal system that will enable or at least aid the identifiability of the true causal model. But for any concrete problem

of causal discovery in application, the search space of candidate causal models will often not include all possible causal structures over the set of variables in the first place, but be highly constrained by available background knowledge concerning e.g. particular causal pathways, time ordering, tier orderings of variables (i.e. that some subsets of variables come before others) or even less specific prior knowledge about, say, the edge density or the connectivity of the true causal structure. This type of background knowledge can similarly aid the identifiability of the causal model, possibly even without making additional assumptions about the functional form of the causal relations.

Recent developments using general constraint satisfaction solvers have enabled the integration of extraordinarily general background information into the causal discovery procedure. The high-level idea of these approaches is to encode (to the extent possible) all the available information as constraints in propositional logic on the underlying causal graph structure. For example, if data was collected and a conditional independence test was performed, then the implications of that test for the d-separation relations in the graph should be encoded in propositional logic. Similarly, if background information concerning specific pathways is available, it should also be translated into a logical constraint. To do so, fundamental propositional variables have to be defined that, if true, state that a particular directed edge is present in the graph. Thus, we might have

$A =$ '$x \to y$ is present in G'
$B =$ '$x \leftarrow y$ is present in G'

If there are only two variables ($\mathbf{V} = \{x, y\}$) then an independence can be encoded as

$$x \perp\!\!\!\perp y \Leftrightarrow \neg A \wedge \neg B$$

When there are more than two variables, the implied logical constraints will become larger. A pathway could be formulated as a conjunction of edges or, if it is only known that there is a causal pathway from x to y, but it is not known which other variables it passes through, it could be formulated as a dependence between x and y in an experiment in which only x is subject to intervention. Such a dependence would in turn be spelled out in terms of a disjunction of possible d-connecting pathways. The key is to find a logical encoding that enables a concise representation of such statements so that one does not have to explicitly state all the possible disjunctions. Hyttinen et al. [18, 16] have experimented with various encodings for a completely general search space that allows for causal models with latent variables and cycles. Triantafillou et al. [46, 45] have developed encodings for the acyclic case.

Once all the information has been encoded in constraints in propositional logic, one can use standard Boolean SAT(isfiability) solvers to determine solutions consistent with the joint set of constraints. The nice feature of using these solvers is that they are entirely domain general and highly optimized. Consequently, with a suitably general encoding one can integrate heterogeneous information from a variety of different sources into the discovery procedure.

A solver will return either one solution consistent with the constraints – that is, one assignment of truth values to the atomic propositional variables, which in turn specify one graph – or it can return only the truth value for those atomic variables that have the same truth value in all the solutions consistent with the constraints. A so-called "backbone" of the constraints specifies those features of the causal graph that are determined in light of the constraints.

However, constraints may conflict, in particular if they are the result of statistical tests. In that case a SAT-solver only returns that there is no solution for the set of constraints. For example, for the following set of independence constraints there is no graph (satisfying Markov and faithfulness) that is consistent with them:

$$x \perp\!\!\!\perp y \qquad x \not\perp\!\!\!\perp z \qquad y \not\perp\!\!\!\perp z \qquad x \perp\!\!\!\perp y \mid z$$

Rejecting the first constraint would make the constraints consistent with the graph $x \to y \to z$ (and its Markov equivalence class). Rejecting the fourth constraint makes the constraints consistent with the graph $x \to z \leftarrow y$. But together they are inconsistent (assuming Markov and faithfulness).

However, if each constraint were accompanied by a weight representing the degree of confidence in the truth of that constraint, then one might have a preference over which constraint should be rejected. In particular, the following optimization used by [16] may seem reasonable: Select a graph that minimizes the sum of the weights of the unsatisfied constraints:

$$\hat{G} \in \min_{G} \sum_{k : G \not\models k} w(k)$$

In this formalization, the causal discovery problem has now been converted into a weighted constrained optimization problem for which off-the-shelf maxSAT solvers can be applied, which guarantee to find the globally optimal solution. We now only have to determine suitable weights for the constraints. Hyttinen et al. [16] have experimented with different weighting schemes, from ones that are motivated by a preference for the simplest model in light of any detected dependencies, to a pseudo-Bayesian weighting scheme. Other weighting schemes, e.g. based on p-values, can be found in [45] and [24]. The more general question of how one should weight background knowledge such that it is well calibrated with any other available information remains an open research challenge, for which even the standard of success remains to be formulated.

