Strong Faithfulness and Uniform Consistency in Causal Inference

Jiji Zhang & Peter Spirtes
Philosophy Department
Carnegie Mellon University
Pittsburgh, PA 15213
{jiji, ps7z}@andrew.cmu.edu

Abstract

A fundamental question in causal inference is whether it is possible to reliably infer the manipulation effects from observational data. There are a variety of senses of asymptotic reliability in the statistical literature, among which the most commonly discussed frequentist notions are pointwise consistency and uniform consistency (see, e.g. Bickel, Doksum [2001]). Uniform consistency is in general preferred to pointwise consistency because the former allows us to control the worst case error bounds with a finite sample size. In the sense of pointwise consistency, several reliable causal inference algorithms have been established under the Markov and Faithfulness assumptions [Pearl 2000, Spirtes et al. 2001]. In the sense of uniform consistency, however, reliable causal inference is impossible under the two assumptions when time order is unknown and/or latent confounders are present [Robins et al. 2000]. In this paper we present two natural generalizations of the Faithfulness assumption in the context of structural equation models, under which we show that the typical algorithms in the literature are uniformly consistent with or without modifications even when the time order is unknown. We also discuss the situation where latent confounders may be present and the sense in which the Faithfulness assumption is a limiting case of the stronger assumptions.

1 INTRODUCTION

1.1 CAUSAL INFERENCE

We consider the kind of causal inference in the literature that intends to predict the manipulation effects based on non-experimental data [Pearl 2000, Spirtes et al. 2001]. Such inference typically involves two steps: discovery of causal structures, represented by directed acyclic graphs (DAGs)[1], and identification of causal parameters. There are two main approaches in the causal discovery step: constraint-based approach and Bayesian approach [Heckerman et al. 1995], of which we focus on the former as we are going to discuss the frequentist notions of consistency. The basic idea of the constraint-based approach is to test the conditional independence relations among the observed variables, which, under certain assumptions, put some graphical constraints on the possible causal structures. The two commonly adopted assumptions are the Markov and Faithfulness assumptions. The **Markov assumption** says that every variable is independent of its non-effects conditional on its direct causes, which is just the (local) Markov property of DAGs. The **Faithfulness assumption** says that no conditional independence relations other than the ones entailed by the Markov assumption are present in the population distribution. Since the conditional independence relations entailed by the Markov assumption correspond exactly to d-separation [Pearl 1988], these two assumptions together translate the conditional independence relations in the population distribution to the d-separation constraints on the possible causal graphs.

Hence the output of the constraint-based algorithms is typically a set of causal graphs compatible with background knowledge that share the same d-separation features (or say, entail the same conditional independence relations) among the observed variables, which is usually called an O-Markov equivalence class. A causal quantity is identifiable with respect to a set of

[1] In general directed (cyclic) graphs can be used to represent causal systems that might have feedback. There are (constraint-based) algorithms of discovering causal graphs with (possibly) cycles assuming no latent confounders [Richardson 1996], of which the results in this paper still hold.

causal graphs if given any graph in the set, the causal quantity can be written uniquely in terms of some estimable statistical quantities. No conclusion can be drawn concerning an unidentifiable causal parameter. When a causal parameter is identified, the inference of this parameter is just the ordinary statistical inference.

In this paper we confine the discussion to one of the most commonly used parameterizations of causal graphs: linear structural equation models (LSEMs), in which the structural parameters can be easily interpreted as the direct manipulation effects. Specifically, given a DAG, each arrow is assigned a coefficient so that every variable can be written as a linear function of its parents plus a Gaussian error[2]. Usually the variables are standardized for the sake of interpreting the structural coefficients. Without loss of generality, we consider standardized LSEMs in what follows.

Finally, there is an important assumption called causal sufficiency that can dramatically simplify causal discovery. A causal system (i.e. a set of observed variables) is causally sufficient if no common cause of any two variables in the system is left out. In more plain words, causal sufficiency assumes that there are no latent confounders of any two observed variables. For simplicity we will present our main results under this assumption, but it is not essential as we will point out later.

1.2 CONSISTENCY OF TESTS

Consistency is a property that corresponds to asymptotic reliability (in the sense of avoiding error). Two notions of consistency are often discussed in the statistical literature — pointwise consistency and uniform consistency (see e.g. Bickel, Doksum [2001]). Here is a generic formal setting that facilitates the formal definitions in the context of causal inference[3]. $\mathbf{O}$ is the set of observed random variables, of which $O^n = \{O_1, ..., O_n\}$ denotes an i.i.d sample. $\mathcal{G}$ is the set of possible causal graphs (defined by our background knowledge) over O and possibly some other latent variables if causal sufficiency is not assumed. Given any $G \in \mathcal{G}$, $\Omega(G)$ is the set of distributions that are legitimate given G according to the assumptions. For example, given the Markov and faithfulness assumptions, $\Omega(G)$ is the set of distributions that are Markov and faithful to G. We use $\Omega_{\mathcal{G}}$ to denote the union of $\Omega(G)$'s: $\Omega_{\mathcal{G}} = \cup_{G\in\mathcal{G}}\Omega(G)$. Finally, for every distribution P, $\tilde{P}$ denotes the marginal distribution of O obtained from P.

