Detection of Unfaithfulness and Robust Causal Inference

Jiji Zhang
Division of Humanities and Social Sciences
California Institute of Technology
jiji@hss.caltech.edu

Peter Spirtes
Department of Philosophy
Carnegie Mellon University
ps7z@andrew.cmu.edu

Abstract

Many algorithms proposed in the machine learning community for inferring causality from data are grounded on two assumptions, known as the Causal Markov Condition and the Causal Faithfulness Condition. Philosophical discussions of the latter condition have focused on how often and in what domains we can expect it to hold or fail. This paper instead investigates to what extent the faithfulness can be tested. The investigation yields a theoretical and a practical result: a strictly weaker Faithfulness condition which is nonetheless sufficient to justify some reliable methods of causal inference, and a way to make some causal inference procedures more robust. The latter, we argue, is related to the possibility of controlling the probability of large errors with finite sample size ("uniform consistency") in causal inference.

1. Introduction

Recent work on causal modeling and reasoning (e.g., Pearl 2000, Spirtes et al. 2000, Dawid 2002) has emphasized an important kind of inductive problem: how to infer what would happen to a unit or a system if the unit or system were intervened upon to change in some way, based on observations of similar units or systems in the absence of the intervention of interest. We encounter this kind of problems when we try, for example, to estimate the outcomes of medical treatments, policy interventions or our own actions before we actually prescribe the treatments, implement the policies or carry out the actions, with the relevant experience being accumulated through passive observations.

Such problems are significantly harder than the typical uniformity-based induction from observed instances to new instances. In the latter situation, we take ourselves to be making an inference about new units in the same population from which the observed samples were drawn. In the former situation, thanks to the intervention under consideration, it is known that the new units do not belong to the same population as the observed samples, and we are making an inference across different populations.

To solve such problems, we need information about the underlying causal structure over relevant attributes (often represented as variables) as well as information about how the causal structure would be modified by the interventions in question. The latter kind of information is usually supplied in the very specification of an intervention, which describes what attributes would be directly affected by the intervention, and what attributes would not be directly affected (and hence would remain governed by their original local mechanisms).

The widely accepted tool for discovering causal structure is of course randomized experiments. But randomized experiments, for a variety of reasons, are not always feasible to carry out. Indeed we would not face the kind of inductive situations described in the first paragraph were randomized experiments always possible. Instead we would face a simpler situation in which observed instances and new instances can be assumed to conform to the same data-generating process, and we can extrapolate observed experimental results to new instances in a fairly straightforward way.

So in the kind of situations that concern us here, we are left with the hope of inferring causal structure from observational data. The task is of course impossible without some assumption connecting causal structure with statistical structure, but is not entirely hopeless given some such assumptions (and possibly limited domain-specific background knowledge). In the past decades, a prominent approach to causal inference based on graphical representations of causal structures has emerged from the artificial intelligence and philosophical literatures, and has drawn wide attention from computer scientists, philosophers, social scientists, statisticians and psychologists. Two assumptions are usually made explicit --- and when not, are usually implicit --- within this framework, known as the Causal Markov Condition (CMC) and the Causal Faithfulness Condition (CFC).

The CMC states that the true probability distribution of a set of variables is *Markov* to the true causal structure in the sense that every variable is independent of its non-effects given its direct causes. The CFC states that the true probability distribution is *faithful* to the true causal structure in the sense that if the true causal structure does not entail a conditional independence relation according to the CMC, then the conditional independence relation does not hold of the true probability distribution.

A considerable philosophical literature is devoted to debating the validity of the CMC, and in particular, the principle of the common cause as an important special case (see e.g. Sober 1987, Artzenius 1992, Cartwright 1999, Hausman and Woodward 1999, among others). The CFC also spurs critical discussions and defenses from philosophers (e.g., Woodward 1998, Cartwright 2001, Hoover 2001, Steel 2006), and despite the fact that published reflections on the CFC are less extensive than those of the CMC, practitioners seem in general to embrace the CMC, but regard the CFC as more liable to failure.

In this paper we propose to examine the CFC from a testing perspective. Instead of inquiring under what conditions and how often should the CFC be expected to hold, we ask whether and to what extent is the CFC testable, assuming the CMC holds. Our purpose is two-fold. First, as a logical or epistemological issue, we hope to understand the minimal core of the untestable part of the CFC, or in other words, the theoretically weakest faithfulness condition one needs to assume in order to employ the graph-based causal inference techniques. Second, and more practically, we want to incorporate necessary checks for the testable part of the CFC into existing causal inference procedures to make them more robust.

The paper is organized as follows. We set up the background in Section **2**, in a slightly different way than what is standard in the literature. In Section **3**, we present a decomposition of the CFC into separate conjuncts, and demonstrate the role each component plays. We consequently show that given one component from the decomposition --- a strictly weaker faithfulness condition --- the other components are either testable or irrelevant. Hence in principle the weaker condition is sufficient to do the job the standard CFC is supposed to do. In Section **4**, we illustrate that even the weaker faithfulness condition identified in Section **3** is more than necessary, and present a more general characterization of what we call undetectable failures of faithfulness. In Section **5**, we discuss how the simple detection of unfaithfulness identified in Section **3** improves the robustness of causal inference procedures. As it turns out, it is not just a matter of guarding against errors that might arise due to unfaithfulness, but also a matter of being cautious about “almost unfaithfulness”. We illuminate the point by connecting it to the interesting issue of uniform consistency in causal inference, which is related to the possibility of estimating the probability of errors as a function of sample size. We end the paper in Section **6** by suggesting how the work can be generalized to the situation where some causally relevant variables are unobserved.

2. Causal Bayes Nets and Causal Inference

2.1 Causal Bayes Nets

Following a recent trend in the philosophical and scientific literature on causation, we focus on causal relations between variables, and adopt a broadly interventionist conception of causation (Woodward 2003). Given a set of random variables **V**, we assume that for any subset **S** of **V** and any vector of values **s** for the variables in **S**, there is a hypothetical (external) intervention that *set*s the value of **S** to be **s** (without disturbing the local mechanisms for variables in **V****S**), associated with which there is a joint probability distribution of **V****S**, P(**V****S || S**:=**s**), denoting the probability distribution the set of members of V that are not members of S, **V\S**, would follow if **S** were intervened on to take value **s**. (Note the double bar in the notation to distinguish it from the ordinary conditional probability.) In the case of null manipulation, i.e., when **S** is the empty set, we simply have the probability distribution **V** follows in the absence of external intervention, P(**V**) = P(**V** || ∅).

For our purposes, the causal structure of **V** is simply regarded as something that enables us to calculate P(**V****S || S**:=**s**) in terms of P(**V**). In other words, we are interested in the aspect or manifestation of the causal structure that supplies information for connecting P(**V****S || S**:=**s**) to P(**V**), because such a connection, if known, would solve the kind of inductive problems we mentioned at the beginning. A well-developed calculus of this sort uses directed acyclic graphs (DAGs) as a representation of causal structures (Spirtes et al. 1993, Pearl 2000). A graph consisting of a set of vertices (representing variables) and a set of edges between variables is *directed* if the edges are all directed arrows (written as →), and is *acyclic* if it contains no directed paths from a vertex back to itself. A simple example we will use to illustrate throughout the paper is given in Figure 1, slightly adapted from an oft cited case in the literature (Hesslow 1976). The graph consists of five vertices, representing five variables, and five arrows between various variables. There is no directed cycle, i.e, no directed path from any variable back to itself.

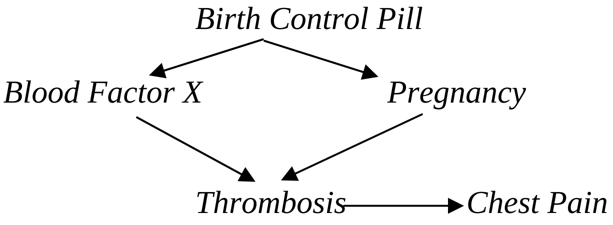


Figure 1: A Causal DAG

Given a DAG representing the causal structure, or a causal DAG, over **V**, we assume the following manipulation principle, versions of which have been formulated by several authors (Strotz and Wold 1960, Robins 1986, Pearl 2000 and Spirtes et al. 1993).

> **Manipulation Principle:** Given a causal DAG *G* over **V**, for any subset **S** of **V** and any value **s** for **S,** $P(\mathbf{V}\backslash\mathbf{S} \;\|\; \mathbf{S}{:=}\mathbf{s}) = \prod_{X\in\mathbf{V}\backslash\mathbf{S}} P(X \mid \mathbf{PA}_{G}(X))$, where $\mathbf{PA}_{G}(X)$ is the set of parents of X in *G*.

One immediate technical problem with this principle is that $P(X \mid \mathbf{PA}_{G}(X))$ may not always be well defined. There are several ways to handle this, but we will adopt the simplest one here by assuming that P(**V**) is positive[1], so that the conditional probabilities are all well defined. This positivity assumption amounts to requiring that there is no deterministic relationship among variables in **V** (and the probability of exogenous variables has full support). We believe that there are interesting extensions of our results when determinism is allowed, but will leave them to another occasion (see Glymour forthcoming for an extension of causal discovery algorithms to allow determinism).

Obviously the manipulation principle postulates a connection between $P(\mathbf{V}\backslash\mathbf{S} \;\|\; \mathbf{S}{:=}\mathbf{s})$ and P(**V**) via a causal DAG, which supplies the information of $\mathbf{PA}_{G}(X)$ for each X in **V**. Note that if we take **S** to be the empty set, the manipulation principle implies that $P(\mathbf{V}) = \prod_{X\in\mathbf{V}} P(X \mid \mathbf{PA}_{G}(X))$, i.e., that the probability distribution of **V** (in the absence of external intervention) "factorizes" according to the causal graph. This gives us what is called a Bayesian network or Bayes net over **V**. The Causal Markov Condition (CMC) is an equivalent of the factorization condition, which we write down here once again with reference to causal DAGs.

[1] One alternative way is to abandon the "ratio definition" of conditional probability (cf. Hajek 2003). Another way is to prohibit manipulations to values that have probability zero in the pre-manipulation setting. We think the first way is particularly promising, but for sake of simplicity will not pursue it in this paper.

> **Causal Markov Condition**: Given a set of variables **V** whose causal structure is represented by a DAG *G*, every variable in **V** is probabilistically independent of its non-descendants in *G* given its parents in *G*.

So, a bit unlike formulations elsewhere (but in line with Tian and Pearl 2002), the present formulation of the manipulation principle includes the CMC as a special case. Our intention is to highlight the close affinity between the CMC and the condition for calculating manipulation effects. The CMC is of a special status not only because it provides a bridge between the (pre-manipulation) probability and causal structure, but, more importantly we think, also because it needs to be posited in the post-manipulation context for the sake of causal reasoning. On the other hand, if we expect the CMC to hold in a post-manipulation context, it seems plausible to extend that expectation to the pre-manipulation context as well (cf. Hausman and Woodward 2004).