While these SAT-based approaches are incredibly versatile in terms of the information they can integrate into the search procedure, and while they can achieve remarkably accurate results, they do not yet scale as well as other causal discovery algorithms. But there are several comments worth making in this regard: (i) The runtime of a constraint optimization using standard SAT-based solvers has a very high variance; many instances can be resolved in seconds while some can take vastly longer. (ii) The runtime is highly dependent on the set of constraints available and the search spaces they are applied to; for example [19] used a SAT-based method for causal discovery in the highly constrained domain of sub-sampled time series and were able to scale to around 70 variables. (iii) We can expect significant improvements in the scalability with the development of more efficient encodings and the parallelization of the computation. (iv) One can always explore the accuracy/speed trade-off and settle for a more scalable method with less accurate or less informative output. And finally, (v) if one is actually doing causal discovery on a specific application, one might be willing to wait for a week for the super-computer to return a good result.

There is another aspect in which the SAT-based approach to causal discovery opens new doors: Previous methods have focused on the identification of the causal structure or some general representation of the equivalence class of causal structures. SAT-based methods do not output the equivalence class of causal structures explicitly, but rather represent it implicitly in terms of the constraints in the solver. So instead of requesting as output a "best" causal structure or an equivalence class, one can also query specific aspects of the underlying causal system. This is particularly useful if one is only interested in a specific pathway or the relations among a subset of variables. In that case one need not compute the entire equivalence class but can query the solver directly to establish what is determined about the question of interest. Magliacane et al. [24] have taken this approach to only investigate the ancestral relations in a causal system and Hyttinen et al. [17] used a query-based approach to check the conditions for the applications of the rules of the *do*-calculus [31] when the true graph is unknown.

7 Outlook

This article has highlighted some of the approaches to causal discovery and attempted to fit them together in terms of their motivations and in light of the formal limits to causal discovery that are known. This article is by no means exhaustive and I encourage the reader to pursue other review articles such as Spirtes & Zhang [42] to gain a more complete overview. Moreover, there are many questions concerning comparative efficiency, finite sample performance, robustness etc. that I have not even touched on. Nevertheless, I hope to have shown that there is a vast array of different methods grounded on a whole set of different assumptions such that the reader may reasonably have some hope to find a method suitable (or adaptable) to their area of application. One almost paradigmatic application of a causal discovery method is illustrated in the article by Stekhoven et al. [43]. It exemplifies how a causal discovery method was applied to observational gene expression data to select candidate causes of the onset of flowering of the plant *Arabidopsis thaliana*. Once candidate causes had been identified, the researchers actually planted specimen, in which the genes, which had been determined to be relevant by the causal discovery method, had been knocked out – the causal hypothesis was put to the experimental test. I think it is fair to say that the results were positive.

Finally, I will highlight a few areas of causal discovery that I think still require a significant development in understanding. Again, the list is not supposed to be exhaustive, it is certainly colored by my own in-

terests and of course there already exists some interesting work in each.

Dynamics and time series. Many areas of scientific investigation describe systems in terms of sets of dynamical equations. How can these results be integrated with the methods for causal discovery in time series? (See e.g. [3,4,48,40,21].)

Variable construction. Standard causal discovery methods (such as the ones discussed in this article) take as input a statistical data set of measurements of well-defined causal variables. The goal is to find the causal relations between them. But how are these causal variables identified or constructed in the first place? Often we have sensor level data but assume that the relevant causal interactions occur at a higher scale of aggregation. Sometimes we only have aggregate measurements of causal interactions at a finer scale. (See e.g. [38,1,2].)

Relational data. In many cases there can be in addition to the causal relation, a dependence structure among the causal variables that is not due to the causal relations, but due to relational features among the causal variables, e.g. whether an actor is in a movie, or which friendship relations are present. In this case we need methods that can disentangle the dependencies due to the relational structure from the dependencies due to causality, and there may be causal effects from the relations to the individuals and vice versa. (See e.g. [37,25].)

In each of these cases the challenge is not simply to develop a new discovery method, but also to first characterize precisely the different concepts and what the goals of causal discovery in these domains are. So while there is a whole set of causal discovery algorithms ready to be applied to different domains, there also remain significant theoretical and conceptual hurdles that need to be addressed.

