EJPS manuscript No.
(will be inserted by the editor)

Identifying intervention variables

Michael Baumgartner, Isabelle Drouet

Received: date / Accepted: date

Abstract

The essential precondition of implementing interventionist techniques of causal reasoning is that particular variables are identified as so-called *intervention variables*. While the pertinent literature standardly brackets the question how this can be accomplished in concrete contexts of causal discovery, the first part of this paper shows that the interventionist nature of variables cannot, in principle, be established based only on an interventionist notion of causation. The second part then demonstrates that standard observational methods that draw on Bayesian networks identify intervention variables only if they also answer all the questions that can be answered by interventionist techniques—which are thus rendered dispensable. The paper concludes by suggesting a way of identifying intervention variables that allows for exploiting the whole inferential potential of interventionist techniques.

Keywords interventionism, causal discovery, causal reasoning, Bayes nets methods, intervention variables, causal assumptions, Markov condition, Faithfulness condition

1 Introduction

Woodward's (2003) interventionist theory of causation has not only stimulated the conceptual literature on causation in recent years, it has also exerted a significant influence on the literature concerned with methods of causal reasoning and discovery. It is widely agreed upon that intervening on causal structures in the manner made precise by Woodward supplies considerable inferential leverage when it comes to uncovering causal structures. If all variables in a causal structure are systematically manipulable—say, in ideal laboratory contexts—that structure can be thoroughly and unambiguously uncovered and it is determined exactly how many experimental manipulations are sufficient to do so (cf. Eberhardt et al. 2006; Eberhardt and Scheines 2007). Moreover, even the manipulability of only some variables in a structure may render it possible to uncover relevant parts of it or to disambiguate causal inferences that would remain ambiguous were it not for the possibility to intervene (cf. Pearl 2000; Spirtes et al. 2000, ch. 4; Korb and Nyberg 2006). Likewise, interventionist techniques can be effectively implemented in the discovery of causal structures that resist analysis by other methods (cf. Nyberg and Korb 2006).

Institute of Cognitive Science, University of Osnabrück, Germany; mbaumgartner@uni-osnabrueck.de
Paris-Sorbonne University, France; isabelle.drouet@paris-sorbonne.fr

All of these powerful applications of interventionism crucially hinge on the possibility to manipulate investigated causal structures in a very particular manner, which Woodward (2003, 130) describes as *surgical*. A surgical manipulation is what Woodward calls an *intervention*. He spells out the conditions an intervention has to satisfy by means of the notion of an *intervention variable*. Roughly, an intervention variable I for X with respect to Y, is a cause of X that is not connected to Y on a causal path that does not go through X and that is independent of any other cause of Y. The essential precondition of fruitfully implementing interventionist means in causal discovery is that particular triples of variables can in fact be established to have these very specific interventionist properties.

While, in the literature on interventionist causal reasoning, it is common to assume that for all variables in an analyzed structure there exist suitable intervention variables, the problem of how these intervention variables can actually be identified is normally bracketed. Still, in ordinary contexts of causal discovery, the important question is not so much whether intervention variables exist for a variable X with respect to another variable Y, but rather which variable is such an intervention variable, or whether a concrete variable I is such an intervention variable or not—and, often, the answers to these questions are far from obvious. For instance, a typical interventionist test design to determine the causal relationship between, say, depression and insomnia would be to administer an anti-depression drug δ to relevant patients and to then investigate whether that treatment is accompanied by reduced (or increased) insomnia. Plainly though, such a test is only revealing if it has been ascertained that δ itself does not have a direct somniferous (or agitating) effect, which calls for an antecedent study on the (side-)effects of δ. Or, important controversies in the history of science can be understood as controversies about whether or not certain experimental actions are interventions in the surgical sense. For example, the controversy between Pasteur and Pouchet concerning the possibility of spontaneous generation is largely about whether specific manipulations (letting air through sulphuric acid, warming air up to a very high temperature, filtering it through cotton…) succeed in eliminating any trace of life it may contain while not at the same time destroying the inert vital component that could produce life according to the proponents of the theory of spontaneous generation (cf. Collins and Pinch 1993, ch. 4).

Accordingly, this paper focuses on the much neglected problem of how intervention variables can be identified in contexts of causal discovery. In order to answer this methodological question, we first answer the following conceptual question: what can justify the belief that a given triple of variables has the interventionist properties? It turns out that such justifications require recourse to non-interventionist causal knowledge. Therefore, the second part of the paper draws on a currently dominant strand of non-interventionist methods of causal reasoning, *viz.* on Bayesian networks methods (cf. Spirtes, Glymour, and Scheines 2000; Pearl 2000), in order to tackle the problem of how intervention triples can actually be identified.

2 Intervention triples

The standard reference for rendering the notion of a (suitable) intervention precise is Woodward's (2003) interventionist theory of causation. The theory is intended to illuminate "how we think about, learn about, and reason with various causal notions" (Woodward 2008, 194). Woodward determines what characteristics interventions must have in order to have maximal inferential purchase in contexts of causal discovery with recourse to the two core conceptual constituents of his interventionist theory of causation: the notion of *type causa-*

tion, which has the two components of *direct* and of *contributing* causation, and the notion of an *intervention variable*. Here are the two corresponding (frequently cited) definitions (M) and (IV):

(M) A necessary and sufficient condition for X to be a (type-level) direct cause of Y with respect to a variable set $\mathbf{V}$ is that there be a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in $\mathbf{V}$. A necessary and sufficient condition for X to be a (type-level) *contributing cause* of Y with respect to variable set $\mathbf{V}$ is that (i) there be a directed path from X to Y such that each link in this path is a direct causal relationship (...), and that (ii) there be some intervention on X that will change Y when all other variables in $\mathbf{V}$ that are not on this path are fixed at some value. (Woodward 2003, 59)

(IV) I is an intervention variable for X with respect to Y iff (i) I causes X; (ii) certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I; (iii) any directed path from I to Y goes through X; (iv) I is statistically independent of any variable Z that causes Y and that is on a directed path that does not go through X. (Woodward 2003, 98)

Relative to the notion of an intervention variable, an intervention on X with respect to Y is then straightforwardly spelled out in terms of an intervention variable I for X with respect to Y taking on some value z_i such that $I = z_i$ causes X to take on some determinate value z_j (Woodward 2003, 98).

It is plain from this that Woodward's interventionism relies on an interdefined conceptual core. Causation is defined in terms of the notion of an intervention which is itself defined based on the notion of an intervention variable which, in turn, is defined in terms of causation. The fact that (M) and (IV) interdefine causation and intervention is not considered to be problematic by Woodward. He maintains that his way of interdefining causation and intervention is not viciously circular (Woodward 2003, 104–105):[1]

> The causal information required to characterize the notion of intervention on X with respect to Y is information about the causal relationship between the intervention variable I and X, information about whether there are other causes of Y that are correlated with I, information about whether there is a causal route from I to Y that does not go through X and so on, *but not information about the presence or absence of a causal relationship between X and Y*.

In a nutshell, thus, the basic idea behind interventionism is that if the triple $\langle I, X, Y\rangle$ satisfies (IV) such that I is an intervention variable for X with respect to Y, interventions on X through I will reveal whether or not X is a cause of Y. According to Woodward, this is possible notwithstanding the interdefined conceptual core of interventionism, because the fact that $\langle I, X, Y\rangle$ is an intervention triple does neither entail nor presuppose anything whatsoever with respect to the relationship between X and Y. That is, interventionism's "primary focus is *methodological*" (Woodward 2008, 194), and interventionist methods aim to modularly uncover causal structures: based on clarity about the causal (in)dependencies between I and X as well as between I and Y, (in)dependencies between X and Y can be revealed. Or more generally, a whole causal structure $\mathcal{G}$ over a set of variables $\mathbf{V}$ is uncovered by first

[1] Even though not all authors agree that interdefining causation and intervention is as unproblematic as Woodward would like to have it (e.g. Strevens 2007, 2008 or Baumgartner 2009), it is clear that Woodward neither aims nor claims to provide a reductive analysis of causation.

identifying substructures in $\mathcal{G}$ that comply with (IV), and by then sequentially uncovering further substructures of $\mathcal{G}$ on the basis of the relevant intervention triples.

Obviously, the crucial prerequisite for implementing this methodological strategy is the availability of suitable intervention variables for an analyzed structure. That is, to get interventionist methods off the ground, some triples of the set of variables for which a causal model is being searched must be known to satisfy (IV). Because of the interdefinition of the notions of an intervention variable and of causation, knowledge about compliance with (IV) is causal knowledge. The next section is going to investigate what this causal knowledge consists in and, more specifically, what can justify that a given triple in fact is an intervention triple.