A test ϕ is a sequence of functions $(\phi_1, \phi_2, ..., \phi_n, ...)$, where each ϕ_i takes data O^i and returns 0, 1, or 2, representing "acceptance", "rejection" or "no conclusion", respectively. Let θ be any causal parameter of interest, which is in general a functional of the probability distribution P and the causal structure G: $\theta = T(P, G)$. With respect to the null hypothesis $H_0 : \theta = \theta_0$ versus the alternative hypothesis $H_1 : \theta \neq \theta_0$, we define

$$\begin{aligned}\Omega_{\mathcal{G}0} &= \{P : \exists G \in \mathcal{G}(P \in \Omega(G) \wedge T(P,G) = \theta_0)\}\\ \Omega_{\mathcal{G}1} &= \{P : \exists G \in \mathcal{G}(P \in \Omega(G) \wedge T(P,G) \neq \theta_0)\}\end{aligned}$$

Intuitively $\Omega_{\mathcal{G}i}$ is the set of distributions that are compatible with H_i, $i = 0, 1$. Usually $\Omega_{\mathcal{G}0}$ and $\Omega_{\mathcal{G}1}$ are not disjoint when the time order between variables is unknown. The distributions in their intersection obviously underdetermine the truth value of the null hypothesis, for which we have to return "no conclusion". Even when they are disjoint, It could also occur, in the presence of latent variables, that a distribution in $\Omega_{\mathcal{G}0}$ shares the same marginal distribution over the observed variables with a distribution in $\Omega_{\mathcal{G}1}$, in which case the hypothesis is still underdetermined by what we can observe. The inclusion of "no conclusion" in the outputs of tests respects the fact that there may exist $P_0 \in \Omega_{\mathcal{G}0}, P_1 \in \Omega_{\mathcal{G}1}$ such that $\tilde{P}_0 = \tilde{P}_1$.

Let P^n denote the n-fold product measure corresponding to P, here are the key definitions:

Definition 1 (pointwise consistency) *A test ϕ is pointwise consistent if*
(i) for every $P \in \Omega_{\mathcal{G}0}$, $\lim_n P^n(\phi_n(O^n) = 1) = 0$ and
(ii) for every $P \in \Omega_{\mathcal{G}1}$, $\lim_n P^n(\phi_n(O^n) = 0) = 0$

Definition 2 (uniform consistency) *A test ϕ is uniformly consistent if*
(i) $\lim_n \sup_{P\in\Omega_{\mathcal{G}0}} P^n(\phi_n(O^n) = 1) = 0$ and
(ii) for every $\delta > 0$, $\lim_n \sup_{P\in\Omega_{\mathcal{G}1\delta}} P^n(\phi_n(O^n) = 0) = 0$
where

$$\Omega_{\mathcal{G}1\delta} = \{P : \exists G \in \mathcal{G}(P \in \Omega(G) \wedge |T(P,G) - \theta_0| \geq \delta)\}$$

It should be clear from the definition that uniform consistency (but not pointwise consistency) allows us to simultaneously control the worst case type I error and type II error with finite sample size (given that the true parameter value is bounded away by a constant from the null value). The error bounds for a merely pointwise consistent procedure depend on the value of θ, which is exactly what we want to figure out in the first place.

An obviously uniformly consistent procedure is to always return 2 in the limit. We exclude such uninfor-

[2] We assume the errors are uncorrelated. Correlated Gaussian errors to a large extent can be dealt with by introducing new latent variables [Spirtes et al. 1996].

[3] Our notations and the subsequent definitions largely follow Robins et al. [2000].

mative tests by considering only non-trivial ones in the following sense:

Definition 3 (non-triviality) *A test ϕ is non-trivial if for some $P \in \Omega_{\mathcal{G}0} \cup \Omega_{\mathcal{G}1}$,*

$$\lim_n P^n(\phi_n(O^n) = 0) = 1 \quad or \quad \lim_n P^n(\phi_n(O^n) = 1) = 1$$

2 UNIFORM CONSISTENCY WITH MORE FAITHFULNESS

We assume causal sufficiency in this section, and discuss the situation without the assumption in the next section.

2.1 A CANONICAL CASE

Consider a canonical case of (constraint-based) causal inference. In Figure 1, all variables are observed, i.e. $\mathbf{O} = \{X_1, X_2, X_3, X_4\}$, and there are no latent variables. Obviously without further background information $\mathcal{G}$ includes all possible DAGs over O. In particular, $G_1, G_2 \in \mathcal{G}$. In the context of standardized structural equation models, X_i's are marginally standard Gaussian variables, and the structural equation models associated with, for example, G_1 and G_2 are M_1 and M_2, respectively, as follows:

$M1$	$M2$
$X_1 = \epsilon_1$	$X_1 = \epsilon_1$
$X_2 = \epsilon_2$	$X_2 = \epsilon_2$
$X_3 = \alpha X_1 + \beta X_2 + \epsilon_3$	$X_3 = fX_1 + gX_2 + hX_4 + \epsilon_3$
$X_4 = \gamma X_3 + \epsilon_4$	$X_4 = mX_1 + nX_2 + \epsilon_4$

where all error terms are uncorrelated Gaussians with zero means. Note that the linear coefficients in the models are naturally interpreted as the direct causal (manipulation) effects of one variable on the other. For example, in M_1 if we manipulate X_3 by one unit (actually one standard deviation in the unstandardized situation) without affecting other variables unless via X_3, the expectation of X_4 will change by γ units. So γ quantifies the direct manipulation effect of X_3 on X_4.

