Journal of Machine Learning Research 21 (2020) 1-38 Submitted 5/16; Revised 2/20; Published 3/20

Learning Causal Networks via Additive Faithfulness

Kuang-Yao Lee KUANG-YAO.LEE@TEMPLE.EDU
Department of Statistical Science, Temple University
1810 Liacouras Walk, Philadelphia, PA 19122

Tianqi Liu TIANQI.TERENCE.LIU@GMAIL.COM
Google LLC
111 8th Ave, New York, NY 10011

Bing Li BXL9@PSU.EDU
Department of Statistics, Pennsylvania State University
326 Thomas Building, University Park, PA 16802

Hongyu Zhao HONGYU.ZHAO@YALE.EDU
Department of Biostatistics, Yale School of Public Health
60 College Street, New Haven, CT 06520

Editor: Peter Spirtes

Abstract

In this paper we introduce a statistical model, called additively faithful directed acyclic graph (AFDAG), for causal learning from observational data. Our approach is based on additive conditional independence (ACI), a recently proposed three-way statistical relation that shares many similarities with conditional independence but without resorting to multi-dimensional kernels. This distinct feature strikes a balance between a parametric model and a fully nonparametric model, which makes the proposed model attractive for handling large networks. We develop an estimator for AFDAG based on a linear operator that characterizes ACI, and establish the consistency and convergence rates of this estimator, as well as the uniform consistency of the estimated DAG. Moreover, we introduce a modified PC-algorithm to implement the estimating procedure efficiently, so that its complexity is determined by the level of sparseness rather than the dimension of the network. Through simulation studies we show that our method outperforms existing methods when commonly assumed conditions such as Gaussian or Gaussian copula distributions do not hold. Finally, the usefulness of AFDAG formulation is demonstrated through an application to a proteomics data set.

Keywords: additive conditional independence, additive reproducing kernel Hilbert space, directed acyclic graph, global Markov property, normalized additive conditional covariance operator, PC-algorithm

1. Introduction

Learning causality is fundamental in many scientific disciplines, such as epidemiology, genetics, sociology, and business. See Pearl (2009), and Spirtes, Glymour, and Scheines (2000). The underlying diagram for the causal structure is often represented as a directed acyclic

graph (DAG). A directed graph consists of a finite set $\mathsf{V} = \{1, \ldots, p\}$ and a subset E of $\{(i,j) \in \mathsf{V} \times \mathsf{V}, i \neq j\}$, where V represents the set of vertices, and E the set of directed edges, with the order in (i,j) indicating $i \to j$. A DAG is a directed graph that contains no directed cycles (page 11, Pearl, 2009). Suppose $X = (X^1, \ldots, X^p)$ is a random vector and, for any $\mathsf{S} \subseteq \mathsf{V}$, let X^{S} denote the subvector $\{X^i : i \in \mathsf{S}\}$. We say that X satisfies the *global Markov property* with respect to a DAG $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ if, for any $(i,j) \in \mathsf{E}$ and any subset $\mathsf{S} \subseteq \mathsf{V} \setminus \{i,j\}$,

$$i \text{ and } j \text{ are d-separated by } \mathsf{S} \text{ under } \mathsf{G} \quad \Rightarrow \quad X^i \perp\!\!\!\perp X^j | X^{\mathsf{S}}. \tag{1}$$

Here, the notation $U \perp\!\!\!\perp V | W$ indicates that random variables U and V are conditionally independent given W, and d-separation means, loosely speaking, that all paths in G between node i and node j are blocked by the set of nodes S; see Lauritzen (1996, page 48) and Section 2.3 for its rigorous definition. An immediate implication of (1) is that two DAGs with the same set of d-separation relations share the same conditional independence structure and therefore cannot be distinguished using conditional independence. For this reason, we regard the class of all DAGs sharing the same d-separation structure as an equivalence class (see, Chickering, 2002, Section 2). It should be mentioned that some other aspects of the statistical information might enable us to further distinguish between DAGs that share the same d-separation relations (see, for example, Shimizu et al., 2006).

