Statistical Science
2014, Vol. 29, No. 4, 662–678
DOI: 10.1214/13-STS429

A Uniformly Consistent Estimator of Causal Effects under the k-Triangle-Faithfulness Assumption

Peter Spirtes and Jiji Zhang

Abstract. Spirtes, Glymour and Scheines [*Causation*, *Prediction*, *and Search* (1993) Springer] described a pointwise consistent estimator of the Markov equivalence class of any causal structure that can be represented by a directed acyclic graph for any parametric family with a uniformly consistent test of conditional independence, under the Causal Markov and Causal Faithfulness assumptions. Robins et al. [*Biometrika* **90** (2003) 491–515], however, proved that there are no uniformly consistent estimators of Markov equivalence classes of causal structures under those assumptions. Subsequently, Kalisch and Bühlmann [*J. Mach. Learn. Res.* **8** (2007) 613–636] described a uniformly consistent estimator of the Markov equivalence class of a linear Gaussian causal structure under the Causal Markov and Strong Causal Faithfulness assumptions. However, the Strong Faithfulness assumption may be false with high probability in many domains. We describe a uniformly consistent estimator of both the Markov equivalence class of a linear Gaussian causal structure and the identifiable structural coefficients in the Markov equivalence class under the Causal Markov assumption and the considerably weaker k-Triangle-Faithfulness assumption.

Key words and phrases: Causal inference, uniform consistency, structural equation models, Bayesian networks, model selection, model search, estimation.

1. INTRODUCTION

A principal aim of many sciences is to model causal systems well enough to provide sound insight into their structures and mechanisms and to provide reliable predictions about the effects of policy interventions. The modeling process is typically divided into two distinct phases: a model specification phase in which some model (with free parameters) is specified, and a parameter estimation and statistical testing phase in which the free parameters of the specified model are estimated and various hypotheses are put to a statistical test. Both model specification and parameter estimation can fruitfully be thought of as search problems.

As pointed out in Robins et al. (2003), common statistical wisdom dictates that causal effects cannot be consistently estimated from observational studies alone unless one observes and adjusts for all possible confounding variables, and knows the time order in which events occurred. However, Spirtes, Glymour and Scheines (1993) and Pearl (2000) developed a framework in which causal relationships are represented by edges in a directed acyclic graph. They also described asymptotically consistent procedures for determining features of causal structure from data even if we allow for the possibility of unobserved confounding variables and/or an unknown time order, under two assumptions: the Causal Markov assumption (roughly, given no unmeasured common causes, each variable is independent of its noneffects conditional on its direct causes) and the Causal Faithfulness assumption

Peter Spirtes is Professor, Department of Philosophy, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA (e-mail: ps7z@andrew.cmu.edu). Jiji Zhang is Associate Professor, Department of Philosophy, Lingnan University, Tuen Mun, N.T., Hong Kong (e-mail: jijizhang@ln.edu.hk).

(all conditional independence relations that hold in the distribution are entailed by the Causal Markov assumption). Under these assumptions, the procedures they propose (e.g., the *SGS* and the *PC* algorithms assuming no unmeasured common causes, and the *FCI* algorithm which does not assume no unmeasured common causes) can infer the existence or absence of causal relationships. In particular, Spirtes et al. (1993), Chapters 5 and 6, proved the Fisher consistency of these procedures. Pointwise consistency follows from the Fisher consistency and the uniform consistency of the test procedures for conditional independence relationships in certain parametric families that the procedures use.

Robins et al. (2003) proved that under the Causal Markov and Faithfulness assumptions made in Spirtes, Glymour and Scheines (1993) there are no uniformly consistent procedures for estimating features of the causal structure from data, even when there are no unmeasured common causes. Spirtes, Glymour and Scheines (2000), Kalisch and Bühlmann (2007) and Colombo et al. (2012) introduced a Strong Causal Faithfulness assumption, which, roughly speaking, assumes that no conditional independence relation not entailed by the Causal Markov assumption "almost" holds. Kalisch and Bühlmann (2007) and Colombo et al. (2012) showed that under this strengthened Causal Faithfulness assumption, some modifications of the pointwise consistent procedures developed in Spirtes, Glymour and Scheines (1993) are uniformly consistent. Maathuis et al. (2010) have also successfully applied these procedures to various biological data sets, experimentally confirming some of the causal inferences made by the procedures.

However, the question remains whether the Strong Causal Faithfulness assumption made by Kalisch and Bühlmann (2007) is too strong. Is it likely to be true? Some analysis done by Uhler et al. (2013) indicates that the strengthened Causal Faithfulness assumption is likely to be false, especially when there are a large number of variables.

In this paper we investigate a number of different ways in which the strengthened Causal Faithfulness assumption can be weakened, while still retaining the guarantees of uniformly consistent estimation by modifying the causal estimation procedures. It is not clear whether the ways we propose to weaken the Strong Causal Faithfulness assumption make it substantially more likely to hold, nor is it clear that all of the modifications that we propose to the estimation procedures make them substantially more accurate in practice. Nevertheless, we believe that the modifications that we propose are a useful first step toward investigating fruitful modifications of the Causal Faithfulness assumption and causal estimation procedures.

In Section 2 we describe the basic setup and assumptions for causal inference. In Section 3 we examine various ways to weaken the Causal Faithfulness assumption and modifications of the estimation procedures that preserve pointwise consistency. In Section 4 we examine weakening the Strong Causal Faithfulness assumption and modification of the estimation procedures that preserves uniform consistency. Finally, in Section 5 we summarize the results and describe areas of future research.

2. THE BASIC ASSUMPTIONS FOR CAUSAL INFERENCE

We first introduce the graph terminology that we will use. Individual variables are denoted with italicized capital letters, and sets of variables are denoted with bold-faced capital letters. A *graph* $G = \langle \mathbf{V}, \mathbf{E} \rangle$ consists of a set of *vertices* $\mathbf{V}$ and a set of *edges* $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$, where for each $\langle X, Y \rangle \in \mathbf{E}, X \neq Y$. If $\langle X, Y \rangle \in \mathbf{E}$ and $\langle Y, X \rangle \in \mathbf{E}$, there is an *undirected edge* between X and Y, denoted by X — Y. If $\langle X, Y \rangle \in \mathbf{E}$ and $\langle Y, X \rangle \notin \mathbf{E}$, there is a directed edge between X and Y, denoted by $X \to Y$. If there is a directed edge from X to Y, or from Y to X, or there is an undirected edge between X and Y, then X and Y are *adjacent* in G. $\mathrm{Adj}(G, X)$ is the set of vertices adjacent to X. If all of the edges in a graph G are directed edges, then G is a *directed graph*. A *path* between X_1 and X_n in G is an ordered sequence of vertices $\langle X_1, \ldots, X_n \rangle$ such that for $1 < i \leq n$, X_{i-1} and X_i are adjacent in G. A path between X_1 and X_n in G is a *directed path* if for $1 < i \leq n$, the edge between X_{i-1} and X_i is a directed edge from X_{i-1} to X_i. A path is *acyclic* if no vertex occurs on the path twice. A directed graph is *acyclic* (DAG) if all directed paths are acyclic. X is a *parent* of Y and Y is a *child* of X if there is an edge $X \to Y$. $\langle X, Y, Z \rangle$ is a *triangle* in G if X is adjacent to Y and Z, and Y is adjacent to Z.

Suppose G is a graph. $\mathbf{Parents}(G, X)$ is the set of parents of X in G. X is an *ancestor* of Y (and Y is a *descendant* of X) if there is a directed path from X to Y. A subset of $\mathbf{V}$ is *ancestral*, if it is closed under the ancestor relation. A triple of vertices $\langle X, Y, Z \rangle$ is *unshielded* if X is adjacent to Y and Y is adjacent to Z, but X is not adjacent to Z. A triple of vertices $\langle X, Y, Z \rangle$ is a *collider* if there are edges $X \to Y \leftarrow Z$. A triple of vertices $\langle X, Y, Z \rangle$ is a *noncollider* if X is adjacent to Y and Y is adjacent to Z, but it is not a collider.

A probability distribution P over a set of variables $\mathbf{V}$ satisfies the (*local directed*) *Markov condition* for a DAG G if and only if each variable V in $\mathbf{V}$ is independent of the set of variables that are neither parents nor descendants of V in G, conditional on the parents of V in G. A *Bayesian network* is an ordered pair $\langle P, G\rangle$ where P satisfies the local directed Markov condition for G. If $M = \langle P, G\rangle$, P_M denotes P and G_M denotes G. Two DAGs G_1 and G_2 over the same set of variables $\mathbf{V}$ are said to be *Markov equivalent* if all of the conditional independence relations entailed by satisfying the local directed Markov condition for G_1 are also entailed by satisfying the local directed Markov condition for G_2, and vice versa. A useful characterization of Markov equivalence between DAGs is that two DAGs are Markov equivalent if and only if they have the same adjacencies and the same unshielded colliders (Verma and Pearl, 1990). A *Markov equivalence class* M is a set of DAGs that contains all DAGs that are Markov equivalent to each other. A Markov equivalence class M can be represented by a graph called a *pattern*; a pattern O is a graph such that (i) if $X \to Y$ in every DAG in M, then $X \to Y$ in O; and (ii) if $X \to Y$ in some DAG in M and $Y \to X$ in some other DAG in M, then X — Y in O. In that case O is said to *represent* M and each DAG in M.

If $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$, we write $I(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$, or if X, Y, and Z are individual variables $I(X, Y|Z)$. In a DAG G, a vertex A is *active* on an acyclic path U between X and Y conditional on set $\mathbf{Z}$ of vertices (not containing X or Y) if $A = X$ or $A = Y$, or A is a noncollider on U and not in $\mathbf{Z}$, or A is a collider on U that is in $\mathbf{Z}$ or has a descendant in $\mathbf{Z}$. An acyclic path U is *active* conditional on a set $\mathbf{Z}$ of vertices if every vertex on the path is active relative to $\mathbf{Z}$. If $X \neq Y$ and $\mathbf{Z}$ does not contain X or Y, X is *d-separated* from Y conditional on $\mathbf{Z}$ if there is no active acyclic path between X and Y conditional on $\mathbf{Z}$; otherwise X and Y are *d-connected* conditional on $\mathbf{Z}$. For three disjoint sets $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, $\mathbf{X}$ is *d-separated* from $\mathbf{Y}$ conditional on $\mathbf{Z}$ if there is no acyclic active path between any member of $\mathbf{X}$ and any member of $\mathbf{Y}$ conditional on $\mathbf{Z}$; otherwise $\mathbf{X}$ and $\mathbf{Y}$ are *d-connected* conditional on $\mathbf{Z}$. If $\mathbf{X}$ is d-separated from $\mathbf{Y}$ conditional on $\mathbf{Z}$ in DAG G, then $I(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ in every probability distribution that satisfies the local directed Markov condition for G (Pearl, 1988). Any conditional independence relation that holds in every distribution that satisfies the local directed Markov condition for DAG G is *entailed by* G. Note, however, that in some distributions that satisfy the local directed Markov condition for G, some conditional independence relation $I(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ may hold even if $\mathbf{X}$ is not d-separated from $\mathbf{Y}$ conditional on $\mathbf{Z}$ in G; such distributions are said to be *unfaithful to* G.