Suppose the parameter of interest, θ, is the direct manipulation effect of X_3 on X_4. In G_1, for example, $\theta = \gamma$, while in G_2 $\theta = 0$. It is easy to verify that the graph G_1 is the only structure that can faithfully generate the distributions such that $X_1 \perp\!\!\!\perp X_2$ and $X_4 \perp\!\!\!\perp \{X_1, X_2\} | X_3$ and no other conditional independence relations hold. In other words, under the Faithfulness assumption, if $X_1 \perp\!\!\!\perp X_2$ and $X_4 \perp\!\!\!\perp \{X_1, X_2\} | X_3$ and no other conditional independence relations hold, G_1 is the true causal structure, in which θ is obviously identified (with the correlation between X_3 and X_4). It follows that there are non-trivial pointwise consistent tests for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ given pointwise consistent tests for correlations and partial correlations. The tests return informative answers in the large sample limit for the distributions faithful to G_1.

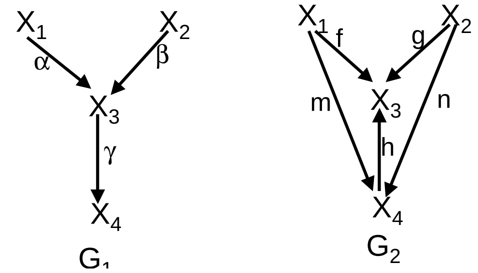

Figure 1: A Canonical Case

However, under the Markov and Faithfulness assumptions, there are no non-trivial uniformly consistent tests of $H_0 : \theta = \theta_0$ for $\theta_0 \neq 0$ [Robins et al. 2000], though we do have uniformly consistent tests for correlations and partial correlations[4]. The proof turns on the fact that for any distribution P compatible with G_1 with $\gamma = \theta_0$, there is a distribution Q compatible with G_2 such that such that Q is arbitrarily close to P, and vice versa.

Below we present two (families of) naturally strengthened versions of Faithfulness — which we call k-Constraint assumption and λ-Strong-Faithfulness assumption respectively — and investigate the consistency property of causal inference in the canonical case and in general under the stronger assumptions. These two families of assumptions are interesting in that they naturally generalize and on the boundary collapse into the usual Faithfulness assumption.

2.2 k-CONSTRAINT

Given a causal graph G, let Ψ_G be the set of linear structural coefficients for G that imply covariance matrices faithful to G. For any (small) fixed positive constant k, we define k-Constraint as below:

Definition 4 (k-Constraint) *The k-Constraint is a subset Ψ_G denoted by Ψ_G^k such that for every $\mu \in \Psi_G$,*

$$\mu \in \Psi_G^k \iff \forall a, b \in \mathbf{O}, C \subseteq \mathbf{O} \backslash \{a, b\} (|\rho_{ab.C}| \geq k|\mu_{ab}|)$$

where $\rho_{ab.C}$ is the partial correlation between a and b given C (or correlation when $C = \emptyset$) in the distribution generated by (μ, G), and μ_{ab} is the arrow coefficient, if any, between a and b.

[4] For multivariate Gaussian distributions, the test of correlation based on Fisher's Z-transformation, for example, is uniformly consistent. There is also a very nice relationship between the inference of partial correlation and the inference of correlation. See, e.g. Anderson [1958]

The k**-constraint assumption** says that for every $G \in \mathcal{G}$, if the true structure is G, then the true structural parameter is in Ψ^k_G.

Note that under the Gaussian parameterizations, the usual Faithfulness assumption entails that for any two variables a and b, if $\rho_{ab.C} = 0$ for some $C \subseteq \mathbf{O}\backslash\{a, b\}$, there cannot be an arrow between a and b, and hence there is no direct causal effect between a and b. In short, vanishing (partial) correlations indicate no (direct) effects. The k-Constraint assumption assumes furthermore that *small* (partial) correlations indicate *small* (direct) causal effects. In this sense it is a natural strengthening of the Faithfulness assumption. In practice, it is not uncommon among social scientists to both interpret the regression coefficients causally and delete insignificant variables based on t test. A charitable interpretation is that they implicitly adopt (something like) the k-Constraint assumption: small correlation means small effect.