Many approaches have been developed to estimate the equivalence class of DAG. For example, Chickering (2002) introduced the greedy equivalence search which provides a BIC-score for each DAG; van de Geer and Bühlmann (2013) proposed likelihood-based approaches combined with non-convex optimization procedures. These methods are intuitively appealing and enjoy good statistical properties, but are computationally intensive, which restrict their use to relatively low-dimensional problems. Another type of approaches are the various forms of the PC-algorithm (Spirtes, Glymour, and Scheines, 2000; Kalisch and Bühlmann, 2007; He and Geng, 2008), which seek to infer the graphical structure by conducting a sequence of conditional independence tests. The PC-algorithm is computationally simpler and can be applied to higher-dimensional problems. As indicated in Kalisch and Bühlmann (2007), the number of tests involved in the PC-algorithm scales up with the maximum degree of the underlying graph, and is only in a polynomial order with respect to the dimensionality when the graph is sparse.

Recently, much work has been done to extend the DAG models to non-Gaussian or nonlinear settings, or both; see Hoyer et al. (2009), Mooij et al. (2009), Tillman et al. (2009), Zhang and Hyvärinen (2009), and Peters et al. (2014). Interestingly, as pointed out in Hoyer et al. (2009), deviation from Gaussianity actually mitigates the identifiability problem in making causal inference, possibly due to the asymmetric structures induced by non-Gaussianity. Although these approaches are freed from the limitations of the strong distribution assumptions, they employ fully-fledged nonparametric methods using multi-dimensional kernels, i.e. kernel functions that take vector-valued inputs. Due to the curse of dimensionality these procedures could have inferior performances with large networks. We should also mention that, Shimizu et al. (2006) proposed to estimate the DAGs via the combination of linear structural equation models and non-Gaussian errors.

1.1. Our proposal and contributions

To take advantage of the flexibility offered by a nonparametric approach without resorting to multi-dimensional kernels, in this paper we propose an alternative theory and estimating procedure for causal learning based on *Additive Conditional Independence* (ACI), a three-way statistical relation recently proposed by Li, Chun, and Zhao (2014) to construct undirected graphs. ACI resembles traditional conditional independence in many ways, but its nonparametric characterization only involves one-dimensional kernel. This feature is particularly attractive when handling high-dimensional data, because the curse of dimensionality caused by multi-dimensional kernel is one of the main hindrances to accuracy in high dimensions. We note that several recent papers have combined the graphical model and univariate transformations under a Gaussian copula assumption; see Liu, Lafferty, and Wasserman (2009), Liu, Han, Yuan, Lafferty, and Wasserman (2012), Xue and Zou (2012), and Harris and Drton (2013). These approaches offer similar advantage of not requiring multi-dimensional kernels. However, as demonstrated in Li, Chun, and Zhao (2014), ACI is capable of detecting intrinsically nonlinear interactions that can elude the Gaussian copula models.

In the classical setting, a DAG is linked to conditional independence through a *faithfulness* condition. Borrowing that idea, we introduce a new condition called *additive faithfulness* to link a DAG with a set of ACI relations, resulting in a new statistical graphical model called additively faithful directed acyclic graph, or AFDAG. We introduce a linear operator, called the *Normalized Additive Conditional Covariance Operator* (NACCO), to characterize ACI. This operator is defined on *reproducing kernel Hilbert spaces* and is equal to the zero operator if and only if ACI holds. The estimation of AFDAG is then based on repeated evaluation of this operator among pairs of nodes in the DAG. To efficiently implement this process we propose a modified PC-algorithm by combing the evaluation of ACI with a standard PC-algorithm, whose computation complexity does not depend on the size of the network but instead on its level of sparseness (Spirtes, Glymour, and Scheines, 2000; Kalisch and Bühlmann, 2007). We investigate the consistency and convergence rate of the proposed estimator. We also study its *uniform consistency* (Zhang and Spirtes, 2002) under a stronger version of additive faithfulness. Through numerical experiments, we show that this condition is weaker than the strong faithfulness condition (Uhler et al., 2013) in the linear setting.

