Nonparametric causal structure learning in high dimensions

Shubhadeep Chakraborty
Department of Biostatistics, University of Washington
e-mail: deep20@uw.edu

Ali Shojaie
Department of Biostatistics, University of Washington
e-mail: ashojaie@uw.edu

Abstract: The PC and FCI algorithms are popular constraint-based methods for learning the structure of directed acyclic graphs (DAGs) in the absence and presence of latent and selection variables, respectively. These algorithms (and their order-independent variants, PC-stable and FCI-stable) have been shown to be consistent for learning sparse high-dimensional DAGs based on partial correlations. However, inferring conditional independences from partial correlations is valid if the data are jointly Gaussian or generated from a linear structural equation model — an assumption that may be violated in many applications. To broaden the scope of high-dimensional causal structure learning, we propose nonparametric variants of the PC-stable and FCI-stable algorithms that employ the conditional distance covariance (CdCov) to test for conditional independence relationships. As the key theoretical contribution, we prove that the high-dimensional consistency of the PC-stable and FCI-stable algorithms carry over to general distributions over DAGs when we implement CdCov-based nonparametric tests for conditional independence. Numerical studies demonstrate that our proposed algorithms perform nearly as good as the PC-stable and FCI-stable for Gaussian distributions, and offer advantages in non-Gaussian graphical models.

Keywords and phrases: Causal Structure Learning, Consistency, FCI algorithm, High Dimensionality, Nonparametric Testing, PC algorithm.

1. Introduction

Directed acyclic graphs (DAGs) are commonly used to represent causal relationships among random variables (Lauritzen, 1996; Spirtes et al., 2000; Maathuis et al., 2019). The PC algorithm (Spirtes et al., 2000) is the most popular constraint-based method for learning DAGs from observational data under the assumption of causal sufficiency, i.e., when there are no unmeasured common causes and no selection variables. It first estimates the skeleton of a DAG by recursively performing a sequence of conditional independence tests, and then uses the information from the conditional independence relations to partially orient the edges, resulting in a completed partially directed acyclic graph (CPDAG). In Section 2, we provide a review of these and other notions commonly used in the graphical modeling literature that are relevant to our work. Also we refer to estimating the CPDAG as structure learning of the underlying DAG throughout the rest of the paper.

Observational studies often involve latent and selection variables, which complicate the causal structure learning problem. Ignoring such unmeasured variables can make the causal inference based on the PC algorithm erroneous; see, e.g., Section 1.2 in Colombo et al. (2012) for some illustrations. The Fast Causal Inference (FCI) algorithm and its variants (Spirtes et al., 2000; Spirtes et al., 2001; Zhang, 2008; Colombo et al., 2012) utilize similar strategies as the PC algorithm to learn the DAG structure in the presence of latent and selection variables.

Both PC and FCI algorithms adopt a hierarchical search strategy — they recursively perform conditional independence tests given subsets of increasingly larger cardinalities in some appropriate search pool. The PC algorithm is usually order-dependent, in the sense that its output depends on the order in which pairs of adjacent vertices and subsets of their adjacency sets are considered.
The FCI algorithm suffers from a similar limitation. To overcome this limitation, Colombo and Maathuis (2014) proposed two variants of the PC and FCI algorithms, namely the PC-stable and FCI-stable algorithms, that resolve the order dependence at different stages of the algorithms.

In general, testing for conditional independence is a problem of central importance in the causal structure learning. The literature on the PC and FCI algorithms predominantly uses partial correlations to infer conditional independence relations. It is well-known that the characterization of conditional independence by partial correlations, or in other words, equivalence between conditional independence and zero partial correlations only holds for multivariate normal variables. Therefore, the high-dimensional consistency results for the PC and FCI algorithms (Kalisch and Bühlmann, 2007; Colombo et al., 2012) are limited to Gaussian graphical models, where the nodes correspond to random variables with a joint Gaussian distribution. Although the Gaussian graphical model is the standard parametric model for continuous data, it may not hold in many real data applications. Although this limitation can be somewhat relaxed by considering linear structural equation models (SEMs) with general noise distributions (Loh and Bühlmann, 2014), linear SEMs and joint Gaussianity are essentially equivalent (Voorman et al., 2014). Moreover, neither approach is appropriate when the observations are categorical, discrete, have heavy-tail distributions, or their support is a subset of the real line. In Section 4.3, for example, we present a real application where all the observed variables are categorical, and therefore far from being Gaussian. As an improvement, Harris and Drton (2013) used rank-based partial correlations to test for conditional independence relations, showing that the high-dimensional consistency of the PC algorithm holds for a broader class of Gaussian copula models. Some nonparametric versions of the PC algorithm have been also proposed in the literature via kernel-based tests for conditional independence (Sun et al., 2007; Zhang et al., 2018); however, they lack theoretical justifications of the correctness of the algorithms, and are not studied in high dimensions.

This work aims to broaden the applicability of the PC-stable and FCI-stable algorithms to general distributions by employing a nonparametric test for conditional independence relationships. To this end, we utilize recent developments on dependence metrics that quantify non-linear and non-monotone dependence between multivariate random variables. More specifically, our work builds on the idea of distance covariance (dCov) proposed by Székely et al. (2007) and its extension to conditional distance covariance (CdCov) by Wang et al. (2015) as a nonparametric measure of non-linear and non-monotone conditional independence between two random vectors of arbitrary dimensions given a third. Utilizing this flexibility, we use the conditional distance covariance (CdCov) to test for conditional independence relationships in the sample versions of the PC-stable and FCI-stable algorithms. The resulting algorithms — which, for distinction, are named ‘nonPC’ and ‘nonFCI’ — facilitate causal structure learning from general distributions over DAGs and are shown to be consistent in sparse high-dimensional settings. Our consistency results only require mild moment and tail conditions on the set of variables, without requiring any strict distributional assumptions. To our knowledge, the proposed generalizations of PC/PC-stable or the FCI/FCI-stable algorithms provide the first general nonparametric framework for causal structure learning with theoretical guarantees in high dimensions.

The rest of the paper is organized as follows. In Section 2, we review the relevant background, including preliminaries on graphical modeling (Section 2.1) and a brief overview of dCov and CdCov (Section 2.2). The nonparametric version of the PC-stable algorithm is presented in Section 3.1. As a key contribution of the paper, we establish that the algorithm consistently estimates the skeleton and the equivalence class of the underlying sparse high-dimensional DAG in a general nonparametric framework. We then present the nonparametric version of the FCI-stable algorithm in Section 3.2 and establish its consistency in sparse high-dimensional settings. As the FCI involves the adjacency search of the PC algorithm, any improvement on the PC/PC-stable directly carries over to the FCI/FCI-stable as well. In Section 4, we compare the performances of our algorithms with the PC-stable and FCI-stable using both simulated datasets (involving both Gaussian and non-Gaussian examples), as well as a real dataset. These numerical studies clearly demonstrate that nonPC and nonFCI algorithms are comparable with PC-stable and FCI-stable for Gaussian data and offer improvements for non-Gaussian data.
2. Background

2.1. Preliminaries on graphical modeling

A graph $G = (V, E)$ consists of a vertex set $V = \{1, \ldots, p\}$ and an edge set $E \subseteq V \times V$. In a graphical model, the vertices or nodes are associated with random variables $X_a$ for $1 \leq a \leq p$. Throughout, we index the nodes by the corresponding random variables. We also allow the edge set $E$ of the graph $G$ to contain (a subset of) the following six types of edges: $\rightarrow$ (directed), $\leftrightarrow$ (bidirected), $\rightarrow \leftarrow$ (undirected), $\circ \circ$ (nondirected), $\circ \rightarrow$ (partially undirected) and $\circ \circ \rightarrow$ (partially directed). The endpoints of an edge are called marks, which can be tails, arrowheads or circles. We use the symbol ‘$\star$’ to denote an arbitrary edge mark. A mixed graph is a graph containing directed, bidirected and undirected edges. A graph containing only directed edges ($\rightarrow$) is called a directed graph, one containing only undirected edges ($\leftrightarrow$) is called an undirected graph, and one containing directed and undirected edges is called a partially directed graph.

The adjacency set of a vertex $X_a$ in the graph $G = (V, E)$, denoted $\text{adj}(G, X_a)$, is the set of all vertices in $V$ that are adjacent to $X_a$, or in other words, are connected to $X_a$ by an edge. The degree of a vertex $X_a$, $|\text{adj}(G, X_a)|$, is defined as the number of vertices adjacent to it. A graph is complete if all pairs of vertices in the graph are adjacent. A vertex $X_b \in \text{adj}(G, X_a)$ is called a parent of $X_a$ if $X_b \rightarrow X_a$, a child of $X_a$ if $X_a \rightarrow X_b$ and a neighbor of $X_a$ if $X_a - X_b$. The skeleton of the graph $G$ is the undirected graph obtained by replacing all the edges of $G$ by undirected edges, in other words, ignoring all the edge orientations. Three vertices $\langle X_a, X_b, X_c \rangle$ are called an unshielded triple if $X_a$ and $X_b$ are adjacent, $X_b$ and $X_c$ are adjacent, but $X_a$ and $X_c$ are not adjacent. A path is a sequence of distinct adjacent vertices. A node $X_b$ is an ancestor of its descendents $X_a$, if $G$ contains a directed path $X_a \rightarrow \cdots \rightarrow X_b$. A non-endpoint vertex $X_a$ on a path is called a collider on the path if both the edges preceding and succeeding it have an arrowhead at $X_a$, or in other words, the path contains $\leftrightarrow X_a \leftrightarrow$. An unshielded triple $\langle X_a, X_b, X_c \rangle$ is called a v-structure if $X_b$ is a collider on the path $\langle X_a, X_b, X_c \rangle$.