3 On what justifies the interventionist nature of a given triple

Knowing that $\langle I, X, Y\rangle$ is an intervention triple amounts to being justified in believing that it satisfies each of the four conjuncts constituting (IV). Of these four conditions, (IV.ii) is the least significant one. Eberhardt and Scheines (2007) refer to a variable I that only complies with (IV.i), (IV.iii), and (IV.iv) as a *soft* or *parametric* intervention variable and show that even soft intervention variables can provide considerable inferential leverage in many contexts of causal discovery (similarly Korb et al. 2004). Moreover, given that a triple $\langle I, X, Y\rangle$ satisfies (IV.i), (IV.iii) and (IV.iv), (IV.ii) is usually taken to be satisfiable by simply choosing a value of I that actually determines a particular value of X. Hence, for simplicity, we subsequently focus on the question of what justifies that triples of variables comply with (IV.i), (IV.iii) and (IV.iv), i.e. we focus on soft intervention variables only.

While the particular manner in which (M) and (IV) interdefine causation and intervention may be claimed not to be viciously circular, the interdefined conceptual core of interventionism triggers infinite regresses when it comes to justifying the interventionist character of variables by application of the definitions (M) and (IV). Based on (M) and (IV), no variable can ever be justified to have the interventionist properties in a finite number of steps. In consequence, it is never possible, in principle, to justify the belief that a particular triple of variables is an intervention triple by applying interventionist definitions. To see this, suppose we want to determine that a variable I_1 is an intervention variable for X with respect to Y. Condition (IV.i) stipulates that a necessary condition for that to be the case is I_1 being a cause of X. According to (M), a necessary condition for I_1 to be a cause of X is that there be a possible intervention on I_1 with respect to X, and hence an intervention variable for I_1 with respect to X, call it I_2. This, in turn, requires I_2 to be a cause of I_1, which again presupposes that there is an intervention variable I_3 for I_2 with respect to I_1, which calls for a further intervention variable I_4 for I_3 with respect to I_2, and so on. Condition (IV.iii) amounts to another necessary condition for I_1 to be an intervention variable for X with respect to Y: there must not be a causal path connecting I_1 and Y that does not go through X. In order to determine whether I_1 satisfies that condition, first, the possible $I_1 - X - Y$ connection must be suppressed (or 'broken') by fixing the value of X by means of a further intervention variable I_5 and, second, it must be established that there is no possible intervention on I_1 that changes Y or the probability distribution of Y when one holds fixed all other variables.[2] Of course, according to (M), I_5 being an intervention variable for X with respect to Y requires there to be another intervention variable I_6 for I_5 with respect to X,

[2] For further details on testing the satisfaction of (IV) cf. Woodward (2003, 99–111).

and so on. In sum, given that causation is defined in terms of (M), justifying that (IV) is satisfied sets off multiple infinite regresses.[3]

In scientific practice, a common way of solving the problem of justificatory regresses that arise when definitions are applied to relevant entities is to draw on some sort of foundationalism. For instance, if we are to determine whether a yellow ring satisfies the definition of gold, we are first going to conduct a chemical analysis that draws on certain causal characteristics of gold. Depending on theoretical preferences, these causal characteristics might then be reduced to, say, probabilistic (in)dependencies, which again depending on theoretical preferences, can be analyzed in terms of a suitable frequency distribution. However, as all scientific theorizing inevitably operates within some conceptual frame that is taken as given and unquestioned, regresses of this ordinary type are *not infinite*, but terminate as soon as some conceptual level is reached that is considered to be primitive by whoever happens to apply a corresponding definition. Applying the definition of gold to a ring induces a progression from one conceptual level to a subsequent *independent* level and stops when some primitive level is reached. By contrast, no primitive conceptual level can ever be reached if (M) is applied in order to justify that a given triple of variables satisfies (IV). The fact that (M) and (IV) are interdefined induces an *infinite oscillation* between two notions, none of which is primitive according to interventionism, which, in turn, yields that no foundationalist solution to the problem of justifying compliance with (IV) is available.

All this demonstrates that the belief that a triple $\langle I, X, Y\rangle$ satisfies (IV) cannot be justified by *direct application* of (M) and (IV) to the triple $\langle I, X, Y\rangle$. Clearly though, in order to determine whether a specific entity satisfies a given definition it is often not necessary to apply the definition itself, rather, *heuristics* will do. If we want to know whether a yellow ring is made of gold, we do not necessarily have to conduct a chemical analysis. Many suitable heuristics are available. The price of the ring will be an indication, or the reputation of the store in which it is sold. Hence, the question arises whether there might exist some heuristic which may warrant, *without direct application* of (M), that $\langle I, X, Y\rangle$ complies with (IV). In general terms, heuristics are experience-based techniques that sufficiently frequently succeed in solving problems in a 'cost-effective' way (cf. Wimsatt 2006, 463-465). Their solutions do not need to be perfect, but the quality of these solutions must be assessable independently of the heuristics, at least in principle. The problem to solve in the present context is determining the satisfaction of definitions. As indicated above, heuristics often render it unnecessary to explicitly apply definitions. For instance, the problem of determining whether a ring complies with the (chemical) definition of gold can be solved by using the ring's price as a heuristic measure. By conducting chemical analyses and, thus, by explicitly applying the definition of gold, it can be seen that gold frequently has a high price. This yields a generalizable constraint on heuristics that evaluate the satisfaction of definitions: in order for a non-definitional criterion Φ to serve as a heuristic measure for whether relevant entities satisfy a definition t, Φ must sufficiently frequently identify such entities as complying and as not complying with t that would be identified as such if t itself were applied and it must be possible to apply t, at least in principle. While that constraint is certainly satisfied in the case of gold and its price, matters are different for intervention variables defined along interventionist lines. Indeed, the considerations of the previous paragraphs show that there is no way to determine for even one single triple of variables whether it satisfies (IV) by applying (M). In view of the lack of a single positive application of (M)

[3] Since the notion of causation is also of crucial importance in (IV.iv) a similar regress is initiated when it comes to determining whether a specific triple $\langle I_1, X, Y\rangle$ satisfies (IV.iv). For brevity, we leave it to the reader to scrutinize that additional regress.

and (IV), there cannot exist a heuristic for assessing the satisfaction of (M) and (IV), for there does not exist an independent gauge for heuristic success.

That means the satisfaction of (IV) cannot be justified on heuristic grounds either. At the same time, however, there exist certain triples of variables $\langle I, X, Y\rangle$ such that I is *de facto* known to be an intervention variable for X with respect to Y. Suppose we want to find out whether treatment with a specific drug (T) is a cause of recovery from a particular disease (R). To answer that question, subjects that suffer from the disease are randomly assigned to treatment and control groups, say, by tossing a coin (C). Everybody will agree that C is an intervention variable for T with respect to R (cf. Woodward 2003, 94–98; 2008, 203–204). C causally determines the assignment to treatment or control group, C does not directly cause R, and C is statistically independent of other causes of R. We have enough background causal knowledge to be reasonably confident that the triple $\langle C, T, R\rangle$ satisfies (IV). We have shown in the previous paragraphs that this causal knowledge cannot be interventionist causal knowledge. Therefore, whoever is convinced that coin tossing has the interventionist properties *cannot* and, as a matter of fact, *does not* understand causation on the sole basis of (M). To justify compliance with (IV) some non-interventionist theory of causation is indispensable (cf. Cartwright 2010).

It turns out that relative to certain non-interventionist theories of causation it can indeed be substantiated that coin tossing is an intervention variable for treatment with respect to recovery. Take for instance an elementary probabilistic account as professed by Suppes (1970). Given a suitable probability distribution over C, T and R, such a theoretical framework determines that C is a direct cause of T if C is positively correlated with T, C temporally precedes T, and C is not screened off from T by any further variable in the structure. Furthermore, if C is screened off from R by T, C can be said not to directly cause R. Finally, if pertaining probabilistic data can be shown not to feature any other (probabilistically defined) causes of R that are correlated with C, it follows that the triple $\langle C, T, R\rangle$ satisfies (IV.i), (IV.iii), (IV.iv) and, thus, that C is an intervention variable for T with respect to R. Of course, such a probabilistic analysis in the vein of Suppes (1970) has long been shown not to adequately capture all causal dependencies, as e.g. causes that lower the probabilities of their effects. Yet, irrespective of whether it successfully accounts for all kinds of causal dependencies, Suppes' theory can be used to straightforwardly justify the belief that coin tosses satisfy (IV) in randomized experiments. Modern and more sophisticated probabilistic analyses as e.g. professed by Kvart (2001) or Eells (1991) could also establish C as intervention variable for T with respect to R in a finite number of steps—even though a corresponding argument would involve more complications.

These considerations show that as long as causation is exclusively understood in the vein of (M) a given triple $\langle I, X, Y\rangle$ cannot be justified to have the interventionist properties. To justify that $\langle I, X, Y\rangle$ complies with (IV.i), (IV.iii), and (IV.iv), recourse must be made to non-interventionist causal knowledge. The interventionist nature of $\langle I, X, Y\rangle$ can, for example, be justified based on a probabilistic account of causation.