1.2. Related work

Nonparametric testing for conditional independence (CI) has gained enormous attention over the past decades. (Linton and Gozalo, 1996; Margaritis, 2005; Su and White, 2007, 2008; Song, 2009; Huang, 2010). Some work has been proposed recently using RKHS operators. For example, Fukumizu et al. (2004, 2009) introduced the *Conditional Covariance Operator* (CCO) and established the equivalence between CCO and CI. Fukumizu et al. (2008) extended CCO to *Normalized Conditional Covariance Operator* (NCCO) by removing the marginal variations from the covariance operators. Moreover, for the purpose of causality learning, Sun et al. (2007) and Tillman et al. (2009) combined the PC algorithm with a permutation-based test of CCO. In a more recent development, Zhang et al. (2011)

proposed to replace the permutation test by the test based on the asymptotic distribution of the empirical CCO.

However, the implementation of CCO or NCCO relies on multi-dimensional kernels, which can be a source of curse of dimensionality. Our solution replaces conditional independence by additive conditional independence as the criterion to construct the DAG. Due to the additive nature of additive conditional independence, we only need one-dimensional kernels to construct the NACCO, avoiding the curse of dimensionality. Thus, our proposal is substantially different from previous work based on conditional independence, in both methodology and theory.

Methodologically, we introduce the *Normalized Additive Conditional Covariance Operator* (NACCO), which is structurally different from and cannot be deduced from NCCO. Moreover, previous works did not study the condition of strong faithfulness, which is often required by methods based on sequential tests like PC algorithm. We introduce the *strong additive faithfulness* (SAF), and show that SAF is weaker than the (linear) strong faithfulness through synthetic examples. Theoretically, since our method is considerably different from the previous ones, we needed to introduce a novel and systematic approach to study the theoretical properties of our new estimator. We derived the consistency of NACCO, which cannot be obtained directly from the consistency result of NCCO in Fukumizu et al. (2008). We also established the uniform consistency of the proposed algorithm—under our framework. This result appears to be the first of its kind, and is useful for attacking broader problems related to nonparametric casual inference.

Our method is also substantially different from the closely related work of Li et al. (2014). First, even though both papers employ the idea of additive conditional independence, the current paper concerns directed acyclic graph, whereas Li et al. (2014) concerns undirected graph. The extension from undirected graphs, also known as conditional independence graphs (CIG), to directed graphs involves a set of completely different methods and theories. Even the skeleton of a DAG, which replaces directed edges with undirected ones and thus is generally less informative than a DAG, is different from a CIG (Kalisch and Bühlmann, 2007). Moreover, a DAG can be converted into a CIG by moralization (Lauritzen, 1996), but cannot be converted from a CIG. Therefore, the methods for estimating a DAG cannot be easily derived from those for estimating CIG. The DAG estimation is often more costly, as it requires evaluation and thesholding of a much larger set of statistics. As a comparison, the algorithm of Li et al. (2014) requires $p(p-1)/2$ evaluations of additive conditional independence, while our algorithm is of order p^{s^*}, where s^* is the maximum degree of the DAG.

Second, to facilitate the theoretical development of our estimators, we modified the definition of ACI so that it is characterized via an additive reproducing kernel Hilbert space, instead of the L_2 space used in Li et al. (2014). For example, Theorem 5 of Li et al. (2014) was based on the L_2-geometry, whereas our Theorem 5 is based on the RKHS-geometry. Finally, we carried out a substantial body of new theoretical work, such as the consistency of various operators, the graph estimation consistency, strong additive faithfulness and uniform consistency, which involves invention of some new techniques and machineries.