There are a number of different parameterizations of a DAG G, which map G onto distributions that satisfy the local directed Markov condition for G. One common parameterization is a recursive linear Gaussian structural equation model. A recursive linear Gaussian structural equation model is an ordered triple $\langle G, Eq, \Sigma\rangle$, where G is a DAG over a set of vertices $X_1, \ldots, X_n$, Eq is a set of equations, one for each X_i such that

$$X_i = \sum_{X_j \in \mathbf{Parents}(G, X_i)} b_{j,i} X_j + \varepsilon_i,$$

where the $b_{j,i}$ are real constants known as the structural coefficients, and the ε_i are multivariate Gaussian that are jointly independent of each other with covariance matrix Σ. The ε_i are referred to as "error terms." In vector notation, where $\mathbf{X}$ is the vector of $X_1, \ldots, X_n$, $\mathbf{B}$ is the matrix of structural coefficients, and $\boldsymbol{\varepsilon}$ is the vector of error terms,

$$\mathbf{X} = \mathbf{B}\mathbf{X} + \boldsymbol{\varepsilon}.$$

The covariance matrix Σ over the error terms, together with the structural equations, determine a distribution over the variables in $\mathbf{X}$, which satisfies the local directed Markov condition for G. Hence, the DAG in a recursive linear Gaussian structural equation model M together with the probability distribution generated by the equations and the covariance matrix over the error terms form a Bayesian network. Because the joint distribution over the nonerror terms of a linear Gaussian structural equation model is multivariate Gaussian, X is independent of Y conditional on $\mathbf{Z}$ in P_M if and only if $\rho_M(X, Y|\mathbf{Z}) = 0$, where $\rho_M(X, Y|\mathbf{Z})$ denotes the conditional or partial correlation between X and Y conditional on $\mathbf{Z}$ according to P_M. Let $e_M(X \to Z)$ denote the structural coefficient of the $X \to Z$ edge in G_M. If there is no edge $X \to Z$ in G_M, then $e_M(X \to Z) = 0$. If X and Z are adjacent in G_M, then $e_M(X$ — $Z) = e_M(X \to Z)$ if there is an $X \to Z$ edge in G_M, and otherwise $e_M(X$ — $Z) = e_M(Z \to X)$.

There is a *causal interpretation* of recursive linear Gaussian structural equation models, in which setting (as in an experiment, as opposed to observing) the value of X_i to the fixed value x is represented by replacing the structural equation for X_i with the equation $X_i = x$. Under the causal interpretation, a recursive linear structural equation model is a *causal model*,

the DAG G_M is a *causal DAG*, and the pattern that represents G_M is a *causal pattern*. A causal model with a set of variables $\mathbf{V}$ is *causally sufficient* when every common direct cause of any two variables in $\mathbf{V}$ is also in $\mathbf{V}$. Informally, under a causal interpretation, an edge $X \to Y$ in G_M represents that X is a direct cause of Y relative to $\mathbf{V}$. A causal model of a population is true when the model correctly predicts the results of all possible settings of any subset of the variables (Pearl, 2000).

There are two assumptions made about the relationship between the causal DAG and the population probability distribution that play a key role in causal inference from observational data. A discussion of the implications of these assumptions, arguments for them, and a discussion of conditions when they should not be assumed are given in Spirtes, Glymour and Scheines (1993), pages 32–42. In this paper, we will consider only those cases where the causal relations in a given population can be represented by a model whose graph is a DAG.

Causal Markov assumption (*CMA*). If the true causal model M of a population is causally sufficient, every variable in $\mathbf{V}$ is independent of the variables that are neither its parents nor descendants in G_M conditional on its parents in G_M.

Causal Faithfulness assumption (*CFA*). Every conditional independence relation that holds in the population probability distribution is entailed by the true causal DAG of the population.

The Causal Markov and Causal Faithfulness assumptions together entail that $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ in the population if and only if $\mathbf{X}$ is d-separated from $\mathbf{Y}$ conditional on $\mathbf{Z}$ in the true causal graph.

3. WEAKENING THE CAUSAL FAITHFULNESS ASSUMPTION

A number of algorithms for causal estimation have been proposed that rely on the assumption of the causal sufficiency of the observed variables, the Causal Markov assumption and the Causal Faithfulness assumption. The *SGS* algorithm (Spirtes, Glymour and Scheines, 1993, page 82), for example, is a Fisher consistent estimator of causal patterns under these assumptions. (This, together with a uniformly consistent test of conditional independence, entails that the *SGS* algorithm is a pointwise consistent estimator of causal patterns.)

In this section we explore ways to weaken the Causal Faithfulness assumption that still allow pointwise consistent estimation of (features of) causal structure, and we illustrate the ideas by going through a sequence of generalizations of the population version of the *SGS* algorithm. None of the results in this section depend upon assuming Gaussianity or linearity. The basic idea is that although the Causal Faithfulness assumption is not fully testable (without knowing the true causal structure), it has testable components given the Causal Markov assumption. Under the Causal Markov assumption, the Causal Faithfulness assumption entails that the probability distribution admits a perfect DAG representation, that is, a DAG that entails all and only those conditional independence relations true of the distribution. Whether there is such a DAG depends only on the distribution, and so is, in theory, testable. In principle, then, one may adopt a weaker-than-faithfulness assumption and test (rather than assume) the testable part of the faithfulness condition.

The *SGS* algorithm takes an oracle of conditional independence as input, and outputs a graph on the given set of variables with both directed edges and undirected edges.

SGS algorithm.

S1. Form the complete undirected graph H on the given set of variables $\mathbf{V}$.

S2. For each pair of variables X and Y in $\mathbf{V}$, search for a subset $\mathbf{S}$ of $\mathbf{V} \setminus \{X, Y\}$ such that X and Y are independent conditional on $\mathbf{S}$. Remove the edge between X and Y in H if and only if such a set is found.

S3. Let K be the graph resulting from S2. For each unshielded triple $\langle X, Y, Z\rangle$ (i.e., X and Y are adjacent, Y and Z are adjacent, but X and Z are not adjacent),
   (i) If X and Z are not independent conditional on any subset of $\mathbf{V} \setminus \{X, Z\}$ that contains Y, then orient the triple as a collider: $X \to Y \leftarrow Z$.
   (ii) If X and Z are not independent conditional on any subset of $\mathbf{V} \setminus \{X, Z\}$ that does not contain Y, then mark the triple as a noncollider (i.e., not $X \to Y \leftarrow Z$).

S4. Execute the following orientation rules until none of them applies:
   (i) If $X \to Y — Z$, and the triple $\langle X, Y, Z\rangle$ is marked as a noncollider, then orient $Y — Z$ as $Y \to Z$.
   (ii) If $X \to Y \to Z$ and $X — Z$, then orient $X — Z$ as $X \to Z$.

(iii) If $X \to Y \leftarrow Z$, another triple $\langle X, W, Z\rangle$ is marked as a noncollider, and $W - Y$, then orient $W - Y$ as $W \to Y$. (This rule was not in the original *SGS* or *PC* algorithm, but added by Meek, 1995.)

Assuming the oracle of conditional independence is perfectly reliable (which we will do throughout this section), the *SGS* algorithm is correct under the Causal Markov and Faithfulness assumptions, in the sense that its output is the pattern that represents the Markov equivalence class containing the true causal DAG (Spirtes, Glymour and Scheines, 1993, page 82; Meek, 1995).

The correctness of *SGS* follows from the following three properties of d-separation (Spirtes, Glymour and Scheines, 1993):

1. X is adjacent to Y in DAG G if and only if X is not d-separated from Y conditional on any subset of the other variables in G.
2. If $\langle X, Y, Z\rangle$ is an unshielded collider in DAG G, then X is not d-separated from Z conditional on any subset of the other variables in G that contains Y.
3. If $\langle X, Y, Z\rangle$ is an unshielded noncollider in DAG G, then X is not d-separated from Z conditional on any subset of the other variables in G that does not contain Y.

We shall not reproduce the full proof here, but a few points are worth stressing. First, S2 is the step of inferring adjacencies and nonadjacencies. The inferred adjacencies, represented by the remaining edges in the graph resulting from S2, are correct because of the Causal Markov assumption alone: every DAG Markov to the given oracle must contain at least these adjacencies. On the other hand, the inferred nonadjacencies (via removal of edges) are correct because of the Causal Faithfulness assumption, or, more precisely, because of the following consequence of the Causal Faithfulness assumption, which we, following Ramsey, Zhang and Spirtes (2006), will refer to as Adjacency-Faithfulness.

Adjacency-Faithfulness assumption. Given a set of variables $\mathbf{V}$ whose true causal DAG is G, if two variables X, Y are adjacent in G, then they are not independent conditional on any subset of $\mathbf{V} \setminus \{X, Y\}$.

Under the Adjacency-Faithfulness assumption, any edge removed in S2 is correctly removed, because any DAG with the adjacency violates the Adjacency-Faithfulness assumption.

Second, the key step of inferring orientations is step S3, in which unshielded colliders and noncolliders are inferred. Given that the adjacencies and nonadjacencies are all correct, the clauses (i) and (ii) in step S3, as formulated here, are justified by the Causal Markov assumption alone. Take clause (i), for example. If the unshielded triple $\langle X, Y, Z\rangle$ is not a collider in the true causal DAG, then the Causal Markov assumption entails that X and Z are independent conditional on some set that contains Y. That is why clause (i) is sound. A similar argument shows that clause (ii) is sound. This does not mean, however, that the Causal Faithfulness assumption does not play any role in justifying S3. Notice that the antecedent of (i) and that of (ii) do not exhaust the logical possibilities. They leave out the possibility that X and Z are independent conditional on some set that contains Y and independent conditional on some set that does not contain Y. This omission is justified by the Causal Faithfulness assumption, or, more precisely, by the following consequence of the Causal Faithfulness assumption (Ramsey, Zhang and Spirtes, 2006):

Orientation-Faithfulness assumption. Given a set of variables $\mathbf{V}$ whose true causal DAG is G, let $\langle X, Y, Z\rangle$ be any unshielded triple in G:

1. If $X \to Y \leftarrow Z$, then X and Z are not independent conditional on any subset of $\mathbf{V} \setminus \{X, Z\}$ that contains Y;
2. Otherwise, X and Z are not independent conditional on any subset of $\mathbf{V} \setminus \{X, Z\}$ that does not contain Y.