Throughout this paper (with an exception at the end), we assume that the set of variables we work with admits a DAG representation of its causal structure, in the sense that *there is a DAG over the variables such that the manipulation principle is satisfied*. We call the representation a *causal Bayes net*, and the assumption the *causal Bayes net assumption*.

> **Causal Bayes Net Assumption**: There is a directed acyclic graph over **V** that satisfies the manipulation principle.

We do not wish to go so far as to elevate the metaphysical status of this assumption by arguing that causal structure and manipulation principle are conceptually or semantically related, in the spirit of Spohn (2000). For all we know, there might be causal structures that cannot be represented by DAGs even in principle. And even assuming a DAG representation, notions like direct cause and causal structure can be defined quite independently of the manipulation principle (e.g., Cooper 1999, Woodward 2003, and also see below). Our reason for making the assumption is simply that it is unclear what a causal structure is good for in quantitative reasoning across different contexts unless the manipulation principle or some surrogate of it holds.

The causal Bayes net assumption does not ensure the uniqueness of the DAG representation of causal structure. In fact, one can show that if a DAG *G* is a causal Bayes net over **V** (i.e., satisfies the manipulation principle), any DAG that is a supergraph of *G* is also a causal Bayes net over **V**. However, all causal Bayes nets over **V** satisfy the following condition:

> **Core of Causal Bayes Nets (CCBN)**: for any two variables X, Y $\in$ **V**, and **Z** = **V** \ {X, Y}, if there exists $x_1 \neq x_2$ and **z**, such that P(Y || X:=x_1, **Z**:=**z**) $\neq$ P(Y || X:=x_2, **Z**:=z), then X is a parent of Y.

Lemma 1: All causal Bayes nets over **V** satisfy CCBN.
Proof: See Appendix C.

The antecedent of CCBN has been used by several authors to define the notion of direct causation relative to **V** (e.g. Woodward 2003, Pearl 2000). Let G_d be the graph over **V** such that there is an arrow from X to Y (X → Y) if and only if X is a direct cause of Y relative to **V** in the sense of CCBN.

Lemma 2: If **V** satisfies the causal Bayes net assumption, then G_d is a correct representation of the causal structure of **V** in the sense that it satisfies the manipulation principle.
Proof: See Appendix C.

We have presented the formalism in a slightly unconventional way. It seems natural to start with the definition of G_d, and call it *the* representation of causal structure. However, our way of presentation seems to have the following purchase. As already intimated, there is no conceptual guarantee, as nearly as we can see, that G_d satisfies the manipulation principle. So if we start with the definition of G_d, we still need to postulate the manipulation principle as an add-on, and the definition of G_d does not seem to give much clue as to what the manipulation principle should be. Now, if the manipulation principle is posited first, as it is done in our setup, G_d naturally falls out as a privileged representation of the causal structure.

G_d naturally falls out because Lemmas 1 and 2 imply that it is the uniquely minimal causal Bayes net for **V** if **V** satisfies the causal Bayes net assumption, that is, if there is a DAG at all that satisfies the manipulation principle for **V**.[2] Thus, besides its intuitive appeal, there is a good pragmatic reason to take G_d as the (representation of) true causal structure, and ignore the possibility that there might be some account of direct causation according to which some variable Z should be regarded as a direct cause of another variable W relative to **V** even though they do not satisfy the antecedent of CCBN. In other words, even if, contrary to our belief, there is some good reason to regard a proper supergraph of G_d, but not G_d, as representing the true causal structure, G_d will fare just as well in getting the manipulation effects right. For these reasons, we will henceforth refer to G_d as the true causal graph.

We also get the assumption of what is called the Causal Minimality Condition for free.

> **Minimality**: No proper subgraph of the true causal graph (G_d) over **V** satisfies the Markov condition with P(**V**), in the sense that P(**V**) factorizes according to the graph.

As we already noted, the true causal graph G_d is the uniquely minimal graph that satisfies the manipulation principle. Since the manipulation principle is stronger than the Markov condition, the latter of which only concerns pre-manipulation probability P(**V**), it is not immediately obvious that G_d is a minimal graph that satisfies the Markov condition with

[2] More specifically, Lemma 1 implies that any causal Bayes net over **V** is a supergraph of G_d, and Lemma 2 implies that if any DAG is a causal Bayes net over **V**, then G_d does. Hence G_d is the uniquely minimal causal Bayes net of **V** unless there is no causal Bayes net for **V** at all.

P(**V**) --- it is certainly not the uniquely minimal one. However, we need just a little extra work to show that:

Lemma 3: No proper subgraph of G_d satisfies the Markov condition with P(**V**).
Proof: See Appendix C.

To summarize, the causal Bayes net assumption gives us both the Causal Markov and the Causal Minimality conditions[3]. These constitute the backdrop of our examination of the testability of faithfulness.

2.2 Causal Faithfulness Condition and Causal Discovery

The Causal Markov and Minimality Conditions do not get us very far in learning causal structure from observational data (i.e., samples from p(**V**)), unless there is strong background knowledge about the causal ordering among the variables. For every ordering of variables in **V**, there is a DAG consistent with that order that satisfies the Markov and minimality conditions with P(**V**). Except in rare cases, these DAGs consistent with different causal orders share little in common, and hence the true causal graph is vastly underdetermined by P(**V**). The Causal Faithfulness Condition (CFC) helps to mitigate the underdetermination problem to a considerable degree. Let us recall what it says.

> **Causal Faithfulness Condition**: Given a set of variables **V** whose true causal DAG is *G*, the joint probability of **V**, P(**V**), is faithful to *G* in the sense that P(**V**) implies no conditional independence relations not already entailed by the CMC.

Clearly the CFC is sort of converse to the CMC. The CMC lists a number of conditional independence relations, which in turn may entail some others via probability calculus, that P(**V**) must satisfy given the true causal graph. The CFC in effect gives a number of conditional dependence relations P(**V**) must satisfy by requiring that those conditional independence relations entailed by the CMC are the only ones that hold.

Indeed there are equivalent formulations of the CMC and CFC that make this converse relationship more explicit. One of the most important formal results in the AI and statistics literature in the past three decades is that a graphical criterion called *d-separation* can capture exactly the conditional independence relations entailed by the Markov condition applied to a graph. A precise definition of d-separation is given in Appendix A. But basically it is a three-place relation, as in "**A** and **B** are d-separated by **C**", where **A, B, C** are three disjoint subsets of **V**.

In terms of d-separation, we can reformulate the CMC as saying that if **A** and **B** are d-separated by **C** in the true causal graph, then **A** and **B** are probabilistically independent conditional on **C** according to the true probability measure. Conversely, the CFC can be reformulated as saying that if **A** and **B** are probabilistically independent conditional on **C**

[3] We hope to have explained clearly the sense in which the causal Bayes net assumption, by itself, gives us the minimality condition. It is not by way of strict entailment --- for that we need the definition of direct cause and true causal graph via CCBN. But that definition, as we argued, is well motivated given the causal Bayes net assumption rather than a pure add-on.

according to the true probability measure, then **A** and **B** are d-separated by **C** in the true causal graph.

The CFC implies the Causal Minimality Condition: if the true causal graph entails exactly the conditional independence relations true of P(**V**), then any proper subgraph of the true graph will entail an additional piece of conditional independence relation that is not true of P(**V**), and hence will fail the Markov condition with P(**V**). Assuming the CMC and CFC, it is usually possible to derive interesting and useful features of the true causal graph from observational data, if the sample size is big enough for reliable inference of conditional independence. The reason is that although the two conditions do not completely dissolve the problem of underdetermination of causal structure by patterns of conditional independence and dependence, the structures that are underdetermined often share interesting common features.

More specifically, the two conditions set up an exact correspondence between conditional independence relations and d-separation relations in the true causal DAG. The conditional independence relations, and hence the d-separation and d-connection relations in the causal DAG, can be determined, in the large sample limit, from the observational data. The question is how much about the structure of the DAG can be inferred from the d-separation and d-connection relations. Quite a bit, usually, because DAGs that share the exact same d-separation features also share the exact same adjacencies and some arrow directions as well.

Take, for instance, the case depicted in Figure 1. Suppose, unknown to us, the DAG in Figure 1 is the true causal graph, and, furthermore, that we have observed a large number of women on these five variables, from which we correctly infer conditional independence relations among the variables. Given this, assuming the CMC and CFC, we can infer that the true causal graph is one of the three in Figure 3.

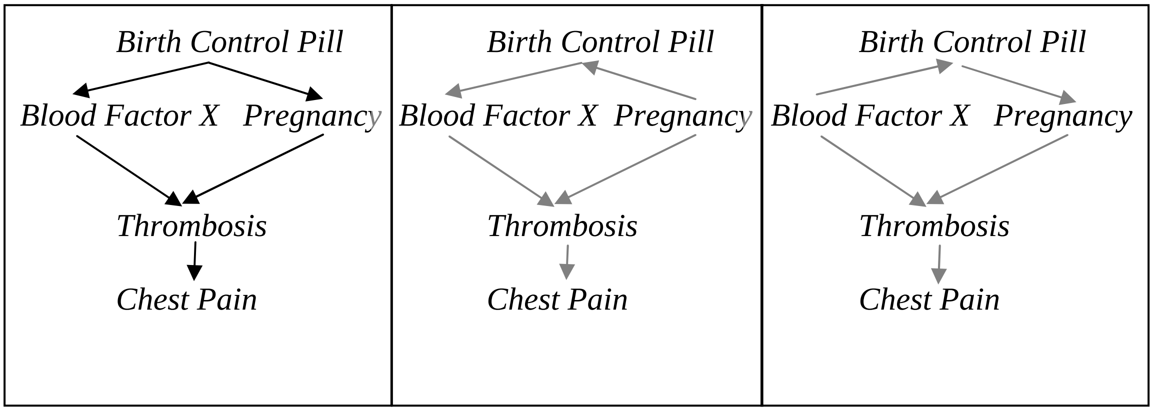


Figure 2: Graphs that are Markov equivalent

These graphs are called *Markov* equivalent because they share the exact same d-separation relations, and hence entail the exact same conditional independence relations. So they cannot be distinguished based on conditional independence facts. The good news is that not every feature of the true causal graph is underdetermined. Notice that all three graphs share the same adjacencies (which is true in general for Markov equivalent DAGs), and share some arrow directions. Oftentimes these common features are sufficient to enable calculations of manipulation effects in terms of pre-manipulation probabilities. For example, in this case, it is true that P(Chest Pain || Thrombosis :=yes) = P(Chest pain | Thrombosis = yes), no matter which of the three is the true causal graph.