1.3. Organization of the paper

The sections of this article are organized as follows. In Section 2, we introduce *additive faithfulness* to relate a DAG to ACI. Based on this relation we propose the AFDAG model. In Section 3, we introduce NACCO to quantify additive conditional independence. An estimating procedure is developed for NACCO in Section 4. In Section 5, we further develop the algorithm for estimating the AFDAG, which involves a PC-type algorithm that efficiently evaluates the NACCO among pairs of nodes. In Sections 6, we study the consistency of NACCO and that of the AFDAG estimator. In Section 7, we introduce *strong additive faithfulness* and establish the uniform consistency of the proposed algorithm. In Section 8, we compare the performance of the proposed estimator with some existing methods by simulation studies and apply it to a pathway analysis. Some concluding remarks are made in Section 9, and all proofs are relegated to Appendix.

2. Directed acyclic graph based on additive conditional independence

We first define *additive faithfulness*, a new concept that connects a DAG with ACI. We begin with additive reproducing kernel Hilbert spaces (RKHS), the platform of all subsequent developments.

2.1. Additive reproducing kernel Hilbert spaces and basic operators

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $X : \Omega \to \mathbb{R}^p$ be a p-dimensional random vector, whose ith component is denoted by X^i. Let Ω_X and Ω_{X^i} be the supports of X and X^i, and assume $\Omega_X = \Omega_{X^1} \times \cdots \times \Omega_{X^p}$. For each i, let $\mathscr{H}_{X^i}$ be an RKHS of functions on Ω_{X^i} to $\mathbb{R}$ defined by a positive definite kernel $\kappa_{X^i} : \Omega_{X^i} \times \Omega_{X^i} \to \mathbb{R}$. For simplicity, we assume these kernels to be the same for $i = 1, \ldots, p$ (which also means that we assume the Ω_{X^i} are the same), and denote the common kernel by κ. Thus, the inner product in $\mathscr{H}_{X^i}$ is determined by $\langle \kappa(\cdot, a), \kappa(\cdot, b) \rangle_{\mathscr{H}_{X^i}} = \kappa(a, b)$ for all $a, b \in \Omega_{X^i}$.

Let $\mathscr{H}_X$ be the direct sum $\oplus_{i=1}^p \mathscr{H}_{X^i}$, that is,

$$\mathscr{H}_X = \{f_1 + \cdots + f_p : f_1 \in \mathscr{H}_{X^1}, \ldots, f_p \in \mathscr{H}_{X^p}\},$$

with inner product defined by

$$\langle f_1 + \cdots + f_p, g_1 + \cdots + g_p \rangle_{\mathscr{H}_X} = \langle f_1, g_1 \rangle_{\mathscr{H}_{X^1}} + \cdots + \langle f_p, g_p \rangle_{\mathscr{H}_{X^p}}.$$

Our construction of $\mathscr{H}_X$ follows that of Aronszajn (1950, page 352). We call $\mathscr{H}_X$ an additive RKHS of functions on Ω_X. Similarly, for any subvector $U = (U^1, \ldots, U^r)$ of X, let $\mathscr{H}_U$ be the direct sum of $\mathscr{H}_{U^i}$, $i = 1, \ldots, r$. The following assumption on the kernel κ guarantees that $\mathscr{H}_{X^i}$ is a subspace of $L_2(P_{X^i})$, the class of all square-integrable functions of X^i. Note that $L_2(P_X)$ is not itself an RKHS, a subspace $\mathscr{K}$ in $L_2(P_X)$ is an RKHS if and only if $\mathscr{K}$ has a reproducing kernel, and an RKHS is a subspace of $L_2(P_X)$ if $E\kappa(X, X)$ is finite.

Assumption 1 $E\kappa(X^i, X^i) < \infty,\ i = 1, \ldots, p.$

This is a mild condition that holds for many commonly used kernels such as the Gaussian radial basis function. See, for example, Fukumizu, Bach, and Jordan (2009).