Obviously, the possibility left out by S3 is indeed ruled out by the Orientation-Faithfulness assumption.

The Orientation-Faithfulness assumption, if true, justifies a much simpler and more efficient step than S3: for every unshielded triple $\langle X, Y, Z\rangle$, we need check only the set found in S2 that renders X and Z independent; the triple is a collider if and only if the set does not contain Y. This simplification is used in the *PC* algorithm, a well-known, more computationally efficient rendition of the *SGS* procedure (Spirtes, Glymour and Scheines, 1993, pages 84–85). Moreover, the Adjacency-Faithfulness condition also justifies a couple of measures to improve the efficiency of S2, used by the *PC* algorithm. Here we are concerned with showing how the basic *SGS* procedure may be modified to be correct under increasingly weaker assumptions of faithfulness, so we will not go into the details of the optimization measures in the *PC* algorithm. Whether these or similar measures are available to the modified algorithms we introduce below is an important question to be addressed in future work.

Let us start with the modification proposed by Ramsey, Zhang and Spirtes (2006), who observed that assuming the Causal Markov and Adjacency-Faithfulness assumptions are true, any failure of the Orientation-Faithfulness assumption is *detectable*, in the sense that the probability distribution in question is not both Markov and Faithful to any DAG (Zhang and Spirtes, 2008). In our formulation of the *SGS* algorithm, it is easy to see how failures of Orientation-Faithfulness can be detected. As already mentioned, the role of the Orientation-Faithfulness assumption in justifying the *SGS* algorithm is to guarantee that at the step S3, either the antecedent of (i) or that of (ii) will obtain. Therefore, if it turns out that for some unshielded triple neither antecedent is satisfied, the Orientation-Faithfulness assumption is detected to be false for that triple.

This suggests a simple modification to S3 in the *SGS* algorithm.

S3*. Let K be the undirected graph resulting from S2. For each unshielded triple $\langle X, Y, Z\rangle$,

(i) If X and Z are not independent conditional on any subset of $\mathbf{V} \setminus \{X, Z\}$ that contains Y, then orient the triple as a collider: $X \to Y \leftarrow Z$.

(ii) If X and Z are not independent conditional on any subset of $\mathbf{V} \setminus \{X, Z\}$ that does not contain Y, then mark the triple as a noncollider.

(iii) Otherwise, mark the triple as ambiguous (or unfaithful).

Ramsey, Zhang and Spirtes (2006) applied essentially this modification to the *PC* algorithm and called the resulting algorithm the *Conservative PC* (*CPC*) algorithm. (Their results show that the main optimization measures used in the *PC* algorithm still apply to this generalization of *SGS* because the Adjacency-Faithfulness condition is still assumed.) We will thus call the algorithm that results from replacing S3 with S3* the *Conservative SGS* (*CSGS*) algorithm.

It is straightforward to prove that the *CSGS* algorithm is correct under the Causal Markov and the Adjacency-Faithfulness assumptions alone, in the sense that if the Causal Markov and Adjacency-Faithfulness assumptions are true and if the oracle of conditional independence is perfectly reliable, then every adjacency, nonadjacency, orientation and marked noncollider in the output of the *CSGS* are correct. As pointed out in Ramsey, Zhang and Spirtes (2006), the output of the *CSGS* can be understood as an *extended pattern* that represents a set of patterns. For example, a sample output used in Ramsey, Zhang and Spirtes (2006) is given in Figure 1(a). There are two ambiguous unshielded triples in the output: $\langle Y, X, Z\rangle$ and $\langle Z, U, Y\rangle$, which are marked by crossing straight lines. Note that there is no explicit mark for noncolliders, with the understanding that all and only unshielded triples that are not oriented as colliders or marked as ambiguous are (implicitly) marked noncolliders. Figure 1(a) represents a set of three patterns, depicted in Figure 1(b)–(d). Each pattern results from some disambiguation of the ambiguous triples in Figure 1(a). The pattern in Figure 1(b), for example, results from taking the triple $\langle Y, X, Z\rangle$ as a noncollider and taking the triple $\langle Z, U, Y\rangle$ as a collider. Note that not every disambiguation results in a pattern. Taking both ambiguous triples as noncolliders would force a directed cycle: $Z \to U \to Y \to X \to Z$, and so would not lead to a pattern. That is why there are only three instead of four patterns in the set represented by Figure 1(a).

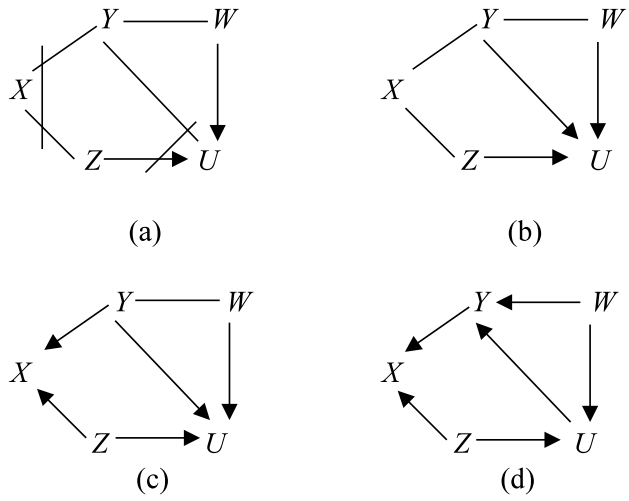

FIG. 1. (a) *is a sample output of the CSGS algorithm. The ambiguous (or unfaithful) unshielded triples are marked by straight lines crossing the two edges. There is no explicit mark for noncolliders, with the understanding that all and only unshielded triples that are not oriented as colliders or marked as ambiguous are (implicitly) marked noncolliders.* (b)–(d) *are the three patterns represented by* (a).

It is easy to see that when the Orientation-Faithfulness assumption happens to hold, the *CSGS* output will be a single pattern (i.e., without ambiguous triples), which is the same as the *SGS* output. In other words, *CSGS* is as informative as *SGS* when the stronger assumption needed for the output of the latter to be guaranteed to be correct happens to be true.

The Adjacency-Faithfulness assumption may be further weakened. In an earlier paper (Zhang and Spirtes, 2008), we showed that some violations of

the Adjacency-Faithfulness assumption are also detectable, and we specified some conditions weaker than the Adjacency-Faithfulness assumption under which any violation of Faithfulness (and so any violation of Adjacency-Faithfulness) is detectable. One of the weaker conditions is known as the Causal Minimality assumption (Spirtes, Glymour and Scheines, 1993, page 31), which states that the true causal DAG is a minimal DAG that satisfies the Markov condition with the true probability distribution, minimal in the sense that no proper subgraph satisfies the Markov condition. This condition is a consequence of the Adjacency-Faithfulness assumption. If the Adjacency-Faithfulness assumption is true, then no edge can be taken away from the true causal DAG without violating the Markov condition.

The other weaker condition is named Triangle-Faithfulness:

Triangle-Faithfulness assumption. Suppose the true causal DAG of $\mathbf{V}$ is G. Let X, Y, Z be any three variables that form a triangle in G (i.e., each pair of vertices is adjacent):

1. If Y is a noncollider on the path $\langle X, Y, Z\rangle$, then X and Z are not independent conditional on any subset of $\mathbf{V} \setminus \{X, Z\}$ that does not contain Y;
2. If Y is a collider on the path $\langle X, Y, Z\rangle$, then X and Z are not independent conditional on any subset of $\mathbf{V} \setminus \{X, Z\}$ that contains Y.

Clearly, the Adjacency-Faithfulness assumption entails the Triangle-Faithfulness assumption, and the latter, intuitively, is much weaker. Our result in Zhang and Spirtes (2008) is that given the Causal Markov, Minimality and Triangle-Faithfulness assumptions, any violation of faithfulness is detectable. But we did not propose any algorithm that is provably correct under the Markov, Minimality and Triangle-Faithfulness assumptions.

What need we modify in the *SGS* algorithm if all we can assume are the Markov, Minimality and Triangle-Faithfulness assumptions? In the step S2, the inferred adjacencies are still correct, which, as already mentioned, is guaranteed by the Causal Markov assumption alone. The inferred nonadjacencies, however, are not necessarily correct, because the Adjacency-Faithfulness assumption might fail. So the first modification we need make is to acknowledge that the nonadjacencies resulting from S2 are only "apparent" but not "definite": there might still be an edge between two variables even though the edge between them was removed in S2 because a screen-off set was found.

Since we do not assume the Orientation-Faithfulness assumption, obviously we need at least modify S3 into S3*. A further worry is that the unshielded triples resulting from S2 are only "apparent": they might be shielded in the true causal DAG but appear to be unshielded due to a failure of Adjacency-Faithfulness. Fortunately, this possibility does not affect the soundness of S3*. Take clause (i) for example. For an apparently unshielded triple $\langle X, Y, Z\rangle$, either X and Z are really nonadjacent in the true DAG or they are adjacent. In the former case, clause (i) is sound by the Markov assumption. In the latter case, clause (i) is still sound by the Triangle-Faithfulness assumption. A similar argument shows that clause (ii) is also sound. So S3* is still sound. Moreover, clause (iii) can now play a bigger role than simply conceding ignorance or ambiguity. If the antecedent of clause (iii) is satisfied, then one can infer that X and Z are really nonadjacent, for otherwise the Triangle-Faithfulness assumption would be violated no matter whether $\langle X, Y, Z\rangle$ is a collider or not.

The soundness of S4 is obviously not affected. Therefore, if we only assume the Causal Markov, Minimality and Triangle-Faithfulness assumptions, the *CSGS* algorithm is still correct if we take the nonadjacencies in its output as uninformative (except for those warranted by S3*).

The question now is whether we can somehow test the Adjacency-Faithfulness assumption in the procedure and confirm the nonadjacencies when the test returns affirmative. The following lemma gives a sufficient condition for verifying the Adjacency-Faithfulness assumption and hence the nonadjacencies in the *CSGS* output. (Recall that the *CSGS* output in general represents a set of patterns, and each pattern represents a set of Markov equivalent DAGs.) A pattern O is *Markov to an oracle* when for every DAG represented by O, each vertex is independent of the set of variables that are neither descendants nor parents in the DAG conditional on the parents in the DAG according to the oracle.