A cycle occurs in a graph when there is a path from $X_a$ to $X_b$, and $X_a$ and $X_b$ are adjacent. A directed path from $X_a$ to $X_b$ forms a directed cycle together with the edge $X_b \rightarrow X_a$, and it forms an almost directed cycle together with the edge $X_b \leftrightarrow X_a$. Three vertices that form a cycle are called a triangle. A directed acyclic graph (DAG) is a directed graph that does not contain any cycle. A DAG entails conditional independence relationships via a graphical criterion called d-separation (Section 1.2.3 in Pearl, 2000). Two vertices $X_a$ and $X_b$ that are not adjacent in a DAG $G$ are d-separated in $G$ by a subset $X_S \subseteq V \{X_a, X_b\}$. A probability distribution $P$ on $\mathbb{R}^p$ is said to be faithful with respect to the DAG $G$ if the conditional independence relationships in $P$ can be inferred from $G$ using d-separation and vice versa, in other words, $X_a \perp \!\!\!\!\!\!| X_b|X_S$ if and only if $X_a$ and $X_b$ are d-separated in $G$ by $X_S$.

A graph that is both (partially) directed and acyclic, is called a partially directed acyclic graph (PDAG). DAGs that encode the same set of conditional independence relations, form a Markov equivalence class (Verma and Pearl, 1990). Two DAGs belong to the same Markov equivalence class if and only if they have the same skeleton and the same v-structures. A Markov equivalence class of DAGs can be uniquely represented by a completed partially directed acyclic graph (CPDAG), which is a PDAG that satisfies the following: i) $X_a \rightarrow X_b$ in the CPDAG if $X_a \rightarrow X_b$ in every DAG in the Markov equivalence class, and, ii) $X_a \rightarrow X_b$ in the CPDAG if the Markov equivalence class contains a DAG in which $X_a \rightarrow X_b$ as well as a DAG in which $X_a \leftarrow X_b$.

Estimation of the CPDAG by the PC algorithm involves two steps: 1) estimation of the skeleton and separating sets (also called the adjacency search step) via recursive conditional independence tests; and, 2) partial orientation of edges; see Algorithms 1 and 2 in Kalisch and Bühlmann (2007) for details.

In presence of latent and selection variables, one needs a generalization of a DAG, called a maximal ancestral graph (MAG). A mixed graph is called an ancestral graph if it contains no directed or almost directed cycles and no subgraph of the type $X_a - X_b \leftrightarrow X_c$. DAGs form a subset of ancestral graphs. A MAG is an ancestral graph in which every missing edge corresponds
to a conditional independence relationship via the m-separation criterion (Richardson and Spirtes, 2002), a generalization of the notion of d-separation. Multiple MAGs may represent the same set of conditional independence relations. Such MAGs form a Markov equivalence class which can be represented by a partial ancestral graph (PAG) (Ali et al, 2009); see Richardson and Spirtes (2002) for additional details.

Under the faithfulness assumption, the Markov equivalence class of a DAG with latent and selection variables can be learned using the FCI algorithm (e.g., Algorithm 3.1 in Colombo et al., 2012), which is a modification of the PC algorithm. The FCI algorithm first employs the adjacency search of the PC algorithm, and then performs additional conditional independence tests because of the presence of latent variables followed by partial orientation of the edges, resulting in an estimated PAG. To estimate the skeletons (of the DAG and the PAG, respectively), both the PC and the FCI algorithms adopt a hierarchical search strategy that starts with a complete undirected graph and recursively removes edges via conditional independence tests given subsets of increasingly larger cardinalities in some appropriate search pool. Colombo and Maathuis (2014) proposed order-independent variants of the PC and the FCI algorithms, namely the PC-stable and FCI-stable algorithms. To make the paper self-contained, we provide the pseudocodes of the oracle proposed order-independent variants of the PC and the FCI algorithms, namely the PC-stable and FCI-stable algorithms in Appendix A.

2.2. Distance covariance and conditional distance covariance

We start by describing the notation used throughout the paper. We denote by $\| \cdot \|_p$ the Euclidean norm of $\mathbb{R}^p$ and use $\| \cdot \|$ when the dimension is clear from the context. We use $X \perp \perp Y$ to denote the independence of $X$ and $Y$ and use $\mathbb{E}_U$ to denote expectation with respect to the probability distribution of the random variable $U$. For any set $S$, we denote its cardinality by $|S|$.

We use the usual asymptotic notation, ‘$O$’ and ‘$o$’, as well as their probabilistic counterparts, $O_p$ and $o_p$, which denote stochastic boundedness and convergence in probability, respectively. For two sequences of real numbers $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, $a_n \asymp b_n$ if and only if $a_n/b_n = O(1)$ as $n \to \infty$. We use the symbol “$a \lesssim b$” to indicate that $a \leq Cb$ for some constant $C > 0$. For a matrix $A = (a_{kl})_{k,l=1}^n \in \mathbb{R}^{n \times n}$, we denote its determinant by $|A|$ and define its $U$-centered version $\tilde{A} = (\tilde{a}_{kl})_{k,l=1}^n$ as

$$\tilde{a}_{kl} = \begin{cases} a_{kl} - \frac{1}{n-2} \sum_{j=1}^n a_{kj} - \frac{1}{n-2} \sum_{i=1}^n a_{il} + \frac{1}{(n-1)(n-2)} \sum_{i,j=1}^n a_{ij}, & k \neq l, \\ 0, & k = l, \end{cases}$$

for $k, l = 1, \ldots, n$. Finally, we denote the integer part of $a \in \mathbb{R}$ by $\lfloor a \rfloor$.

Székely et al. (2007), in their seminal paper, introduced the notion of distance covariance (dCov, henceforth) to quantify non-linear and non-monotone dependence between two random vectors of arbitrary dimensions. Consider two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with $\mathbb{E}\|X\|_p < \infty$ and $\mathbb{E}\|Y\|_q < \infty$. The distance covariance between $X$ and $Y$ is defined as the positive square root of

$$\text{d Cov}^2(X, Y) = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t, s) - f_X(t) f_Y(s)|^2 \frac{dt ds}{\|t\|_p^{p+} \|s\|_q^{q+}},$$

where $f_X, f_Y$ and $f_{X,Y}$ are the individual and joint characteristic functions of $X$ and $Y$ respectively, and $c_p = \pi^{(1+p)/2} / \Gamma((1+p)/2)$ is a constant with $\Gamma(\cdot)$ being the complete gamma function.

The key feature of dCov is that it completely characterizes the independence between two random vectors, or in other words dCov$(X, Y) = 0$ if and only if $X \perp \perp Y$. According to Remark 3 in Székely et al. (2007), dCov can be equivalently expressed as

$$\text{d Cov}^2(X, Y) = \mathbb{E}\|X - X'||_p\|Y - Y'||_q + \mathbb{E}\|X - X'||_p \mathbb{E}\|Y - Y'||_q - 2 \mathbb{E}\|X - X'||_p\|Y - Y''\|_q.$$

This alternate expression comes handy in constructing V or U-statistic type estimators for the quantity. For an observed random sample $(X_i, Y_i)_{i=1}^n$ from the joint distribution of $X$ and $Y$, 

define the distance matrices \( d^X = (d^X_{ij})_{i,j=1}^n \) and \( d^Y = (d^Y_{ij})_{i,j=1}^n \in \mathbb{R}^{n \times n} \) where \( d^X_{ij} := \| X_i - X_j \|_p \) and \( d^Y_{ij} := \| Y_i - Y_j \|_q \). Following the U-centering idea in Székely and Rizzo (2014), an unbiased U-statistic type estimator of \( d\text{Cov}^2(X,Y) \) can be expressed as

\[
d\text{Cov}^2_n(X,Y) := (\tilde{d}^X \cdot \tilde{d}^Y) := \frac{1}{n(n-3)} \sum_{i \neq j} \tilde{d}^X_{ij} \tilde{d}^Y_{ij},
\]

where \( \tilde{d}^X = (\tilde{d}^X_{ij})_{i,j=1}^n \) and \( \tilde{d}^Y = (\tilde{d}^Y_{ij})_{i,j=1}^n \) are the \( \mathcal{U} \)-centered versions of the matrices \( d^X \) and \( d^Y \), respectively, as defined in (2.1).

Wang et al. (2015) recently generalized the notion of \( d\text{Cov} \) and introduced the conditional distance covariance (\( d\text{Cov} \), henceforth) as a measure of conditional dependence between two random vectors of arbitrary dimensions given a third. \( d\text{Cov} \) essentially replaces the characteristic functions used in the definition of \( d\text{Cov} \) by conditional characteristic functions. Consider a third random vector \( Z \in \mathbb{R}^r \) with \( \mathbb{E}(\|X\|_p + \|Y\|_q \mid Z) < \infty \). Denote by \( f_{X,Y \mid Z} \) the conditional joint characteristic function of \( X \) and \( Y \) given \( Z \), and by \( f_{X\mid Z} \) and \( f_{Y\mid Z} \) the conditional marginal characteristic functions of \( X \) and \( Y \) given \( Z \), respectively. Then \( d\text{Cov} \) between \( X \) and \( Y \) given \( Z \) is defined as the positive square root of

\[
d\text{Cov}^2(X,Y \mid Z) = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y\mid Z}(t,s) - f_{X\mid Z}(t)f_{Y\mid Z}(s)|^2}{\|t\|_p^{1+p} \|s\|_q^{1+q}} \, dt \, ds.
\]

The key feature of \( d\text{Cov} \) is that \( d\text{Cov}(X,Y\mid Z) = 0 \) almost surely if and only if \( X \perp \!\!\!\perp Y \mid Z \), which is quite straightforward to see from the definition.