Under the Markov and k-Constraint assumptions (no matter how small k is), we can construct non-trivial uniformly consistent procedures to test $H_0 : \theta = \theta_0$ in the canonical case by modifying the typical constraint-based algorithms slightly. We will describe the modification after we introduce the λ-Strong-Faithfulness[5], where more intuition can be gained. To see why the existing algorithms in the literature have to be modified to be uniformly consistent, here we present a negative result that may give a hint. Suppose, in the canonical case, the background knowledge is sufficient for us to conclude that the true causal graph is either G_1 or G_2, i.e. $\mathcal{G} = \{G_1, G_2\}$. Under this circumstance, $U = \Omega_{\mathcal{G}0} \cap \Omega_{\mathcal{G}1} = \emptyset$, that is, there is no issue of underdetermination. Intuitively we do not need the answer of "no conclusion" at all. However, there are no uniformly consistent tests that do not return "no conclusion" under the Markov and k-Constraint assumptions, which is a direct consequence of the following theorem.

Theorem 1 *Given the Markov assumption and the k-Constraint assumption, for any $\theta_0 \neq 0$, if $G_1, G_2 \in \mathcal{G}$ in case 2 , there is no uniformly consistent test ϕ for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ such that*

$$\lim_{n\to\infty} \sup_{P\in\Omega_{G_1 0}} P^n(\phi_n(O^n) = 1 \vee \phi_n(O^n) = 2) = 0$$

where $\Omega_{G_1 0}$ is the set of all legitimate Gaussian distributions generated by G_1 with $\gamma = \theta_0$

Obviously if there were a uniformly test that does not return 2, the condition in Theorem 1 would be satisfied. This result applies to any test procedures, including, for example, the tests based on the various model selection scores. In the context of constraint-based causal inference, the intuition behind it is that the arrow between X_1 and X_3 and the one between X_2 and X_3 in G_1 can be arbitrarily weak, and hence the detection of the correlation between X_1 and X_3 and that between X_2 and X_3 can become arbitrarily difficult as α, β approach zero. In short, in the canonical case, α and β can become arbitrarily small, which is responsible for the lack of uniform consistency of the typical algorithms in the literature that only test whether the correlations are vanishing. We have to modify those algorithms to control the nuisance parameters α, β somehow in order to obtain uniform consistency.

[5] The formal proof can be found in Zhang [2002], which does not fit in the room here.

2.3 λ-STRONG-FAITHFULNESS

A perhaps more direct strengthening of Faithfulness is the following:

Definition 5 (λ-Strong-Faithfulness) *A multivariate Gaussian distribution P is said to be λ-strong-faithful to a DAG G with observed variables O if for any $a, b \in O$ and $C \subseteq O\backslash\{a, b\}$,*

$$a \text{ is d-connected to } b \text{ given } C \text{ in } G \iff |\rho_{ab.C}| > \lambda$$

where $\lambda \in (0, 1)$ is a fixed (small) constant.

Here is the λ**-Strong-Faithfulness assumption**: the Gaussian distributions generated by a causal graph are λ-Strong-Faithful to the graph. Intuitively the difference between λ-Strong-Faithfulness and k-Constraint is that λ-Strong-Faithfulness further rules out the possibility of weak arrows in the graph. For example, it entails that $|\alpha| > \lambda, |\beta| > \lambda$ in G_1. Under the λ-Strong-Faithfulness assumption (no matter how small λ is), the inference of causal structure in general can be uniformly consistent.

Theorem 2 *Let $\mathcal{M}$ be an arbitrary Markov equivalence class. Consider the null hypothesis H_0: data O^n are generated from a structure in $\mathcal{M}$. Given the Markov and λ-strong-faithfulness assumptions, there exists a test ϕ of the null hypothesis such that ϕ only returns 0 and 1 and*

$$\lim_n \sup_{P\in\Gamma_0} P^n(\phi_n(O^n) = 1) = 0$$
$$\text{and} \quad \lim_n \sup_{P\in\Gamma_1} P^n(\phi_n(O^n) = 0) = 0$$

where

$$\Gamma_0 = \bigcup_{G\in\mathcal{M}} \Omega(G) \qquad \Gamma_1 = \bigcup_{G\notin\mathcal{M}} \Omega(G)$$

The test constructed in the proof (in the appendix) is a combination of a series of tests of vanishing partial correlations, which is exactly what the constraint-based algorithms suggest. In other words, those algorithms are uniformly consistent without further modification under the λ-Strong-Faithfulness assumption. It follows from Theorem 2 that as long as a causal parameter can be identified in some Markov equivalence class, there exist non-trival uniformly consistent tests for that parameter provided that there are uniformly consistent tests for the corresponding statistical quantity.

It follows from Theorem 1 that Theorem 2 cannot be true if the λ-Strong-Faithfulness assumption is replaced by the k-Constraint assumption. As noted earlier, unlike the λ-Strong-Faithfulness assumption, the k-Constraint assumption allows the possibility of arbitrarily weak arrows, which usually act as sort of nuisance parameters when they are not of direct interest. In the canonical case, for instance, α and β are nuisance parameters, which we have to control somehow in order to get uniformly consistent tests. This requires modifying the existing algorithms that only involve testing vanishing partial correlations. One way to modify the algorithm is to further test the null hypotheses $|\rho_{X_1X_3}| \geq \alpha_0$ and $|\rho_{X_2X_3}| \geq \beta_0$ for fixed α_0, β_0 and return "no conclusion" if any of the tests rejects the null hypotheses (see Zhang [2002] for details). Intuitively this modification amounts to building some λ-Strong-Faithfulness into the algorithm, as the modified algorithm refuses to give informative answers when the correlations are below some threshold. We do not yet have a general proof of the existence of uniformly consistent procedures under the k-Constraint assumption, but in view of the general result on the λ-Strong-Faithfulness, it is quite intuitive that this kind of modification should work in general to obtain uniform consistency under the k-Constraint assumption.