4 Identifying intervention triples

In the previous section, we have seen how the belief that coin tossing is an intervention variable for treatment with respect to discovery can be justified based on a (rough) probabilistic notion of causation. This, however, does not tell us how coin tossing has been selected as a candidate intervention variable in the first place. To unfold the whole inferential potential of interventionist methodologies we not only need to be able to justify the belief that

a given triple complies with (IV). We also require a methodically guided manner of identifying triples that justifiably comply with (IV) within a given set $\mathbf{V}$ of causally modeled variables. The previous section has shown that a method for identifying intervention triples within analyzed variable sets cannot rely on an interventionist conception of causation only. A currently dominant framework of causal inference that searches for causal dependencies as defined by a non-interventionist theory of causation is constituted by the procedures of causal discovery that draw on Bayesian networks (cf. Spirtes, Glymour, and Scheines 2000; Pearl 2000). These *BN methods* do not implement an interventionist notion of causation (rather, they stipulate a connection between probabilities and causation which is very similar to the one stated by Suppes' probabilistic analysis of causation). Moreover, they process observational (non-experimental) data. Accordingly, this section investigates whether and how intervention variables can be identified by means of BN methods.

4.1 Standard BN algorithms

Before we can address this question, the basic ideas behind BN methods must be briefly reviewed. To this end, some definitional preliminaries are called for. A Bayesian network (BN) over a set of variables $\mathbf{V}$ is a pair $\langle G, p\rangle$ such that: (i) G is a directed acyclic graph (DAG) over $\mathbf{V}$, (ii) p is a probability distribution over $\mathbf{V}$, and (iii) $\langle G, p\rangle$ satisfies the *Markov* condition (MC):

(MC) Any V in $\mathbf{V}$ is independent in p of all its non-descendants in G conditional on its parents in G.[4]

BNs appeared in artificial intelligence in the 1980s as a means to represent and handle uncertainty (cf. e.g. Pearl 1988). More recently, they have come to be employed for the purpose of causal inference. Algorithms for causal reasoning that rely on BNs were developed chiefly by two groups. Their main results are presented in Spirtes et al. (2000) and Pearl (2000), respectively.

BN causal inference methods aim to infer causal structures from probabilistic dependencies and independencies over a set of variables $\mathbf{V}$. To explain how these methods work the notions of a causal graph and of causal sufficiency are required. For any set of variables $\mathbf{V}$, the *causal graph* over $\mathbf{V}$ is the directed acyclic graph over $\mathbf{V}$ in which there is an arrow from a variable X to another variable Y if and only if X is a direct cause of Y relative to $\mathbf{V}$. BN methods operate under the assumption that the causal structures they analyze can indeed be modeled by causal graphs in this sense. Among other things, this means that BN methods employ the assumption that the structures to which they are applied are causally acyclic, i.e. that they do not involve any causal cycles or feedbacks. Moreover, a set of variables $\mathbf{V}$ is said to be *causally sufficient* if and only if for every common cause C of two variables X and Y in $\mathbf{V}$ it holds that C is in $\mathbf{V}$, or a cause of C is in $\mathbf{V}$, or an effect of C which is located on all directed paths from C to X and on all directed paths from C to Y is in $\mathbf{V}$. Against this conceptual background, the so-called *causal Markov condition* (CMC) which is assumed by all BN methods can be stated as follows (cf. e.g. Glymour 1997, 206; Ramsey et al. 2006; Zhang and Spirtes 2008):

(CMC) For any set of variables $\mathbf{V}$, if $\mathbf{V}$ is causally sufficient, then any probability distribution p over $\mathbf{V}$ that is generated by the causal graph $\mathcal{G}$ over $\mathbf{V}$ is such that $\langle \mathcal{G}, p\rangle$ satisfies (MC).

[4] Definitions of all subsequently used graph-theoretical notions can be found in Spirtes et al. (2000, 5-10) or in Pearl (2000, 12-13).

(CMC) is a universal assumption. For expositional purposes it will be useful to also have a label for the predicate which (CMC) claims to be satisfied by all causally sufficient variable sets. We shall hence say that a sufficient set of variables $\mathbf{V}$ satisfies ($\text{CMC}_\mathbf{V}$) if and only if any probability distribution p over $\mathbf{V}$ that is generated by the causal graph $\mathcal{G}$ over $\mathbf{V}$ is such that $\langle \mathcal{G}, p \rangle$ satisfies (MC).

In addition to (CMC), BN procedures of causal inference usually also exploit the assumption that the *causal Faithfulness condition* (CFC) holds (cf. Ramsey et al. 2006):

(CFC) For any set of variables $\mathbf{V}$, if $\mathbf{V}$ is causally sufficient, then all independencies in probability distributions over $\mathbf{V}$ that are generated by the causal graph $\mathcal{G}$ over $\mathbf{V}$ are implied by ($\text{CMC}_\mathbf{V}$) for $\mathbf{V}$.

Analogously to ($\text{CMC}_\mathbf{V}$), we label the predicate that (CFC) universally applies to causally sufficient variable sets ($\text{CFC}_\mathbf{V}$). A sufficient set of variables $\mathbf{V}$ satisfies ($\text{CFC}_\mathbf{V}$) if and only if all independencies in any probability distribution generated by the causal graph $\mathcal{G}$ over $\mathbf{V}$ are implied by ($\text{CMC}_\mathbf{V}$) for $\mathbf{V}$.

There is a certain variance in the literature on BN methods as to what exactly the logical forms of (CMC) and (CFC) are. While most authors explicitly subscribe to the universal version of (CMC) given above, (CFC) is often not explicitly stated in the universal form advanced here. Rather, authors frequently content themselves with defining the predicate ($\text{CFC}_\mathbf{V}$) in order to then (implicitly) assume that particular variable sets that are relevant to a pertinent causal analysis satisfy ($\text{CFC}_\mathbf{V}$). However, as these latter presentations of causal Faithfulness are normally not very clear about exactly which variable sets must be assumed to satisfy ($\text{CFC}_\mathbf{V}$) in order for BN methods to be correct, we, for the time being, settle for the universal version (CFC). We will consider weakenings of (CFC) in section 4.2. Although they do not explicitly give necessary and sufficient probabilistic conditions for a variable X to cause a variable Y, (CMC) and (CFC) taken together constitute the theory of causation that is conveyed by BN procedures of causal inference. They define a connection between causation and probabilities which is very similar to the one stipulated by Suppes and which entails necessary or sufficient conditions for various causal notions (cf. Pearl 2000, §2.7). These conditions are put to work in BN algorithms of causal inference.

There exist two types of such algorithms. To state the prime difference between these types, let $\mathbf{O}$ be the set of variables that are *observed* (or *measured*) in a given study: while algorithms of the first type employ the assumption that $\mathbf{O}$ is causally sufficient, algorithms of the second type do not rely on that assumption. Among the algorithms of the first type are PC, SGS (Spirtes et al. 2000) or IC (Pearl 2000). Algorithms of the second type are, for instance, CI, FCI (Spirtes et al. 2000) or IC* (Pearl 2000). For brevity, we label algorithms of the first type *$\mathcal{S}$-algorithms* and algorithms of the second type *$\mathcal{NS}$-algorithms*.

$\mathcal{S}$-algorithms take as input the set $\mathbf{I}$ of conditional independencies over a set of variables $\mathbf{O}$ which is assumed to be causally sufficient, and they output a graphical pattern, to which we shall refer as an *$\mathcal{S}$-pattern*, for short.[5] An $\mathcal{S}$-pattern represents a set of DAGs over $\mathbf{O}$ all of which constitute a BN in combination with all and only the probability distributions featuring exactly the independencies recorded in $\mathbf{I}$. The set of DAGs represented by an $\mathcal{S}$-pattern is called a *Markov equivalence class*, because for all its members (MC) entails the same conditional independence relations among the variables in $\mathbf{O}$. $\mathcal{S}$-algorithms are correct in the following sense (cf. e.g. Spirtes, Glymour, and Scheines 1991): if two variables X and

[5] The idea of treating the set $\mathbf{I}$ of all conditional independencies over $\mathbf{O}$ as a premise is not unproblematic in contexts of causal discovery, since what is usually available in such contexts are mere statistical data. How conditional independencies in $\mathbf{O}$ are hypothesized based on corresponding statistical data is irrelevant for our current purposes though.