Similar to \( d\text{Cov} \), an equivalent alternative expression can be established for \( d\text{Cov} \) that avoids complicated integrations involving conditional characteristic functions. Let \( W_i = (X_i, Y_i, Z_i)_{i=1}^n \) be an i.i.d. sample from the joint distribution of \( W := (X,Y,Z) \). Define \( \tilde{d}_{ijkl} := (d^X_{ij} + d^X_{kl} - d^X_{ik} - d^X_{jl}) (d^Y_{ij} + d^Y_{kl} - d^Y_{ik} - d^Y_{jl}) \), which is not symmetric with respect to \( \{i,j,k,l\} \), and therefore necessitates defining the following symmetric form: \( d^{S_{ijkl}} := d_{ijkl} + d_{iklj} + d_{jikl} \). Lemma 1 in Wang et al. (2015) establishes an equivalent representation of \( d\text{Cov}^2(X,Y \mid Z = z) \) as

\[
d\text{Cov}^2(X,Y \mid Z = z) = \frac{1}{12} \mathbb{E} \left[ d^{S_{1234}}_{ijkl} : Z_1 = z, Z_2 = z, Z_3 = z, Z_4 = z \right].
\]  

Remark 2.1. In a recent work, Sheng and Sriperumbudur (2019) explore the connection between conditional independence measures induced by distances on a metric space and reproducing kernels associated with a reproducing kernel Hilbert space (RKHS). They generalize \( d\text{Cov} \) to arbitrary metric spaces of negative type — termed generalized \( d\text{Cov} \) (\( g\text{Cov} \)) — and develop a kernel-based measure of conditional independence, namely the Hilbert-Schmidt conditional independence criterion (\( H\text{SCIC} \)). Theorem 1 in their paper establishes an equivalence between \( g\text{Cov} \) and \( H\text{SCIC} \), or in other words, between distance and kernel-based measures of conditional independence.

For \( w \in \mathbb{R}^r \), let \( K_H(w) := |H|^{-1} K(H^{-1}w) \) be a kernel function where \( H \) is the diagonal matrix \( \text{diag}(h, \ldots, h) \) determined by a bandwidth parameter \( h \). \( K_H \) is typically considered to be the Gaussian kernel \( K_H(w) = (2\pi)^{-\frac{r}{2}} |H|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} w^T H^{-2} w \right) \), where \( w \in \mathbb{R}^r \).

Let \( K_{iu} := K_H(Z_i - Z_u) = |H|^{-1} K(H^{-1}(Z_i - Z_u)) \) and \( K_i(Z) := K_H(Z - Z_i) \) for \( 1 \leq i, u \leq n \). Then by virtue of the equivalent representation of \( d\text{Cov} \) in (2.3), a V-statistic type estimator of \( d\text{Cov}^2(X,Y \mid Z) \) can be constructed as

\[
d\text{Cov}^2_n(X,Y \mid Z) := \sum_{i,j,k,l} K_i(Z) K_j(Z) K_k(Z) K_l(Z) \frac{12}{\left( \sum_{i=1}^n K_i(Z) \right)^4} d^{S_{ijkl}},
\]

Under certain regularity conditions, Theorem 4 in Wang et al. (2015) shows that conditioned on \( Z \), \( d\text{Cov}^2_n(X,Y \mid Z) \xrightarrow{p} d\text{Cov}^2(X,Y \mid Z) \) as \( n \to \infty \).
3. Methodology and Theory

3.1. The Nonparametric PC Algorithm in High Dimensions

To get a measure of conditional independence between $X$ and $Y$ given $Z$ that is free of $Z$, we define

$$\rho_0^*(X,Y|Z) := \mathbb{E} \left[ \text{CdCov}_n^2(X,Y|Z) \right].$$

Clearly $\rho_0^*(X,Y|Z) = 0$ if and only if $X \perp Y \mid Z$. Consider a plug-in estimate of $\rho_0^*(X,Y|Z)$ as

$$\hat{\rho}^*(X,Y|Z) := \frac{1}{n} \sum_{u=1}^{n} \text{CdCov}_n^2(X,Y|Z_u) = \frac{1}{n} \sum_{u=1}^{n} \Delta_{i,j,k,l;u},$$

where

$$\Delta_{i,j,k,l;u} := \sum_{i,j,k,l} \frac{K_{iu} K_{ju} K_{ku} K_{lu}}{12 \left( \sum_{i=1}^{n} K_{iu} \right)^4} d_{ijkl}^S.$$

We reject $H_0 : X \perp Y | Z$ vs $H_A : X \not\perp Y | Z$ at level $\alpha \in (0,1)$ if $\hat{\rho}^*(X,Y|Z) > \xi_{n,\alpha}$, where the threshold $\xi_{n,\alpha}$ is typically obtained by a local bootstrap procedure (see Section 4.3 in Wang et al., 2015). Henceforth we will often denote $\rho_0^*(X,Y|Z)$ simply by $\rho_0^*$ for notational simplicity, whenever there is no confusion.

In view of the complete characterization of conditional independence by $\rho_0^*$, we propose testing for conditional independence relations nonparametrically in the sample version of the PC-stable algorithm based on $\rho_0^*$, rather than partial correlations. We coin the resulting algorithm the ‘nonPC’ algorithm, to emphasize that it is a nonparametric generalization of parametric PC-stable algorithms.

The oracle version of the first step of nonPC, or the skeleton estimation step, is exactly the same as that of the PC-stable algorithm (Algorithm 1 in Appendix A). The second step which extends the skeleton estimated in the first step to a CPDAG (Algorithm 2 in Appendix A) comprises of some purely deterministic rules for edge orientations, and is exactly the same for both the nonPC and PC-stable as well. The only difference lies in the implementation of the tests for conditional independence relationships in the sample versions of the first step. Specifically, we replace all the conditional independence queries in the first step by tests based on $\rho_0^* \mid X,Y \mid Z$. At some pre-specified significance level $\alpha$, we infer that $X_a \perp X_b \mid X_S$ when $\hat{\rho}^*(X_a,X_b|X_S) \leq \xi_{n,\alpha}$, where $a, b \in V$ and $S \subseteq V, |S| \neq \phi$. When $|S| = \phi$, $\hat{\rho}^*(X_a,X_b|X_S) = \text{dCov}_n^2(X_a,X_b)$ and $\rho_0^*(X,Y|Z) = \text{dCov}_n^2(X,Y)$. The critical value $\xi_{n,\alpha}$ in this case is obtained by a bootstrap procedure (see, e.g., Section 4 in Chakraborty and Zhang, 2019 with $d = 2$).

Given that the equivalence between conditional independence and zero partial correlations only holds for multivariate normal random variables, our generalization broadens the scope of applicability of causal structure learning by the PC/PC-stable algorithm to general distributions over DAGs. This nonparametric approach is thus a natural extension of Gaussian and Gaussian copula models. It enables capturing non-linear and non-monotone conditional dependence relationships among the variables, which partial correlations fail to detect.

Next, we establish theoretical guarantees on the correctness of the nonPC algorithm in learning the true underlying causal structure in sparse high-dimensional settings. Our consistency results only require mild moment and tail conditions on the set of variables, without making any strict distributional assumptions. Denote by $m_p$ the maximum cardinality of the conditioning sets considered in the adjacency search step of the PC-stable algorithm. Clearly $m_p \leq q$, where

$q := \max_{1 \leq a \leq p} |\text{adj}(G, a)|$ is the maximum degree of the DAG $G$. For a fixed pair of nodes $a, b \in V$, the conditioning sets considered in the adjacency search step are elements of

$$J_{a,b}^{m_p} := \{ S \subseteq V \setminus \{a, b\} : |S| \leq m_p \}.$$

We first establish a concentration inequality that gives the rate at which the absolute difference of $\rho_0^*(X_a, X_b|X_S)$ and its plug-in estimate $\hat{\rho}_n^*(X_a, X_b|X_S)$ decays to zero, for any fixed pair of nodes $a$ and $b \in V$ and a fixed conditioning set $S$. Towards that, we impose the following regularity conditions.
The maximum degree of the DAG is a commonly used condition, for example, in the high-dimensional feature screening literature (see, for example, Liu et al., 2014). Condition (A2) is a mild condition on the kernel function \( K(\cdot) \) that is guaranteed by many commonly used kernels, including the Gaussian kernel. Under conditions (A1) and (A2), the next result shows that the plug-in estimate \( \hat{\rho}(X_a, X_b | X_S) \) converges in probability to its population counterpart \( \rho^*(X_a, X_b | X_S) \) exponentially fast.

**Theorem 3.1.** Under conditions (A1) and (A2), for any \( \epsilon > 0 \) there exist positive constants \( A, B \) and \( \gamma \in (0, 1/4) \) such that

\[
P \left( |\hat{\rho}^*(X_a, X_b | X_S) - \rho^*_0(X_a, X_b | X_S)| > \epsilon \right) \leq O \left( 2 \exp \left( -A n^{1-2\gamma} \epsilon^2 \right) + n^4 \exp \left( -B n^{\gamma} \right) \right).
\]

The proof of Theorem 3.1 is long and somewhat technical; it is thus relegated to Section 5. Theorem 3.1 serves as the main building block towards establishing the consistency of the nonPC algorithm in sparse high-dimensional settings.

For notational convenience, henceforth we denote \( \rho^*_0(X_a, X_b | X_S) \) and \( \hat{\rho}^*(X_a, X_b | X_S) \) by \( \rho^*_0; ab; S \) and \( \hat{\rho}^*_0; ab; S \), respectively. In Theorem 3.2 below, we establish a uniform bound for the errors in inferring conditional independence relationships using the \( \rho^*_0 \)-based test in the skeleton estimation step of the sample version of the nonPC algorithm.

**Theorem 3.2.** Under conditions (A1) and (A2), for any \( \epsilon > 0 \) there exist positive constants \( A, B \) and \( \gamma \in (0, 1/4) \) such that

\[
\sup_{a,b \in V} P \left( |\hat{\rho}^*_0; ab; S - \rho^*_0; ab; S| > \epsilon \right) \leq P \left( \sup_{a,b \in V} \sup_{S \in J^{mp}_a, b} |\hat{\rho}^*_0; ab; S - \rho^*_0; ab; S| > \epsilon \right)
\leq O \left( p^{m_p+2} \left( 2 \exp \left( -A n^{1-2\gamma} \epsilon^2 \right) + n^4 \exp \left( -B n^{\gamma} \right) \right) \right).
\]

Next, we turn to proving the consistency of the nonPC algorithm in the high-dimensional setting where the dimension \( p \) can be much larger than the sample size \( n \), but the DAG is considered to be sparse. We impose the following regularity conditions, which are similar to the assumptions imposed in Section 3.1 of Kalisch and Bühlmann (2007) in order to prove the consistency of the PC algorithm for Gaussian graphical models.