2.4 ESTIMATORS AND CONFIDENCE REGIONS

There is nothing special about tests. All the previous results on consistency can be formulated in terms of point estimators and confidence regions[6]. Since a causal parameter is only sometimes estimable, we need also generalize the notions of point estimators and confidence regions, just as we include an uninformative answer in the possible outputs of tests. Robins et al. [2000] defined a generalized estimator $\hat{\theta}$ of θ as a sequence $(\hat{\theta}_1, ..., \hat{\theta}_n, ...)$, where each $\hat{\theta}_i$ is a function of O^i that returns a non-empty subset of the parameter space Θ. A non-singleton set (e.g. Θ) indicates that no estimation can be made. Let $\Omega\mathcal{G} = \{(P, G) : G \in \mathcal{G}, P \in \Omega(G)\}$. $\hat{\theta}$ is **pointwise consistent** if for every $(P, G) \in \Omega\mathcal{G}$, $\hat{\theta}$ converges in probability to $\theta = T(P, G)$, namely, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P^n(d[\hat{\theta}_n(O^n), T(P, G)] > \epsilon) = 0$$

$\hat{\theta}$ is said to be **uniformly consistent** if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \sup_{(P,G)\in\Omega\mathcal{G}} P^n(d[\hat{\theta}_n(O^n), T(P, G)] > \epsilon) = 0$$

where the distance d between a set S and a real number r is defined as the shortest Euclidean distance between r and the elements in S: $d[S, r] = \inf_{s\in S} |s-r|$. Finally $\hat{\theta}$ is **non-trivial** if

$$\lim_{n\to\infty} P^n(\hat{\theta}_n(O^n) \text{ is a singleton}) = 1$$

In the constraint-based causal inference, a typical estimator of a causal parameter first pins down a O-Markov equivalence class via a series of tests of independence and conditional independence relations, and then estimates the parameter if it is estimable in the resulting equivalence class or returns the whole sample space otherwise. In view of Theorem 2, it is not hard to see that this estimator is uniformly consistent under the Markov and λ-strong-faithfulness assumptions, provided that the estimator of the statistical quantity, with which the causal parameter is identified, is uniformly consistent.

We define the generalized confidence regions in a way that maintains the well known duality between tests and confidence regions (see e.g. Casella and Berger [1990]). Let $\mathbf{R}_\alpha$ denote a sequence: $(\mathbf{R}_{\alpha,1}, ..., \mathbf{R}_{\alpha,n}...)$, where each $\mathbf{R}_{\alpha,i}$ is a function of O^i which returns a triple partition $(S^0_{\alpha,i}, S^1_{\alpha,i}, S^2_{\alpha,i})$ of the parameter space Θ. $\mathbf{R}_\alpha$ is called a $1-\alpha$ **confidence region** of $\theta = T(P, G)$ if

$$\liminf_n \inf_{(P,G)\in\Omega\mathcal{G}} P^n(T(P, G) \notin S^1_{\alpha,n}) \geq 1 - \alpha$$

The definition is obviously given with test inversion in mind. We can easily invert a test (actually a family of tests) into a confidence region thus defined: $S^0_{\alpha,n}$ contains the values of the parameter that are accepted, $S^1_{\alpha,n}$ contains the values rejected and $S^2_{\alpha,n}$ contains the rest, for which "no conclusion" are returned.

$\mathbf{R}_\alpha$ is said to be **pointwise consistent** if for every $(P, G), (Q, H) \in \Omega\mathcal{G}$ such that $T(P, G) \neq T(Q, H)$,

$$\lim_{n\to\infty} P^n(T(Q, H) \in S^0_{\alpha,n}) = 0$$

it is **uniformly consistent** if for every $\delta > 0$

$$\lim_{n\to\infty} \sup_{|T(Q,H)-T(P,G)|>\delta} P^n(T(Q, H) \in S^0_{\alpha,n}) = 0$$

[6] Many details are left out in this section, which can be found in Zhang [2002]

it is **non-trivial** if at least for some $(P, G) \in \Omega\mathcal{G}$,

$$\lim_{n\to\infty} P^n(S^2_{\alpha,n} \neq \Theta) = 1$$

It is not hard to verify the usual duality between tests and confidence regions in this generalized setting. In particular, a family of uniformly consistent tests can be inverted into a uniformly consistent $1-\alpha$ confidence region for any $\alpha \in (0,1)$, at least in principle. Hence under the strong-faithfulness assumptions, we can hope for uniformly consistent confidence regions of causal parameters.