LEMMA 1. *Suppose the Causal Markov, Minimality and Triangle-Faithfulness assumptions are true, and E is the output of CSGS given a perfectly reliable oracle of conditional independence. If every pattern in the set represented by E is Markov to the oracle, then the true causal DAG has exactly those adjacencies present in E.*

PROOF. As we already pointed out, the true causal DAG, G_T, must have at least the adjacencies in E (in

order to satisfy the Causal Markov assumption), and must have the colliders and noncolliders in E (in order to satisfy the Causal Markov and the Triangle-Faithfulness assumptions). Now suppose every pattern in the set represented by E is Markov to the oracle, and suppose, for the sake of contradiction, that G_T has still more adjacencies. Let G be the proper subgraph of G_T with just the adjacencies in E. Then every unshielded collider and every unshielded noncollider in E are also present in G, and other unshielded triples in G, if any, are ambiguous in E. Thus, the pattern that represents the Markov equivalence class of G is in the set represented by E. It follows that G is Markov to the oracle, which shows that G_T is not a minimal graph that is Markov to the oracle. This contradicts the Causal Minimality assumption. Therefore, G_T has exactly the adjacencies present in E. □

So we have the following *Very Conservative SGS* (*VCSGS*):

VCSGS algorithm.

V1. Form the complete undirected graph H on the given set of variables $\mathbf{V}$.

V2. For each pair of variables X and Y in $\mathbf{V}$, search for a subset $\mathbf{S}$ of $\mathbf{V} \setminus \{X, Y\}$ such that X and Y are independent conditional on $\mathbf{S}$. Remove the edge between X and Y in H and mark the pair $\langle X, Y\rangle$ as "apparently nonadjacent," if and only if such a set is found.

V3. Let K be the graph resulting from V2. For each apparently unshielded triple $\langle X, Y, Z\rangle$ (i.e., X and Y are adjacent, Y and Z are adjacent, but X and Z are apparently nonadjacent),

(i) If X and Z are not independent conditional on any subset of $\mathbf{V} \setminus \{X, Z\}$ that contains Y, then orient the triple as a collider: $X \rightarrow Y \leftarrow Z$.

(ii) If X and Z are not independent conditional on any subset of $\mathbf{V} \setminus \{X, Z\}$ that does not contain Y, then mark the triple as a noncollider.

(iii) Otherwise, mark the triple as ambiguous (or unfaithful), and mark the pair $\langle X, Z\rangle$ as "definitely nonadjacent."

V4. Execute the same orientation rules as in S4, until none of them applies.

V5. Let M be the graph resulting from V4. For each consistent disambiguation of the ambiguous triples in M (i.e., each disambiguation that leads to a pattern), test whether the resulting pattern satisfies the Markov condition. If every pattern does, then mark all the "apparently nonadjacent" pairs as "definitely nonadjacent."

[An obvious way to test the Markov condition in V5 on a given pattern is to extend the pattern to a DAG and test the local Markov condition. That is, we need to test, for each variable X, whether X is independent of the variables that are neither its descendants nor its parents conditional on its parents. In linear Gaussian models, this can be done by regressing X on its nondescendants and testing whether the regression coefficients are zero for its nonparents. More generally, assuming composition, we need only run a conditional independence test for each nonadjacent pair, and, thus, in the worst case the number of conditional independence tests is $O(n^2)$, where n is the number of vertices. The number of patterns to be tested in V5 is $O(2^a)$, where a is the number of ambiguous unshielded triples.]

As we already explained, steps V1–V4 are sound under the Causal Markov, Minimality and Triangle-Faithfulness assumptions. Lemma 1 shows that V5 is also sound. Hence, the *VCSGS* algorithm is correct under the Causal Markov, Minimality and Triangle-Faithfulness assumptions, in the sense that given a perfectly reliable oracle of conditional independence, all the adjacencies, definite nonadjacencies, directed edges and marked noncolliders are correct. Moreover, when the Causal Faithfulness assumption happens to hold, the *CSGS* output will be a single pattern and this single pattern will satisfy the Markov condition; hence, the *VCSGS* algorithm will return a single pattern with full information about nonadjacencies. Therefore, *VCSGS* is also as informative as *SGS* when the Causal Faithfulness assumption happens to be true.

One might think (or hope) that the *VCSGS* algorithm is as informative as the *CSGS* algorithm when Adjacency-Faithfulness (but not Orientation-Faithfulness) happens to hold. Unfortunately this is not true in general because the sufficient condition given in Lemma 1 (and checked in V5) is not necessary for the Adjacency-Faithfulness assumption.

To illustrate, consider the following example. Suppose the true causal DAG is the one given in Figure 2(a). Suppose the causal Markov assumption and the Adjacency-Faithfulness assumption are satisfied. And suppose that, besides the conditional independence relations entailed by the graph, the true distribution features one and only one extra conditional independence: $I(X, Z|Y)$, due, for example, to some sort of balancing-out of the path $\langle X, Y, Z\rangle$ (active conditional on $\{Y\}$) and the path $\langle X, W, Z\rangle$ (active conditional on $\{Y\}$). This violates the Orientation-Faithfulness assumption. The *CSGS* output will thus be the graph in

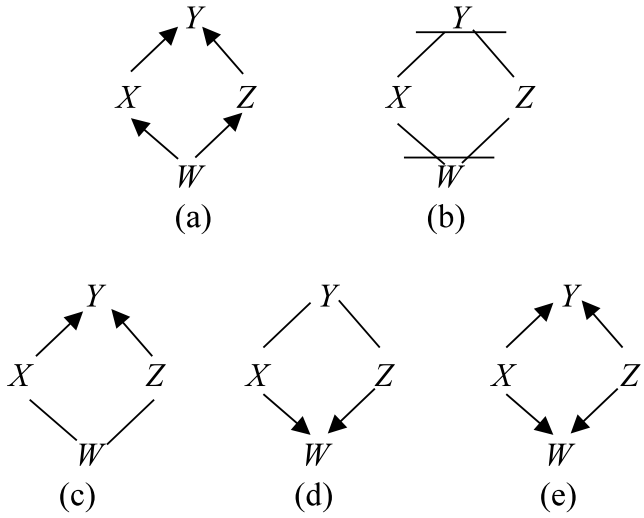

FIG. 2. *An example in which the test in step V5 of VCSGS does not confirm the nonadjacencies even though the nonadjacencies are correct.*

Figure 2(b), in which both the triple $\langle X, Y, Z\rangle$ and the triple $\langle X, W, Z\rangle$ are ambiguous. This output represents a set of three patterns, as shown in Figure 2(c)–(e). (Again, the two ambiguous triples cannot be noncolliders at the same time.) However, only the patterns in Figure 2(c) and 2(d) satisfy the Markov condition. The pattern in Figure 2(e) violates the Markov condition because it entails that $I(X, Z|\varnothing)$, which is not true.

For this example, then, the *VCSGS* will not return the full information of nonadjacencies, even though the Adjacency-Faithfulness assumption is true.

In light of this example, it is natural to consider the following variant of step V5 in *VCSGS*:

V5*. Let M be the graph resulting from V4. If *some* disambiguation of the ambiguous triples in M leads to a pattern that satisfies the Markov condition, then mark all remaining "apparently nonadjacent" pairs as "definitely nonadjacent."

We suspect that V5* is also sound under the Causal Markov, Minimality and Triangle-Faithfulness assumptions, but we have not found a proof. In other words, we conjecture that the sufficient condition presented in Lemma 1 can be weakened to that *some* pattern in the set represented by the *CSGS* output satisfies the Markov condition. (This conjecture is a consequence of the following plausible conjecture: Suppose a DAG G and a probability distribution P satisfy the Markov, Minimality and Triangle-Faithfulness conditions. Then no DAG with strictly fewer adjacencies than in G is Markov to P. We thank an anonymous referee for making the point and the conjecture.) Note that if the Adjacency-Faithfulness assumption happens to hold, then at least one pattern (i.e., the pattern representing the true causal DAG) satisfies the Markov condition. Therefore, if our conjecture is true, we can replace V5 with V5* in the *VCSGS* algorithm, and the condition tested in V5* is both sufficient and necessary for Adjacency-Faithfulness. The resulting algorithm will then be as informative as the *CSGS* algorithm whenever the Adjacency-Faithfulness assumption happens to hold, and as informative as the *SGS* algorithm whenever both the Adjacency-Faithfulness assumption and the Orientation-Faithfulness assumption happen to hold.

It is worth noting that if we adopt a natural, interventionist conception of causation (e.g., Woodward, 2003), the Causal Minimality assumption is guaranteed to be true if the probability distribution is positive (Zhang and Spirtes, 2011). Since positivity is a property of the probability distribution alone, we may also try to incorporate a test of positivity at the beginning of *VCSGS*, and proceed only if the test returns affirmative. We then need not assume the Causal Minimality assumption in order to justify the procedure.

4. WEAKENING THE STRONG CAUSAL FAITHFULNESS ASSUMPTION

In this section we consider sample versions of the *CSGS* and *VCSGS* algorithms, assuming Gaussianity and linearity, and prove some positive results on uniform consistency, under a generalization and strengthening of the Triangle-Faithfulness assumption, which we call the k-Triangle-Faithfulness assumption.

If a model M does not satisfy the Causal Faithfulness assumption, then M contains a zero partial correlation $\rho_M(X, Y|\mathbf{W})$ even though the Causal Markov assumption does not entail that $\rho_M(X, Y|\mathbf{W})$ is zero. If $\rho_M(X, Y|\mathbf{W}) = 0$ but is not entailed to be zero for all values of the parameters, the parameters of the model satisfy an algebraic constraint. A set of parameters that satisfies such an algebraic constraint is a "surface of unfaithfulness" in the parameter space that is of a lower dimension than the full parameter space. Lying on such a surface of unfaithfulness is of Lebesgue measure zero. For a Bayesian with a prior probability over the parameter space that is absolutely continuous with Lebesgue measure, the prior probability of unfaithfulness is zero.

However, in practice, the *SGS* (or *PC*) algorithm does not have access to the population correlation coefficients. Instead it performs statistical tests of whether a partial correlation is zero. If $|\rho_M(X, Y|\mathbf{W})|$ is small

enough, then with high probability a statistical test of whether $\rho_M(X, Y|\mathbf{W})$ equals zero will not reject the null hypothesis. If $\rho_M(X, Y|\mathbf{W}) = 0$ fails to be rejected, this can lead to some edges that occur in the true causal DAG not appearing in the output of *SGS* and to errors in the orientation of edges in the output of *SGS*. (Such errors can also lead to the output of the *SGS* algorithm to fail to be a pattern, either because it contains double-headed edges or undirected nonchordal cycles.) Robins et al. (2003) showed that even if it is assumed that there are no unfaithful models, there are always models so "close to unfaithful" [i.e., with $|\rho_M(X, Y|\mathbf{W})|$ nonzero but small enough that a statistical test will probably fail to reject the null hypothesis] that there is no algorithm that is a uniformly consistent estimator of the pattern of a causal model.

Kalisch and Bühlmann (2007) showed that under a strengthened version of the Causal Faithfulness assumption, the *PC* algorithm is a uniformly consistent estimator of the pattern that represents the true causal DAG. Let n be the sample size. Their strengthened set of assumptions were as follows:

(A1) The distribution P_n is multivariate Gaussian and faithful to the DAG G_n for all n.