We let the number of variables \( p \) grow with the sample size \( n \) and consider \( p = p_n \), and also the DAG \( G = G_n := (V_n, E_n) \) and the distribution \( P = P_n \).

(A3) The dimension \( p_n \) grows at a rate such that the right hand side of (3.3) tends to zero as \( n \to \infty \). In particular this is satisfied when \( p_n = O(n^r) \) for any \( 0 \leq r < \infty \).

(A4) The maximum degree of the DAG \( G_n \), denoted by \( q_n := \max_{1 \leq a \leq p_n} |adj(G_n, a)| \), grows at the rate of \( O(n^{1-b}) \), where \( 0 < b \leq 1 \).

(A5) The distribution \( P_n \) is faithful to the DAG \( G_n \) for all \( n \). In other words, for any \( a, b \in V_n \) and \( S \in J^{mp}_a, b \),

\[
X_a \text{ and } X_b \text{ are d-separated by } X_S \iff X_a \independent X_b | X_S \iff \rho^*_0; ab; S = 0.
\]

Moreover, \( \rho^*_0; ab; S \) values are uniformly bounded both from above and below. Formally,

\[
C_{min} := \inf_{a,b \in V_n, \ S \in J^{mp}_a, b} \rho^*_0; ab; S \geq \lambda_{min}, \ 0 < \lambda_{min} = O(n^r)
\]

and

\[
C_{max} := \sup_{a,b \in V_n, \ S \in J^{mp}_a, b} \rho^*_0; ab; S \leq \lambda_{max}.
\]
where $\lambda_{\text{min}}, \lambda_{\text{max}}$ are positive constants and $0 < \nu < 1/4$.

Condition (A3) allows the dimension to grow at any arbitrary polynomial rate of the sample size. Condition (A4) is a sparsity assumption on the underlying true DAG, allowing the maximum degree of the DAG to also grow, but at a slower rate than $n$. Since $m_p \leq q_n$, we also have $m_p = O(n^{1-b})$.

Finally, Condition (A5) is the strong faithfulness assumption (Definition 1.3 in Uhler et al., 2013) on $P_n$ and is similar to condition (A4) in Kalisch and Buhlmann (2007). This essentially requires $\rho_{a,b|S}^*$ to be bounded away from zero when the vertices $X_a$ and $X_b$ are not d-separated by $X_S$.

It is worth noting that the faithfulness assumption alone is not enough to prove the consistency of the PC/PC-stable/nonPC algorithms in high-dimensional settings, and the more stringent strong faithfulness condition is required.

**Remark 3.1.** For notational convenience, treat $X_a, X_b$ and $X_S$ as $X, Y$ and $Z$, respectively, for any $a, b \in V_n$ and $S \in J_{a,b}^{m,n}$. From Equation (2.3) we have

$$\text{CdCov}^2(X, Y|Z) = \frac{1}{12} \mathbb{E} \left[ d_{1234}^S \right] = \frac{1}{12} \mathbb{E} \left[ d_{1234} \right] + \frac{1}{12} \mathbb{E} \left[ d_{1234} + d_{1243} + d_{1432} \right].$$

which implies

$$\rho_0^* = \mathbb{E} \left[ \text{CdCov}^2(X, Y|Z) \right] = \frac{1}{12} \mathbb{E} \left[ d_{1234}^S \right] = \frac{1}{12} \mathbb{E} \left[ d_{1234} + d_{1243} + d_{1432} \right].$$

Condition (A1) implies $\sup_{1 \leq n \leq p} \max_{1 \leq a \leq p} \mathbb{E} X_a^2 < \infty$. With this and the definition of $d_{ijkl}$ in Section 2.2, it follows from some simple algebra and the Cauchy-Schwarz inequality that $\rho_0^* < \infty$. This provides a justification for the second part of Assumption (A5) that $\sup_{a, b \in V_n \atop S \in J_{a,b}^{m,n}} \rho_{0|ab|S}^* \leq \lambda_{\text{max}}$ for some positive constant $\lambda_{\text{max}}$.

The next theorem establishes that the nonPC algorithm consistently estimates the skeleton of a sparse high-dimensional DAG, thereby providing the necessary theoretical guarantees to our proposed methodology. It is worth noting that in the sample version of the PC-stable and hence the nonPC algorithm, all the inference is done during the skeleton estimation step. The second step that involves appropriately orienting the edges of the estimated skeleton, is purely deterministic (see Sections 4.2 and 4.3 in Colombo and Maathuis, 2014). Therefore to prove the consistency of the nonPC algorithm in estimating the equivalence class of the underlying true DAG, it is enough to prove the consistency of the estimated skeleton.

**Theorem 3.3.** Assume that Conditions (A1)–(A5) hold. Let $G_{\text{skel}, n}$ be the true skeleton of the graph $G_n$, and $\hat{G}_{\text{skel}, n}$ be the skeleton estimated by the nonPC algorithm. Then as $n \to \infty$, $\mathbb{P}(\hat{G}_{\text{skel}, n} = G_{\text{skel}, n}) \to 1$.

### 3.2. The Nonparametric FCI Algorithm in High Dimensions

The FCI is a modification of the PC algorithm that accounts for latent and selection variables. Thus generalizations of the PC algorithm naturally extend to the FCI as well. Similar to nonPC, we propose testing for conditional independence relations nonparametrically in the sample version of the FCI-stable algorithm (Algorithm 3 in Appendix A) based on $\rho_0^*$, instead of partial correlations. We coin the resulting algorithm the ‘nonFCI’ algorithm, to emphasize that it is a generalization of parametric FCI-stable algorithms. Again, the oracle version of the nonFCI is exactly the same as that of the FCI-stable algorithm. The difference is in the implementation of the tests for conditional independence relationships in their sample versions. This broadens the scope of the FCI algorithm in causal structural learning for observational data in the presence of latent and selection variables when Gaussianity is not a viable assumption. More specifically, it enables capturing non-linear and non-monotone conditional dependence relationships among the variables that partial correlations would fail to detect.
Equipped with the theoretical guarantees we established for the nonPC in Section 3.1, we establish below in Theorem 3.5 the consistency of the nonFCI algorithm for general distributions in sparse high-dimensional settings. Let $\mathcal{H} = (V, E)$ be a DAG with the vertex set partitioned as $V = V_X \cup V_L \cup V_T$, where $V_X$ indexes the set of $p$ observed variables, $V_L$ denotes the set of latent variables and $V_T$ stands for the set of selection variables. Let $\mathcal{M}$ be the unique MAG over $V_X$. We let $p = p_n$, $\mathcal{H} = \mathcal{H}_n$ and $Q = Q_n$, where $Q$ is the distribution of $(U_1, \ldots, U_p) := (X_1|V_T, \ldots, X_p|V_T)$. We provide below the definition of possible-D-SEP sets (Definition 3.3 in Colombo et al., 2012).

Definition 3.4. Let $\mathcal{C}$ be a graph with any of the following edge types: $\circ \rightarrow$, $\circ \leftarrow$ and $\leftrightarrow$. A possible-D-SEP $(X_a, X_b)$ in $\mathcal{C}$, denoted $\text{pds}(\mathcal{C}, X_a, X_b)$, is defined as follows: $X_c \in \text{pds}(\mathcal{C}, X_a, X_b)$ if and only if there is a path $\pi$ between $X_a$ and $X_c$ in $\mathcal{C}$ such that for every subpath $(X_e, X_f, X_g)$ of $\pi$, $X_f$ is a collider on the subpath in $\mathcal{C}$ or $(X_e, X_f, X_g)$ is a triangle in $\mathcal{C}$.

To prove the consistency of the nonFCI algorithm in sparse high-dimensional settings, we impose the following regularity conditions, which are similar to the assumptions imposed in Section 4 in Colombo et al. (2012).

(C3) The distribution $Q_n$ is faithful to the underlying MAG $\mathcal{M}_n$ for all $n$.

(C4) The maximum size of the possible-D-SEP sets for finding the final skeleton in the FCI-stable algorithm (Algorithm 3 in Appendix A), $q'_n$, grows at the rate of $O(n^{1-b})$, where $0 < b \leq 1$.

(C5) For any $U_i, U_j \in \{U_1, \ldots, U_{p_n}\}$ and $U_S \subseteq \{U_1, \ldots, U_{p_n}\}\{U_i, U_j\}$ with $|U_S| \leq q'_n$, assume

\[
\inf \{|\rho_0(U_i, U_j|U_S)\} : \rho_0(U_i, U_j|U_S) \neq 0\} \geq \lambda'_{\min}, \quad (\lambda'_{\min})^{-1} = O(n^r)
\]

and

\[
\sup |\rho_0(U_i, U_j|U_S)| \leq \lambda'_{\max},
\]

where $\lambda'_{\min}, \lambda'_{\max}$ are positive constants and $0 < r < 1/4$.

Theorem 3.5. Suppose conditions (A1)–(A3) and (C3)–(C5) hold. Denote by $\mathcal{C}_n$ and $\mathcal{C}_n^*$ the true underlying FCI-PAG and the output of the nonFCI algorithm, respectively. Then as $n \to \infty$, $\mathbb{P}(\mathcal{C}_n^* = \mathcal{C}_n) \to 1$.

4. Numerical Studies

4.1. Performance of the nonPC Algorithm

In this subsection, we compare the performances of the nonPC and the PC-stable algorithms in finding the skeleton and the CPDAG for various simulated datasets. We simulate random DAGs in the following examples and sample from probability distributions faithful to them.

Example 4.1 (Linear SEM). We first fix a sparsity parameter $s \in (0, 1)$ and enumerate the vertices as $V = \{1, \ldots, p\}$. We then construct a $p \times p$ adjacency matrix $\Lambda$ as follows. First initialize $\Lambda$ as a zero matrix. Next, fill every entry in the lower triangle (below the diagonal) of $\Lambda$ by independent realizations of Bernoulli random variables with success probability $s$. Finally replace each nonzero entry in $\Lambda$ by independent realizations of a Uniform(0,1) random variable.