It is worth noting that in most cases, the test inversion will lead to a confidence region such that either $S^2_{\alpha,n} = \emptyset$ when θ is identifiable, or $S^0_{\alpha,n} = S^1_{\alpha,n} = \emptyset$ when θ is not identifiable (and hence we are totally ignorant of θ). For example, in case 2, either the causal structure suggested by data is G_1 in which case we can calculate an informative confidence interval, or the structure indicated by data is not G_1 in which case the confidence interval is the uninformative one, the whole set of legitimate values. Either way the resulted confidence interval looks just like the ordinary confidence interval. Only in some special cases, say in case 1, where some value of θ (0 in case 1) is of special status, the generalized confidence regions do look weird.

3 DISCUSSION

3.1 WITHOUT CAUSAL SUFFICIENCY

It is clear that the proof of Theorem 2 does not depend on the assumption of causal sufficiency at all. So, even in the presence of latent confounders, the typical causal inference algorithms, such as FCI in Spirtes et al. [1993], are uniformly consistent under the Markov and λ-Strong-Faithfulness assumptions. Under the k-Constraint assumption as currently defined, however, causal sufficiency is in general necessary to guarantee the possibility of uniformly consistent causal inference[7]. It is possible nonetheless to define the k-Constraint with respect to the parameterization of Maximal Ancestral Graphs (MAGs) [Richardson, Spirtes 2000] so that uniform consistency can be established without causal sufficiency. A problem with such a definition is that the nice intuitive explanation of k-Constraint — small (partial) correlation indicates small (direct) effect — is no longer available, as the parameters in MAGs do not always correspond to direct causal effects.

[7] In the canonical case, for example, if we allow the possibility that there might be a latent confounder between X_1 and X_4, it can be shown fairly easily that there is no non-trivial uniformly consistent test even under the k-Constraint assumption.

3.2 REFLECTION ON STRONG-FAITHFULNESS

The two Strong-Faithfulness assumptions laid out in this paper are both indexed by a positive constant. No matter how small the constant is, the assumptions entail the possibility of uniformly consistent causal inference. Both assumptions collapse into the faithfulness assumption on the boundary: 0-constraint is a vacuous constraint and 0-strong-faithfulness is essentially the same as faithfulness[8].

The perhaps most powerful and frequently used defense for the faithfulness assumption is that under the Gaussian or the multinomial parameterization, given a causal structure G, the set of parameters that lead to distributions unfaithful to G has zero Lebesgue measure (Spirtes et al. 1993, Meek 1995). A nice consequence of this fact is that for any causal structure G, any prior that is absolutely continuous with respect to Lebesgue measure will assign 0 probability to the unfaithful distributions. It is not necessarily the case, as it is tempting to conjecture, that the Lebesgue measure of the set of parameters ruled out by the λ-strong-faithfulness assumption can be made arbitrarily small (by decreasing λ), unless the parameter space is bounded.[9] But it is true that given a causal structure G and a prior absolutely continuous with respect to Lebesgue measure, for any $\epsilon > 0$, there exists a λ such that the prior probability assigned to the set of distributions that are not λ-strong-faithful to G is less than ϵ. This readily follows from the continuity of the probability measure.

An implication of the "measure 0" result is the existence of faithful multivariate Gaussian distributions (and multinomial distributions) for every causal structure, which is also cited fairly often in the literature. It is certainly not the case that for every $\lambda \in (0,1)$ and every causal structure G, there exists a distribution λ-strong-faithful to G.[10] For example, if A, B, C are three independent Gaussian parents of D, it is impossible that the correlations between D and each of the parents are all greater than $\sqrt{3}/3$. On the other hand, it is trivial to see that for any causal structure G, there exists a multivariate Gaussian distribution λ-strong-faithful to G for some λ. More interestingly, it can be

[8] The part of the faithfulness assumption we really use in causal inference only involves the observed variables.

[9] Here is a simple illustration of the remark. In the two-dimensional plane, the line $x = y$ has zero Lebesgue measure, but the region $S = \{(x,y) : x - k \leq y \leq x + k\}$ has infinite Lebesgue measure no matter how small k is, though it becomes the line when $k = 0$.

[10] It is not yet clear whether for every $k \in (0,1)$ and every causal structure G, there exists a Gaussian distribution that satisfies the k-constraint with G.

shown that given a fixed set of observed variables O, we can find a small λ such that for every causal structure G with O as the observed variables, there exists a multivariate Gaussian distribution λ-strong-faithful to G. The magnitude of λ depends on the number of variables in O.

Another popular interpretation of the faithfulness assumption appeals to the notion of "stableness". [Pearl 2000] The faithful distributions are stable in the sense that the independence and conditional independence relations associated with the distributions cannot be destroyed by the waggling of parameters. Similarly, a faithful but close to unfaithful distribution may be said to be unstable in the sense that some dependence relations may be destroyed by a slight change in parameterization. In this sense, the λ in the λ-strong-faithfulness serves as an rough index of stableness.