(A2) The dimension $p_n = O(n^a)$ for some $0 \le a < \infty$.

(A3) The maximal number of neighbors in the DAG G_n is denoted by

$$q_n = \max_{1 \le j \le p_n} \bigl|\mathrm{adj}(G, j)\bigr|$$

$$\text{with } q_n = O\bigl(n^{1-b}\bigr) \text{ for some } 0 < b \le 1.$$

(A4) The partial correlations between $\mathbf{X}(i)$ and $\mathbf{X}(j)$ given $\{\mathbf{X}(r); r \in \mathbf{k}\}$ for some set $\mathbf{k} \subseteq \{1, \ldots, p_n\} \setminus \{i, j\}$ are denoted by $\rho_{n;i,j|\mathbf{k}}$. Their absolute values are bounded from below and above:

$$\inf\bigl\{|\rho_{i,j|\mathbf{k}}|; i, j, \mathbf{k} \text{ with } \rho_{i,j|\mathbf{k}} \ne 0\bigr\} \ge c_n,$$

$$c_n^{-1} = O\bigl(n^d\bigr),$$

$$\text{for some } 0 < d < b/2,$$

$$\sup_{n;i,j,\mathbf{k}} |\rho_{i,j|\mathbf{k}}| \le M < 1,$$

$$\text{where } 0 < b \le 1 \text{ is as in (A3).}$$

We will refer to the assumption that all nonzero partial correlations are bounded below in absolute value by a number greater than zero [as in the first part of (A4)] as the Strong Causal Faithfulness assumption. Uhler et al. (2013) provide some reason to believe that unless c_n is quite small, the probability of violating Strong Causal Faithfulness assumption is high, especially when the number of variables is large. [This problem with assumption (A4) is somewhat mitigated by the fact that the size of c_n can decrease with increasing sample size. But see Lin et al. (2012), for an interesting analysis of the asymptotics when c_n approaches zero.]

It is difficult to see how a uniformly consistent estimator of a causal pattern would be possible without assuming something like the Strong Causal Faithfulness assumption. However, what we will show is that it is possible to weaken the Strong Causal Faithfulness assumption in several ways as long as the standard of success is not finding a uniformly consistent estimator of the causal pattern, but is instead finding a uniformly consistent estimator of (some of) the structural coefficients in a pattern. The latter standard is compatible with missing some edges that are in the true causal graph, as long as the edges that have not been included in the output have sufficiently small structural coefficients.

We propose to replace the faithfulness assumption in (A1), and the Strong Faithfulness assumption with the following assumption, where $e_M(X \text{ — } Z)$, as we explained in Section 2, denotes the structural coefficient associated with the edge between X and Z.

k-Triangle-Faithfulness assumption. Given a set of variables $\mathbf{V}$, suppose the true causal model over $\mathbf{V}$ is $M = \langle P, G\rangle$, where P is a Gaussian distribution over $\mathbf{V}$, and G is a DAG with vertices $\mathbf{V}$. For any three variables X, Y, Z that form a triangle in G (i.e., each pair of vertices is adjacent),

1. If Y is a noncollider on the path $\langle X, Y, Z\rangle$, then $|\rho_M(X, Z|\mathbf{W})| \ge k \times |e_M(X \text{ — } Z)|$ for all $\mathbf{W} \subseteq \mathbf{V}$ that do not contain Y; and
2. If Y is a collider on the path $\langle X, Y, Z\rangle$, then $|\rho_M(X, Z|\mathbf{W})| \ge k \times |e_M(X \text{ — } Z)|$ for all $\mathbf{W} \subseteq \mathbf{V}$ that do contain Y.

As k approaches 0, the k-Triangle-Faithfulness assumption approaches the Triangle-Faithfulness assumption. For (small) $k > 0$, the k-Triangle-Faithfulness assumption prohibits not only exact cancellations of active paths in a triangle, but also *almost* cancellations.

The k-Triangle-Faithfulness assumption is a weakening of the Strong Causal Faithfulness assumption in two ways. First, Triangle-Faithfulness is significantly weaker than Faithfulness. Second, it does not entail a lower limit on the size of nonzero partial correlations; it

only puts a limit on the size of a nonzero partial correlation in relation to the size of the structural coefficient of an edge that occurs in a triangle.

The Strong Causal Faithfulness assumption entails that there are no very small structural coefficients (which, if present, entail the existence of some partial correlation that is very small). In contrast, the k-Triangle-Faithfulness assumption does not entail that there are no nonzero but very small structural coefficients. However, there is a price to be paid for weakening the Strong Causal Faithfulness assumption; the estimator we propose is both computationally more intensive than the PC algorithm used in Kalisch and Bühlmann (2007) and also requires testing partial correlations conditional on larger sets of variables, which means some of the tests performed have lower power than the tests performed in the PC algorithm.

Our results also depend on the following assumptions. First, we assume a fixed upper bound to the size of the set of variables that does not change as sample size increases. We have no reason to think that there are not analogous results that would hold even if, as in Kalisch and Bühlmann (2007), the number of variables and the degree of the graph increased with the sample size; however, we have not proved any such results yet. We also make the assumption of nonvanishing variance (NVV) and the assumption of upper bound for partial correlations (UBC):

Assumption NVV(J).

$$\inf_{X_i \in \mathbf{V}} \operatorname{var}_M\big(X_i | \mathbf{V} \setminus \{X_i\}\big) \geq J$$

for some (small) $J > 0$.

Assumption UBC(C).

$$\sup_{X_i, X_j \in \mathbf{V}, \mathbf{W} \subseteq \mathbf{V} \setminus \{X_i, X_j\}} \big|\rho_M(X_i, X_j | \mathbf{W})\big| \leq C$$

for some $C < 1$.

The assumption NVV is a slight strengthening of the positivity requirement, which, as we noted in the previous section, is needed to guarantee the Causal Minimality assumption. Uniform consistency requires that the distributions be bounded away from nonpositivity.

The assumption UBC [cf. the second part of assumption (A4)] is used to guarantee that sample partial correlations are uniformly consistent estimators of population partial correlations (Kalisch and Bühlmann, 2007).

We now proceed to establish two positive results about uniform consistency. In Section 4.1 we show that the *Conservative SGS* (*CSGS*) algorithm, using uniformly consistent tests of partial correlations, is uniformly consistent in inferring certain features of the causal structure. In Section 4.2 we show that the *Very Conservative SGS* (*VCSGS*) algorithm, when combined with a uniformly consistent procedure for estimating structural coefficients, provides a uniformly consistent estimator of structural coefficients (that returns "Unknown" in some, but not all cases).

4.1 Uniform Consistency in the Inference of Structure

Recall that the *CSGS* algorithm, given a perfect oracle of conditional independence, is correct under the Causal Markov, Minimality and Triangle-Faithfulness assumptions, in the sense that the adjacencies, orientations and marked noncolliders in the output are all correct. In Gaussian models, we can implement the oracle with tests of zero partial correlations. A test φ of $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ is a family of functions: $\varphi_1, \ldots, \varphi_n, \ldots$, one for each sample size, that takes an i.i.d. sample V_n from the joint distribution over $\mathbf{V}$ and returns 0 (acceptance of H_0) or 1 (rejection of H_0). Such a test is *uniformly consistent* with respect to a set of distributions Ω if and only if

1. $\lim_{n\to\infty} \sup_{P \in \Omega \wedge \rho(P)=0} P^n(\varphi_n(V_n) = 1) = 0$, and
2. for every $\delta > 0$,

$$\lim_{n\to\infty} \sup_{P \in \Omega \wedge |\rho(P)| \geq \delta} P^n\big(\varphi_n(V_n) = 0\big) = 0.$$

For simplicity, we assume the variables in $\mathbf{V}$ are standardized. Under the assumption UBC, there are uniformly consistent tests of partial correlations based on sample partial correlations, such as Fisher's z test (Robins et al., 2003; Kalisch and Bühlmann, 2007). We consider a sample version of the *CSGS* algorithm in which the oracle is replaced by uniformly consistent tests of zero partial correlations in the adjacency step S2. In the orientation phase, the step S3* is refined as follows, based on a user chosen parameter L.

S3* (sample version). Let K be the undirected graph resulting from the adjacency phase. For each unshielded triple $\langle X, Y, Z \rangle$,

1. If there is a set $\mathbf{W}$ not containing Y such that the test of $\rho(X, Z|\mathbf{W}) = 0$ returns 0 (i.e., accepts the hypothesis), and for every set $\mathbf{U}$ that contains Y, the test of $|\rho(X, Z|\mathbf{U})| = 0$ returns 1 (i.e., rejects the hypothesis), and the test of $|\rho(X, Z|\mathbf{U}) - \rho(X, Z|\mathbf{W})| \geq L$ returns 0 (i.e., accepts the hypothesis), then orient the triple as a collider: $X \to Y \leftarrow Z$.
2. If there is a set $\mathbf{W}$ containing Y such that the test of $\rho(X, Z|\mathbf{W}) = 0$ returns 0 (i.e., accepts the hypothesis), and for every set $\mathbf{U}$ that does not contain Y, the test

of $|\rho(X, Z|\mathbf{U})| = 0$ returns 1 (i.e., rejects the hypothesis), and the test of $|\rho(X, Z|\mathbf{U}) - \rho(X, Z|\mathbf{W})| \geq L$ returns 0 (i.e., accepts the hypothesis), then mark the triple as a noncollider.

3. Otherwise, mark the triple as ambiguous.

Larger values of L return "Unknown" more often than smaller values of L, but reduce the probability of an error in orientation at a given sample size.

Step S4 remains the same as in the population version.

Given any causal model $M = \langle P, G\rangle$ over $\mathbf{V}$, let $C(L, n, M)$ denote the (random) output of the *CSGS* algorithm with parameter L, given an i.i.d. sample of size n from the distribution P_M. Say that $C(L, n, M)$ errs if it contains (i) an adjacency not in G_M, or (ii) a marked noncollider not in G_M, or (iii) an orientation not in G_M.[1]

Let $\psi^{k,J,C}$ be the set of causal models over $\mathbf{V}$ that respect the k-Triangle-Faithfulness assumption and the assumptions of NVV(J) and UBC(C). We shall prove that given the causal sufficiency of the measured variables $\mathbf{V}$ and the causal Markov assumption,

$$\lim_{n\to\infty} \sup_{M\in\psi^{k,J,C}} P_M^n\big(C(L, n, M) \text{ errs}\big) = 0.$$

In other words, given the causal sufficiency of $\mathbf{V}$, the Causal Markov, k-Triangle-Faithfulness, NVV(J) and UBC(C) assumptions, the *CSGS* algorithm is uniformly consistent in that the probability of it making a mistake uniformly converges to zero in the large sample limit.