Assuming the CMC and CFC, various algorithms have been developed in the artificial intelligence community to learn a set of Markov equivalent graphs from data and extract the common features (e.g., Verma and Pearl 1990, Spirtes et al. 2000, Chickering 2002). The PC algorithm (Spirtes et al. 1993), for example, is provably correct for learning from an oracle of conditional independence relations[4] the Markov equivalence class to which the true causal DAG belongs. We will end this section by introducing the basics of the PC algorithm, because it will help to illustrate our points later. The PC algorithm assumes that every variable in **V** is observed, which is also what we will assume throughout, until in the end when we will briefly discuss the case where only some variables in **V** are observed.

The PC algorithm contains two major steps. The first major step determines which variables are adjacent to each other in the causal DAG. It is motivated by the following lemma about d-separation due to Pearl (1988).

Lemma 4 (Pearl)**:** Two variables are adjacent in a DAG if and only if they are not d-separated by any subset of other variables in the DAG.

In the first step of determining adjacencies, the PC algorithm essentially searches for a conditioning set for each pair of variables that renders them independent, which is called a *screen-off* conditioning set. Given Lemma 4, two variables are not adjacent if and only if such a screen-off set is found. What distinguishes the PC algorithm is the way it performs search, in which some tricks are employed to increase both computational and statistical efficiency. The details of the tricks are not important for our purpose, and we include the pseudo-code in Appendix B for interested readers.

For example, if we apply the PC algorithm to the case in Figure 1 with a correct oracle of conditional independence as input, we get the adjacency graph (an undirected graph) in Figure 3a after the first step.

[4] In practice the oracle is of course implemented with statistical tests, which are reliable only when the sample size is sufficiently large (and the distributional assumptions are satisfied for parametric tests). We will return to the sample size issue in Section **5**.

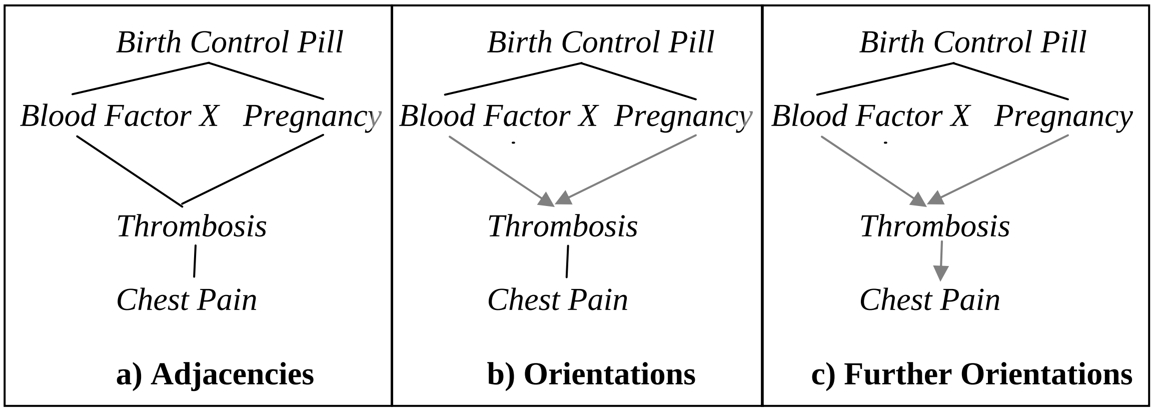


Figure 3: Phases of the PC Algorithm

The second major step of the PC algorithm seeks to derive as many arrow orientations as possible. Call a triple of variables <*X*, *Y*, *Z*> in a DAG an *unshielded triple* if *X* and *Z* are both adjacent to *Y*, but *X* and *Z* are not adjacent. It is called an unshielded *collider* if the two edges are both into *Y*: $X \rightarrow Y \leftarrow Z$; otherwise it called an unshielded non-collider.

The following fact about d-separation is crucial for deriving arrow orientations.

Lemma 5 (Pearl)**:** In a DAG, any unshielded triple <*X*, *Y*, *Z*> is a collider if and only if all sets that d-separate *X* from *Z* do not contain *Y*; it is a non-collider if and only if all sets that d-separate *X* from *Z* contain *Y*.

In light of Lemma 5, the PC algorithm simply looks at every unshielded triple <*X*, *Y*, *Z*> in the adjacency graph resulting from the first step, and orients the triple as a collider if and only if the screen-off set for *X* and *Z* found in the first step does not contain *Y*. For example, the PC algorithm will produce Figure 3b from Figure 3a after this operation. Finally, some logical consequences of the orientation information discovered so far are made explicit. The exact inference rules are not important for the present purpose. In our example, the final output from the PC algorithm is in Figure 3c.

The output of the PC algorithm is called a pattern (a.k.a. PDAG or essential graph) that contains both undirected and directed edges. The undirected edges indicate ambiguity regarding arrow orientation. Meek (1995a) presented a version of the PC algorithm such that the output is complete in the sense that if an edge $A \rightarrow B$ occurs in every DAG in the Markov equivalence class, and hence is not underdetermined, then it is oriented as $A \rightarrow B$ in the output pattern.

With this basic understanding of how causal inference goes, we are ready to examine what role the CFC plays, and to what extent the CFC is testable.

3. A Decomposition of CFC[5]

Although the CFC looks like a standard methodological assumption of simplicity, the standard asymptotical justifications of causal discovery procedures do take the CFC as a substantive posit about the world (Spirtes et al. 2000). As a substantive assumption, the CFC incurs criticisms that cite versions of Simpson's paradox or more generally cases where multiple causal pathways exactly cancel each other, examples of homeostatic systems, and cases where causal transitivity fails, etc. Relevant discussions have thus focused on how often or rarely cancellation of multiple pathways or failure of transitivity could occur, or in what domains would the CFC be particularly shaky or safe (See e.g. Meek 1995b, Woodward 1998, Pearl 2000, Spirtes et al. 2000, Cartwright 2001, Hoover 2001, Steel 2006). There is certainly more to say along this line, but we will leave it aside. Instead we examine the condition from a testing perspective.

The issue of testing the CFC has got very little mention in the literature, not surprisingly, because a short answer to our question is obviously no. The CFC is in general not a testable assumption, at least in the context of inferring causal structure from patterns of probabilistic associations, because the CFC does not simply specify a property of the background probability distribution for a set of variables, but rather specifies a relationship between the probability distribution and the underlying (unknown) causal structure. To test the CFC, intuitively, one needs information about the causal structure in the first place (see Spanos 2006 for an example of testing the CFC with assumptions about causal structure).

This general consideration, sound as it is, overlooks a simple point that turns out to be both theoretically interesting and practically fruitful. Although not every aspect of the CFC is testable, some kinds of failure may be detectable. In specifying a relationship between the background probability distribution and the underlying causal structure, the CFC also specifies a testable property of the probability distribution --- that it is faithful to *some* causal DAG. In principle, assuming the CMC holds, whether a probability distribution is faithful to any causal structure is testable.

Thus there is a distinction to draw between violations of the CFC that are not detectable and violations of the CFC that are in principle detectable using the probabilistic information alone. The idea is simple. If the true probability distribution is not faithful to any DAG, then it is a ***detectable*** failure of faithfulness. By contrast, if the true probability distribution is not faithful to the true causal DAG, but is nonetheless faithful to some other DAG, then it is a case of ***undetectable*** violation of faithfulness (see the examples in Section 4).

We now suggest a decomposition of the CFC that gives us a simple but nice result on detecting unfaithfulness. Recall that the CFC can be formulated as saying that for any three disjoint subsets of **V**, **X, Y,** and **Z**, if **X** and **Y** are not d-separated by **Z**, then **X** are

[5] Part of this section was included in a joint paper with Joseph Ramsey (Ramsey, Spirtes and Zhang 2006).

Y are probabilistically dependent conditional on **Z**. In view of Lemma 4 in Section **2.2**, it is easy to see that the CFC implies the following:

> **Adjacency-Faithfulness Condition:** Given a set of variables **V** whose true causal DAG is *G*, if two variables *X*, *Y* are adjacent in *G*, then they are dependent conditional on any subset of **V**\{*X*,*Y*}.

Another part of the CFC needed to justify the PC algorithm is this:

> **Orientation-Faithfulness Condition:** Given a set of variables **V** whose true causal DAG is *G*, let <*X*,*Y*,*Z*> be any unshielded triple in *G*.
> 1. if $X \rightarrow Y \leftarrow Z$, then *X* and *Z* are dependent given any subset of **V**\{*X*,*Z*} that contains *Y*;
> 2. otherwise, *X* and *Z* are dependent conditional on any subset of **V**\{*X*,*Z*} that does not contain *Y*.

That the Orientation-Faithfulness condition is also a consequence of the CFC is evident given Lemma 5 in Section **2.2**.

The Adjacency-Faithfulness and the Orientation-Faithfulness do not constitute an exhaustive decomposition of the CFC. Both of them are consequences of the CFC, but they together do not imply the CFC. Consider, for instance, a causal graph consisting of a simple chain $X \rightarrow Y \rightarrow Z \rightarrow W$. We can easily cook up a case/parameterization for this causal structure in which the causal influence along the chain fails to be transitive (more on this below), and as a result X is probabilistically independent of W, which violates the CFC because they are obviously d-connected. But the distribution does not have to violate the Adjacency-Faithfulness or the Orientation-Faithfulness. We can easily make the case such that the only independence relations that hold are $X \perp W$, $X \perp W|Y$, $X \perp W|Z$ and $X \perp W|\{Y,Z\}$.[6] It is easy to check that Adjacency-Faithfulness and Orientation-Faithfulness are both satisfied, whereas the CFC is violated due to the independence between X and W.

However, the leftover of the CFC apart from Adjacency-Faithfulness and Orientation-Faithfulness is irrelevant to the correctness of causal discovery procedures like PC. The correctness of the PC algorithm only depends on the truth of the Adjacency-Faithfulness and Orientation-Faithfulness conditions. As long as these two components of the CFC hold, the PC algorithm will not err given the right oracle of conditional independence. In the above four-variable chain, for instance, the PC algorithm will output X — Y — Z — W, with <X, Y, Z> and <Y, Z, W> being unshielded non-colliders, which is obviously correct. It is in general true that only Adjacency-Faithfulness and Orientation-Faithfulness play a role in justifying causal inference procedures like PC.

[6] $\perp$ is a symbol that denotes probabilistic independence introduced by Dawid (1979). The vertical bar | denotes conditioning.