In this scheme, each node has the same expected degree $E(m) = (p-1)s$, where $m$ is the degree of a node and follows a Binomial $(p-1, s)$ distribution. Using the adjacency matrix $\Lambda$, the data are then generated from the following linear structural equation model (SEM):

\[X = \Lambda X + \epsilon,
\]

where $\epsilon = (\epsilon_1, \ldots, \epsilon_p)$ and $\epsilon_1, \ldots, \epsilon_p$ are jointly independent. To obtain samples $\{X_1^k, \ldots, X_p^k\}_{k=1}^n$ on $\{X_1, \ldots, X_p\}$, we first sample $\{\epsilon_1^k, \ldots, \epsilon_p^k\}_{k=1}^n$ from the three following data-generating schemes. For $1 \leq k \leq n$ and $1 \leq i \leq p$,

1. Normal: Generate $\epsilon_i^k$’s independently from a standard normal distribution.
2. Copula: Generate $\epsilon_i^k$’s as in (1) and then transform the marginals to a $F_{1,1}$ distribution.
Table 1
Comparison of the average structural Hamming distances (SHD) of nonPC and PC-stable algorithms across simulation studies.

| n  | p  | $E(m)$ | Normal   | Copula   | Mixture  | Nonlinear SEM |
|----|----|--------|----------|----------|----------|----------------|
|    |    |        | nonPC | PC-stable | nonPC | PC-stable | nonPC | PC-stable | nonPC | PC-stable | nonPC | PC-stable |
| 50 | 9  | 1.4    | 3.35   | 3.05     | 5.55    | 5.75       | 3.8    | 3.5       | 2.9    | 3.7       |
| 100| 27 | 2.0    | 14.55  | 11.00    | 25.6    | 28.6       | 17.75  | 18.00     | 15.05  | 20.05     |
| 150| 81 | 2.4    | 53.70  | 43.45    | 97.3    | 121.3      | 69.05  | 77.75     | 62.583 | 95.083    |
| 200| 243| 2.8    | 186.2  | 183.4    | 331.0   | 471.45     | 250.3  | 336.1     | 213.70 | 375.45    |

3. Mixture: Generate $\epsilon_i^k$’s independently from a 50-50 mixture of a standard normal and a standard Cauchy distribution.

**Example 4.2 (Nonlinear SEM).** In this example, we first generate a $p \times p$ adjacency matrix $\Lambda$ in the similar way as in Example 4.1 and then generate the data from the following nonlinear SEM (similar to Voorman et al., 2014): $X_i = \sum_{j: \Lambda_{ij} \neq 0} f_{ij}(X_j) + \epsilon_i$ with $\epsilon_i \overset{i.i.d.}{\sim} N(0, 1)$, where $1 \leq j < i \leq p$. If the functions $f_{ij}$’s are chosen to be nonlinear, then the data will typically not correspond to a well-known multivariate distribution. We consider $f_{ij}(x_j) = b_{ij}^1 x_j + b_{ij}^2 x_j^2$, where $b_{ij}^1$ and $b_{ij}^2$ are independently sampled from $N(0, 1)$ and $N(0, 0.5)$ distributions, respectively.

With the exception of Example 4.1.1, the above examples are all non-Gaussian graphical models. We would thus expect the nonPC to perform better than the PC-stable in learning the unknown causal structure in these examples. For each of the four data generating methods considered above, we compare the Structural Hamming Distance (SHD) (Tsamardinos et al., 2006) between the estimated and the true skeletons of the underlying DAGs using the nonPC and PC-stable algorithms. The SHD between two undirected graphs is the number of edge additions or deletions necessary to make the two graphs match. Therefore larger SHD values between the estimated and the true skeleton correspond to worse estimates.

We consider 199 bootstrap replicates for the CdCov-based conditional independence tests in the implementation of our nonPC algorithm and the significance level $\alpha = 0.05$. Table 1 presents the average SHD for the different data generating schemes over 20 simulation runs, for different choices of $n, p$ and $E(m)$.

The results in Table 1 demonstrate that the nonPC performs nearly as good as the PC-stable for the Gaussian data example, in terms of the average SHD. But for each of the non-Gaussian data examples the nonPC performs better than the PC-stable in estimating the true skeleton of the underlying DAGs. The improvement in SHD becomes more substantial as the dimension grows. The superior performance of the nonPC over PC-stable for the non-Gaussian graphical models is expected, as the characterization of conditional independence by partial correlations is only valid under the assumption of joint Gaussianity.

### 4.2. Performance of the nonFCI algorithm

In this subsection, we compare the performances of the nonFCI and the FCI-stable algorithms over various simulated datasets. We first generate random DAGs as in Examples 4.1 and 4.2. To assess the impact of latent variables, we randomly define half of the variables with no parents and at least one child as latent. We do not consider selection variables. We run both the nonFCI and the FCI-stable algorithms on the above data examples with $n = 200$, $p = \{10, 20, 30, 100, 200\}$ and $\alpha = 0.01$, using 199 bootstrap replicates for the CdCov-based conditional independence tests. We consider 20 simulation runs for each of the data generating models. Table 2 reports the average SHD between the estimated and true PAG skeleton by the nonFCI and FCI-stable algorithms.

The results in Table 2 demonstrate that in both the Gaussian and non-Gaussian examples the nonFCI algorithm outperforms the FCI-stable in estimating the true PAG skeleton.
Table 2
Comparison of the average structural Hamming distances (SHD) of nonFCI and FCI-stable algorithms across simulation studies.

| p  | E.N | Normal    | Copula    | Mixture   | Nonlinear SEM |
|----|-----|-----------|-----------|-----------|---------------|
|    |     | nonFCI    | FCI-stable| nonFCI    | FCI-stable    |
|----|-----|-----------|-----------|-----------|---------------|
| 10 | 2.0 | 7.15      | 7.60      | 1.3       | 1.8           | 5.65          | 6.80          | 7.15          | 8.20          |
| 20 | 2.0 | 14.55     | 17.60     | 4.55      | 6.85          | 13.65         | 18.55         | 19.0          | 20.8          |
| 30 | 2.0 | 27.65     | 33.95     | 5.25      | 10.15         | 19.3          | 27.8          | 33.40         | 37.85         |
| 100| 3.0 | 109.30    | 150.35    | 26.95     | 60.05         | 62.25         | 111.10        | 115.2         | 149.0         |
| 200| 3.0 | 287.75    | 371.40    | 76.733    | 157.267       | 136.05        | 255.10        | 289.6         | 354.1         |

4.3. Real data example

To demonstrate the flexibility of the proposed framework, we apply the nonPC algorithm to the Montana Economic Outlook Poll dataset. The poll was conducted in May 1992 where a random sample of 209 Montana residents were asked whether their personal financial status was worse, the same or better than a year ago, and whether they thought the state economic outlook was better than the year before. Accompanying demographic information on the respondents’ age, income, political orientation, and area of residence in the state were also recorded. We obtained the dataset from the Data and Story Library (DASL), available at https://math.tntech.edu/e-stat/DASL/page4.html. The study comprises of the following 7 categorical variables: AGE = 1 for under 35, 2 for 35-54, 3 for 55 and over; SEX = 0 for male, 1 for female; INC = yearly income: 1 for under $20K, 2 for $20-35K, 3 for over $35K; POL = 1 for Democrat, 2 for Independent, 3 for Republican; AREA = 1 for Western, 2 for Northeastern, 3 for Southeastern Montana; FIN (Financial status): 1 for worse, 2 for same, 3 for better than a year ago; and STAT (State economic outlook): 1 for better, 0 for not better than a year ago.

After removing the cases with missing values, we are left with n = 163 samples. Since all the variables are categorical, the Gaussianity assumption is outrightly violated. Thus we would expect the nonPC to perform better than the PC-stable in learning the true causal structure among the variables in this case. Figure 1 below presents the CPDAGs estimated by the nonPC and PC-stable algorithms at a significance level α = 0.1 with 199 bootstrap replicates for the CdCov-based conditional independence tests.

It is quite intuitive that age and sex are likely to affect the income; one’s financial status and the area of residence might also influence their political inclination; and improvements or downturns in the state economic outlook might impact an individual’s financial status. The CPDAG estimated by the nonPC algorithm in Figure 1a affirms such common-sense understandings of these causal influences. However, in the CPDAG estimated by the PC-stable in Figure 1b, the edge between age and income is missing. Also the directed edges POL → AREA and POL → FIN seem to make little sense in this case.

5. Proofs of the Theoretical Results

In this section, we provide detailed technical proofs of the theoretical results presented in the paper. We first state and prove a concentration inequality that will be used in the proof of Theorem 3.1.