First of all, we prove a useful lemma:

LEMMA 2. *Let $M \in \psi^{k,J,C}$. For any X_i and X_j such that X_j is not an ancestor of X_i, if $e_M(X_i \to X_j) = b_{j,i}$, then*

$$\frac{|b_{j,i}|}{\sqrt{J}} \geq \big|\rho_M\big(i, j|\mathbf{X}[1, \ldots, j-1] \setminus \{X_i\}\big)\big| \geq |b_{j,i}|\sqrt{J},$$

where $\mathbf{X}[1, \ldots, j]$ is an ancestral set that contains X_i but does not contain any descendant of X_j.

PROOF. Let Σ be the correlation matrix for the set of variables $\{X_1, \ldots, X_j\}$, and $\mathbf{R} = \Sigma^{-1}$. Let $\mathbf{B}$ be the (lower-triangular) matrix of structural coefficients in M restricted to $\{X_1, \ldots, X_j\}$, and $\text{var}(\mathbf{E})$ be the (diagonal) covariance matrix for the error terms $\{\varepsilon_1, \ldots, \varepsilon_j\}$. Then

$$\mathbf{R} = (\mathbf{I} - \mathbf{B})^T \,\text{var}(\mathbf{E})^{-1}(\mathbf{I} - \mathbf{B}).$$

Note that

$$(\mathbf{I} - \mathbf{B}) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -b_{2,1} & 1 & \cdots & 0 \\ \vdots & \cdots & \cdots & 0 \\ -b_{j,1} & \cdots & -b_{j,j-1} & 1 \end{bmatrix} \text{var}(\mathbf{E})^{-1}$$
$$= \begin{bmatrix} 1/\varepsilon_1 & 0 & \cdots & 0 \\ 0 & 1/\varepsilon_2 & \cdots & 0 \\ \vdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1/\varepsilon_j \end{bmatrix},$$

where the b's are the corresponding structural coefficients in M, and the ε's are the variances of the corresponding error terms. Thus, $\mathbf{R}[j, j] = 1/\varepsilon_j$, and $\mathbf{R}[i, j] = -b_{j,i}/\varepsilon_j$. So we have (Whittaker, 1990)

$$\rho_M\big(X_i, X_j|\mathbf{X}[1, \ldots, j-1] \setminus \{X_i\}\big) = -\frac{\mathbf{R}[i, j]}{(\mathbf{R}[i, i] \cdot \mathbf{R}[j, j])^{1/2}} = \frac{b_{j,i}}{\mathbf{R}[i, i]^{1/2}\varepsilon_j^{1/2}}.$$

Since $\mathbf{R}[i, i]^{-1}$ is the variance of X_i conditional on all of the other variables in $\{X_1, \ldots, X_j\}$, which is a subset of $\mathbf{V} \setminus \{X_i\}$, $\mathbf{R}[i, i]^{-1} \geq \text{var}_M(X_i|\mathbf{V} \setminus \{X_i\}) \geq J$. Since the variables are standardized and the residual of X_i regressed on the other variables is uncorrelated with X_i, $\mathbf{R}[i, i]^{-1} \leq 1$. Similarly, $1 \geq \varepsilon_j \geq J$. Thus,

$$\frac{|b_{j,i}|}{\sqrt{J}} \geq \big|\rho_M\big(i, j|\mathbf{X}[1, \ldots, j-1] \setminus \{X_i\}\big)\big| \geq |b_{j,i}|\sqrt{J}.$$

□

We now categorize the mistakes $C(L, n, M)$ can make into three kinds. $C(L, n, M)$ *errs in kind I* if $C(L, n, M)$ has an adjacency that is not present in G_M; $C(L, n, M)$ *errs in kind II* if every adjacency in $C(L, n, M)$ is in G_M but $C(L, n, M)$ contains a marked noncollider that is not in G_M; $C(L, n, M)$ errs in *kind III* if every adjacency in $C(L, n, M)$ is in G_M, every marked noncollider in $C(L, n, M)$ is in G_M, but $C(L, n, M)$ contains an orientation that is not in G_M. Obviously if $C(L, n, M)$ errs, it errs in at least one of the three kinds.

The following three lemmas show that for each kind, the probability of $C(L, n, M)$ erring in that kind uniformly converges to zero.

[1] Note that at this stage we are taking non-adjacencies as uninformative, and not counting any missing edge as an error. So an algorithm that always returns a structure with no edges is treated as totally uninformative and hence trivially consistent, in the sense of triviality defined in Robins et al. (2003). The CSGS algorithm is obviously nontrivial in that it does not always return a completely uninformative answer.

LEMMA 3. *Given causal sufficiency of the measured variables* **V**, *the Causal Markov*, *k-Triangle-Faithfulness*, *NVV*(*J*) *and UBC*(*C*) *assumptions*,

$$\lim_{n\to\infty} \sup_{M\in\psi^{k,J,C}} P_M^n\big(C(L,n,M) \text{ errs in kind } I\big) = 0.$$

PROOF. $C(L,n,M)$ has an adjacency not in G_M only if some test of zero partial correlation falsely rejects its null hypothesis. Since uniformly consistent tests are used in *CSGS*, for every $\varepsilon > 0$, for every test of zero partial correlation t_i, there is a sample size N_i such that for all $n > N_i$ the supremum (over $\psi^{k,J,C}$) of the probability of the test falsely rejecting its null hypothesis is less than ε. Given **V**, there are only finitely many possible tests of zero partial correlations. Thus, for every $\varepsilon > 0$, there is a sample size N such that for all $n > N$, the supremum (over $\psi^{k,J,C}$) of the probability of any of the tests falsely rejecting its null hypothesis is less than ε. The lemma then follows. □

LEMMA 4. *Given causal sufficiency of the measured variables* **V**, *the Causal Markov*, *k-Triangle-Faithfulness*, *and NVV*(*J*) *and UBC*(*C*) *assumptions*,

$$\lim_{n\to\infty} \sup_{M\in\psi^{k,J,C}} P_M^n\big(C(L,n,M) \text{ errs in kind } II\big) = 0.$$

PROOF. For any $M \in \psi^{k,J,C}$, if $C(L,n,M)$ errs in kind II, then $C(L,n,M)$ contains a marked noncollider, say, $\langle X,Y,Z\rangle$ which is not in G_M, but every adjacency in $C(L,n,M)$ is also in G_M, including the adjacency between X and Y, and that between Y and Z. It follows that $\langle X,Y,Z\rangle$ is a collider in G_M. Since *CSGS* marks a triple as a noncollider only if the triple is unshielded, X and Z are not adjacent in $C(L,n,M)$. Hence, errors of kind II can be further categorized into two cases: (II.1) $C(L,n,M)$ contains an unshielded noncollider that is an *unshielded* collider in G_M, and (II.2) $C(L,n,M)$ contains an unshielded noncollider that is a *shielded* collider in G_M. We show that the probability of either case uniformly converges to zero.

For case (II.1) there is an unshielded collider $\langle X,Y,Z\rangle$ in G_M, so X and Z are independent conditional on some set of variables **W** that does not contain Y, by the Causal Markov assumption. Then the *CSGS* algorithm (falsely) marks $\langle X,Y,Z\rangle$ as a noncollider only if the test of $\rho_M(X,Z|\mathbf{W}) = 0$ (falsely) rejects its null hypothesis. Therefore, the *CSGS* algorithm gives rise to case (II.1) only if some test of zero partial correlation falsely rejects its null hypothesis. Then, by essentially the same argument as the one used in proving Lemma 3, the probability of case (II.1) uniformly converges to zero as sample size increases.

For case (II.2), suppose for the sake of contradiction that the probability of *CSGS* making such a mistake does not uniformly converge to zero. Then there exists $\varepsilon > 0$, such that for every sample size n, there is a model $M(n)$ such that the probability of $C(L,n,M(n))$ contains an unshielded noncollider that is a shielded collider in $M(n)$ is greater than ε.

Now, $C(L,n,M(n))$ contains an unshielded noncollider that is a shielded collider in $G_{M(n)}$, say $\langle X^{M(n)}, Y^{M(n)}, Z^{M(n)}\rangle$, only if there is a set $\mathbf{W}^{M(n)}$ that contains Y such that the test of $\rho(X^{M(n)}, Z^{M(n)}|\mathbf{W}^{M(n)}) = 0$ returns 0 (i.e., accepts the hypothesis).

Without loss of generality, suppose $Z^{M(n)}$ is not an ancestor of $X^{M(n)}$. Let $\mathbf{U}^{M(n)} = \mathbf{A}^{M(n)} \setminus \{X^{M(n)}, Z^{M(n)}\}$, where $\mathbf{A}^{M(n)}$ is an ancestral set that contains $X^{M(n)}$ and $Z^{M(n)}$ but no descendent of $Z^{M(n)}$. Since $Y^{M(n)}$ is a child of $Z^{M(n)}$ in $G_{M(n)}$, $\mathbf{U}^{M(n)}$ does not contain $Y^{M(n)}$. Then, $\langle X^{M(n)}, Y^{M(n)}, Z^{M(n)}\rangle$ is marked as a noncollider in $C(L,n,M(n))$ only if the test of $|\rho(X^{M(n)}, Z^{M(n)}|\mathbf{U}^{M(n)}) - \rho(X^{M(n)}, Z^{M(n)}|\mathbf{W}^{M(n)})| \geq L$ returns 0 (i.e., accepts the hypothesis).

The test of $|\rho(X^{M(n)}, Z^{M(n)}|\mathbf{U}^{M(n)}) - \rho(X^{M(n)}, Z^{M(n)}|\mathbf{W}^{M(n)})| \geq L$ will be denoted by $\varphi_{n(L)}$ and $\varphi_{n(0)}$ denotes the test of $\rho(X^{M(n)}, Z^{M(n)}|\mathbf{W}^{M(n)}) = 0$. By our supposition, $P^n_{M(n)}(\varphi_{n(0)} = 0 \text{ and } \varphi_{n(L)} = 0) > \varepsilon$. It follows that for all n,

(1) $P^n_{M(n)}(\varphi_{n(0)} = 0) > \varepsilon$,
(2) $P^n_{M(n)}(\varphi_{n(L)} = 0) > \varepsilon$.