We now lay out some notations that will be used in the rest of the article. For two Hilbert spaces $\mathscr{H}$ and $\mathscr{K}$, let $\mathscr{B}(\mathscr{H}, \mathscr{K})$ denote the class of bounded linear operators from $\mathscr{H}$ to $\mathscr{K}$. The class $\mathscr{B}(\mathscr{H}, \mathscr{K})$ is a Banach space, endowed with the operator norm, which is denoted by $\|\cdot\|$. When $\mathscr{H} = \mathscr{K}$, we use $\mathscr{B}(\mathscr{H})$ to denote $\mathscr{B}(\mathscr{H}, \mathscr{H})$. For a linear operator A, let $\text{null}(A)$ denote the null space of A, $\text{ran}(A)$ denote the range of A, and $\overline{\text{ran}}(A)$ denote the closure of $\text{ran}\, A$. Further information about bounded operators can be found in Weidmann (1980, Chapters 6 and 7).

Under Assumption 1, one can show that, for any $(i, j) \in \mathsf{V} \times \mathsf{V}$ there exists an operator $\Sigma_{X^iX^j} \in \mathscr{B}(\mathscr{H}_{X^j}, \mathscr{H}_{X^i})$ such that

$$\langle f, \Sigma_{X^iX^j} g \rangle = \text{cov}\left[f(X^i), g(X^j)\right], \text{ for any } f \in \mathscr{H}_{X^i} \text{ and } g \in \mathscr{H}_{X^j}.$$

This operator is also called the *covariance operator* (see Baker, 1973; Fukumizu, Bach, and Jordan, 2009). Let P_{X^i,X^j}, P_{X^i}, and P_{X^j} be the distributions of (X^i, X^j), X^i, and X^j, respectively. As shown by Sejdinovic et al. (2013), the Hilbert-Schmidt norm of $\Sigma_{X^iX^j}$ is identical to the distance between the RKHS embeddings of P_{X^i,X^j} and $P_{X^i} \times P_{X^j}$, also known as the *maximum mean discrepancy* (MMD). See Gretton et al. (2012) and Muandet et al. (2014, 2016) for more details on embedding RKHS elements and MMD.

Let $\Sigma_{XX} : \mathscr{H}_X \to \mathscr{H}_X$ be the matrix of operators whose (i, j)th element is the operator $\Sigma_{X^iX^j}$. That is, for any $f = f_1 + \cdots + f_p \in \mathscr{H}_X$,

$$\Sigma_{XX} f = \textstyle\sum_{j=1}^{p} \sum_{i=1}^{p} \Sigma_{X^jX^i} f_i.$$

This structure was also used in Bach (2008) and Lee, Li, and Zhao (2016). We call Σ_{XX} the additive covariance operator of X. For subvectors U, V of X, we similarly define matrices of operators $\Sigma_{UV} : \mathscr{H}_V \to \mathscr{H}_U$ and $\Sigma_{UU} : \mathscr{H}_U \to \mathscr{H}_U$.

By Baker (1973), for any $\Sigma_{U^jU^i}$, there exists $R_{U^jU^i} \in \mathscr{B}(\overline{\text{ran}}(\Sigma_{U^iU^i}), \overline{\text{ran}}(\Sigma_{U^jU^j}))$ such that

$$\Sigma_{U^jU^i} = \Sigma_{U^jU^j}^{1/2} R_{U^jU^i} \Sigma_{U^iU^i}^{1/2}.$$

The operator $R_{U^jU^i}$ is called the correlation operator from $\mathscr{H}_{U^i}$ to $\mathscr{H}_{U^j}$. When a characteristic kernel is used, $R_{U^jU^i}$ can capture all the nonlinear information between U^i and U^j. Let D_{UU} denote the $r \times r$ diagonal matrix of operators whose diagonal elements are $\Sigma_{U^iU^i}$, and R_{UU} the $r \times r$ matrix operator whose (i, j)th element is $R_{U^iU^j}$. Then we have $\Sigma_{UU} = D_{UU}^{1/2} R_{UU} D_{UU}^{1/2}$. We call R_{UU} the additive correlation operator of U.