Y are connected by a graphical feature ϕ in an $\mathcal{S}$-pattern, X and Y are related in the causal graph $\mathcal{G}$ over $\mathbf{O}$ in the way represented by ϕ; and if X and Y are not connected by any graphical feature in an $\mathcal{S}$-pattern, X and Y are causally unrelated. That is, assuming that causation and probabilities are connected in the way that (CMC) and (CFC) define, $\mathcal{G}$ is among the graphs represented by the $\mathcal{S}$-pattern that an $\mathcal{S}$-algorithm outputs when given the conditional independencies over $\mathbf{O}$ as input. For there to be an edge between X and Y in an $\mathcal{S}$-pattern, it is necessary and sufficient that X and Y are probabilistically dependent and that no subset of $\mathbf{O}\backslash\{X, Y\}$ screens off between them (i.e. renders them probabilistically independent). In light of the correctness of $\mathcal{S}$-algorithms, that means that the impossibility to screen off between X and Y in $\mathbf{O}\backslash\{X, Y\}$ is necessary and sufficient for there to be a direct causal relation between X and Y. Finally, $\mathcal{S}$-algorithms are *independence complete*: they output all the causal conclusions that can be drawn from their probabilistic input (cf. Meek 1995; Verma and Pearl 1992; Glymour 2010, 183).

Contrary to $\mathcal{S}$-algorithms, $\mathcal{NS}$-algorithms do not employ the assumption that $\mathbf{O}$ is causally sufficient. From the set $\mathbf{I}$ of conditional independencies over $\mathbf{O}$, they build an $\mathcal{NS}$-*pattern* which is a graphical representation of a class of DAGs each of which is defined over the union of $\mathbf{O}$ and of a set of so-called *latent* variables. In combination with (MC), each of the graphs represented by an $\mathcal{NS}$-pattern entails exactly the independencies recorded in $\mathbf{I}$ for the variables in $\mathbf{O}$. $\mathcal{NS}$-algorithms are correct in the following sense (cf. e.g. Spirtes, Meek, and Richardson 1995): if two variables X and Y are connected by a graphical feature ϕ in an $\mathcal{NS}$-pattern, X and Y are related in the way represented by ϕ in the causal graph $\mathcal{G}$ over $\mathbf{O} \cup \mathbf{CC_O}$, where $\mathbf{CC_O}$ designates a set such that for any common cause C of two variables X and Y in $\mathbf{O}$ it holds that C, or a cause of C, or an effect of C which is located on all directed paths from C to X and on all directed paths from C to Y is in $\mathbf{CC_O}$. Moreover, if X and Y are not connected by a graphical feature in an $\mathcal{NS}$-pattern, X and Y are causally unrelated. That means that the causal graph $\mathcal{G}$ over the causally sufficient set $\mathbf{O} \cup \mathbf{CC_O}$ is among the graphs represented by an $\mathcal{NS}$-pattern. As a result of not assuming that $\mathbf{O}$ is causally sufficient, the so-called *observational equivalence classes* that $\mathcal{NS}$-algorithms infer each contain an infinite number of DAGs—contrary to the Markov equivalence classes inferred by $\mathcal{S}$-algorithms. Moreover, $\mathcal{NS}$-patterns are more ambiguous than $\mathcal{S}$-patterns: while there is an arrow between X and Y in an $\mathcal{S}$-pattern only if this arrow appears in all the DAGs it represents, an analogous arrow in an $\mathcal{NS}$-pattern only means that all the represented DAGs feature either a path directed from X to Y or (at least) one subgraph featuring a path from a variable C to X and a path from C to Y where C is not in $\mathbf{O}$. That is, whereas in case of the assumed causal sufficiency of $\mathbf{O}$ a probabilistic dependency of X and Y conditional on every subset of $\mathbf{O} \setminus \{X, Y\}$ is necessary and sufficient for a direct causal relation between X and Y, such a probabilistic dependency of X and Y, when $\mathbf{O}$ is not assumed to be causally sufficient, only determines that X is a direct cause of Y in $\mathbf{O} \cup \mathbf{CC_O}$, *or* Y is a direct cause of X in $\mathbf{O} \cup \mathbf{CC_O}$, *or* X and Y have (at least) one common cause outside of $\mathbf{O}$. $\mathcal{NS}$-algorithms infer that X is a direct cause of Y only under very special probabilistic conditions (cf. Pearl 2000, 55). When these conditions are met, the arrow from X to Y in the corresponding $\mathcal{NS}$-pattern is marked with a special graphical feature. Finally, at least some $\mathcal{NS}$-algorithms, e.g. FCI, have also been proven to be independence complete (cf. Zhang 2008; Claassen and Heskes 2011; Glymour 2010, 186).[6]

[6] Strictly speaking, the algorithm that Zhang (2008) proves to be independence complete is not FCI as e.g. presented in Spirtes et al. (2000), but an extended version of FCI that implements two additional tail inference rules. Moreover, note that when we subsequently talk about $\mathcal{NS}$-algorithms we are only referring to algorithms that are provably independence complete.

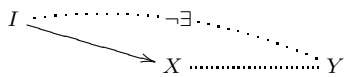

Fig. 1 A diagram that represents relevant features of $\mathbf{SP_1}$: there is a directed path from I to X, any sort of path or none at all between X and Y, and no path between I and Y that does not go through X.

Certain patterns output by BN algorithms are sufficient to very straightforwardly conclude that a triple of variables $\langle I, X, Y\rangle$ is an intervention triple. More precisely, certain patterns enable to identify triples of variables as being such that, based on the notion of causation that is employed by BN algorithms, the probabilistic input that these algorithms take is enough to justify their interventionist nature. For instance, take one of the simplest BN algorithms, the $\mathcal{S}$-algorithm IC (Pearl 2000, 50), and suppose that it outputs an $\mathcal{S}$-pattern $\mathbf{SP_1}$ which has the features depicted in figure 1.[7] Against the background of (CMC), (CFC), and the assumed causal sufficiency of $\mathbf{O}$, it can be concluded from $\mathbf{SP_1}$ that $\langle I, X, Y\rangle$ is a (soft) intervention triple, i.e. that $\langle I, X, Y\rangle$ satisfies (IV.i), (IV.iii), and (IV.iv). To see this, note first that in light of the correctness of IC the directed path from I to X in figure 1 entails that I is a cause of X: it is visible in $\mathbf{SP_1}$ that $\langle I, X, Y\rangle$ satisfies (IV.i). Second, the correctness of IC also yields that the absence in $\mathbf{SP_1}$ of a path between I and Y that does not go through X, i.e. the absence of what we shall henceforth call an *$\overline{X}$-path* between I and Y, warrants the conclusion that I is not a cause of Y along an $\overline{X}$-path. Therefore, (IV.iii) is also satisfied in $\mathbf{SP_1}$. Finally, to see that from the features of the triple $\langle I, X, Y\rangle$ in figure 1 it can likewise be inferred that (IV.iv) holds, suppose that—in violation of (IV.iv)—I is correlated with a cause Z of Y that is located on an $\overline{X}$-path. (CMC) and the correlation of I and Z entail that either (a) I is a cause of Z, or (b) Z is a cause of I, or (c) I and Z have a common cause (cf. Williamson 2005, 51-52). None of these, however, can in fact be the case. (a) is excluded because, in combination with Z being a cause of Y along an $\overline{X}$-path, (a) yields that I is a cause of Y on an $\overline{X}$-path. Yet, in light of the fact that $\langle I, X, Y\rangle$ complies with (IV.iii), this cannot be the case. Furthermore, it follows from both (b) and (c) that I and Y have a common cause on an $\overline{X}$-path. The sufficiency assumption exploited by $\mathcal{S}$-algorithms, in turn, implies that this common cause, or one of its causes, or one of its effects that is also a common cause of I and Y, is contained in the set $\mathbf{O}$ of analyzed variables. However, due to the correctness of IC, that I and Y have a common cause in $\mathbf{O}$ on an $\overline{X}$-path is incompatible with the absence of an $\overline{X}$-path between I and Y in $\mathbf{SP_1}$. (b) and (c) are thereby excluded. We can, hence, conclude that there indeed does not exist a variable Z that causes Y along an $\overline{X}$-path and that is correlated with I. Thus, $\langle I, X, Y\rangle$ satisfies (IV.iv).

Yet, identifying variables that can be justified to be intervention variables on the basis of BN methods, notwithstanding the ease with which this can be done, comes with a catch. For, as we shall see in the remainder of this section, whenever outputs of BN algorithms are determinate enough to establish the interventionist nature of a triple $\langle I, X, Y\rangle$, these outputs also determine whether X is a cause of Y or not. That is, BN methods, when identifying I as an intervention variable for X with respect to Y, also render it expendable to actually intervene on X via I to test whether X causes Y. We first show this for outputs of BN procedures that have the properties depicted in figure 1 and then demonstrate that the same holds for BN outputs in general.