(1) implies that there exists δ_n such that $|\rho(X^{M(n)}, Z^{M(n)}|\mathbf{W}^{M(n)})| < \delta_n$, and $\delta_n \to 0$ as $n \to \infty$ since the tests are uniformly consistent. $|e_M(X^{M(n)} — Z^{M(n)})| \leq |\rho(X^{M(n)}, Z^{M(n)}|\mathbf{W}^{M(n)})|/k < \delta_n/k$ by *k*-Triangle-Faithfulness. By Lemma 2, $|\rho(X^{M(n)}, Z^{M(n)}|\mathbf{U}^{M(n)})| \leq J^{-1/2}|e_M(X^{M(n)} — Z^{M(n)})| < \delta_n J^{-1/2}/k$. Thus, $|\rho(X^{M(n)}, Z^{M(n)}|\mathbf{U}^{M(n)}) - \rho(X^{M(n)}, Z^{M(n)}|\mathbf{W}^{M(n)})| < \delta_n(1 + J^{-1/2}/k) \to 0$ as $n \to \infty$. Therefore, it is not true that (2) holds for all n, which is a contradiction. So the initial supposition is false. The probability of case (II.2) uniformly converges to zero as sample size increases. □

LEMMA 5. *Given causal sufficiency of the measured variables* **V**, *the Causal Markov*, *k-Triangle-Faithfulness*, *NVV*(*J*) *and UBC*(*C*) *assumptions*,

$$\lim_{n\to\infty} \sup_{M\in\psi^{k,J,C}} P_M^n\big(C(L,n,M) \text{ errs in kind } III\big) = 0.$$

PROOF. Given that all the adjacencies and marked noncolliders in $C(L, n, M)$ are correct, there is a mistaken orientation if and only if there is an unshielded collider in $C(L, n, M)$ which is not a collider in G_M, for the other orientation rules in step S4 would not lead to any mistaken orientation if all the unshielded colliders were correct. Thus, $C(L, n, M)$ errs in kind III only if there is a noncollider $\langle X, Y, Z\rangle$ in G_M that is marked as an unshielded collider in $C(L, n, M)$.

There are then two cases to consider: (III.1) $C(L, n, M)$ contains an unshielded collider that is an *unshielded* noncollider in G_M, and (III.2) $C(L, n, M)$ contains an unshielded collider that is a *shielded* noncollider in G_M. The argument for case (III.1) is extremely similar to that for (II.1) in the proof of Lemma 4, and the argument for case (III.2) is extremely similar to that for (II.2) in the proof of Lemma 4. □

THEOREM 1. *Given causal sufficiency of the measured variables* **V**, *the Causal Markov*, *k-Triangle-Faithfulness*, *NVV*(*J*) *and UBC*(*C*) *assumptions*, *the CSGS algorithm is uniformly consistent in the sense that*

$$\lim_{n\to\infty} \sup_{M\in\psi^{k,J,C}} P_M^n\big(C(L, n, M) \text{ errs}\big) = 0.$$

PROOF. It follows from Lemmas 3–5 [and the fact that $C(L, n, M)$ errs if and only if it errs in one of the three kinds]. □

4.2 Uniform Consistency in the Inference of Structural Coefficients

We now combine the structure search with estimation of structural coefficients, when possible.

Edge Estimation algorithm.

E1. Run the *CSGS* algorithm on an i.i.d. sample of size n from P_M.

E2. Let the output from E1 be $C(L, n, M)$. Apply step V5 in the *VCSGS* algorithm (from Section 3), using tests of zero partial correlations.

E3. If the nonadjacencies in $C(L, n, M)$ are not confirmed in E2, return "Unknown" for every pair of variables.

E4. If the nonadjacencies in $C(L, n, M)$ are confirmed in E2, then

(i) For every nonadjacent pair $\langle X, Y\rangle$, let the estimate $\hat{e}(X \text{ — } Y)$ be 0.

(ii) For each vertex Z such that all of the edges containing Z are oriented in $C(L, n, M)$, if Y is a parent of Z in $C(L, n, M)$, let the estimate $\hat{e}(Y \text{ — } Z)$ be the sample regression coefficient of Y in the regression of Z on its parents in $C(L, n, M)$.

(iii) For any of the remaining edges, return "Unknown."

The basic idea is that we first run the Very Conservative SGS (*VCSGS*) algorithm, which, recall, is the *CSGS* algorithm (E1) plus a step of testing whether the output satisfies the Markov condition (E2). If the test does not pass, we do not estimate any edge; if the test passes, we estimate those edges that are into a vertex that is not part of any unoriented edge.

Let M_1 be an output of the Edge Estimation algorithm, and M_2 be a causal model. We define the *structural coefficient distance*, $d[M_1, M_2]$, between M_1 and M_2 to be

$$d[M_1, M_2] = \max_{i,j} |\hat{e}_{M_1}(X_i \to X_j) - e_{M_2}(X_i \to X_j)|,$$

where by convention $|\hat{e}_{M_1}(X_i \to X_j) - e_{M_2}(X_i \to X_j)| = 0$ if $\hat{e}_{M_1}(X_i \to X_j) =$ "Unknown."

Intuitively, the structural coefficient distance between the output and the true causal model measures the (largest) estimation error the Edge Estimation algorithm makes. Our goal is to show that under the specified assumptions, the Edge Estimation algorithm is uniformly consistent, in the sense that for every $\delta > 0$, the probability of the structural coefficient distance between the output and the true model being greater than δ uniformly converges to zero.

Obviously, by placing no penalty on the uninformative answer of "Unknown," there is a trivial algorithm that is uniformly consistent, namely, the algorithm that always returns "Unknown" for every structural coefficient. For this reason, Robins et al. (2003) also requires any admissible algorithm to be nontrivial in the sense that it returns an informative answer (in the large sample limit) for some possible joint distributions. The Edge Estimation algorithm is clearly nontrivial in this sense. There is no guarantee that it will always output an informative answer for some structural coefficient, and rightly so, because there are cases—for example, when the true causal graph is a complete one and there is no prior information about the causal order—in which every structural coefficient is truly underdetermined or unidentifiable. An interesting question, however, is whether a given algorithm is maximally informative or complete in the sense that it returns (in the large sample limit) "Unknown" only on

those structural coefficients that are truly underdetermined. The condition in question is of course much stronger than Robins et al.'s condition of nontriviality. We suspect that the Edge Estimation algorithm is not maximally informative in this sense. (We thank an anonymous referee for raising this issue.)

THEOREM 2. *Given causal sufficiency of the measured variables* **V**, *the Causal Markov*, *k-Triangle-Faithfulness*, *NVV*(J) *and UBC*(C) *assumptions*, *the Edge Estimation algorithm is uniformly consistent in the sense that for every* $\delta > 0$

$$\lim_{n\to\infty} \sup_{M\in\psi^{k,J,C}} P_M^n\big(d\big[\hat{O}(M), M\big] > \delta\big) = 0,$$

where $\hat{O}(M)$ *is the output of the algorithm given an i.i.d. sample from* P_M.

PROOF. Let **O** be the set of possible graphical outputs of the *CSGS* algorithm. Given **V**, there are only finitely many graphs in **O**. So it suffices to show that for each $O \in \mathbf{O}$,

$$\lim_{n\to\infty} \sup_{M\in\psi^{k,J,C}} P_M^n\big(d\big[\hat{O}(M), M\big] > \delta| C(L,n,M) = O\big) \cdot P_M^n\big(C(L,n,M) = O\big) = 0.$$

Given O, $\psi^{k,J,C}$ can be partitioned into the following three sets:

$\Psi_1 = \{M|$All adjacencies, nonadjacencies and orientations in O are true of $M\}$;

$\Psi_2 = \{M|O$ contains an adjacency or an orientation not true of $M\}$;

$\Psi_3 = \{M|$All adjacencies and orientations in O are true of M, but some nonadjacencies are not true of $M\}$.

It suffices to show that for each Ψ_i,

$$\lim_{n\to\infty} \sup_{M\in\psi_i} P_M^n\big(d\big[\hat{O}(M), M\big] > \delta|C(L,n,M) = O\big) \cdot P_M^n\big(C(L,n,M) = O\big) = 0.$$

Consider Ψ_1 first. Given any $M \in \Psi_1$, the zero estimates in $\hat{O}(M)$ are all correct (since all nonadjacencies are true). For each edge $Y \to Z$ that is estimated, the true structural coefficient $e_M(Y \to Z)$ is simply $r_M(Y, Z, \mathbf{Parents}(O, Z))$, the population regression coefficient for Y when Z is regressed on its parents in O, because the set of Z's parents in O is the same as the set of Z's parents in G_M.

The sampling distribution of the estimate of an edge $X \to Y$ in O is given by

$$\hat{r}_M\big(Y, Z, \mathbf{Parents}(O, Z), n\big) \sim \mathcal{N}\bigg(r_M\big(Y, Z, \mathbf{Parents}(O, Z)\big), \frac{\sigma_e^2}{n\,\mathrm{var}(Y|\mathbf{Parents}(O, Z) \setminus \{Y\})}\bigg),$$

where σ_e^2 is the variance of the residual for Z when regressed upon $\mathbf{Parents}(O, Z)$ in P_M, and $\mathrm{var}(Y|\mathbf{Parents}(O, Z) \setminus \{Y\})$ is the variance of Y conditional on $\mathbf{Parents}(O, Z) \setminus Y$ in P_M (Whittaker, 1990). The numerator of the variance is bounded above by 1, since the variance of each variable is 1, and the residual is independent of the set of variables regressed on. The denominator is bounded away from zero by assumption NVV(J). Hence, sample regression coefficients are uniformly consistent estimators of population regression coefficients under our assumptions, and we have

$$\begin{aligned}&\lim_{n\to\infty} \sup_{M\in\psi_1} P_M^n\big(d\big[\hat{O}(M), M\big] > \delta|C(L,n,M) = O\big) \cdot P_M^n\big(C(L,n,M) = O\big)\\ &\le \lim_{n\to\infty} \sup_{M\in\psi_1} P_M^n\big(d\big[\hat{O}(M), M\big] > \delta| C(L,n,M) = O\big)\\ &= 0.\end{aligned}$$

For Ψ_2, note that given any $M \in \Psi_2$, the *CSGS* algorithm errs if it outputs O. Thus, by Theorem 1,

$$\begin{aligned}&\lim_{n\to\infty} \sup_{M\in\psi_2} P_M^n\big(d\big[\hat{O}(M), M\big] > \delta|C(L,n,M) = O\big) \cdot P_M^n\big(C(L,n,M) = O\big)\\ &\le \lim_{n\to\infty} \sup_{M\in\psi_2} P_M^n\big(C(L,n,M) = O\big) = 0.\end{aligned}$$

Now consider Ψ_3. Let $O(M)$ be the population version of $\hat{O}(M)$, that is, all the sample regression coefficients in $\hat{O}(M)$ are replaced by the corresponding population coefficients. Since sample regression coefficients are uniformly consistent estimators of population regression coefficients under our assumptions, and there are only finitely many regression coefficients to consider, for every $\varepsilon > 0$, there is a sample size N_1, such that for all $n > N_1$, and all $M \in \Psi_3$,

$$P_M^n\big(d\big[\hat{O}(M), O(M)\big] > \delta/2|C(L,n,M) = O\big) < \varepsilon.$$