It is not the main purpose of the reflection to argue for the plausibility of the strong-faithfulness assumptions. Rather the discussion is to illustrate the close relation between the usual faithfulness condition and the stronger faithfulness conditions laid out in the paper. Clearly in several important respects, the faithfulness assumption is just a limiting case of the λ-strong-faithfulness assumption (or the k-constraint assumption). This suggests that the stronger assumptions are not only sufficient but also close to necessary to entail the existence of uniformly consistent causal inference procedures without substantial background knowledge.

Reference

[1] Anderson, T.W. (1958) *An Introduction to Multivariate Statistical Analysis.* New York: Wiley.

[2] Bickel, P.J., Doksum, K.A. (2001) *Mathematical Statistics - Basic Ideas and Selected Topics, VOL. 1* (2nd ed.). New Jersey: Prentice Hall.

[3] Heckerman, D., Geiger, D., Chickering, D. (1995) Learning Bayesian networks: the combination of knowledge and statistical data, in *Machine Learning* 20(3):197-243.

[4] Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems.* San Mateo, Calif.: Morgan Kaufmann.

[5] Pearl, J. (2000) *Causality: Models, Reasoning, and Inference.* Cambridge, U.K.:Cambridge University Press.

[6] Richardson, T. (1996) A Discovery Algorithm for Directed Cyclic Graphs, in *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*, 454-461. San Francisco: Morgan Kaufmann.

[7] Richardson, T., and Spirtes, P. (2000) Ancestral Markov Graphical Models, University of Washington Department of Statistics Technical Report 375.

[8] Robins, J.M., Scheines, R., Spirtes, P., Wasserman, L. (2000) Uniform Consistency in Causal Inference. Carnegie Mellon University Statistics Department Technical Report 725. Forthcoming in *Biometrica*

[9] Spirtes, P., Glymour, C., Scheines, R. (2000) *Causation, Prediction and Search* (2nd ed.) Cambridge, MA: MIT Press. (1993, 1st ed.) New York: Springer-Verlag.

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

[11] Zhang, J. (2002) Consistency In Causal Inference Under A Variety Of Assumptions. Carnegie Mellon University Philosophy Department, Master Thesis.

Appendix

Proof of Theorem 1.

Lemma Consider the linear structural equation models (with independent error terms and no latent variables) associated with G_1 and G_2 in Figure 2:

$$\begin{array}{ll} M1 & M2 \\ X_1 = \epsilon_1 & X_1 = \epsilon_1 \\ X_2 = \epsilon_2 & X_2 = \epsilon_2 \\ X_3 = \alpha X_1 + \beta X_2 + \epsilon_3 & X_3 = fX_1 + gX_2 + hX_4 + \epsilon_3 \\ X_4 = \gamma X_3 + \epsilon_4 & X_4 = mX_1 + nX_2 + \epsilon_4 \end{array}$$

X_1, X_2, X_3, X_4 are all standard Gaussians. For every $0 < k < 1$, $\theta_0 \neq 0$ and every $\epsilon > 0$, there are $\mu_1 = (\alpha, \beta, \gamma = \theta_0)$ and $\mu_2 = (f, g, h, m, n)$ such that $\mu_i \in \Psi^k_{G_i}$ and $KL(P_1, P_2) < \epsilon$, where P_i is the distribution generated by (μ_i, G_i), and $KL(P_1, P_2)$ is the Kullback-Leibler divergence.

Proof The correlation matrix generated by M_1 is

$$\Sigma_1 = \begin{pmatrix} 1 & 0 & \alpha & \alpha\gamma \\ & 1 & \beta & \beta\gamma \\ & & 1 & \gamma \\ & & & 1 \end{pmatrix}$$

It is easy to verify that all legitimate parameters of M_1 are in $\Psi^k_{G_1}$, i.e. no (more) constraints are placed on the parameters in M_1 by the k-constraint. The correlation matrix generated by M_2 is

$$\Sigma_2 = \begin{pmatrix} 1 & 0 & f + mh & m \\ & 1 & g + nh & n \\ & & 1 & fm + gn + h \\ & & & 1 \end{pmatrix}$$

The k-constraint assumption puts the following constraints on μ_2:

$$\begin{aligned}
k|m| &\leq \left|\frac{m-(f+mh)(fm+gn+h)}{\sqrt{(1-(f+mh)^2)(1-(fm+gn+h)^2)}}\right| \\
&= |\rho_{X_1X_4.X_3}| \\
k|n| &\leq \left|\frac{n-(g+nh)(fm+gn+h)}{\sqrt{(1-(g+nh)^2)(1-(fm+gn+h)^2)}}\right| \\
&= |\rho_{X_2X_4.X_3}| \\
k|f| &\leq |f+mh| = |\rho_{X_1X_3}| \\
k|g| &\leq |g+nh| = |\rho_{X_2X_3}| \\
k|h| &\leq |fm+gn+h| = |\rho_{X_3X_4}|
\end{aligned}$$