Acknowledgements I am very grateful to the organizers of the 2016 KDD Causal Discovery Workshop for encouraging me to put together and write up this overview. I am also very grateful to two anonymous reviewers who made several suggestions to improve the presentation and who alerted me to additional important papers that I was not aware of before.

References

1. Chalupka, K., Perona, P., Eberhardt, F.: Visual causal feature learning. In: Proceedings of UAI (2015)
2. Chalupka, K., Perona, P., Eberhardt, F.: Multi-level cause-effect systems. In: Proceedings of AISTATS (2016)
3. Dash, D.: Restructuring dynamic causal systems in equilibrium. In: Proceedings of AISTATS (2005)
4. Dash, D., Druzdzel, M.: Caveats for causal reasoning with equilibrium models. In: European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty, pp. 192–203. Springer (2001)
5. Eberhardt, F., Scheines, R.: Interventions and causal inference. Philosophy of Science **74**(5), 981–995 (2007)
6. Fisher, R.: The design of experiments. Hafner (1935)
7. Frydenberg, M.: The chain graph markov property. Scandinavian Journal of Statistics pp. 333–353 (1990)
8. Geiger, D., Pearl, J.: On the logic of causal models. In: Proceedings of UAI (1988)
9. Gillispie, S., Perlman, M.: The size distribution for Markov equivalence classes of acyclic digraph models. Artificial Intelligence **141**(1), 137–155 (2002)
10. Glymour, C.: Markov properties and quantum experiments. In: W. Demopoulos, I. Pitowsky (eds.) Physical Theory and Its Interpretation: Essays in Honor of Jeffrey Bub. Springer (2006)
11. He, Y., Jia, J., Yu, B.: Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs. Journal of Machine Learning Research **16**, 2589–2609 (2015)
12. Hitchcock, C.: Causation. In: S. Psillos, M. Curd (eds.) The Routledge Companion to Philosophy of Science. Routledge (2008)
13. Hitchcock, C.: Probabilistic causality. In: Stanford Encyclopedia of Philosophy. The Metaphysics Research Lab (2010)
14. Hoyer, P., Janzing, D., Mooij, J., Peters, J., Schölkopf, B.: Nonlinear causal discovery with additive noise models. In: D. Koller, D. Schuurmans, Y. Bengio, L. Bottou (eds.) Advances in Neural Information Processing Systems 21, pp. 689–696 (2008)
15. Hoyer, P., Shimizu, S., Kerminen, A., Palviainen, M.: Estimation of causal effects using linear non-Gaussian causal models with hidden variables. International Journal of Approximate Reasoning **49**, 362–378 (2008)
16. Hyttinen, A., Eberhardt, F., Järvisalo, M.: Constraint-based causal discovery: Conflict resolution with Answer Set Programming. In: Proceedings of UAI (2014)
17. Hyttinen, A., Eberhardt, F., Järvisalo, M.: Do-calculus when the true graph is unknown. In: Proceedings of UAI (2015)
18. Hyttinen, A., Hoyer, P., Eberhardt, F., Järvisalo, M.: Discovering cyclic causal models with latent variables: A general SAT-based procedure. In: Proceedings of UAI, pp. 301–310. AUAI Press (2013)
19. Hyttinen, A., Plis, S., Järvisalo, M., Eberhardt, F., Danks, D.: Causal discovery from subsampled time series data by constraint optimization. In: Proceedings of PGM (2016)
20. Hyvärinen, A., Karhunen, J., Oja, E.: Independent component analysis, vol. 46. John Wiley & Sons (2004)
21. Jantzen, B.: Projection, symmetry, and natural kinds. Synthese **192**(11), 3617–3646 (2015)