For any $M \in \Psi_3$, there are some edges in G_M missing in O. Let $\mathbf{E}(M)$ be the set of edges missing in O. Let M' be the same as M except that the structural coefficients associated with the edges in $\mathbf{E}(M)$ are set to zero. Let $O(M')$ be the same as $O(M)$ except that for each edge with an identified coefficient, the coefficient in $O(M')$ is the relevant regression coefficient derived from P'_M [whereas that in $O(M)$ is derived from P_M]. By the setup of M', the identified edge coefficients in $O(M')$ are equal to the corresponding edge coefficients in M', which are the same as the corresponding edge coefficients in M. Thus, the structural coefficient distance between $O(M')$ and M is simply

$$d\big[O(M'), M\big] = \max_{\langle i,j\rangle \in \mathbf{E}(M)} \big|e_M(X_i \to X_j)\big|.$$

For any edge $Y \to Z$ in O that has a different edge coefficient in $O(M)$ than that in $O(M')$, the edge coefficients are both derived from a regression of Z on $\mathbf{Parents}(O, Z)$, but one is based on P_M and the other is based on $P_{M'}$. The regression coefficient $r(Y, Z, \mathbf{Parents}(O, Z))$ is equal to the Y component of the vector $\operatorname{cov}(Z, \mathbf{Parents}(O, Z)) \operatorname{var}^{-1}(\mathbf{Parents}(O, Z))$ (Whittaker, 1990), which, given the structure G_M, is a rational function of the structural coefficients in M. Since $M \in \psi^{k,J,C}$, every submatrix of the covariance matrix for P_M is invertible, and so $r_M(Y, Z, \mathbf{Parents}(O, Z))$ is defined. For M', $r_{M'}(Y, Z, \mathbf{Parents}(O, Z)) = r_{M'}(Y, Z, \mathbf{A})$, where $\mathbf{A}$ is the smallest ancestral set that contains $\mathbf{Parents}(O, Z)$ in G_M. $\operatorname{var}(\mathbf{A})^{-1} = (\mathbf{I} - \mathbf{B})^T \operatorname{var}(\mathbf{E})^{-1}(\mathbf{I} - \mathbf{B})$, where $\mathbf{B}$ is the submatrix of structural coefficients in M' for variables in $\mathbf{A}$, and $\operatorname{var}(\mathbf{E})$ is the diagonal covariance matrix of error terms for variables in $\mathbf{A}$, which is a submatrix of Σ_M. Since $M \in \psi^{k,J,C}$, the variance of every error term is bounded from below by J. Thus, $\operatorname{var}(\mathbf{A})^{-1}$ is defined and so is $r_{M'}(Y, Z, \mathbf{Parents}(O, Z))$. Therefore, $r_M(Y, Z, \mathbf{Parents}(O, Z))$ and $r_{M'}(Y, Z, \mathbf{Parents}(O, Z))$ are values of a rational function of the structural coefficients.

A continuous function is uniformly continuous on a closed, bounded interval anywhere that it is defined. A rational function is continuous at every point of its domain where its denominator is not zero, that is, where the function value is defined. By Lemma 2 and assumption UBC(C), every structural coefficient $b_{j,i}$ in M lies in the closed bounded interval from $-C/J^{1/2}$ to $C/J^{1/2}$. Obviously the coefficients in M' still lie in this interval. Hence, given G_M, the difference between $r_{M'}(Y, Z, \mathbf{Parents}(O, Z))$ and $r_M(Y, Z, \mathbf{Parents}(O, Z))$ can be arbitrarily small if the differences between the structural coefficients in M' and those in M are sufficiently small. Given the set of variables $\mathbf{V}$, there are only finitely many structures and finitely many relevant regressions to consider. Therefore, there is a $\gamma \in (0, \delta/4)$ such that for every $M \in \psi_3$, $d[O(M), O(M')] < \delta/4$ if

$$\max_{\langle i,j\rangle \in \mathbf{E}(M)} \big|e_M(X_i \to X_j)\big| < \gamma.$$

Consider then the partition of Ψ_3 into

$$\Psi_{3.1} = \Big\{M \in \Psi_3 \Big| \max_{\langle i,j\rangle \in \mathbf{E}(M)} \big|e_M(X_i \to X_j)\big| < \gamma\Big\}$$

and

$$\Psi_{3.2} = \Big\{M \in \Psi_3 \Big| \max_{\langle i,j\rangle \in \mathbf{E}(M)} \big|e_M(X_i \to X_j)\big| \geq \gamma\Big\}.$$

It follows from the previous argument that for every $M \in \Psi_{3.1}, d[O(M), M] \leq d[O(M), O(M')] + d[O(M'), M] < \delta/4 + \gamma < \delta/2$. Then there is a sample size N_1, such that for all $n > N_1$ and all $M \in \Psi_{3.1}$,

$$\begin{aligned} &P_M^n\big(d\big[\hat{O}(M), M\big] > \delta | C(L, n, M) = O\big) \\ &\quad \leq P_M^n\big(d\big[\hat{O}(M), O(M)\big] > \delta/2 | C(L, n, M) = O\big) \\ &\quad < \varepsilon. \end{aligned}$$

For every $M \in \Psi_{3.2}$, there is at least one edge, say, $X \to Y$ missing from O such that $|e_M(X \to Y)| \geq \gamma$. Then by Lemma 2, there is a set $\mathbf{U}$ such that $|\rho(X, Y|\mathbf{U})| \geq \gamma J^{1/2}$, but O entails that $\rho(X, Y|\mathbf{U}) = 0$. Thus, the test of the Markov condition in step E2 is passed only if the test of $\rho(X, Y|\mathbf{U}) = 0$ returns 0 (i.e., accepts the null hypothesis). Note that if the test is not passed, then every structural coefficient is "Unknown," and so by definition the structural coefficient distance is zero. Therefore, the distance is greater than δ (and so nonzero) only if the test of $\rho(X, Y|\mathbf{U}) = 0$ returns 0 while $|\rho(X, Y|\mathbf{U})| \geq \gamma J^{1/2}$. Since tests are uniformly consistent, it follows that there is a sample size N_2, such that for all $n > N_2$ and all $M \in \Psi_{3.2}$,

$$P_M^n\big(d\big[\hat{O}(M), M\big] > \delta | C(L, n, M) = O\big) < \varepsilon.$$

Let $N = \max(N_1, N_2)$. Then for all $n > N$,

$$\begin{aligned} &\sup_{M \in \psi_3} P_M^n\big(d\big[\hat{O}(M), M\big] > \delta | C(L, n, M) = O\big) \\ &\qquad \cdot P_M^n\big(C(L, n, M) = O\big) \\ &\quad \leq \sup_{M \in \psi_3} P_M^n\big(d\big[\hat{O}(M), M\big] > \delta | C(L, n, M) = O\big) \\ &\quad < \varepsilon. \end{aligned}$$

□

5. CONCLUSION

We have shown that there is a pointwise consistent estimator of causal patterns and a uniformly consistent estimator of some of the structural coefficients in causal patterns, even when the Causal Faithfulness assumption and Strong Causal Faithfulness assumptions are substantially weakened. The *k*-Triangle Faithfulness assumption is a restriction on many fewer partial correlations than the Causal Faithfulness assumption and the Strong Causal Faithfulness assumptions, and does not entail that there are no edges with very small but nonzero structural coefficients.

There are a number of open problems associated with the Causal Faithfulness assumption:

1. Is it possible to speed up the *Very Conservative SGS* algorithm to make it applicable to data sets with large numbers of variables?
2. If unfaithfulness is detected, is it possible to reduce the number of structural coefficients where the algorithm returns "Unknown?"
3. In practice, on realistic sample sizes, how does the *Very Conservative SGS* algorithm perform? [Ramsey, Zhang and Spirtes (2006), have already shown that the *Conservative PC* algorithm is more accurate and not significantly slower than the *PC* algorithm].
4. Is the *k*-Triangle Faithfulness assumption unlikely to hold for reasonable values of *k* and large numbers of variables?
5. Is there an assumption weaker than the *k*-Triangle Faithfulness assumption for which there is a uniformly consistent estimator for structural coefficients in a causal pattern?
6. Are there analogous results that apply when the number of variables and the maximum degree of a vertex increases and the size of *k* decreases with increasing sample size [as in the Kalisch and Bühlmann (2007), results]?
7. Are there analogous results that apply when the assumption of causal sufficiency is abandoned?
8. Are there analogous results that apply for other families of distributions or for nonparametric tests of conditional independence?

ACKNOWLEDGMENTS

We thank two anonymous referees for helpful comments. Zhang's research was supported in part by the Research grants Council of Hong Kong under the General Research Fund LU341910.

REFERENCES

COLOMBO, D., MAATHUIS, M. H., KALISCH, M. and RICHARDSON, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.* **40** 294–321. MR3014308

KALISCH, M. and BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8** 613–636.

LIN, S., UHLER, C., STURMFELS, B. and BÜHLMANN, P. (2012). Hypersurfaces and their singularities in partial correlation testing. Available at arXiv:1209.0285.

MAATHUIS, M. H., COLOMBO, D., KALISCH, M. and BÜHLMANN, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nat. Methods* **7** 247–248.

MEEK, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* 403–411. Morgan Kaufmann, San Francisco, CA.

PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*: *Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA. MR0965765

PEARL, J. (2000). *Causality*: *Models*, *Reasoning*, *and Inference*. Cambridge Univ. Press, Cambridge. MR1744773

RAMSEY, J., ZHANG, J. and SPIRTES, P. (2006). Adjacency-faithfulness and conservative causal inference. In *Proceedings of the* 22*nd Conference on Uncertainty in Artificial Intelligence* 401–408. AUAI Press, Arlington, VA.

ROBINS, J. M., SCHEINES, R., SPIRTES, P. and WASSERMAN, L. (2003). Uniform consistency in causal inference. *Biometrika* **90** 491–515. MR2006831

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (1993). *Causation*, *Prediction*, *and Search*. *Lecture Notes in Statistics* **81**. Springer, New York. MR1227558

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. MIT Press, Cambridge, MA. MR1815675

UHLER, C., RASKUTTI, G., BÜHLMANN, P. and YU, B. (2013). Geometry of faithfulness assumption in causal inference. *Ann. Statist.* **41** 436–463.

VERMA, T. and PEARL, P. (1990). Equivalence and synthesis of causal models. In *Proceedings of the* 6*th Conference on Uncertainty in Artificial Intelligence* 220–227. AUAI Press, Corvallis, OR.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester. MR1112133

WOODWARD, J. (2003). *Making Things Happen*: *A Theory of Causal Explanation*. Oxford Univ. Press, London.

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

ZHANG, J. and SPIRTES, P. (2011). Intervention, determinism, and the causal minimality condition. *Synthese* **182** 335–347. MR2833024