We now turn to the two relevant components of the CFC. The four-variable chain example shows that in general there exist cases where Adjacency-Faithfulness and Orientation Faithfulness are both satisfied but the standard CFC is violated. It is of course equally obvious that there exist cases where the Adjacency-Faithfulness condition holds but the Orientation-Faithfulness condition fails. Again, this can be illustrated with a simple chain $X\rightarrow Y \rightarrow Z$ where the causal influence fails to be transitive along the chain. McDermott (1995) gave a well-known example of the sort. The story goes roughly like this: a right-handed terrorist is about to press a detonation button to explode a building when a dog bites his right hand, so he uses his left hand instead to press the button and triggers the explosion. Intuitively, the dog-bite causes the terrorist pressing the button with his left hand, which in turn causes the explosion, but the dog-bite does not cause the explosion.

Let X be the variable that takes two values: 'yes' if dog bites, and 'no' otherwise; Y be the variable that takes three values: 'right' if the terrorist presses the button with his right hand, 'left' if he does it with his left hand, and 'none' if he does not press the button at all; and Z be the variable that takes two values: 'yes' if explosion occurs, and 'no' otherwise. In line with McDermott's story, X is a direct cause of Y, and Y is a direct cause of Z, relative to {X, Y, Z}, but there is no direct causal relationship between X and Z. So the causal graph is $X\rightarrow Y \rightarrow Z$. Moreover, $P(Z \,\|\, X{:=}yes) = P(Z \,\|\, X{:=}no)$, and there is no counterfactual dependence of Z on X of any sort. This kind of failure of causal transitivity is peculiar[7], but it takes some formalism to make it precise, and is not important for our present purpose. Suffice it to say that, in such cases, we have $X\perp\!\!\!\perp Z$ and $X\perp\!\!\!\perp Z|Y$.[8] Checking against the causal graph, we observe that the Adjacency-Faithfulness condition holds, but the Orientation-Faithfulness condition is violated.[9]

[7] It is of course old news that counterfactual dependence can fail to be transitive, which motivated David Lewis's earliest attempt to define causation in terms of ancestral of counterfactual dependence. And no one expect the relation of direct cause to be transitive either. What is peculiar about this case is that it is a failure of transitivity along a single path, and thus it is case of intransitivity of what is called *contributing cause* (Pearl 2000, Hitchcock 2001b). Most counterexamples to causal transitivity in the literature are either cases of intransitivity of what is called *total cause* or cases of intransitivity of probability-increasing, which involve multiple causal pathways (Hitchcock 2001a). It seems to us that intransitivity of contributing cause is more surprising.

[8] McDermott's story does not give us a strictly positive joint distribution over the three variables. But it is easy to modify the story in order to meet the assumption of positivity we made in Section **2**. For example, we can imagine that the terrorist is not so resolute as to admit no positive probability of not pressing the button, and there are some other factors that render a positive probability of explosion even in the absence of the terrorist's action. As long as whether dog bits or not does not affect the (non-zero) probability of the terrorist abstaining, and which hand the terrorist uses does not affect the probability of explosion, we have our case.

Note, however, that the assumption of positivity, or alternatively, the assumption of no determinism is not relevant to the result presented in this section. It will be relevant to the more general result in Section 4, but only because we need the causal minimality condition there, and our justification of the causal minimality condition in Section **2** uses, though not in an essential way, the assumption of positivity.

[9] A technical point: for restricted class of causal structures and family of probability distributions, the adjacency-faithfulness condition may imply the orientation-faithfulness condition. In other words, there is no probability from the given family that is adjacent-faithful but not orientation-faithful to a causal structure in the given class. For example, in the case of simple chains, if we restrict to binary variables or Gaussian variables that bear linear relationships, there do not exist distributions that are adjacency-faithful

But this case of unfaithfulness is obviously detectable. It is easy enough to check that the distribution of which only $X \perp\!\!\!\perp Z$ and $X \perp\!\!\!\perp Z|Y$ is not faithful to any DAG over {X, Y, Z}. And the point is a general one: any failure of the Orientation-Faithfulness condition alone is detectable. In other words, if we assume the Adjacency-Faithfulness condition, the Orientation-Faithfulness condition is testable. The argument is quite simple, and reveals how the test could be done. Suppose the CMC and the Adjacency-Faithfulness condition hold. As we explained by way of the PC algorithm, the two assumptions imply that out of a correct oracle of conditional independence, one can construct the correct adjacency graph, and thus obtain unshielded triples in the true causal graph. For any such unshielded triple, say, <X, Y, Z>, recall what the Orientation-Faithfulness requires: if the triple is a collider in the true causal graph, no screen-off set of X and Z includes Y; otherwise, every screen-off set of X and Z includes Y. How could this condition fail? By the CMC, if the triple is a collider, then there exists some screen-off set of X and Z that does not include Y (either the set of X's parents or the set of Z's parents in the true causal graph). So it cannot be the case that the triple is a collider but every screen-off set of X and Z includes Y. Likewise, it cannot be the case that the triple is a non-collider but no screen-off set of X and Z includes Y, as again implied by the CMC. Therefore, the Orientation-Faithfulness fails of the triple <X, Y, Z> if and only if Y is included in some screen-off set of X and Z, and not in others. For example, in the simple dog-bite case, the Orientation-Faithfulness condition fails because one screen-off set of X and Z, i.e., the empty set does not include Y ($X \perp\!\!\!\perp Z$), and another screen-off set of X and Z, i.e., {Y}, includes Y ($X \perp\!\!\!\perp Z|Y$).

Since this sufficient and necessary condition for the failure of Orientation-Faithfulness does not refer to the unknown graph, whether it is true or not is answerable by the oracle of conditional independence, and hence is in principle testable. Again, the reason why we can test it without knowing whether the triple is a collider or a non-collider, is because any distribution that is Markov and Adjacency-Faithful to the true causal DAG is either Orientation-Faithful to the true causal DAG, or not Orientation-Faithful to *any* DAG. So we have just established the following simple but useful theorem:

Theorem 1: Assuming the CMC and the Adjacency-Faithfulness condition, any violation of the Orientation-Faithfulness condition is detectable.

As intimated earlier, the standard CFC is in a sense stronger than necessary to justify some standard causal inference procedures. All that matters are the two components:

but not orientation-faithful. (The dog-bite case obviously involves a variable with three categories.) More generally there are known results (e.g., Becker et al. 2000) that imply that in binary tree-like networks adjacency-faithfulness implies orientation-faithfulness. This result can be generalized to Gaussian tree-like networks as well. If we do not restrict to tree-like causal structures and consider general DAGs, as we are concerned with in this paper, both binary and linear Gaussian networks admit failure of orientation-faithfulness but not adjacency-faithfulness. The simplest example is cancellation of two causal pathways, as, for example, Birth Control Pill → Blood Factor X → Thrombosis, and Birth Control Pill → Pregnancy → Thrombosis in Figure 1. So in general, assuming the adjacency-faithfulness condition does not imply orientation-faithfulness at all.

Adjacency-Faithfulness and Orientation-Faithfulness. But this observation does not have any implication for actual methodology. Theorem 1, by contrast, has a methodological overtone. It suggests that we can further relax the Faithfulness assumption to Adjacency-Faithfulness alone, and empirically test the Orientation-Faithfulness part rather than simply assuming it.

This motivates a simple twist to the PC algorithm. As we briefly described in Section **2.2**, a key step for deriving orientations in the PC algorithm is to check, for any unshielded trip <X, Y, Z>, whether Y is contained in the screen-off set of X and Z found in the earlier stage of inferring adjacencies. Under the Orientation-Faithfulness assumption, this single check is enough to determine whether the triple is a collider or a non-collider. Without the assumption of Orientation-Faithfulness condition, however, this single check can lead us astray.

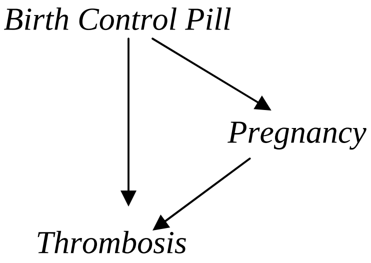
For example, the case depicted in Figure 1 has appealed to philosophers due to the fact that taking birth control pills, on the one hand, raises the chance of thrombosis via, in our modified case, the route Birth Control Pill → Blood Factor X → Thrombosis, and, on the other hand, decreases the chance of thrombosis via the route Birth Control Pill → Pregnancy → Thrombosis (taking birth control pills prevents pregnancy, which is itself a positive factor for thrombosis). Suppose the chance raising route and the chance lowering route cancel each other exactly, and as a result, whether a woman takes birth control pills is probabilistically independent of whether she suffers thrombosis (conditional on the empty set). This violates the Orientation-Faithfulness assumption, and the PC algorithm, given a correct oracle of conditional independence, will wrongly infer that <Birth Control Pill, Pregnancy, Thrombosis> is a collider, because Pregnancy is not included in the screen-off set of Birth Control Pill and Thrombosis it checks, i.e., the empty set.

A simple remedy is to test the Orientation-Faithfulness condition by also checking whether Pregnancy is included in some other screen-off set of Birth Control Pill and Thrombosis. If it is, which means that the Orientation-Faithfulness fails, then one cannot infer whether the triple is a collider or not, and should rightly remain silent on this matter. This leads to what we call the Conservative PC algorithm. It is labeled conservative because it avoids making definite inference when it detects failure of Orientation-Faithfulness.

More details of the Conservative PC algorithm are described in Appendix B. The procedure is provably correct under the assumptions of CMC and Adjacency-Faithfulness. By incorporating a test of Orientation-Faithfulness, the procedure is, not surprisingly, computationally more expensive than the PC algorithm. But extensive simulations have shown that the extra computational burden is negligible (Ramsey et al. 2006). More interestingly, simulations strongly suggest that the Conservative PC algorithm returns significantly more accurate result than the PC algorithm on moderate sample sizes, even when the sampling distribution is in fact faithful to the true causal structure. We will return to this interesting issue in Section **5**, but before that there is more to say about detectable unfaithfulness.

4. A Further Characterization of Undetectable Failure of Faithfulness

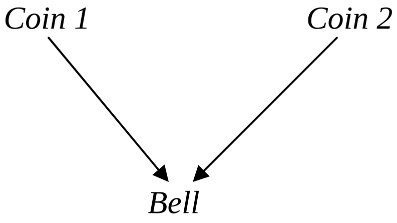
Theorem 1 isolates the Orientation-Faithfulness part of the CFC as testable given that the Adjacency-Faithfulness part of the CFC is assumed. What about violations of the Adjacency-Faithfulness condition? Certainly not every violation of the Adjacency-Faithfulness condition is detectable. For example, consider the version of the Thrombosis case usually discussed in the literature, where only three variables are considered, Bill Control Pill, Pregnancy and Thrombosis, with the causal structure as in Figure 4. Again, if the two causal paths from Birth Control Pill to Thrombosis cancel off exactly, Birth Control Pill is probabilistically independent of Thrombosis, which fails the Adjacency-Faithfulness condition because the two variables are adjacent in the graph. This case of unfaithfulness, however, is not detectable. Because the distribution is faithful to an alternative structure: Birth Control Pill → Pregnancy ← Thrombosis.