If, as illustrated in figure 1, none of the DAGs represented by the $\mathcal{S}$-pattern $\mathbf{SP_1}$ features an $\overline{X}$-path between I and Y and, accordingly, the causal graph over $\mathbf{O}$ does not feature such a path, (CMC) entails that either (d) I and Y are unconditionally independent or (e) they

[7] Note that figure 1 does not depict $\mathbf{SP_1}$ itself. Rather, it represents relevant properties of $\mathbf{SP_1}$.

are unconditionally dependent but independent conditional on X. Against the background of (CFC), (d) implies that I and Y are causally independent, which yields that there does not exist a directed causal path from I to Y. As figure 1 ensures that there is a directed path from I to X in the causal graph over $\mathbf{O}$, it follows from the correctness of IC that there cannot exist a directed causal path from X to Y (for otherwise, contra (d), there would be a path from I via X to Y after all). That is, X is not a cause of Y. By contrast, (e) entails that there exists at least one directed or back-door[8] causal path between I and Y and that every such path goes through X. More explicitly, (e) entails that in the causal graph over $\mathbf{O}$, there exists at least one path of one of the following forms and that all paths between I and Y are of one of those types:

- $I \longrightarrow X \longrightarrow Y$
- $I \longleftarrow X \longleftarrow Y$
- $I \longleftarrow X \longrightarrow Y$
- $I \longleftarrow X \longleftarrow Z \longrightarrow Y$
- $I \longleftarrow Z \longrightarrow X \longrightarrow Y$

Only two of these possibilities are compatible with the acyclicity assumption that is embedded in BN methodologies. To see this, note again that $\mathbf{SP_1}$ guarantees that the causal graph $\mathcal{G}$ over $\mathbf{O}$ features a directed path from I to X. Combining the possible paths that follow from (e) with the directed path from I to X contained in $\mathcal{G}$ yields the following graphs:

- $I \longrightarrow X \longrightarrow Y$
- $I \rightleftarrows X \longleftarrow Y$
- $I \rightleftarrows X \longrightarrow Y$
- $I \rightleftarrows X \longleftarrow Z \longrightarrow Y$
- $I \longleftarrow Z \longrightarrow X \longrightarrow Y,\ I \longrightarrow X$

Only the first and the last of the graphs in that list are DAGs; the others contain a feedback structure that violates the acyclicity restriction. That is, if (e) holds, there is either a directed path from I via X to Y in $\mathcal{G}$ or there is a common common cause Z of I and X which, in turn, is a cause of Y. In both cases, X is a cause of Y. Therefore, (e) implies that X is a cause of Y.

Overall, the probabilistic (in)dependencies represented by $\mathbf{SP_1}$, in combination with the assumptions employed by $\mathcal{S}$-algorithms, entail that X is not a cause of Y (if I and Y are unconditionally independent) or they entail that X is a cause of Y (if I and Y are not unconditionally independent). Since it is either given with the input data fed into IC that I and Y are unconditionally independent or that they are not and since IC is independence complete, whichever is the case will be visible in $\mathbf{SP_1}$. If X is a cause of Y, there is a directed path from X to Y in $\mathbf{SP_1}$; and if X is not a cause of Y, none of the DAGs that are represented by $\mathbf{SP_1}$ features a directed path from X to Y.

A natural reaction to this finding will be to conjecture that the features depicted in figure 1, though sufficient, are not necessary to establish the interventionist nature of $\langle I, X, Y\rangle$, based on the probabilistic input taken by BN algorithms in combination with the account of the connection between causation and probabilities that these algorithms rely on. Other types of $\mathcal{S}$-patterns might reveal that I is an intervention variable for X with respect to Y

[8] For the notion of a *back-door path* cf. Pearl (1993).

without, at the same time, rendering it superfluous to actually intervene on X through I in order to uncover the causal relationship between X and Y. That, however, is not the case. Even though an $\mathcal{S}$-pattern may indeed warrant the interventionist nature of $\langle I, X, Y\rangle$ without all the features in figure 1, it nonetheless holds that all $\mathcal{S}$-patterns that substantiate that I is an intervention variable for X with respect to Y also determine whether X is a cause of Y or not. In other words, whenever an $\mathcal{S}$-pattern validates the interventionist nature of $\langle I, X, Y\rangle$, it renders actual interventions on X dispensable when it comes to investigating the causal relation between X and Y.

To show this, we establish that, together with the assumptions employed by $\mathcal{S}$-algorithms and with the conditional (in)dependencies that constitute the input of these algorithms, the fact that $\langle I, X, Y\rangle$ is an intervention triple either implies that X is a cause of Y or that X is not a cause of Y. It then follows, by independence completeness, that whenever an $\mathcal{S}$-pattern validates the interventionist nature of $\langle I, X, Y\rangle$ it also validates "X is a cause of Y" or it validates "X is not a cause of Y". We demonstrate this on a case-by-case basis where we distinguish between the three following cases: (1) I and Y are unconditionally independent, (2a) I and Y are unconditionally dependent and they are independent conditional on X, (2b) I and Y are unconditionally dependent and they are dependent conditional on X. In case (1), (CFC) implies that I and Y are *causally independent* in the following sense: neither of the two variables causes the other one nor do they have a common cause—be it inside or outside of $\mathbf{O}$. When $\langle I, X, Y\rangle$ satisfies (IV.i), i.e. when I is a cause of X, I and Y are causally independent in this sense only if there is no directed path from X to Y in the causal graph over $\mathbf{O}$, that is, if X does not cause Y (for, otherwise, there would be a directed path from I via X to Y after all). In case (2a), (CMC) and (CFC) yield that, in the causal graph over $\mathbf{O}$, I and Y are dependent and X is located on all the directed or back-door paths between I and Y. Using the example of $\mathbf{SP_1}$, we have shown above that all the causal graphs that comply with the acyclicity assumption and that feature a dependency among I and Y that is mediated by X feature either a directed causal path from I via X to Y or a common cause Z of I and X which is, in turn, a cause of Y. In both of those cases, X is a cause of Y. Finally, in case (2b), (CMC) implies that I and Y are causally dependent and that in the causal graph over $\mathbf{O}$ there exists at least one directed or back-door path between I and Y that does not go through X. If $\langle I, X, Y\rangle$ is an intervention triple, (IV.iii) entails that no $\overline{X}$-path is directed from I to Y. Analogously, given (CFC), (IV.iv) implies that there does not exist an $\overline{X}$-back-door path between I and Y in the causal graph over $\mathbf{O}$. Therefore, (2b) must be realized by an $\overline{X}$-path that is directed from Y to I in the causal graph over $\mathbf{O}$.[9] In combination with the fact that I causes X and with the acyclicity assumption that is used by $\mathcal{S}$-algorithms, the existence of an $\overline{X}$-path that is directed from Y to I implies that X does not cause Y (for, otherwise, there would be a circle from Y to I and via X back to Y).

In sum, if $\langle I, X, Y\rangle$ is an intervention triple, the probabilistic (in)dependencies that constitute the input of $\mathcal{S}$-algorithms determine either that X causes Y or that X does not cause Y. As a consequence, if these probabilistic (in)dependencies entail that $\langle I, X, Y\rangle$ is an intervention triple, then they also entail that X causes Y or that X does not cause Y. Thus, the independence completeness of $\mathcal{S}$-algorithms yields the following: whenever an $\mathcal{S}$-pattern is determinate enough to ensure that $\langle I, X, Y\rangle$ is an intervention triple, it is also determinate

[9] It is an open question whether I should be considered an intervention variable on X with respect to Y if there is a directed path from Y to I, i.e. if Y is a cause of I. On the one hand, Y being a cause of I is compatible with $\langle I, X, Y\rangle$ being an intervention triple in the sense provided by Woodward's definition (IV), which supplies the notion of an intervention variable we are concerned with here. On the other hand, Scheines (2005, 932) argues that Y being a direct cause of I gives rise to a form of treatment-bias that can induce fallacies in causal reasoning. Settling this issue must be left to another occasion.

enough to ensure that X causes Y or that X does not cause Y. Whether X causes Y can therefore be read off the $\mathcal{S}$-pattern itself—independently of any actual interventions on X.

It remains to be seen whether the same holds for $\mathcal{NS}$-algorithms. As a result of not assuming the causal sufficiency of $\mathbf{O}$, $\mathcal{NS}$-patterns represent classes of DAGs which are not defined over $\mathbf{O}$, but over supersets of $\mathbf{O}$. The (true) causal graph over $\mathbf{O} \cup \mathbf{CC_O}$—a common cause completion of $\mathbf{O}$—is among the graphs represented by an $\mathcal{NS}$-pattern. It turns out that the exact same case-by-case reasoning we have just applied to the causal graph over $\mathbf{O}$ can be repeated for the causal graph over $\mathbf{O} \cup \mathbf{CC_O}$. Irrespective of whether $\mathbf{O}$ is causally sufficient or not it holds that in case (1) (CFC) implies that I and Y are causally independent. This is only compatible with I being a cause of X (as entailed by IV.i) if there is no directed path from X to Y in the causal graph over $\mathbf{O} \cup \mathbf{CC_O}$. Hence, X is not a cause of Y. Similarly, whether $\mathbf{O}$ is causally sufficient has no bearing whatsoever on the fact that (2a)—in combination with (CMC) and (CFC)—implies that I and Y are causally dependent and that, in the causal graph over $\mathbf{O} \cup \mathbf{CC_O}$, X lies on all directed or back-door causal paths between I and Y. Since $\mathcal{NS}$-algorithms employ the same acyclicity assumption as $\mathcal{S}$-algorithms, it still holds that all the DAGs that can realize this consequence of (2a) feature a directed path from X to Y. Accordingly, there exists a directed path from X to Y in the causal graph over $\mathbf{O} \cup \mathbf{CC_O}$. Finally, what can be inferred from (2b) based on (CMC) and (CFC) is not affected by the causal insufficiency of $\mathbf{O}$ either. (2b) entails that I and Y are causally dependent via at least one $\overline{X}$-path. All such $\overline{X}$-paths that are compatible with (IV.iii) and (IV.iv) are directed from Y to I which, in combination with (IV.i) and the acyclicity restriction, yields that there is no directed path from X to Y in the causal graph over $\mathbf{O} \cup \mathbf{CC_O}$. Hence, X does not cause Y. As $\mathcal{NS}$-algorithms are also independence complete, it holds that whether X causes Y or not is visible in every $\mathcal{NS}$-pattern that entails that $\langle I, X, Y\rangle$ is an intervention triple. Whenever an $\mathcal{NS}$-pattern is determinate enough to warrant that $\langle I, X, Y\rangle$ is an intervention triple, it is also determinate enough to warrant that X causes Y or that X does not cause Y.