2.2. Regression operator and additive conditional independence

Building on the additive covariance and correlation operators we now introduce the *regression operator*, and thereby additive conditional independence. We first make the following assumption.

Assumption 2 *There exists $T_{WU} \in \mathscr{B}(\mathscr{H}_U, \mathscr{H}_W)$ such that*

$$R_{WU} D_{UU}^{1/2} = R_{WW} D_{WW}^{1/2} T_{WU}.$$

Lee, Li, and Zhao (2016) imposed a similar (but stronger) assumption, which assumes T_{WU} is a Hilbert-Schmidt operator. Here we only need it to be bounded. Because $\Sigma_{WW} = D_{WW}^{1/2} R_{WW} D_{WW}^{1/2}$ and $\Sigma_{WU} = D_{WW}^{1/2} R_{WU} D_{UU}^{1/2}$, the operator T_{WU} can be formally represented as $T_{WU} = \Sigma_{WW}^{-1} \Sigma_{WU}$, which resembles the regression coefficient matrix in multivariate linear regression. For this reason we refer to T_{WU} as the *regression operator*.

Definition 1 *Let U, V, and W be subvectors of X. We say that U and V are additively conditionally independent given W* (denoted by $U \perp\!\!\!\perp_A V|W$) *iff, for all $f \in \mathscr{H}_U$ and $g \in \mathscr{H}_V$,*

$$\operatorname{cov}[f(U) - (T_{WU} f)(W), g(V) - (T_{WV} g)(W)] = 0. \tag{2}$$

Li, Chun, and Zhao (2014) originally defined ACI in terms of the $L_2(P)$-geometry. Specifically, for a generic subvector $S = (S^1, \ldots, S^r)$ of X, let $\mathscr{H}_{S^i}$ be a subspace of $L_2(P_{S^i})$, the collection of square-integrable functions of S^i, and let

$$\mathscr{H}_S \triangleq \textstyle\sum_{i=1}^r \mathscr{H}_{S^i} = \{\sum_{i=1}^r f_i : f_i \in \mathscr{H}_{S^i}\}.$$

For any subvectors U, V, W of X, we say $U \perp\!\!\!\perp_A V|W$ iff

$$(\mathscr{H}_U + \mathscr{H}_W) \ominus \mathscr{H}_W \perp (\mathscr{H}_U + \mathscr{H}_V) \ominus \mathscr{H}_W, \tag{3}$$

where $\ominus$ is defined via $A \ominus B = A \cap B^\perp$, and orthogonality is defined by $f \perp g$ iff $\operatorname{cov}[f(X), g(X)] = 0$. Li, Chun, and Zhao (2014) showed that the three-way relation $U \perp\!\!\!\perp_A V|W$ as defined by (3) satisfies the conditions of a *semi-graphoid* (Lauritzen, 1996), a set of four axioms extracted from conditional independence to convey the idea that U and V are separated by W. See Pearl and Verma (1987) and Pearl, Geiger, and Verma (1989) for more details. Although ACI defined by (3) requires weaker assumptions and is more intuitive than Definition 1, the latter is useful because it allows us to borrow some of the recently developed asymptotic tools for linear operators in RKHS (Fukumizu, Bach, and Gretton, 2007; Bach, 2008). Because the regression operators T_{WU} and T_{WV} are the projections from $\mathscr{H}_U$ onto $\mathscr{H}_W$, and $\mathscr{H}_V$ onto $\mathscr{H}_W$, respectively, (2) in Definition 1 implies $(I - T_{WU})\mathscr{H}_U \perp (I - T_{WU})\mathscr{H}_U$, which further implies the relation in (2) is also a semi-graphoid, see Li, Chun, and Zhao (2014, Theorem 1).