$\textit{Birth Control Pill} \perp\!\!\!\perp \textit{Thrombosis}$

Figure 4: Undetectable Failure of Adjacency-Faithfulness

Nonetheless, there are detectable violations of Adjacency-Faithfulness. Consider the following case adapted from Pearl (1988). Two fair coins are flipped independently. If the two coins both turn up heads or both turn up tails, a bell rings with probability 0.2, and otherwise the bell rings with probability 0.8. The causal structure is depicted in Figure 5. It is easy to calculate that P(*Bell* =1 | *Coin1* = H) = (*Bell* =1 | *Coin1* = T) = 0.5, and hence $\textit{Bell} \perp\!\!\!\perp \textit{Coin1}$. (The same goes with *Bell* and *Coin 2*.) The distribution and the causal structure clearly violate the Adjacency-Faithfulness, because *Bell* and *Coin 1* are adjacent in the graph. However, the distribution is not faithful to any DAG over the three variables, unless the CMC is violated. So, assuming the CMC, the unfaithfulness in this case is detectable.

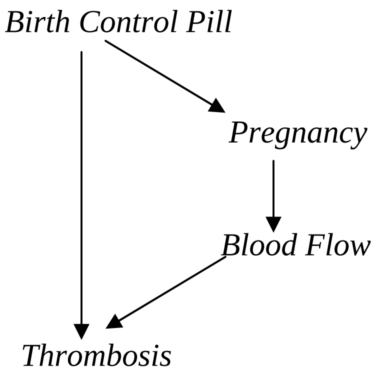


P(*Coin1* = H) = 0.5
P(*Coin2* = H) = 0.5
P(*Bell* = 1 | H, H) = 0.2
P(*Bell* = 1 | T, T) = 0.2
P(*Bell* = 1 | H, T) = 0.8
P(*Bell* = 1 | T, H) = 0.8

Figure 5: Detectable Failure of Adjacency-Faithfulness

Here is one notable difference between the two examples. In the undetectable case, the failure of Adjacency-Faithfulness is due to the fact that there is another pathway that causally connects the two variables, besides the direct connection between them, such that the two pathways cancel out each other. In the detectable case, the failure of Adjacency-Faithfulness is not due to cancellation of multiple paths. As we will see presently, all undetectable cases of unfaithfulness involve some sort of cancellation of multiple causal connections between two variables.

Another relevant feature of the case in Figure 4 is that the graph contains a triangle, three variables that are adjacent to one another. To see this, consider a modification of the case by adding an intermediate variable between Pregnancy and Thrombosis, say, the speed of blood flow --- pregnancy increases the chance of thrombosis by reducing the speed of blood flow. Suppose Figure 6 represents the true causal structure, and suppose again that the two causal pathways between Birth Control Pill and Thrombosis exactly cancel each other, resulting in a failure of Adjacency-Faithfulness. It is easy to check that the resulting distribution is not faithful to *any* DAG over the four variables (unless the CMC is violated). Hence the failure of Adjacency-Faithfulness in this case is detectable, even though it arises out of cancellation. Breaking the triangle makes the unfaithfulness detectable.



$\textit{Birth Control Pill} \perp \textit{Thrombosis}$

Figure 6: Detectable Failure of Adjacency-Faithfulness

We are thus motivated to define Triangle-Faithfulness as follows:

> **Triangle-Faithfulness Condition**: Given a set of variables **V** whose true causal DAG is *G*, let *X, Y*, *Z* be any three variables that form a triangle in *G* (i.e., they are adjacent to one another):
>
> (1) If *Y* is a non-collider on the path <*X*, *Y*, *Z*>, then *X*, *Z* are dependent conditional on any subset of **V**\{*X, Z*} that does not include *Y*.
>
> (2) If *Y* is a collider on the path <*X, Y, Z*>, then *X*, *Z* are dependent conditional on any subset of **V**\{*X, Z*} that includes *Y*

Despite the somewhat complicated formulation, the Triangle-Faithfulness Condition is obviously a consequence of the Adjacency-Faithfulness condition. It is strictly weaker than the latter, because the examples in Figure 4 and Figure 6 are clearly cases in which the Adjacency-Faithfulness condition fails but the Triangle-Faithfulness condition still holds.

To appreciate what the Triangle-Faithfulness condition requires, it is best to consider what a violation of the condition involves. It involves a triangle <X, Y, Z> such that X and Z are independent conditional on some subset **S** of **V**, and moreover, the conditioning set does not block (see the definition of d-connecting path in Appendix A) the path <X, Y, Z>. So at least two paths (<X, Y, Z> and <X, Z>) are active relative to the conditioning set **S**, and some sort of cancellation (which may involve more triangles) takes place to produce the independence. This point can be made more precise in linear models, but we will content ourselves with this informal remark here.

The main result is that if the CMC and the Minimality condition hold, as the causal Bayes net assumption gave us, then any failure of the CFC is detectable as long as the Triangle-Faithfulness condition is not violated.

Theorem 2: Under the assumptions of CMC and Minimality, if the CFC fails and the failure is undetectable, then the Triangle-Faithfulness condition fails.

Proof: See Appendix C.

Given Theorem 2, we can make better sense of two common remarks in the literature. One remark is that the CFC is plausible because exact cancellation rarely occurs, as if there are no other ways to fail the CFC than cancellation of multiple causal paths. There are other ways, but there is now a sense in which the more serious violations are all due to cancellations. The other remark is that causal inference procedures are most hopeful when the underlying causal structure is sparse. This remark of course already makes a lot of sense from a computational and statistical point of view. And now we have yet another perspective to make sense of the remark. Since triangles are needed for undetectable failures of the CFC, the sparser the causal structure, the more unlikely to have triangles in the structure, and the more unlikely for undetectable unfaithfulness.

Theorem 2 immediately entails Theorem 1. But Theorem 1 is special in that the argument the lead to it in the last section was constructive and readily presented a concrete check of unfaithfulness to be incorporated in standard inference procedures. The added check turns out sufficiently efficient, and is sufficiently localized so that when one triple is detected to be unfaithful, suspense of judgment only applies there, and informative inference may still be made about other parts of the structure. We are currently exploring analogous detectives based on Theorem 2.

More importantly, it seems that incorporating a test of faithfulness in the inference procedure not only guards against detectable unfaithfulness, it actually improves performance even when the CFC actually holds. To this interesting issue we now turn.

5. More Robust Causal Inference with a Check of Unfaithfulness

As we briefly mentioned in Section **3**, the PC algorithm is modified to incorporate a test of Orientation-Faithfulness. The resulting algorithm is labeled Conservative PC. Both algorithms are described in Appendix B. Since the Conservative PC algorithm, but not PC, is provably correct asymptotically under a strictly weaker assumption (i.e., the Adjacency-Faithfulness condition) than the standard CFC, it is, in a clear theoretical sense, more robust than the PC algorithm. One may worry, however, that the theoretical robustness not only comes with a computational cost, but, more importantly, may not cash out in practice if the situations where the Orientation-Faithfulness fails do not arise often. After all isn't a usual defense of the CFC simply that it will rarely fail? When the CFC actually holds, wouldn't the Conservative PC algorithm be unnecessarily conservative?

With these questions in mind, Joseph Ramsey did extensive simulations comparing the two algorithms on moderate sample sizes, with samples coming from a distribution faithful to the data-generating process. In other words, it is a comparison of the finite-sample performance of the two algorithms when the CFC actually holds in the population. It turns out, as reported in Ramsey et al. (2006), the Conservative PC algorithm runs almost as fast as the PC algorithm, and is significantly more reliable than the standard PC algorithm.

Why is this so? It is intuitively clear that the answer is to be sought in a largely vague concept of "almost unfaithfulness" or "close-to-unfaithfulness" in the literature (Meek 1995b, Robins et al. 2003, Zhang and Spirtes 2003, Steel 2006). If the true population distribution is available, the PC and the CPC algorithm will give the exact same output as long as the distribution satisfies the CFC (and more accurately, the Adjacency-Faithfulness and Orientation-Faithfulness conditions)[10], and will diverge if the Orientation-Faithfulness condition fails. There is no issue of close-to-unfaithfulness in that theoretical result; all that matters is the black-and-white matter of whether the Orientation-Faithfulness holds. In practice, however, we do not have direct access to the true population distribution, and need to do statistical inference based on finite sample. Here, it becomes very relevant to causal inference whether the probability distribution, though faithful to the true causal structure, is far from or close to being unfaithful.

Intuitively, a population distribution is close-to-unfaithful to a causal structure, if the structure does not entail some conditional independence relation according to the CMC, but the conditional independence almost holds, or in other words, the conditional dependence is by some measure very weak in the population. Exactly how weak counts as "close to independence" is a matter of degree and, properly speaking, a matter relative

[10] By the way, this is another virtue of the Conservative PC procedure. At least in theory, it is appropriately conservative in that it only suspends judgment when the input distribution is truly compatible with multiple alternatives, where PC would make a definite choice.

to sample size.[11] But it is clear that at every finite sample size, there are distributions that are faithful to the true causal structure but are so close to being unfaithful that they may make trouble for inference at that sample size, just as a strict failure of faithfulness may cause trouble even with infinite sample size.

So the reason why the Conservative PC algorithm is more robust than the standard PC algorithm, even when the Orientation-Faithfulness holds in the population, is that the detective of unfaithfulness inserted there also guards against "almost failure" of the Orientation-Faithfulness at finite sample size. When the sample size is not enough to distinguish between a given unshielded triple being a collider and it being a non-collider, the Conservative PC procedure suspends judgment, and returns "don't know" for that triple. It is quite analogous to the fact that the Conservative PC procedure will suspect judgment in the large sample limit, if the Orientation-Faithfulness strictly fails of that triple, because even an infinite amount of data cannot distinguish between the two alternatives.

To further illustrate the point, let us connect the issue to some interesting formal work. An important impossibility result on causal inference using statistical data is proved in Robins et al. (2003), stating that under the assumptions of CMC and CFC, causal inference from statistical data can only be *pointwise consistent*, but not *uniformly consistent*. Their argument essentially turns on the fact that the CFC allows the possibility of distributions that are arbitrarily close to being unfaithful to the true causal structure. Zhang and Spirtes (2003) highlight the point by showing that slight strengthening of the CFC that rules out some close-to-unfaithful situations defeats the impossibility theorem. Given our forgoing discussion, it is not surprising that the issue is closely related to the comparison between Conservative PC and PC.