All of this shows that the problem we diagnosed for $\mathbf{SP_1}$ based on the features depicted in figure 1 generalizes to all BN-based identifications of intervention variables. Even though BN methods can identify, in a given set $\mathbf{O}$ of analyzed variables, triples of variables whose interventionist nature is justified by probabilistic information in combination with (CMC) and (CFC), this BN-guided identification of intervention variables always overshoots its mark. Whenever the output of a BN procedure—be it an algorithm of the $\mathcal{S}$- or of the $\mathcal{NS}$-type—determines that a particular triple $\langle I, X, Y\rangle$ in $\mathbf{O}$ is an intervention triple, this same output either establishes that X is a cause of Y or that X is not a cause of Y.[10] Thereby, identifying I as an intervention variable on X with respect to Y by means of BN algorithms answers the very question that an interventionist method of causal reasoning is designed to answer by intervening on X with respect to Y via I. BN methods not only identify $\langle I, X, Y\rangle$ triples as intervention triples but, when they do so, they also render interventionist techniques dispensable when it comes to resolving whether X causes Y.

[10] Scheines (2005, 928) makes a sort of converse observation: "Although we have no general characterization of the conditions under which a causal inference to $X \rightarrow Y$ can be made in observational studies, it turns out that when the inference is possible it is often driven by the existence of what I call a *detectible instrumental variable* that stands in the same relationship to X and Y in the observational study as does the ideal intervention on X in the experimental study." While Scheines says that if there is an arrow from X to Y in a BN pattern, this is *often* induced by the fact that the corresponding BN algorithm has detected an intervention variable for X with respect to Y in $\mathbf{O}$, we have proven that if $\langle I, X, Y\rangle$ is entailed to be an intervention triple, the corresponding BN pattern *always* determines whether X causes Y.

It might be objected that this finding is at odds with the popular idea of using interventions as a means to disambiguate causal inferences that remain ambiguous based on BN procedures alone (cf. Pearl 2000; Spirtes et al. 2000; Eberhardt et al. 2006; Eberhardt and Scheines 2007). However, as anticipated in the introduction, the literature on disambiguating causal inferences on the basis of interventionist techniques does not address the problem of identifying intervention variables. Rather, it simply presupposes that for all variables in an analyzed set of variables **O** suitable intervention variables are available *outside* of **O**. The suggestion then is that if the causal structure regulating the behavior of the members of **O** is not unambiguously determinable based on available probabilistic data, corresponding causal inferences can be disambiguated by means of those intervention variables that are assumed to exist outside of **O**. Yet, the question as to how the interventionist nature of these latent variables is detected or warranted is sidestepped. By contrast, this is exactly the question addressed in this paper. It is clear from the outset that BN methods can only be serviceable in answering this question as far as the identification of intervention variables *inside* a given **O** is concerned. And here, as this section has shown, BN methods overshoot the mark.

4.2 BN algorithms with weakened assumptions

The general methodological strategy employed by interventionist techniques is to modularly uncover a whole causal structure $\mathcal{G}$ based on substructures of $\mathcal{G}$ that comply with (IV). So far, this paper has shown that identifying these interventionist substructures in a methodologically guided manner that paves the way for unfolding the whole inferential potential of interventionist techniques is more problematic than might have been expected.

The main reason why BN methods overshoot the mark when it comes to identifying substructures complying with (IV) is that those methods rely on what might be called *total* causal assumptions. On the one hand, (CMC) and (CFC) are universal in the sense that they relate the probability distributions and the causal graphs over *all* (causally sufficient) variable sets. On the other hand, (CMC) and (CFC) are holistic in the sense that they connect features of probability distributions to relationships among all variables in a graph, rather than to relationships among a proper subset of those variables. As such, BN assumptions equally guarantee the causal interpretability of all graphical features in patterns output by BN algorithms. And the previous subsection has shown that both $\mathcal{S}$- and $\mathcal{NS}$-patterns have graphical features which, when interpreted causally, entail that a triple $\langle I, X, Y\rangle$ complies with (IV) only if these patterns also have graphical features which (again when interpreted causally) determine whether X is a cause of Y or not. That means, both in the case of $\mathcal{S}$- and $\mathcal{NS}$-algorithms, the total assumptions employed by BN methods support an inference to the interventionist nature of a triple $\langle I, X, Y\rangle$ only if they likewise support an inference to "X is a cause of Y" or to "X is not a cause of Y".

At the same time, the total nature of BN assumptions yields that they are very strong. Accordingly, they have been much criticized and various counterexamples to both (CMC) and (CFC) have been offered.[11] Moreover, in light of their strength, BN assumptions are difficult to justify. Indeed, the justifications that have been advanced in the literature have

[11] Concerning (CMC), see in particular Cartwright (1999a; 2001); Freedman and Humphreys (1999); Williamson (2005, 4.2). As to (CFC), see Freedman (1997) and Cartwright (2001), as well as Spirtes et al. (2000, 38-42) and Pearl (2000, 63-64) for typologies of relevant counterexamples.

given rise to some controversies.[12] This is consequential for our present purposes because BN-based identifications of interventionist triples provide a proper justification of the belief that the corresponding triples indeed have the interventionist properties only if the assumptions employed by BN algorithms can themselves be justified.

A natural reaction to the justificatory questions associated with the strength of BN assumptions is to weaken those assumptions: the weaker they are, the easier it is to justify them. For quite some time, therefore, proponents of BN methods have undertaken considerable efforts to exploit causal assumptions that are weaker than (CMC) and (CFC). In particular, as indicated before, they sometimes do not rely on the universal version of (CFC), but content themselves with assuming that particular variable sets that are relevant for a given causal analysis satisfy ($\text{CFC}_{\mathbf{V}}$). Or, Ramsey et al. (2006) and Zhang and Spirtes (2008) identify two components of ($\text{CFC}_{\mathbf{V}}$) called *Adjacency-* and *Orientation-Faithfulness*, only the former of which must be assumed to be satisfied by particular variable sets because the latter is empirically testable. Still, though, tests that determine whether a given variable set satisfies Orientation-Faithfulness are only correct under the assumption that this set satisfies both Adjacency-Faithfulness and ($\text{CMC}_{\mathbf{V}}$).

Here we can confine ourselves to noting that the total assumptions presented in the previous subsection, though sufficient, are *not necessary* for the causal interpretation of the output of a BN algorithm when run on a particular **O**—let alone for the causal interpretation of the specific graphical features that identify intervention variables. To establish the interventionist nature of a triple $\langle I, X, Y\rangle$, it is not necessary to assume anything about how, in general, probability distributions relate to whole causal graphs generated by (causally sufficient) variable sets. Roughly, all that is required for that purpose are assumptions about how the probabilistic (in)dependencies among the variables in that particular triple relate to the corresponding causal (in)dependencies. Hence, the assumptions needed to infer from a BN-output that $\langle I, X, Y\rangle$ complies with (IV) are (much) weaker than the assumptions advanced in the previous subsection. Moreover, as we shall see shortly, there exist patterns output by BN algorithms which, in combination with those weaker BN assumptions, yield that $\langle I, X, Y\rangle$ is an intervention triple without, at the same time, determining whether X is a cause of Y.

Whenever a concrete context of causal discovery allows for justifying the weaker but not the total BN assumptions, it may happen that $\langle I, X, Y\rangle$ is identifiable as intervention triple without it being equally determinable whether X causes Y. More precisely, whenever causal assumptions that suffice to infer from the pattern output by a BN algorithm that $\langle I, X, Y\rangle$ is an intervention triple do not suffice to infer (from the same pattern) that X causes Y or that X does not cause Y, a researcher can be in an epistemic position to justifiably draw the conclusion that I is an intervention variable for X with respect to Y while not being in a position to decide whether X causes Y. In such cases, intervening on X through I is non-redundant in order to determine whether X causes Y. That is, BN methodologies do not generally make modular interventionist techniques of causal reasoning dispensable when it comes to assessing whether one variable causes another. There exist epistemic contexts in which BN methods identify intervention variables and which are such that this identification does not *ipso facto* render the corresponding interventions expendable.