Lemma 5.1. Consider a U-statistic $U_n = U(X_1, \ldots, X_n) = \binom{n}{m}^{-1} \sum_{i_1 < \cdots < i_m} h(X_{i_1}, \ldots, X_{i_m})$ with a symmetric kernel $h$ such that $E U_n = Eh(X_1, \ldots, X_m) = \theta$. Further suppose $|h(X_1, \ldots, X_m)| \leq M$ for some $M > 0$. Then for any $\epsilon > 0$, we have

$$
P(\left| U_n - \theta \right| > \epsilon) \leq 2 \exp \left( -\frac{\epsilon^2 k}{2M^2} \right),$$

where $k := \left\lfloor \frac{n}{m} \right\rfloor$.
Proof of Lemma 5.1. Define

\[ W(X_1, \ldots, X_n) := \frac{1}{k} \left[ h(X_1, \ldots, X_m) + h(X_{m+1}, \ldots, X_{2m}) + \cdots + h(X_{km-m+1}, \ldots, X_{km}) \right] . \]

Then, following Section 5.1.6 in Serfling (1980), we can write

\[ U_n = \frac{1}{n!} \sum_{\pi} W(X_{i_1}, \ldots, X_{i_n}), \quad (5.1) \]

where \( \sum_{\pi} \) denotes summation over all \( n! \) permutations \((i_1, \ldots, i_n)\) of \( (1, 2, \ldots, n) \). Thus \( U_n \) can be expressed as an average of \( n! \) terms, each of which is an average of \( k \) i.i.d. random variables. Using Markov’s inequality, convexity of the exponential function and Jensen’s inequality, we have, for any \( t > 0 \),

\[
P(U_n - \theta > \epsilon) = \mathbb{P}\left( \exp \left( t(U_n - \theta) \right) > \exp \left( t\epsilon \right) \right) \leq \exp(-t\epsilon) \mathbb{E} \left[ \exp \left( tU_n \right) \right] \]

\[
= \exp(-t\epsilon) \exp(-t\theta) \mathbb{E} \left[ \exp \left( \frac{1}{n!} \sum_{p} W(X_{i_1}, \ldots, X_{i_n}) \right) \right] \]

\[
\leq \exp(-t\epsilon) \exp(-t\theta) \frac{1}{n!} \sum_{\pi} \mathbb{E} \left[ \exp \left( tW(X_{i_1}, \ldots, X_{i_n}) \right) \right] \]

\[
= \exp(-t\epsilon) \exp(-t\theta) \left[ \mathbb{E} \left( \exp \left( \frac{t}{k} h \right) \right) \right]^k \]

\[
= \exp(-t\epsilon) \mathbb{E}^k \left[ \exp \left( \frac{t}{k} (h - \theta) \right) \right] , \quad (5.2)
\]

where, for notational simplicity, we use \( h \) to denote \( h(X_1, \ldots, X_m) \). Using Hoeffding’s Lemma, we have from (5.2)

\[
P(U_n - \theta > \epsilon) \leq \exp \left( -t\epsilon + k \frac{1}{8} \frac{t^2}{k^2} (2M)^2 \right) = \exp \left( -t\epsilon + \frac{t^2M^2}{2k} \right) .
\]

Symmetrically, we obtain

\[
P(|U_n - \theta| > \epsilon) \leq 2 \exp \left( -t\epsilon + \frac{t^2M^2}{2k} \right) . \quad (5.3)
\]
The right hand side of (5.3) is minimized at \( t = \epsilon k/M^2 \). Therefore, choosing \( t = \epsilon k/M^2 \), we get

\[
P(\{|U_n - \theta| > \epsilon\}) \leq 2 \exp \left( -\frac{\epsilon^2 k}{2M^2} \right).
\]

\( \square \)

**Proof of Theorem 3.1.** When \(|S| = 0\), it can be shown in similar lines of Theorem 1 in Li et al. (2012) that for any \( \epsilon > 0 \), there exist positive constants \( A, B \) and \( \gamma \in (0,1/4) \) such that

\[
P(\{\hat{\rho}^*(X_a, X_b|X_S) - \rho_0^*(X_a, X_b|X_S) > \epsilon\}) \leq O \left( 2 \exp \left( -A n^{1-2\gamma} \epsilon^2 \right) + n \exp \left( -B n^\gamma \right) \right).
\]

Now consider the case \( 0 < |S| \leq m_p \). For notational convenience, we treat \( X_a, X_b \) and \( X_S \) as \( X, Y \) and \( Z \), respectively.

Denote \( \delta_Z := \text{CdCov}(X, Y|Z) \). Then, \( \rho_0^* = \mathbb{E}[\delta_Z] \). Recall that

\[
\hat{\rho}^*(X, Y|Z) := \frac{1}{n} \sum_{i=1}^{n} \text{CdCov}_n^2(X, Y|Z_i) := \frac{1}{n} \sum_{i=1}^{n} \Delta_{i,j,k,l,u}, \quad \text{(5.4)}
\]

where

\[
\Delta_{i,j,k,l,u} := \sum_{i,j,k,l} \frac{K_{iu} K_{ju} K_{ku} K_{iu}}{12 \left( \sum_{i=1}^{n} K_{iu} \right)^4} d_{ijkl}^s.
\]

From (5.4), we have

\[
\mathbb{E} \left[ \text{CdCov}_n^2(X, Y|Z_u)|Z \right] = \frac{1}{12} \mathbb{E} \left[ d_{1234}^s | Z_1 = Z_u, \ldots, Z_4 = Z_u \right] \sum_{i,j,k,l} K_{iu} K_{ju} K_{ku} K_{iu} / \left( \sum_{i=1}^{n} K_{iu} \right)^4 = \frac{1}{12} \mathbb{E} \left[ d_{1234}^s | Z_1 = Z_u, \ldots, Z_4 = Z_u \right] = \delta_Z, \quad \text{(5.5)}
\]

where the last equality follows from Lemma 1 in Wang et al. (2015). Together, (5.4) and (5.5) imply

\[
\mathbb{E}[\hat{\rho}^*] = \rho_0^*.
\]

Now consider the truncation

\[
\rho_0^* = \rho_0^{*1} + \rho_0^{*2}
\]

\[
:= \mathbb{E} \left[ \frac{1}{12} d_{i,j,k,l}^s 1 \left( \left| \frac{1}{12} d_{i,j,k,l}^s \right| \leq M \right) \right] + \mathbb{E} \left[ \frac{1}{12} d_{i,j,k,l}^s 1 \left( \left| \frac{1}{12} d_{i,j,k,l}^s \right| > M \right) \right], \quad \text{(5.6)}
\]

where \( M > 0 \) will be specified later. Then, using triangle inequality,

\[
P(\{|\hat{\rho}^* - \rho_0^*| > \epsilon\}) = P \left( \left| \frac{1}{n} \sum_{i,j,k,l} \Delta_{i,j,k,l,u} - \rho_0^* \right| > \epsilon \right)
\]

\[
\leq P \left( \left| \frac{1}{n} \sum_{i,j,k,l} \Delta_{i,j,k,l,u} 1 \left( \left| \frac{1}{12} d_{i,j,k,l}^s \right| \leq \rho_0^{*1} \right) \right| > \epsilon/2 \right) + P \left( \left| \frac{1}{n} \sum_{i,j,k,l} \Delta_{i,j,k,l,u} 1 \left( \left| \frac{1}{12} d_{i,j,k,l}^s \right| > \rho_0^{*2} \right) \right| > \epsilon/2 \right)
\]

\[
:= I + II. \quad \text{(5.7)}
\]

Clearly from (5.4) we have \( |\Delta_{i,j,k,l,u}| \leq M \) when \( \left| \frac{1}{12} d_{i,j,k,l}^s \right| \leq M \). With this observation, we have

\[
I \leq 2 \exp \left( -n \frac{\epsilon^2}{8M^2} \right), \quad \text{(5.8)}
\]
which follows from Lemma 5.1 by setting \( m = 1, k = \lfloor n \rfloor \) and \( \epsilon = \epsilon/2 \). Choosing \( M = cn^\gamma \) for \( \gamma \in (0, 1/4) \) and some positive constant \( c \), it follows from (5.8) that

\[
I \leq 2 \exp(-C_1 n^{1-2\gamma} \epsilon^2),
\]

for some \( C_1 > 0 \).

Now to find a suitable upper bound for II, note that a simple application of triangle inequality yields

\[
\frac{\epsilon}{2} < \left| \frac{1}{n} \sum_{u=1}^{n} \sum_{i,j,k,l} \Delta_{i,j,k,l;u} 1 \left( \frac{1}{12} d^S_{i,j,k,l} > M \right) - \rho^*_0 \right| \leq \left| \frac{1}{n} \sum_{u=1}^{n} \sum_{i,j,k,l} \Delta_{i,j,k,l;u} 1 \left( \frac{1}{12} d^S_{i,j,k,l} > M \right) \right| + |\rho^*_0|.
\]

(5.10)

For the choice of \( M = cn^\gamma \), we have

\[
\rho^*_0 = \mathbb{E} \left[ \frac{1}{12} d^S_{i,j,k,l} 1 \left( \frac{1}{12} d^S_{i,j,k,l} > M \right) \right] < \frac{\epsilon}{4}
\]

(5.11)

for sufficiently large \( n \) (see, for example, Exercise 6 in Chapter 5, Resnick (1999)). Combining (5.10) and (5.11), we get

\[
\left\{ \left| \frac{1}{n} \sum_{u=1}^{n} \sum_{i,j,k,l} \Delta_{i,j,k,l;u} 1 \left( \frac{1}{12} d^S_{i,j,k,l} > M \right) - \rho^*_0 \right| > \epsilon/2 \right\} \subseteq \left\{ \left| \frac{1}{n} \sum_{u=1}^{n} \sum_{i,j,k,l} \Delta_{i,j,k,l;u} 1 \left( \frac{1}{12} d^S_{i,j,k,l} > M \right) \right| > \epsilon/4 \right\} \subseteq \left\{ \left| \frac{1}{12} d^S_{i,j,k,l} > M \right| \text{ for some } 1 \leq i, j, k, l \leq n \right\},
\]

which implies

\[
\mathbb{P} \left( \left| \frac{1}{n} \sum_{u=1}^{n} \sum_{i,j,k,l} \Delta_{i,j,k,l;u} 1 \left( \frac{1}{12} d^S_{i,j,k,l} > M \right) - \rho^*_0 \right| > \epsilon/2 \right) \leq \mathbb{P} \left( \left| \frac{1}{n} \sum_{u=1}^{n} \sum_{i,j,k,l} \Delta_{i,j,k,l;u} 1 \left( \frac{1}{12} d^S_{i,j,k,l} > M \right) \right| > \epsilon/4 \right) \leq n^4 \mathbb{P} \left( \left| \frac{1}{12} d^S_{i,j,k,l} > M \right| \right).
\]

(5.12)

This is because if \( \left| \frac{1}{12} d^S_{i,j,k,l} \right| \leq M \) for all \( 1 \leq i, j, k, l \leq n \), then

\[
n^{-1} \sum_{u=1}^{n} \sum_{i,j,k,l} \Delta_{i,j,k,l;u} 1 \left( \frac{1}{12} d^S_{i,j,k,l} > M \right) = 0.
\]