Let us explore a little bit. For that we need some formalism. Let V^n denotes a random sample from the distribution P(**V**) with sample size *n*. A *statistical test* of a null hypothesis H_0 versus alternative H_1 is a function ϕ that takes V^n as input, and returns one of three possible answers: 0, 1, or 2, representing ``acceptance", ``rejection" or ``no conclusion", respectively.[12] Notice that we allow a test to return an uninformative answer, which is needed especially in the context of causal inference, where alternative hypotheses may be underdetermined by a sampling distribution.

In fact we are interested in hypothesis testing as a special case of model selection. From our point of view, the purpose of a statistical test is to decide whether the observed data came from a probability distribution compatible with the null hypothesis or from a probability distribution compatible with the alternative hypothesis. So a statistical test amounts to a procedure to discriminate between two sets of probabilities --- one corresponding to the null hypothesis and the other corresponding to the alternative hypothesis --- based on the observed sample. Let $\mathbf{P}_0$ be the set of probability distributions

[11] In Zhang and Spirtes (2003), we defined stronger versions of the faithfulness condition to exclude close-to-unfaithful parameterizations in linear Gaussian models. One defect of our definitions there is that it is uniform across all sample sizes rather than being adaptive to sample size.

[12] Strictly speaking, ϕ denotes a sequence of functions (ϕ_1, ϕ_2, …, ϕ_n, …), one for each sample size.

compatible with the null hypothesis H_0, and $\mathbf{P}_1$ the set of probability distributions compatible with the alternative hypothesis H_1. For all we know, $\mathbf{P}_0$ and $\mathbf{P}_1$ may not be disjoint, and the intersection of them underdetermines the hypotheses of interest in an obvious sense. Here is what pointwise consistency means:

> **Pointwise Consistency**: A test ϕ of H_0 versus H_1 is pointwise consistent, if
> (1) for every $P \in \mathbf{P}_0$, $\lim_n P(\phi(V^n) = 1) = 0$;
> (2) for every $P \in \mathbf{P}_1$, $\lim_n P(\phi(V^n) = 0) = 0$; and
> (3) for some $P \in \mathbf{P}_0 \cup \mathbf{P}_1$, $\lim_n P(\phi(V^n) = 0) = 1$ or $\lim_n P(\phi(V^n) = 1) = 1$

Pointwise consistency requires that the probability of the test making an error converges to zero, as the sample size increases without limit, no matter what the true distribution is. The last clause in the definition is a requirement of non-triviality, because one can trivially avoid error by always suspending judgments. The non-triviality requirement imposed here is a very weak one: a test is non-trivial as long as it gives a definite answer eventually for some distribution. But it suffices for our purpose.[13]

Uniform consistency is a stronger criterion:

> **Uniform Consistency**: A test ϕ of H_0 versus H_1 is uniformly consistent, if
> (1) $\lim_n \sup_{p\in\mathbf{P}0} P(\phi(V^n) = 1) = 0$;
> (2) $\lim_n \sup_{p\in\mathbf{P}1} P(\phi(V^n) = 0) = 0$; and
> (3) for some $P \in \mathbf{P}_0 \cup \mathbf{P}_1$, $\lim_n P(\phi(V^n) = 0) = 1$ or $\lim_n P(\phi(V^n) = 1) = 1$

So the difference is that uniform consistency requires that the supremum of error probabilities converges to zero. If we want to control the error probability with some big enough sample size, uniform consistency implies that there is a *single* sample size that can do the job for all possible underlying probability distributions. By contrast, pointwise consistency only implies that there is a sample size for each underlying distribution that can control the error probability to a certain level. But how big the sample size should be depends on the unknown true distribution.

Hence uniform consistency is a more useful property from the practical perspective of finite-sample inference. With uniform consistency, it is in principle possible to provide a bound on worst-case error rate at a finite sample size, whereas it is not possible with mere pointwise consistency.

To see how this is related to the difference between the PC procedure and its conservative, empirically more robust variant, consider the simplest case on which they differ, the case of deciding whether an unshielded triple is a collider or a non-collider. That is, suppose it is our background knowledge that three variables X, Y, Z form an unshielded triple in their causal graph. In other words, it is known that there is no direct causal relationship

[13] Zhang (2006a) considers a stronger and more reasonable alternative. Our definitions and lemmas here are drawn follow Zhang (2006a), except that we call pointwise consistency and uniform consistency here are referred to as *weak pointwise consistency* and weak *uniform consistency* in Zhang (2006a).

between X and Z relative to {X, Y, Z}, and there is direct causal relationship between X and Y, and between Y and Z relative to {X, Y, Z}, but we do not know the direction. Our two hypotheses are H_0: the triple is a collider, i.e., X and Z are both direct causes of Y, and H_1: not H_0, i.e., the triple is a non-collider.

Assuming the CMC and CFC hold of the true population distribution, from which samples are drawn, is there a uniformly consistent procedure to test H_0 versus H_1? The impossibility result of Robins et al. (2003) does not apply here, because the essential condition for their argument, what Zhang (2006a) calls *strong inseparability* of H_0 from H_1, does not hold in this particular case. We think that there is a uniformly consistent procedure for this simple task, and the Conservative PC procedure, with a proper schedule of changing its parameters in conditional independence tests with sample size, is such a procedure. We are working on a formal argument for this. On the other hand, there is already a formal argument to show that the PC procedure is definitely not uniformly consistent.

The argument is based on the following lemma, adapted from Zhang (2006a):

Lemma 6: There is no uniformly consistent test of H_0 versus H_1 that does not return 2 ("don't know") if $\mathbf{P_0}$ and $\mathbf{P_1}$ are inseparable in the sense that for every $\varepsilon > 0$, there are $P_0 \in \mathbf{P_0}$ and $P_1 \in \mathbf{P_1}$ such that the total variation distance[14] between P_0 and P_1 is less than ε.
Proof: See Appendix C.

In the simple case of deciding whether an unshielded triple is a collider or a non-collider, it is easy to check that P_0 and P_1 are disjoint given the CMC and CFC assumptions, which implies that a pointwise consistent procedure does not need to use the answer of "don't know" at any time. Indeed, the PC procedure is such a pointwise consistent test that always returns a definite answer. But exactly because of this feature of always being definitive, we know from Lemma 6 that PC is not uniformly consistent, because P_0 and P_1, though disjoint, are still inseparable in this case. As we have demonstrated in Section **3**, there are distributions that violate the Orientation-Faithfulness condition in that both $X \perp\!\!\!\perp Z$ and $X \perp\!\!\!\perp Z \mid Y$ hold. The CFC assumption rules out such distributions as impossible, so they are in neither P_0 nor P_1. However, in both P_0 and P_1 there are distributions arbitrary close to being unfaithful, and that is why P_0 and P_1 are inseparable.

Lemma 6 implies that in situations where P_0 and P_1 are inseparable, a uniformly consistent procedure, if any, has to be cautious at finite sample sizes and be prepared to return "don't know". A test that always decides the matter cannot be uniformly consistent, even though it might be pointwise consistent. The PC algorithm, like many other algorithms in the literature including Bayesian and likelihood-based algorithms, is not uniformly consistent in our simple case because it will always make a definite choice as to whether the unshielded triple is a collider or not.

The Conservative PC algorithm, on the other hand, is not disqualified by Lemma 6 as a candidate for achieving uniform consistency in our simple case. More generally, we

[14] In probability theory, the total variation distance between two probability measures P_0 and P_1 is defined as $\sup_E|P_0(E) - P_1(E)|$, with E ranging over all events in the algebra.

conjecture that, given the right adjacency graph, a properly tuned Conservative PC algorithm can make uniformly consistently inference of orientations, under the assumptions of CMC and CFC. Of course it cannot be in general uniformly consistent under CMC and CFC, in view of Robins et al.'s impossibility theorem. However, some version of it may be uniformly consistent under the strong-faithfulness assumptions given in Zhang and Spirtes (2003). All these await further investigations.

6. Conclusion

We have examined the controversial Causal Faithfulness Condition from a perspective, different than and supplementary to the standard one in the philosophical literature. Our perspective is the empiricists' favorite: testing. It is evident that the condition specifies a relationship between the probability distribution of a set of random variables and the underlying causal structure, and hence in general is not testable without knowing the causal structure in the first place. But, as we hope to have shown, this is far from the end of the story. The condition has a testable consequence to exploit, given what we call the causal Bayes net assumption (which gave us the Causal Markov Condition and the Minimality condition).

The testable consequence is that the probability distribution is faithful to *some* causal structure. This suggests a distinction between detectable violations of the Causal Faithfulness Condition and undetectable ones. We have in this paper provided some general characterization of this distinction.

One reason we chose the testing perspective is that the testability results, besides their theoretical interest, may have implications for actual methodology. Indeed the theorem presented in Section 3 results from a close examination of the role the Causal Faithfulness Condition plays in justifying causal inference methods, and, in turn, makes constructive recommendations to improve the existing methods. There are both theoretical reasons, as argued in Section 5, and strong empirical evidence, as reported in Ramsey et al. (2006), for believing that the improvement is significant. We hope that the more general result presented in Section 4 will bear similar practical fruits.

Our testability results are based on the assumption that every variable in **V** is observed, so that in principle we have access to the joint distribution over **V**. When, as a more realistic scenario, some variables in **V** are not observed, we can only rely on data from the marginal distribution over a proper subset of **V**, and the testability results have to be altered. An extension of the DAG machinery, known as the *ancestral graphical models,* has been developed in the statistics literature to can represent situations where some variables in **V** are unobserved, or phrased in another way, the set of observed variables is not causally sufficient (Richardson and Spirtes 2002), and causal inference procedures analogous to PC have also been designed (Spirtes et al. 2000, Zhang 2006b). It is quite conceivable that our work in this paper can be extended to cover those situations. We have some preliminary results on that, but reporting them will make the current paper more complicated.

We have also assumed the Causal Markov Condition throughout the paper, as a consequence of the more general causal Bayes net assumption. As we argued in Section 2, the causal Bayes net assumption, and hence the CMC, is crucial for the usefulness of causal structure in solving the kind of inductive problems posed at the beginning. But this of course does not help the committed skeptics. If we do not assume the CMC, we can only detect in principle the failure of the conjunction of the CMC and CFC, and the familiar Duhemian problem surfaces. We take comfort in the thought that one cannot really test or discover anything without making some assumptions.

Appendix A Basic Graph-Theoretical Notions

In this Appendix, we provide definitions of the graphical theoretical notions we used, in particular, the definition of active or d-connecting path and that of d-separation.