To conclude this paper we substantiate this claim by way of a (very) simple example. Let **O** be $\{I, J, X, Y\}$ and suppose that, when given the probabilistic independencies among

[12] For justifications of (CMC), see e.g. Pearl (2000, 44, 61-62), Spirtes et al. (2000, 33-40). As to (CFC), see in particular Spirtes et al. (2000, 41-42). The argument in favor of (CFC) that is given in this passage was rejected by Freedman (1997) and Cartwright (1999b, ch. 5; 2001).

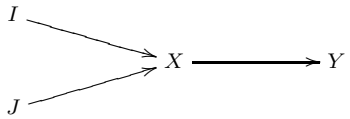

Fig. 2 Exemplary $\mathcal{S}$-pattern output by IC for $\mathbf{O} = \{I, J, X, Y\}$.

the members of $\mathbf{O}$ as an input, the $\mathcal{S}$-algorithm IC outputs the pattern depicted in figure 2. This pattern is itself a DAG and, accordingly, it is the only element of the set of DAGs that it represents. It indicates that I and X, J and X, X and Y cannot be screened off in $\mathbf{O}$, that I and J are unconditionally independent, that I and Y and J and Y are screened off by $\{X\}$, and that I and J are dependent conditional on X. Moreover, the subgraph over the triple $\langle I, X, Y\rangle$ in that $\mathcal{S}$-pattern has all of the features of figure 1, for which we showed in the previous subsection that they—in combination with (CMC), (CFC) and the assumption that $\mathbf{O}$ is causally sufficient—entail that I is a cause of X, that there is no causal $\overline{X}$-path from I to Y, and that I is not correlated with any other cause Z of Y that is located on an $\overline{X}$-path, i.e. that $\langle I, X, Y\rangle$ satisfies (IV.i), (IV.iii), and (IV.iv). That is, figure 2 has graphical features which, when interpreted against the background of the usual BN assumptions, are sufficient to establish that I is an intervention variable for X with respect to Y.

However, there exist weaker assumptions that also warrant the inference from figure 2 to the interventionist nature of I. To show this, we introduce the following predicates that are jointly weaker than the combination of predicates usual BN assumptions claim to be satisfied by all pairs or triples of variables in the corresponding sets:

Orientation_Markov_or_Common_Cause (OMCC): $\langle A, B_1, B_2\rangle$ satisfies (OMCC) relative to variable set $\mathbf{V}$ iff:
if
(1) A and B_1 cannot be screened off in $\mathbf{V}\backslash\{A, B_1\}$ and
(2) A and B_2 cannot be screened off in $\mathbf{V}\backslash\{A, B_2\}$ and
(3) B_1 and B_2 can be screened off in $\mathbf{V}\backslash\{B_1, B_2, A\}$
then
(4) A is an unshielded collider, i.e.
(a) A is an effect of B_1 or A and B_1 have a common cause outside of $\mathbf{V}$, and
(b) A is an effect of B_2 or A and B_2 have a common cause outside of $\mathbf{V}$
, and
(5) B_1 and B_2 are not causally related in $\mathbf{V}$, and
(6) B_1 and B_2 do not have a common cause outside of $\mathbf{V}$.

Causal_Sufficiency (CS): $\{A, B\}$ satisfies (CS) relative to $\mathbf{V}$ iff: every common cause C of A and B (if there is any) either is in $\mathbf{V}$, or has a cause which is in $\mathbf{V}$, or has an effect which is located on all directed paths from C to A and on all directed paths from C to B and which is in $\mathbf{V}$.

Path_Faithfulness (PF): $\{A, B\}$ satisfies (PF) relative to $\mathbf{V}$ iff: if $\mathbf{V}$ screens off between A and B, then there do not exist directed or back-door causal paths between A and B that do not go through $\mathbf{V}$.

Asymmetry (AS): $\langle A, B\rangle$ satisfies (AS) iff: if A causes B, then B does not cause A.

Existential_Common_Cause_Principle (ECCP): $\{A, B\}$ satisfies (EPCC) iff: if A and B are probabilistically dependent, then either A is a cause of B, or B is a cause of A, or they have a common cause.

If suitable pairs and triples of variables in **O** satisfy these predicates, it can be concluded from figure 2 that $\{I, X, Y\}$ is an intervention triple. To see this, consider the following five causal assumptions:

(i) $\langle I, X, J\rangle$ satisfies (OMCC) relative to **O**;
(ii) $\{I, X\}$ satisfies (CS) relative to **O**;
(iii) $\{I, Y\}$ satisfies (PF) relative to $\{X\}$;
(iv) $\langle I, X\rangle$ satisfies (AS);
(v) for every Z that is a cause of Y along an $\overline{X}$-path, $\{I, Z\}$ satisfies (ECCP).

(i) to (v) provide a connection between causal facts and probabilistic facts. Unlike the connection between causal facts and interventionist facts that is stipulated by (M) or the connection between causation and probabilities that is established by Suppes' theory or by (CMC) and (CFC) taken together, the connection that is provided by (i) to (v) is particular rather than universal and local rather than global. The only consequence this has in regard to justifying that $\langle I, X, Y\rangle$ is an interventionist triple consists in simplifying that justificatory task.

Together with the probabilistic (in)dependencies that are represented by figure 2, it can be inferred that $\langle I, X, Y\rangle$ is an interventionist triple. From (i) and the probabilistic (in)dependencies in **O** that are represented by figure 2 it follows that I is a cause of X or I and X have a common cause outside of **O**. Yet, in light of (ii) the latter can be excluded, to the effect that $\langle I, X, Y\rangle$ satisfies (IV.i). Moreover, (iii) in combination with the fact that X screens off between I and Y entails:

(†) There do not exist directed or back-door causal $\overline{X}$-paths between I and Y.

(†) immediately yields that there does not exist a causal $\overline{X}$-path that is directed from I to Y, i.e. that $\langle I, X, Y\rangle$ satisfies (IV.iii). Finally, under the assumptions listed above, $\langle I, X, Y\rangle$ also satisfies (IV.iv). To see this, note that (v) excludes that I is accidentally correlated with any other cause Z of Y that is located on an $\overline{X}$-path. As a consequence, I can only be correlated with such a cause Z of Y—in violation of (IV.iv)—if I is a cause of Z, or Z is a cause of I, or I and Z have a common cause. Yet, as we shall see shortly, all of these possible violations of (IV.iv) contradict (†). First, suppose that I is a cause of Z (which has been defined to be a cause of Y on an $\overline{X}$-path). It follows that I is a cause of Y on an $\overline{X}$-path, which is incompatible with (†). Second, assume that Z is a cause of I. It follows that either a directed causal path from Z to Y contains I but not X, or Z is a common cause of I and of Y while causing Y along an $\overline{X}$-path. The first disjunct entails that there exists an $\overline{X}$-path that is directed from I to Y, which is ruled out by (†). The second disjunct is compatible with (†) only if X is an intermediate between Z and I; but (IV.i), which we validated above, in combination with (iv) excludes that X is located on the directed path from Z to I. Therefore, both disjuncts of the disjunction entailed by Z being a cause of I are incompatible with (†). Third, suppose that Z and I have a common cause C. (IV.i) and (iv) again exclude that X is an intermediate on the path from C to I. Furthermore, as Z is defined to be located on an $\overline{X}$-path to Y, X cannot be an intermediate between C and Z either. Therefore, (†) equally excludes that Z and I have a common cause C. That is, if (IV.iv) is violated by any of the three causal structurings that can possibly lead to a correlation of I and another cause Z of Y, (†) is also violated. But we have already established (†) based on (iii) and figure 2. By contraposition we thus get that $\langle I, X, Y\rangle$ satisfies (IV.iv). In sum, in combination with the assumptions (i) to (v), the $\mathcal{S}$-pattern of figure 2 establishes that I is an intervention variable for X with respect to Y.

The assumptions (i) to (v), though sufficient, are not necessary for the validity of the inference to the interventionist nature of the triple $\langle I, X, Y\rangle$, i.e. they could be further weakened. Yet, all that is important for us here is that they are already weak enough *not* to validate

the inference from figure 2 to "X is a cause Y". Or differently put, assumptions (i) to (v) do not warrant the causal interpretation of the arrow $X \longrightarrow Y$ in figure 2. To see this, note that (i) to (v) do not entail any of the following:

(vi) if X and Y cannot be screened off in $\mathbf{O}$, then X is a direct cause of Y in $\mathbf{O}$, or Y is a direct cause of X in $\mathbf{O}$, or X and Y have a common cause outside of $\mathbf{O}$;
(vii) if there exists a subset of $\mathbf{O} \setminus \{I, Y\}$ containing X that screens off between I and Y, then $\langle I, X, Y\rangle$ does not form an unshielded collider at X;
(viii) if there exists a subset of $\mathbf{O} \setminus \{J, Y\}$ containing X that screens off between J and Y, then $\langle J, X, Y\rangle$ does not form an unshielded collider at X.