Under Condition (A1), Lemma 2 in the supplementary materials of Wen et al. (2018) proves that there exists \( s > 0 \) for which \( \mathbb{E} [\exp(s \mid d^S_{234})] \) is finite. Using Markov’s inequality, we have

\[
\mathbb{P} \left( \left| \frac{1}{12} d^S_{i,j,k,l} \right| > M \right) \leq \mathbb{P} \left( \exp \left( s \left| \frac{1}{12} d^S_{i,j,k,l} \right| > \exp(sM) \right) \right) \leq \exp(-sM) \mathbb{E} \left[ \exp \left( s \left| \frac{1}{12} d^S_{i,j,k,l} \right| \right) \right] \leq C_2 \exp(-sM) \leq C_2 \exp(-s_1 n^\gamma),
\]

(5.13)
for some positive constants \( C_2 \) and \( s_1 \), where last line uses the fact that \( M = c n^{\gamma} \). Combining (5.12) and (5.13), we have

\[
\Pi \leq C_2 n^4 \exp(-s_1 n^{\gamma}). \tag{5.14}
\]

Finally, combining (5.7), (5.9) and (5.14), we get

\[
P(|\hat{\rho}^* - \rho_0^*| > \epsilon/2) \leq 2 \exp\left(-C_1 n^{1-2\gamma} \epsilon^2\right) + C_2 n^4 \exp\left(-s_1 n^{\gamma}\right),
\]

for some positive constants \( \gamma, C_1, C_2 \) and \( s_1 \). This completes the proof of the theorem.

\[\square\]

**Proof of Theorem 3.2.** The first inequality in Theorem 3.2 simply follows by observing the fact that for any generic random sequence \( \{X_n\}_{n=1}^\infty \) and any \( \epsilon > 0 \),

\[
P(|X_n| > \epsilon) \leq P(\sup_n |X_n| > \epsilon)
\]

for all \( n \geq 1 \), which in turn implies

\[
\sup_n P(|X_n| > \epsilon) \leq P(\sup_n |X_n| > \epsilon).
\]

The second inequality follows from union bound and Theorem 3.1.

\[\square\]

**Proof of Theorem 3.3.** Denote by \( E_{ab|S} \) the event that “an error occurs while testing for \( X_a \perp X_b | X_S \)” for \( a, b \in V \) and \( S \in J_{a,b}^{m_{pa}} \). Then

\[
P(\text{an error occurs in the nonPC algorithm}) \leq P\left(\bigcup_{a, b \in V} E_{ab|S}\right) \lesssim p_{m_{pa}}^{n} + 2 P(E_{ab|S}), \tag{5.15}
\]

which is essentially due to the union bound. Now, we can write \( E_{ab|S} = E_{ab|S}^I \cup E_{ab|S}^{II} \), where

\[
\begin{align*}
\text{(Type I error)} & \quad E_{ab|S}^I : |\hat{\rho}_{ab|S}^*| > \xi_n \quad \text{when } \rho_{0;ab|S}^* = 0 \\
\text{and} & \quad \text{(Type II error)} \quad E_{ab|S}^{II} : |\hat{\rho}_{ab|S}^*| \leq \xi_n \quad \text{when } \rho_{0;ab|S}^* > 0.
\end{align*}
\]

Then by the using triangle inequality

\[
P(E_{ab|S}^I) = P(|\hat{\rho}_{ab|S}^*| > \xi_n) = P(|\hat{\rho}_{ab|S}^* - \rho_{0;ab|S}^* + \rho_{0;ab|S}^*| > \xi_n)
\leq P(|\hat{\rho}_{ab|S}^* - \rho_{0;ab|S}^*| > \xi_n - C_{max}) \tag{5.16}
\lesssim 2 \exp\left(-A n^{1-2\gamma}(\xi_n - C_{max})^2\right) + n^4 \exp\left(-B n^{\gamma}\right)
\]

for positive constants \( A, B \) and \( \gamma \in (0, 1/4) \), where the last inequality follows from Theorem 3.2. Similarly, using the definition of \( C_{min} \) and the identity \( |a| - |b| \leq |a - b| \) for \( a, b \in \mathbb{R} \), we have

\[
P\left(E_{ab|S}^{II}\right) = P(|\hat{\rho}_{ab|S}^*| \leq \xi_n) = P(-|\hat{\rho}_{ab|S}^*| \geq -\xi_n)
= P(|\rho_{0;ab|S}^*| - |\hat{\rho}_{ab|S}^*| \geq \xi_n)
\leq P(|\rho_{0;ab|S}^*| - |\hat{\rho}_{ab|S}^*| \geq C_{min} - \xi_n)
\lesssim 2 \exp\left(-A n^{1-2\gamma}(\xi_n - C_{min})^2\right) + n^4 \exp\left(-B n^{\gamma}\right). \tag{5.17}
\]

Again the last inequality follows from Theorem 3.2. Combining equations (5.15), (5.16) and (5.17), we have

\[
P(\text{an error occurs in the nonPC algorithm})
= O\left(p_{m_{pa}}^{n} + 2 \exp\left(-A n^{1-2\gamma}(\xi_n - C_{max})^2\right) + 2 \exp\left(-A n^{1-2\gamma}(\xi_n - C_{min})^2\right)
+ n^4 \exp\left(-B n^{\gamma}\right)\right)
= o(1),
\]
where the last step follows from the fact that $\gamma \in (0, 1/4)$ and Assumption (A5). This implies that as $n \to \infty$,
\[
P\left( \hat{G}_{\text{skel}, n} = G_{\text{skel}, n} \right) = 1 - P(\text{an error occurs in the nonPC algorithm}) \to 1.\]

\[\diamond\]

Proof of Theorem 3.5. The proof follows similar lines of the proof of Theorem 4.2 in Colombo et al. (2012), replacing Lemma 1.4 in their supplement by Theorem 3.2 in our paper.

\[\diamond\]

6. Discussion

We proposed nonparametric variants of the widely popular PC-stable and FCI-stable algorithms, which employ conditional distance covariance (CdCov) to test for conditional independence relationships in their sample versions. Our proposed algorithms broaden the applicability of the PC/PC-stable and FCI/FCI-stable algorithms to general distributions over DAGs, and enable taking into account non-linear and non-monotone conditional dependence among the random variables, which partial correlations fail to capture. We show that the high-dimensional consistency of the PC-stable and FCI-stable algorithms carry over to general distributions over DAGs when we implement CdCov-based nonparametric tests for conditional independence. Our consistency results only require mild moment and tail conditions on the set of variables, without requiring any strict distributional assumptions.

There are several intriguing potential directions for future research. First, it is generally difficult to select the tuning parameter (i.e., the significance threshold for the CdCov test) in causal structure learning. One possible strategy is to use ideas based on stability selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013). By assessing the stability of the estimated graphs in multiple subsamples, this strategy allows us to choose the tuning parameter in order to control the false positive error. However, the repeated subsampling increases the computational burden. Second, the computational and sample complexities of the PC and FCI algorithms (and hence those of the nonPC and nonFCI) scale with the maximum degree of the DAG, which is assumed to be small relative to the sample size. However, in many applications one encounters sparse graphs containing a small number of highly connected ‘hub’ nodes. In such cases, Sondhi and Shojaie (2019) proposed a low-complexity variant of the PC algorithm, namely the reduced PC (rPC) algorithm, that exploits the local separation property of large random networks (Anandkumar et al., 2012). The rPC is shown to consistently estimate the skeleton of a high-dimensional DAG by conditioning only on sets of small cardinality. In this light, it would be intriguing to develop computationally faster variants of the nonPC and nonFCI in future by exploiting the idea of local separation.

Appendix A: Preliminaries

For the sake of completeness, we illustrate in this section the pseudocodes of the oracle versions of the PC-stable and FCI-stable algorithms.

Algorithms 1 presents the pseudocode of the oracle version of Step 1 of the PC-stable algorithm (Algorithm 4.1 of Colombo and Maathuis, 2014), which estimates the skeleton of the underlying DAG. Algorithm 2 presents the pseudocode of Step 2 of the PC-stable algorithm (Algorithm 2 of Kalisch and Bühlmann, 2007), that extends the skeleton estimated in Step 1 to the CPDAG. Algorithm 3 presents the pseudocode of the FCI-stable algorithm (Section 4.4 in Colombo and Maathuis, 2014). It implements Algorithm 4 to obtain an initial skeleton of the underlying PAG, Algorithm 5 to orient the v-structures, and finally Algorithm 6 to obtain the final skeleton that the FCI-stable returns.
Algorithm 1 Step 1 of the PC-stable algorithm (oracle version)

Require: Conditional independence information among all variables in \( V \), and an ordering \( \text{order}(V) \) on the variables.
Form the complete undirected graph \( C \) on the vertex set \( V \).
Let \( l = -1 \);
repeat
    \( l = l + 1 \);
    for all vertices \( X_a \) in \( C \) do
        let \( u(X_a) = \text{adj}(C, X_a) \)
    end for
repeat
    Select a (new) ordered pair of vertices \( (X_a, X_b) \) that are adjacent in \( C \) such that \(|u(X_a) \setminus \{X_b\}| \geq l\), using \text{order}(V);
repeat
    Choose a (new) set \( S \subseteq u(X_a) \setminus \{X_b\} \) with \(|S| = l\), using \text{order}(V);
    if \( X_a \not\perp_{C} X_b | S \) then
        Delete the edge \( X_a - X_b \) from \( C \);
        Let \( \text{sepset}(X_a, X_b) = \text{sepset}(X_b, X_a) = S \);
    end if
until \( X_a \) and \( X_b \) are no longer adjacent in \( C \) or all \( S \subseteq u(X_a) \setminus \{X_b\} \) with \(|S| = l\) have been considered
until all ordered pairs of adjacent vertices \( (X_a, X_b) \) in \( C \) with \(|u(X_a) \setminus \{X_b\}| \geq l\) have been considered
until all pairs of adjacent vertices \( (X_a, X_b) \) in \( C \) satisfy \(|u(X_a) \setminus \{X_b\}| \leq l\)
Output: The estimated skeleton \( C \), separation sets \( \text{sepset} \).