A ***directed graph*** is a mathematical object consisting of a pair <**V, E**>, where **V** is a set of vertices and **E** is a set of arrows. An arrow is an ordered pair of vertices, <*X*, *Y*>, represented visually by $X \rightarrow Y$. Given a graph G(**V, E)**, if <*X*, *Y*>$\in$ **E**, then *X* and *Y* are said to be ***adjacent***, and *X* is called a ***parent*** of *Y*, and *Y* a ***child*** of *X*. We usually denote the set of *X*'s parents in *G* by $\mathbf{PA}_G(X)$. A ***path*** in G is a sequence of distinct vertices <V_1, ..., V_n>, such that for $0\leq i \leq n\text{-}1$, V_i and V_{i+1} are adjacent in G. A ***directed path*** in G from *X* to *Y* is a sequence of distinct vertices <V_1,...,V_n>, such that $V_1=X$, $V_n=Y$ and for $0\leq i \leq n\text{-}1$, V_i is a parent of V_{i+1} in G, i.e., all arrows on the path point in the same direction. *X* is called an ***ancestor*** of *Y*, and *Y* a ***descendant*** of *X* if $X=Y$ or there is a directed path from *X* to *Y*. *X* is called a ***proper ancestor*** of *Y*, and *Y* a **proper descendant** of *X* if $X \neq Y$ and *X* is an ancestor of *Y*. ***Directed acyclic graphs (DAGs)*** are simply directed graphs in which there are no directed cycles, or in other words, there are no two distinct vertices in the graph that are ancestors of each other.

Given two directed graphs *G* and *H* over the same set of variables **V**, *G* is called a ***(proper) subgraph*** of *H*, and *H* a ***(proper) supergraph*** of *G* if the set of arrows of *G* is a (proper) subset of the set of arrows of *H*. An arrow $X \rightarrow Y$ in *G* is ***covered*** if $\mathbf{PA}_G(Y) = \mathbf{PA}_G(X) \cup \{X\}$.

Given a path *p* in a DAG, a non-endpoint vertex *V* on *p* is called a ***collider*** if the two edges incident to *V* on *p* are both into *V* (i.e., $\rightarrow V \leftarrow V$), otherwise *V* is called a ***non-collider***. Here are the key definitions:

Active Path: In a directed graph, a path *p* between vertices *A* and *B* is ***active*** (or ***d-connecting***) relative to a set of vertices **Z** ($A,B \notin \mathbf{Z}$) if
(i) every non-collider on *p* is not a member of **Z**; and
(ii) every collider on *p* is an ancestor of some member of **Z**.

D-separation: *A* and *B* are said to be ***d-separated*** by **Z** if there is no active path between *A* and *B* relative to **Z**. Two disjoint sets of variables **A** and **B** are d-separated by **Z** if every vertex in **A** and every vertex in **B** are d-separated by **Z.**

Appendix B PC and Conservative PC

The PC algorithm (Spirtes et al. 2000) is probably the best known representative of what is called constraint-based causal discovery algorithms. It is reproduced here, in which **ADJ**(G, *X*) denotes the set of nodes adjacent to *X* in a graph *G*:

PC Algorithm

```
[S1]  Form the complete undirected graph U on the set of variables V;

[S2]  n=0
      repeat
          For each pair of variables X and Y that are adjacent in (the current) U such
          that ADJ(U, X) \ {Y} or ADJ(U, Y) \ {X} has at least n elements, check
          through the subsets of ADJ(U, X) \ {Y} and the subsets of ADJ(U, Y) \ {X}
          that have exactly n variables. If a subset S is found conditional on which X
          and Y are independent, remove the edge between X and Y in U, and record S
          as Sepset(X, Y);
          n = n+1;
     until for each ordered pair of adjacent variables X and Y in U, $ ADJ(U, X) \ {Y}
     has less than $n$ elements.

[S3]  Let P be the graph resulting from step [S2]. For each unshielded triple <A, B, C> in
      P, orient it as A → B ← C if and only if B is not in Sepset(A,C).

[S4]  Execute the following orientation rules until none of them applies:
      (a) If A → B — C, and A, C are not adjacent, orient as B → C.
      (b) If A → B → C and A — C orient as A → C.
      (c) If A → B ← C, A —  D — C, B — D, and A, C are not adjacent,
          orient B — D as B ← D.
```

In the PC algorithm, [S2] constitutes the adjacency stage; [S3] and [S4] constitute the orientation stage. In [S2], the PC algorithm essentially searches for a conditioning set for each pair of variables that renders them independent. What distinguishes the PC algorithm from other constraint-based algorithms is the way it performs search. As we can see, two tricks are employed: (1) it starts with the conditioning set of size 0 (i.e., the empty set) and gradually increases the size of the conditioning set; and (2) it confines the search of a screen-off conditioning set for two variables within the potential parents -- i.e., the currently adjacent nodes -- of the two variables, and thus systematically narrows down the space of possible screen-off sets as the search goes on. These two tricks increase both computational and statistical efficiency in most real cases.

In [S3], the PC algorithm uses a very simple criterion to identify unshielded colliders or non-colliders. [S4] consists of orientation propagation rules based on information about non-colliders obtained in S3 and the assumption of acyclicity. These rules are shown to be both sound and complete in Meek (1995a).

The Conservative PC (CPC) algorithm, replaces [S3] in PC with the following [S3'], and otherwise remains the same.

CPC Algorithm

[S1’]: Same as [S1] in PC.

[S2’]: Same as [S2] in PC.

[S3'] Let P be the graph resulting from step [S2’]. For each unshielded triple <*A*, *B*, *C*> in P, check all subsets of *A*'s potential parents (vertices that are adjacent to *A* but are not *A*’s children) and of *C*'s potential parents:

(a) If *B* is NOT in any such set conditional on which *A* and *C* are independent, orient the triple as a collider: $A \rightarrow B \leftarrow C$;

(b) If *B* is NOT in all such set conditional on which *A* and *C* are independent, leave the triple as it is, i.e., a non-collider;

(c) Otherwise, mark the triple as “ambiguous” (or “don’t know”) by an underline.

[S4’] Same as [S4] in PC. (Of course a triple marked ``ambiguous" does not count as a non-collider in [S4](a) and [S4](c).)

Proposition (**Correctness of CPC**): Under the CMC and Adjacency-Faithfulness assumptions, the CPC algorithm is asymptotically correct in the sense that given a perfect conditional independence oracle, the algorithm returns a graphical object such that (1) it has the same adjacencies as the true causal graph does; and (2) all arrowheads and unshielded non-colliders in it are also in the true graph.

Proof: Suppose the true causal graph is G, and all conditional independence judgments are correct. The Markov and Adjacency-Faithfulness assumptions imply that the undirected graph P resulting from step [S2’] has the same adjacencies as G does (Spirtes et al. 2000). Now consider [S3']. If [S3'](a) obtains, then $A \rightarrow B \leftarrow C$ must be a subgraph of G, because otherwise by the CMC, either *A*'s parents or *C*'s parents d-separate *A* and *C*, which means that there is a subset **S** of either *A*'s potential parents or *C*'s potential parents containing *B* such that $A \perp C \mid \mathbf{S}$, contradicting the antecedent in [S3'](a). If [S3'](b) obtains, then $A \rightarrow B \leftarrow C$ cannot be a subgraph of G (and hence the triple must be an unshielded non-collider), because otherwise by the Markov assumption, there is a subset **S** of either *A*'s potential parents or *C*'s potential parents not containing *B* such that $A \perp C \mid \mathbf{S}$, contradicting the antecedent in [S3'](b). So neither [S3'](a) nor [S3'](b) will introduce an orientation error. Trivially [S3'](c) does not produce an orientation error, and it has been proven (in e.g., Meek 1995a) that [S4’] will not produce any, which completes the proof. Q.E.D.

Appendix C Omitted Proofs

Lemma 1: All causal Bayes nets over **V** satisfy CCBN.

Proof: Suppose, for sake of contradiction, that G is a causal Bayes net over **V** but does not satisfy CCBN. This means that there are two variables $X, Y \in \mathbf{V}$ such that X is not a parent of Y, but there exists $x_1 \neq x_2$ and **z** such that $P(Y \parallel X{:=}x_1, \mathbf{Z}{:=}\mathbf{z}) \neq P(Y \parallel X{:=}x_2, \mathbf{Z}{:=}z)$, where $\mathbf{Z} = \mathbf{V} \setminus \{X, Y\}$. However, since G is a causal Bayes net over **V**, it satisfies the manipulation principle, and hence

$$P(Y \parallel X{:=}x_1, \mathbf{Z}{:=}\mathbf{z}) = P(Y \parallel X{:=}x_2, \mathbf{Z}{:=}\mathbf{z}) = P(Y|\mathbf{PA}_{\mathrm{G}}(Y))$$

Because $X \notin \mathbf{PA}_{\mathrm{G}}(Y)$, there is a contradiction. Therefore, all causal Bayes nets over **V** satisfy CCBN. Q.E.D.

Lemma 2: If **V** satisfies the causal Bayes net assumption, then G_{d} is a correct representation of the causal structure of **V** in the sense that it satisfies the manipulation principle.

Proof: Since **V** satisfies the causal Bayes net assumption, there is a DAG over **V** that satisfies the manipulation principle. Let G be a minimal such DAG, that is, be a DAG that satisfies the manipulation principle such that no proper subgraph of G satisfies the manipulation principle. By Lemma 1, G satisfies CCBN, and it follows that G is a supergraph of G_{d}. We now show that $G= G_{\mathrm{d}}$. Suppose not. Then G is a proper supergraph of G_{d}. So there is an edge $X \rightarrow Y$ in G which is not in G_{d}. Let G' be the same graph as G except that the edge $X \rightarrow Y$ is removed. We claim that G' also satisfies the manipulation principle. To show this, it suffices to show that $P(Y \mid \mathbf{PA}_G(Y)) = P(Y \mid \mathbf{PA}_{G'}(Y), X) = P(Y \mid \mathbf{PA}_{G'}(Y))$, because other vertices have the exact same parents in G' as they do in G. But we know that for any value pa for $\mathbf{PA}_{G'}(X))$, and any $x_1 \neq x_2$,

$$\begin{aligned} P(Y \mid \mathbf{PA}_{G'}(Y) = \mathrm{pa}, X=x_1) &= P(Y \parallel \mathbf{PA}_{G'}(Y) := \mathrm{pa}, X{:=}x_1, \mathbf{Z}{:=}\mathbf{z}) \\ &= P(Y \parallel \mathbf{PA}_{G'}(Y) := \mathrm{pa}, X{:=}x_2, \mathbf{Z}{:=}\mathbf{z}) = P(Y \mid \mathbf{PA}_{G'}(Y) = \mathrm{pa}, X=x_2) \end{aligned}$$

where $\mathbf{Z} = \mathbf{V} \setminus (\{X, Y\} \cup \mathbf{PA}_{G'}(Y))$, and **z** is any value for **Z**. This is true because otherwise X would be a parent of Y in G_{d}. It follows that $P(Y \mid \mathbf{PA}_G(Y)) = P(Y \mid \mathbf{PA}_{G'}(Y))$, which implies that G' also satisfies the manipulation principle. But G' is a proper subgraph of G, a contradiction. So $G= G_{\mathrm{d}}$. Q.E.D.