Given any $\epsilon > 0$, there exists $\delta_{\alpha,\beta} > 0$ ($\delta_{\alpha,\beta}$ depends on α and β) such that if

$$\begin{aligned}
f+mh &= \alpha & (1)\\
g+nh &= \beta & (2)\\
fm+gn+h &= \gamma = \theta_0 & (3)\\
|m-\alpha\theta_0| &< \delta_{\alpha,\beta} & (4)\\
|n-\beta\theta_0| &< \delta_{\alpha,\beta} & (5)
\end{aligned}$$

then $KL(P_1, P_2) < \epsilon$. Solve (6),(7),(8) for f, g, h, we get

$$\begin{aligned}
f &= \frac{(1-n^2)\alpha + mn\beta - m\theta_0}{1-m^2-n^2} \\
g &= \frac{(1-m^2)\beta + mn\alpha - n\theta_0}{1-m^2-n^2} \\
h &= \frac{\theta_0 - m\alpha - n\beta}{1-m^2-n^2}
\end{aligned}$$

It is not hard to check that we can choose appropriate (small) m, n, α, β to satisfy (9), (10) and all the constraints. **Q.E.D**

Proof of Theorem 1 For the sake of contradiction, suppose there is such a test ϕ. Choose $\delta < \theta_0$ so that $\Omega_{\mathcal{G}1\delta}$ includes all the distributions in Ω_{G_2}. Now given any $\epsilon > 0$, by the above Lemma, there are $P_1 \in \Omega_{G_1}$ and $P_2 \in \Omega_{G_2}$ such that $KL(P_1^n, P_2^n) < 4\epsilon^2$. Hence

$$\sup_E |P_1^n(E) - P_2^n(E)| \leq 1/2\sqrt{KL(P_1^n, P_2^n)} = \epsilon$$

The supremum is over all the events in the sample space. Therefore,

$$\begin{aligned}
&\sup_{R\in\Omega_{\mathcal{G}1\delta}} R^n(\phi_n(O^n) = 0) \\
\geq\ & P_2^n(\phi_n(O^n) = 0) \\
\geq\ & P_1^n(\phi_n(O^n) = 0) - \epsilon \\
=\ & 1 - P_1^n(\phi_n(O^n) = 1 \vee \phi_n(O^n) = 2) \\
\geq\ & 1 - \sup_{P\in\Omega_{G_1}} P^n(\phi_n(O^n) = 1 \vee \phi_n(O^n) = 2) - \epsilon \\
\rightarrow\ & 1 - \epsilon
\end{aligned}$$

Note that the selection of ϵ is arbitrary, which implies

$$\lim_{n\rightarrow\infty} \sup_{R\in\Omega_{\mathcal{G}1\delta}} R^n(\phi_n(O^n) = 0) = 1$$

Hence a contradiction. **Q.E.D**

Proof of Theorem 2.

$\mathcal{M}$ entails a unique set of zero partial correlations (we treat correlations as partial correlations where the set of variables to control for is empty), and there are altogether a finite number of partial correlations to consider among the observed variables. We construct a test ϕ as below:

$$\phi_n(O^n) = \begin{cases} 0 & \text{if } T_n^1(O^n) = \ldots = T_n^l(O^n) = 0 \\ & \text{and } T_n^{l+1}(O^n) = \ldots = T_n^m(O^n) = 1 \\ 1 & \text{otherwise} \end{cases}$$

where $T^1, \ldots, T^l$ are the uniformly consistent tests of vanishing partial correlations for the ones entailed to be zero by $\mathcal{M}$, and $T^{l+1}, \ldots, T^m$ are the tests of vanishing partial correlations for the rest. We show that ϕ satisfies the requirement. Obviously it only returns 0 and 1. Let

$$\begin{aligned}
p_i^n &= \sup_{P\in\mathcal{H}_{1\lambda}^i} P^n(T^i(O^n) = 0), i = 1, \ldots, l \\
p_j^n &= \sup_{P\in\mathcal{H}_0^j} P^n(T^j(O^n) = 1), j = l+1, \ldots, m
\end{aligned}$$

where $\mathcal{H}_{1\lambda}^i$ and $\mathcal{H}_0^j$ have the obvious definitions. Since $T^1, \ldots, T^m$ are uniformly consistent, we have

$$\lim_{n\rightarrow\infty} p_i^n = 0, i = 1, \ldots, m.$$

Hence

$$\lim_{n\rightarrow\infty} \max\{p_i^n : i = 1, \ldots, m\} = 0,$$

Under the λ-strong-faithfulness assumption, it is quite obvious that,

$$P^n(\phi_n^\lambda(O^n) = 0) \leq \max\{p_i^n : i = 1, \ldots, m\}$$

Therefore

$$\sup_{P\in\Gamma_1} P^n(\phi_n^\lambda(O^n) = 0) \leq \max\{p_i^n : i = 1, \ldots, m\} \rightarrow 0$$

To show that

$$\lim_n \sup_{P\in\Gamma_0} P^n(\phi_n(O^n) = 1) = 0$$

it suffices to replace max in the foregoing argument with summation, because the worst case error rate is bounded above by the sum of the error rates of all the partial correlation tests, which converges to 0 because the number of the tests is finite. **Q.E.D**