Algorithm 2 Step 2 of the PC-stable algorithm

Require: Skeleton \( C \), separation sets \( \text{sepset} \).
for all pair of nonadjacent vertices \( X_a, X_c \) with common neighbor \( X_b \) in \( C \) do
    if \( X_b \not\in \text{sepset}(X_a, X_c) \) then
        Replace \( X_a - X_b - X_c \) in \( C \) by \( X_a \rightarrow X_b \leftarrow X_c \);
    end if
end for
In the resulting PDAG, try to orient as many undirected edges as possible by repeated applications of the following rules:
(R1) Orient \( X_b - X_c \) into \( X_b \rightarrow X_c \) whenever there is an arrow \( X_a \rightarrow X_b \) such that \( X_a \) and \( X_c \) are nonadjacent (otherwise a new v-structure is created).
(R2) Orient \( X_a - X_c \) into \( X_a \rightarrow X_c \) whenever there is a chain \( X_a \rightarrow X_b \rightarrow X_c \) (otherwise a directed cycle is created).
(R3) Orient \( X_b - X_c \) into \( X_b \rightarrow X_c \) whenever there are two chains \( X_a - X_b \rightarrow X_c \) and \( X_a - X_d \rightarrow X_c \) such that \( X_b \) and \( X_d \) are nonadjacent (otherwise a new v-structure or a directed cycle is created).

Algorithm 3 The FCI-stable algorithm (oracle version)

Require: Conditional independence information among all variables in \( Y_X \) given \( Y_T \).
Use Algorithm 4 to find an initial skeleton \( (C) \), separation sets \( \text{sepset} \) and unshielded triple list \( (M) \); Use Algorithm 5 to orient v-structures (update \( C \)); Use Algorithm 6 to find the final skeleton (update \( C \) and \( \text{sepset} \)); Use Algorithm 5 to orient v-structures (update \( C \)); Use rules (R1)-(R10) of Zhang (2008) to orient as many edge marks as possible (update \( C \));
Output: \( C \), \( \text{sepset} \).
Algorithm 4 Obtaining an initial skeleton in the FCI-stable algorithm (Algorithm 4.1 in the supplement of Colombo et al., 2012)

Require: Conditional independence information among all variables in $V_X$ given $V_T$, and an ordering order($V_X$) on the variables.
Form the complete undirected graph $C$ on the vertex set $V_X$ with edges o-o.
Let $l = -1$;
repeat
  $l = l + 1$;
  for all vertices $X_a$ in $C$ do
    let $u(X_a) = \text{adj}(C, X_a)$
  end for
repeat
  Select a (new) ordered pair of vertices $(X_a, X_b)$ that are adjacent in $C$ such that
  $|u(X_a) \setminus \{X_b\}| \geq l$, using order $V_X$;
repeat
  Choose a (new) set $Y \subseteq u(X_a) \setminus \{X_b\}$ with $|Y| = l$, using order($V_X$);
  if $X_a \perp X_b | Y \cup V_T$ then
    Delete the edge $X_a \leftrightarrow X_b$ from $C$;
    Let sepset($X_a, X_b$) = sepset($X_b, X_a$) = $Y$;
  end if
until $X_a$ and $X_b$ are no longer adjacent in $C$ or all $Y \subseteq u(X_a) \setminus \{X_b\}$ with $|Y| = l$ have been considered
until all ordered pairs of adjacent vertices $(X_a, X_b)$ in $C$ with $|u(X_a) \setminus \{X_b\}| \geq l$ have been considered
 until all pairs of adjacent vertices $(X_a, X_b)$ in $C$ satisfy $|u(X_a) \setminus \{X_b\}| \leq l$
Form a list $M$ of all unshielded triples $\langle X_c, \cdot, X_d \rangle$ (i.e., the middle vertex is left unspecified) in $C$ with $c < d$.
Output: $C$, sepset, $M$.

Algorithm 5 Orienting v-structures in the FCI-stable algorithm (Algorithm 4.2 in the supplement of Colombo et al., 2012)

Require: Initial skeleton ($C$), separation sets (sepset) and unshielded triple list ($M$).
for all elements $(X_a, X_b, X_c)$ of $M$ do
  if $X_b \notin \text{sepset}(X_a, X_c)$ then Orient $X_a \leftrightarrow X_b \leftrightarrow X_c$ as $X_a \leftrightarrow X_b \leftrightarrow X_c$
end if
end for
Output: $C$, sepset.

Algorithm 6 Obtaining the final skeleton in the FCI-stable algorithm (Algorithm 4.3 in the supplement of Colombo et al., 2012)

Require: Partially oriented graph ($C$) and separation sets (sepset).
for all vertices $X_a$ in $C$ do
  let $v(X_a) = \text{pds}(C, X_a, \cdot)$;
  for all vertices $X_b \in \text{adj}(C, X_a)$ do
    Let $l = -1$;
    repeat
      $l = l + 1$;
      repeat
        Choose a (new) set $Y \subseteq v(X_a) \setminus \{X_b\}$ with $|Y| = l$;
        if $X_a \perp X_b | Y \cup V_T$ then
          Delete the edge $X_a \leftrightarrow X_b$ from $C$;
          Let sepset($X_a, X_b$) = sepset($X_b, X_a$) = $Y$;
        end if
      until $X_a$ and $X_b$ are no longer adjacent in $C$ or all $Y \subseteq v(X_a) \setminus \{X_b\}$ with $|Y| = l$ have been considered
      until $X_a$ and $X_b$ are no longer adjacent in $C$ or $|v(X_a) \setminus \{X_b\}| < l$
    end for
  end for
Reorient all edges in $C$ as o-o.
Form a list $M$ of all unshielded triples $\langle X_c, \cdot, X_d \rangle$ in $C$ with $c < d$.
Output: $C$, sepset, $M$. 
References

[1] Anandkumar, A., Tan, V.Y.F., Huang, F. and Willsky, A.S. (2012). High-Dimensional Gaussian Graphical Model Selection: Walk Summability and Local Separation Criterion. *Journal of Machine Learning Research*, 13, 2293-2337.

[2] Ali, R.A., Richardson, T.S. and Spirtes, P. (2009). Markov equivalence for ancestral graphs. *Annals of Statistics*, 37(5B) 2808-2837.

[3] Chakraborty, S. and Zhang, X. (2019). Distance Metrics for Measuring Joint Dependence with Application to Causal Inference. *Journal of the American Statistical Association*, 114(528), 1638-1650.

[4] Colombo, D., Maathuis, M.H., Kalisch, M. and Richardson, T.S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, 40(1) 294-321.

[5] Colombo, D. and Maathuis, M.H. (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15 3921-3962.

[6] Harris, N. and Drton, M. (2013). PC Algorithm for Nonparanormal Graphical Models. *Journal of Machine Learning Research*, 14 3365-3383.

[7] Kalisch, M. and Bühlmann, P. (2007). Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research*, 8 613-636.

[8] Lauritzen, S.L. (1996). Graphical models . Oxford University Press.

[9] Li, R., Zhong, W. and Zhu, L. (2012). Feature selection via distance correlation learning. *Journal of the American Statistical Association*, 107(499) 1129-1139.

[10] Liu, J., Li, R. and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, 109(505) 266-274.

[11] Loh, P-L. and Bühlmann, P. (2014). High-Dimensional Learning of Linear Causal Networks via Inverse Covariance Estimation. *Journal of Machine Learning Research*, 15 3065-3105.

[12] Maathuis, M., Drton, M., Lauritzen, S. and Wainwright, M. (2019). Handbook of graphical models. CRC Press.

[13] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society*, 72(4) 417–473.

[14] Pearl, J. (2000). Causality. Cambridge University Press.

[15] Resnick, S. I. (1999). A Probability Path. Springer.

[16] Richardson, T.S. and Spirtes, P. (2002). Ancestral graph markov models. *Annals of Statistics*, 30(4) 962-1030.

[17] Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics . Wiley, New York.

[18] Shah, R.D. and Samworth, R.J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society*, 75(1) 55–80.

[19] Sheng, T. and Sriperumbudur, B.K. (2019). On distance and kernel measures of conditional independence. arXiv:1912.01103.

[20] Sondhi, A. and Shojaie, A. (2019). The Reduced PC-Algorithm: Improved Causal Structure Learning in Large Random Networks. *Journal of Machine Learning Research*, 20, 1-31.

[21] Spirtes, P., Glymour, C. and Scheines, R. (2000). Causation, Prediction, and Search. The MIT Press, 2nd edition.

[22] Spirtes, P. (2001). An anytime algorithm for causal inference. *Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics*, 213-221.

[23] Sun, X., Janzing, D., Schölkopf, B. and Fukumizu, K. (2007). A kernel-based causal learning algorithm. *Proceedings of the 24th International Conference on Machine Learning*, 855-862.

[24] Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing independence by correlation of distances. *Annals of Statistics*, 35(6) 2709-2794.

[25] Székely, G. J. and Rizzo, M. L. (2014). Partial distance correlation with methods for
dissimilarities. *Annals of Statistics, 42*(6) 2382-2412.

[26] Tsamardinos, I., Brown, L.E. and Aliferis, C.F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning, 65* 31–78.

[27] Uhler, C., Raskutti, G., Bühlmann, P. and Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *Annals of Statistics, 41*(2) 436-463.

[28] Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, 255-270.*

[29] Voorman, A., Shojaie, A. and Witten, D. (2014). Graph estimation with joint additive models. *Biometrika, 99*(1) 1-25.

[30] Wang, X., Wenliang, P., Hu, W., Tian, Y. and Zhang, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association, 110*(512) 1726-1734.

[31] Wen, C., Wenliang, P., Huang, M. and Wang, X. (2018). Sure Independence Screening Adjusted for Confounding Covariates with Ultrahigh Dimensional Data. *Statistica Sinica, 28* 293-317.

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

[33] Zhang, K., Peters, J., Janzing, D. and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. arXiv:1202.3775.