However, violations of these conditions render the probabilistic (in)dependencies represented in figure 2 compatible with X *not* causing Y. For instance, if condition (vi) is violated, the fact that X and Y cannot be screened off in $\mathbf{O}$ does not entail that X is a cause of Y, or Y is a cause of X, or X and Y have a common cause outside of $\mathbf{O}$. Accordingly, the arrow from X to Y in figure 2 is compatible with the causal independence of X and Y. Similarly, if conditions (vii) and (viii) are violated, the probabilistic (in)dependencies represented by figure 2 are compatible with $\langle I, X, Y\rangle$ and $\langle J, X, Y\rangle$ forming unshielded colliders at X, to the effect that Y or a cause of Y located outside of $\mathbf{O}$ causes X. In other words, the causal interpretation of the arrow from X to Y in figure 2 requires that conditions (vi) to (viii) are satisfied. Yet, this is not ascertained by the causal assumptions (i) to (v). That is, (i) to (v) are causal assumptions which are sufficient to infer from figure 2 that I is an intervention variable for X with respect to Y, but at the same time (i) to (v) do not validate the inference from figure 2 to "X is a cause of Y".

All of this illustrates that BN methodologies, when employed in combination with assumptions that are weaker than usual BN assumptions, do not necessarily render interventionist techniques dispensable. A researcher can be in an epistemic position to justifiably infer from a BN output that I is an intervention variable on X with respect to Y while not being in a position to infer from that output that X causes Y or that X does not cause Y. In such an epistemic context, intervening on X through I is a relevant and non-redundant way of investigating the causal relation between X and Y.

5 Conclusion

The availability of intervention variables is the essential prerequisite for unfolding the inferential potential of interventionist methods of causal reasoning. The first part of this paper has shown that the belief that a triple $\langle I, X, Y\rangle$ is an intervention triple can only be justified if recourse is made to a non-interventionist account of causation. Moreover, we have seen that, depending on the background knowledge at hand about a given $\langle I, X, Y\rangle$, its interventionist nature is more or less easily established on non-interventionist grounds. Determining whether administering a particular anti-depression drug counts as an intervention variable for depression with respect to insomnia presupposes extended studies on the side-effects of that drug. By contrast, that coin tossing is an intervention variable for treatment with respect to recovery is justifiable based on an elementary probabilistic account of causation because we have the relevant background knowledge concerning the probabilistic dependencies and independencies among these variables.

Consequently, the second part of the paper has then turned to the most standard procedures of causal discovery that process observational data, *viz.* BN methods. We have shown that methodically identifying interventionist triples by means of these procedures renders

interventionist techniques dispensable. BN methods rely on causal assumptions that are so strong that they substantiate the interventionist nature of $\langle I, X, Y \rangle$ only if they support the inference to "X causes Y" or to "X does not cause Y". We have also seen, though, that relevant BN assumptions can be weakened—and thus rendered more easily justifiable—to the effect that outputs of BN procedures may identify I as intervention variable for X with respect to Y without, at the same time, deciding whether X causes Y. This finding suggests that whether interventionist techniques are fruitfully applicable in a given context of causal discovery, in the end, is determined by the epistemic background of that discovery context. If BN methods are used to identify intervention variables, it is profitable to implement interventionist techniques only if that epistemic background allows for justifying BN assumptions which, like (i) to (v), authorize the inference to the interventionist nature of $\langle I, X, Y \rangle$, but does not allow for justifying BN assumptions which, like (vi), (vii) and (viii), support the causal interpretation of the graphical feature relating X and Y in a corresponding BN output.

Acknowledgements We are grateful to Alessio Moneta and Wolfgang Spohn for helpful comments and discussions, and to audiences at the universities of Geneva, Konstanz, Madrid and Rotterdam where previous versions of this paper have been presented. Moreover, we thank the anonymous referees of this journal for their valuable comments and suggestions. Finally, Michael Baumgartner is indebted to the Deutsche Forschungsgemeinschaft (DFG) and Isabelle Drouet to the Agence Nationale de la Recherche (ANR) for generous support of this work (project CAUSAPROBA).

References

Baumgartner, M. (2009). Interdefining causation and intervention. *Dialectica 63*, 175–194.

Cartwright, N. (1999a). Causal diversity and the Markov condition. *Synthese 121*, 3–27.

Cartwright, N. (1999b). *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

Cartwright, N. (2001). What is wrong with Bayes nets? *The Monist 84*(2), 242–264.

Cartwright, N. (2010). What are randomised controlled trials good for? *Philosophical Studies 147*(1), 59–70.

Claassen, T. and T. Heskes (2011, April). Arrowhead completeness from minimal conditional independencies. Technical report, Radboud University Nijmegen, iCIS/IS/ML.

Collins, H. M. and T. J. Pinch (1993). *The Golem: what you should know about science* (2 ed.). Cambridge: Cambridge University Press.

Eberhardt, F., C. Glymour, and R. Scheines (2006). $n - 1$ experiments suffice to determine the causal relations among n variables. In D. Holmes and L. Jain (Eds.), *Innovations in Machine Learning*, pp. 97–112. Berlin: Springer.

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

Eells, E. (1991). *Probabilistic Causality*. Cambridge: Cambridge University Press.

Freedman, D. (1997). From association to causation via regression. In V. R. McKim and S. P. Turner (Eds.), *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*, pp. 113–161. Notre Dame: University of Notre Dame Press.

Freedman, D. and P. Humphreys (1999). Are there algorithms that discover causal structure? *Synthese 121*(1), 29–54.

Glymour, C. (1997). A review of recent work on the foundations of causal inference. In V. R. McKim and S. P. Turner (Eds.), *Causality in Crisis?*, pp. 201–248. Notre Dame: University of Notre Dame Press.

Glymour, C. (2010). What is right with 'Bayes net methods' and what is wrong with 'hunting causes and using them'? *British Journal for the Philosophy of Science 61*(1), 161–211.

Korb, K. B., L. R. Hope, A. E. Nicholson, and K. Axnick (2004). Varieties of causal intervention. *Lecture Notes in Computer Science ISSU 3157*, 322–331.

Korb, K. B. and E. Nyberg (2006). The power of intervention. *Minds and Machines 16*, 289–302.

Kvart, I. (2001). Causal relevance. In B. Brown (Ed.), *New Studies in Exact Philosophy: Logic, Mathematics and Science*, pp. 59–90. Oxford: Hermes Scientific Publications.

Meek, C. (1995). Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, San Francisco, USA, pp. 411–418. Morgan Kaufmann.

Nyberg, E. and K. Korb (2006). Informative interventions. In F. Russo and J. Williamson (Eds.), *Causality and Probability in the Sciences*, London. College Publications.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann Publishers, Inc.

Pearl, J. (1993). Graphical models, causality, and intervention. *Statistical Science 8*, 266–269.

Pearl, J. (2000). *Causality. Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Ramsey, J., J. Zhang, and P. Spirtes (2006). Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, Arlington, Virginia, pp. 401–408. AUAI Press.

Scheines, R. (2005). The similarity of causal inference in experimental and non-experimental studies. *Philosophy of Science 72*(5), 927–940.

Spirtes, P., C. Glymour, and R. Scheines (1991). From probability to causality. *Philosophical Studies 64*, 1–36.

Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search* (2 ed.). Cambridge: MIT Press.

Spirtes, P., C. Meek, and T. Richardson (1995). Causal inference in the presence of latent variables and selection bias. In *In Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 499–506. Morgan Kaufmann.

Strevens, M. (2007). Review of Woodward, Making Things Happen. *Philosophy and Phenomenological Research LXXIV*, 233–249.

Strevens, M. (2008). Comments on Woodward, Making Things Happen. *Philosophy and Phenomenological Research LXXVII*, 171–192.

Suppes, P. (1970). *A Probabilistic Theory of Causality*. Amsterdam: North Holland.

Verma, T. and J. Pearl (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proc. of the Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 323–330. Morgan Kaufmann.

Williamson, J. (2005). *Bayesian Nets and Causality* (Oxford ed.). Oxford University Press.

Wimsatt, W. C. (2006). Reductionism and its heuristics: Making methodological reductionism honest. *Synthese 151*.

Woodward, J. (2003). *Making Things Happen*. Oxford: Oxford University Press.

Woodward, J. (2008). Response to Strevens. *Philosophy and Phenomenological Research LXXVII*, 193–212.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence 172*(16-17), 1873 – 1896.

Zhang, J. and P. Spirtes (2008). Detection of unfaithfulness and robust causal inference. *Minds and Machines 18*(2), 239–271.