22. Klatsky, A.: Drink to your health? Scientific American **Feb.**, 75–81 (2003)
23. Lacerda, G., Spirtes, P., Ramsey, J., Hoyer, P.O.: Discovering cyclic causal models by independent components analysis. In: Proceedings of UAI, pp. 366–374 (2008)
24. Magliacane, S., Claassen, T., Mooij, J.: Ancestral causal inference. arXiv:1606.07035 (2016)
25. Maier, M., Marazopoulou, K., Arbour, D., Jensen, D.: A sound and complete algorithm for learning causal models from relational data. Proceedings of UAI (2013)
26. Meek, C.: Strong completeness and faithfulness in bayesian networks. In: Proceedings of UAI, pp. 411–418. Morgan Kaufmann Publishers Inc. (1995)
27. Mooij, J., Janzing, D., Peters, J., Schölkopf, B.: Regression by dependence minimization and its application to causal inference in additive noise models. In: Proceedings of ICML, pp. 745–752 (2009)
28. Nyberg, E., Korb, K.: Informative interventions. In: J. Williamson (ed.) Causality and Probability in the Sciences. College Publications (2006)
29. Park, G., Raskutti, G.: Learning large-scale poisson dag models based on overdispersion scoring. In: Advances in Neural Information Processing Systems, pp. 631–639 (2015)
30. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann (1988)
31. Pearl, J.: Causality. Oxford University Press (2000)
32. Pearl, J., Verma, T.: Equivalence and synthesis of causal models. In: Proceedings of Sixth Conference on Uncertainty in Artijicial Intelligence, pp. 220–227 (1991)
33. Peters, J., Janzing, D., Schölkopf, B.: Identifying cause and effect on discrete data using additive noise models. In: Proceedings of AISTATS, pp. 597–604 (2010)
34. Peters, J., Janzing, D., Schölkopf, B.: Causal inference on discrete data using additive noise models. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(12), 2436–2450 (2011)
35. Peters, J., Mooij, J., Janzing, D., Schölkopf, B.: Identifiability of causal graphs using functional models. In: Proceedings of UAI, pp. 589–598. AUAI Press (2011)
36. Richardson, T.: Feedback models: Interpretation and discovery. Ph.D. thesis, Carnegie Mellon University (1996)
37. Schulte, O., Khosravi, H., Kirkpatrick, A., Gao, T., Zhu, Y.: Modelling relational statistics with Bayes nets. Machine Learning **94**(1), 105–125 (2014)
38. Shalizi, C., Moore, C.: What is a macrostate? Subjective observations and objective dynamics. arXiv preprint cond-mat/0303625 (2003)
39. Shimizu, S., Hoyer, P., Hyvärinen, A., Kerminen, A.: A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research **7**, 2003–2030 (2006)
40. Sokol, A., Hansen, N.: Causal interpretation of stochastic differential equations. Electronic Journal of Probability **19**(100), 1–24 (2014)
41. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction and Search, 2 edn. MIT Press (2000)
42. Spirtes, P., Zhang, K.: Causal discovery and inference: concepts and recent methodological advances. In: Applied Informatics, vol. 3, p. 1. Springer Berlin Heidelberg (2016)
43. Stekhoven, D., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M., Bühlmann, P.: Causal stability ranking. Bioinformatics **28**(21), 2819–2823 (2012)
44. Tillman, R., Eberhardt, F.: Learning causal structure from multiple datasets with similar variable sets. Behaviormetrika **41**(1), 41–64 (2014)
45. Triantafillou, S., Tsamardinos, I.: Constraint-based causal discovery from multiple interventions over overlapping variable sets. Journal of Machine Learning Research **16**, 2147–2205 (2015)
46. Triantafillou, S., Tsamardinos, I., Tollis, I.G.: Learning causal structure from overlapping variable sets. In: Proceedings of AISTATS, pp. 860–867. JMLR (2010)
47. Uhler, C., Raskutti, G., Bühlmann, P., Yu, B.: Geometry of the faithfulness assumption in causal inference. The Annals of Statistics **41**(2), 436–463 (2013)
48. Voortman, M., Dash, D., Druzdzel, M.: Learning why things change: the difference-based causality learner. arXiv preprint arXiv:1203.3525 (2012)
49. Zhang, J., Spirtes, P.: The three faces of faithfulness. Synthese **193**(4), 1011–1027 (2016)
50. Zhang, K., Chan, L.W.: Extensions of ica for causality discovery in the hong kong stock market. In: International Conference on Neural Information Processing, pp. 400–409. Springer (2006)
51. Zhang, K., Hyvärinen, A.: On the identifiability of the post-nonlinear causal model. In: Proceedings of UAI, pp. 647–655. AUAI Press (2009)