Lemma 3: No proper subgraph of G_{d} satisfies the Markov condition with P(**V**).

Proof: Suppose, for sake of contradiction, G is a proper subgraph of G_{d}, and also satisfies the Markov condition with P(**V**). We claim that G also satisfies the manipulation principle. Let X be any variable in **V**. Because G is a subgraph of G_{d}, $\mathbf{PA}_G(X) \subseteq \mathbf{PA}_{G\mathrm{d}}(X)$. Let $\mathbf{R} = \mathbf{PA}_{G\mathrm{d}}(X) \setminus \mathbf{PA}_G(X)$. If **R** is empty, then trivially $P(X| \mathbf{PA}_G(X) = P(X \mid \mathbf{PA}_{G\mathrm{d}}(X))$. If **R** is not empty, every variable in **R** is a non-descendant of X in G, because it is a parent of X in G_d, an acyclic graph. Since G by supposition satisfies the Markov condition with P(**V**), we have $X \perp\!\!\!\perp \mathbf{R} \mid \mathbf{PA}_G(X)$, which implies that $P(X| \mathbf{PA}_G(X) = P(X \mid \mathbf{PA}_{Gd}(X))$. Since this is true for every vertex in **V**, and G_d satisfies the manipulation principle, it is easy to see that G also satisfies the manipulation principle. But G is a proper subgraph of G_d, and hence does not satisfy CCBN, which contradicts Lemma 1. Q.E.D.

Theorem 2: Under the assumptions of CMC and Minimality, if the CFC fails and the failure is undetectable, then the Triangle-Faithfulness condition fails.

Proof: Let P is the true probability distribution of **V**, and *G* is the true causal DAG. By assumption, P is not faithful to *G*, but the unfaithfulness is undetectable, which means that P is faithful to some DAG *H*. But P is Markov to *G*, so *G* entails strictly fewer conditional independence relations than *H* does. It follows that the adjacencies in G form a proper superset of adjacencies in *H*. But *H* is not a proper subgraph of *G*, for otherwise the Minimality condition fails.

Let *G'* be the subgraph of *G* with the same adjacencies as *H*. G' and *H* are not Markov equivalent because otherwise minimality would be violated for *G*. So *G'* has an unshielded collider X → Y ← Z where *H* has unshielded non-collider X – Y – Z, or vice-versa. Suppose the former. Since the distribution is Markov and faithful to *H*, all independencies between X and Z are conditional on subsets containing Y, and there is an independence between X and Z conditional on some subset containing Y. If *G* does not contain an edge between X and Z, then *G* entails that X and Z are independent conditional on some set not containing Y – but there is no such conditional independence, and hence P would not be Markov to *G*. So *G* contains an edge between X and Z, and the triangle faithfulness is violated. The case where *G'* contains an unshielded non-collider where *H* has an unshielded collider is similar. Q.E.D.

Lemma 6: There is no uniformly consistent test of H_0 versus H_1 that does not return 2 ("don't know") if $\mathbf{P_0}$ and $\mathbf{P_1}$ are inseparable in the sense that for every $\varepsilon > 0$, there are $P_0 \in \mathbf{P_0}$ and $P_1 \in \mathbf{P_1}$ such that the total variation distance between P_0 and P_1 is less than ε.

Proof: Suppose for sake of contradiction that there is a uniformly consistent test ϕ that does not return 2. Let $\varepsilon > 0$ be any positive real number. By assumption we can choose $P_0 \in \mathbf{P_0}$ and $P_1 \in \mathbf{P_1}$ such that the total variation distance between P_0 and P_1 is less than ε. Since ϕ does not return 2, it follows that

$$\begin{aligned}\sup_{p\in\mathbf{P1}} P(\phi(V^n) = 0) &\geq P_1(\phi(V^n) = 0)\\ &\geq P_0(\phi(V^n) = 0) - \varepsilon\\ &= 1- P_0(\phi(V^n) = 1) - \varepsilon\\ &\geq 1 - \sup_{p\in\mathbf{P0}} P(\phi(V^n) = 1) - \varepsilon\end{aligned}$$

However, by clause (i) in the definition of uniform consistency

$$\lim_n \sup_{p\in\mathbf{P0}} P(\phi(V^n) = 1) = 0$$

which implies that

$$\lim_n \sup_{p\in\mathbf{P}1} P(\phi(V^n) = 0) \geq 1 - \lim_n \sup_{p\in\mathbf{P}0} P(\phi(V^n) = 1) - \varepsilon = 1 - \varepsilon$$

Since this is true for any ε > 0, it follows that

$$\lim_n \sup_{p\in\mathbf{P}1} P(\phi(V^n) = 0) = 1$$

which contradicts clause (ii) in the definition of uniform consistency. Q.E.D.

References

Artzenius, F. (1992) The Common Cause Principle. *PSA Procceding*, Eds. D. Hull and K. Okruhlik, Vol.2, East Lansing, MI: PSA, pp. 227-37.

Becker, A., D. Geiger, and C. Meek (2000) Perfect Tree-like Markovian Distributions. *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence,* 19-23 Morgan Kaufmann.

Cartwright, N. (1999) *The Dappled World*. Cambridge: Cambridge University Press.

Cartwright, N. (2001). What is Wrong with Bayes Nets? *The Monist, pp* 242-264

Chickering, D.M. (1995) A transformational characterization of equivalent Bayesian network structures. *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence,* 87-98. Morgan Kaufmann.

Chickering, D.M. (2002) Optimal Structure Identification with Greedy Search. *Journal of Machine Learning Research,* 3:507-54.

Cooper, G. (1999) An Overview of the Representation and Discovery of Causal Relationships Using Bayesian Networks, in *Computation, Causation, and Discovery*, Eds. C. Glymour and G.F. Cooper. Cambridge, MA: MIT Press, 3-62.

Dawid, P. (1979) Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society,* Series B 41: 1-31.

Dawid, P. (2002) Influence Diagrams for Causal Modelling and Inference. *International Statistical Review* 70: 161-189.

Glymour, C. (forthcoming)

Hajek, A. (2003) What Conditional Probability Could Not Be. *Synthese,*137, pp: 273-323

Hausman, D. M., and J. Woodward (1999) Independence, Invariance and the Causal Markov Condition. *British Journal for the Philosophy of Science* 50, pp. 521-83.

Hausman, D. M., and J. Woodward (2004) Manipulation and Causal Markov Condition. *Philosophy of Science* 71: 846-856.

Hesslow, G. (1976) Two Notes on the Probabilistic Approach to Causality. *Philosophy of Science* 43: 290-92.

Hitchcock, C. (2001a) The intransitivity of Causation Revealed in Equations and Graphs. *Journal of Philosophy*, 98, pp. 273-299.

Hitckcock, C. (2001b) A Tale of Two Effects. *Philosophical Review* 110: 361-96.

Hoover, K.D. (2001) *Causality in Macroeconomics*. Cambridge: Cambridge University Press.

McDermott, M. (1995) Redundant Causation. *British Journal for the Philosophy of Science*, 40, pp. 523-544.

Meek, C. (1995a) Causal Inference and Causal Explanation with Background Knowledge. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence,* 403-411. Morgan Kaufmann.

Meek, C. (1995b) Strong Completeness and Faithfulness in Bayesian Networks, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 411-418. San Francisco: Morgan Kaufmann.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligence Systems.* San Mateo, California: Morgan Kaufmann.

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

Ramsey, J., P. Spirtes, and J. Zhang (2006) Adjacency-Faithfulness and Conservative Causal Inference. *Proceedings of 22nd Conference on Uncertainty in Artificial Intelligence*, 401-408, Oregon, AUAI Press.

Richardson, T., and P. Spirtes (2002) Ancestral Markov Graphical Models. *Annals of Statistics* 30(4): 962-1030.

Robins, J. (1986) A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods -Applications to Control of the Healthy Worker Survivor Effect. *Mathematical Modeling* 7: 1393-1512.

Robins, J.M., Scheines, R., Spirtes, P., Wasserman, L. (2003) Uniform Consistency in Causal Inference. *Biometrika* 90(3):491-515.

Sober, E. (1987) The Principle of the Common Cause, in *Probability and Causation: Essays in Honor of Wesley Salmon*, Eds. J. Fetzer, Dordrecht:Redel, pp. 211-28.

Spanos, A. (2006) Revisiting the Omitted Variables Argument: Substantive vs. Statistical Adequacy, forthcoming in *Journal of Economic Methodology*.

Spirtes, P., C. Glymour, and R. Scheines (2000) *Causation,Prediction and Search.* New York: Springer-Verlag. (2000, 2nd ed.) Cambridge, MA: MIT Press.

Spohn, W. (2000) Bayesian Nets Are All There Is To Causal Dependence, in *Stochastic Dependence and Causality*, Eds. M.C. Galavotti et al. (eds.), CSLI Publications, pp. 157-172.

Steel, D. (2006) Homogeneity, Selection, and the Faithfulness Condition. *Minds and Machines,* 16: 303-17

Strotz, R.H., and H.A. Wold (1960) Recursive versus NonrecursiveSystems: An Attempt at Synthesis. *Econometrica* 28: 417-427

Tian, J., and J. Pearl (2002) A New Characterization of the Experimental Implications of Causal Bayesian Networks, in *Proceddings of the National Conference on Artificial Intelligence (AAAI)*

Verma, T., and J. Pearl (1990) Equivalence and Synthesis of Causal Models. *Proceedings of 6th Conference on Uncertainty in Artificial Intelligence*, 220-227.

Woodward, J. (1998) Causal Independence and Faithfulness. *Multivariate Behavioral Research* 33: 129-48.

Woodward, J (2003) *Making Things Happen: A Theory of Causal Explanation*. Oxford and New York: Oxford University Press.

Zhang, J. (2006a) Underdetermination of Causal Hypotheses by Statistical Data. Technical Report, Department of Philosophy, Carnegie Mellon University.

Zhang, J. (2006b) Causal Inference and Reasoning in Causally Insufficient Systems. PhD dissertation, Department of Philosophy, Carnegie Mellon University, available at www.hss.caltech.edu/jiji/dissertation.pdf.

Zhang, J. and P. Spirtes (2003) Strong Faithfulness and Uniform Consistency in Causal Inference. *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*. 632-639. Morgan Kaufmann.