2.3. Additive faithfulness

The traditional DAG models are based on conditional independence, i.e. a random vector X is said to be *faithful* with respect to a DAG G if for any $i, j \in \mathsf{V}$, and any subset $\mathsf{S} \subseteq \mathsf{V} \backslash \{i, j\}$,

$$X^i \perp\!\!\!\perp X^j | X^{\mathsf{S}} \quad \Leftrightarrow \quad i \text{ and } j \text{ are d-separated by } \mathsf{S} \text{ under } \mathsf{G}. \tag{4}$$

The precise definition of d-separation is as follows: suppose we are given a DAG G; then, for two nodes $i, j \in \mathsf{V}$, a subset S of $\mathsf{V} \setminus \{i, j\}$ d-connects i and j if there exists a path L between i and j such that every collider in L either belongs to S or has a descendent in S, and no other node in L belongs to S. If S does not d-connect i and j, then it d-separates i and j.

The implication $\Leftarrow$ in (4) is called the global Markov condition, and the equivalence $\Leftrightarrow$ is the faithfulness condition; see Pearl (2009, Chapter 2). The relation in (4) means that

the set of all conditional independence has a one-to-one correspondence with the set of all d-separations. More precisely, let $\mathscr{T}$ denote the set of all possible triplets $\{(i, j, \mathsf{S}) : \ i, j \in \mathsf{V},\ i \neq j,\ \mathsf{S} \subseteq \mathsf{V} \setminus \{i, j\}\}$, and then let $\mathscr{D}$ and $\mathscr{C}$ be two subsets of $\mathscr{T}$

$$\mathscr{D} = \{(i, j, \mathsf{S}) \in \mathscr{T} : \ \mathsf{S} \text{ d-separates } i, j\}, \quad \mathscr{C} = \{(i, j, \mathsf{S}) \in \mathscr{T} : \ X^i \perp\!\!\!\perp X^j | X^{\mathsf{S}}\}.$$

Then the global Markov condition means $\mathscr{D} \subseteq \mathscr{C}$, and the faithfulness condition means $\mathscr{D} = \mathscr{C}$. When the faithfulness condition is satisfied, X is said to be faithful with respect to the DAG G. For more information about faithfulness and DAG, see Spirtes, Glymour, and Scheines (2000), Kalisch and Bühlmann (2007), and Harris and Drton (2013).

Our idea is to associate X with a DAG not by conditional independence but by additive conditional independence, so as to avoid the caveat mentioned in Section 1.

Definition 2 *We say a random vector X is additively faithful with respect to a direct acyclic graph G if the following equivalence holds*

$$X^i \perp\!\!\!\perp_A X^j | X^{\mathsf{S}} \quad \Leftrightarrow \quad i \text{ and } j \text{ are d-separated by } \mathsf{S} \text{ under } \mathsf{G}. \tag{5}$$

If this occurs we say X follows an additively faithful directed acyclic graphical model (AFDAG) with respect to G. Furthermore, the "$\Leftarrow$" and "$\Leftrightarrow$" of (5) are called the Additive Global Markov condition and Additive Faithfulness, respectively.

Besides its satisfying the semi-graphoid axioms, an important rationale for us to use ACI to replace CI in both Markov and faithful conditions, is that ACI is a good approximation of CI. Suppose, for any pair $(i, j) \in \mathsf{V} \times \mathsf{V}$, we let $C^{i,j} = \{\mathsf{S} : X^i \perp\!\!\!\perp X^j \mid X^{\mathsf{S}}\}$, and $C_A^{i,j} = \{\mathsf{S} : X^i \perp\!\!\!\perp_A X^j \mid X^{\mathsf{S}}\}$. If CI and ACI are very close to being equivalent, then we can expect that the estimates of $C^{i,j}$ and $C_A^{i,j}$ are practically identical. First of all, it is shown in Li et al. (2014) that, under the copula Gaussian assumption, ACI and CI are practically equivalent. That is, ACI implies CI mathematically, and although CI does not imply ACI mathematically, the numerical evidence in Li et al. (2014) shows that the difference is vanishingly small. We then summarize this result in the following proposition.

Proposition 3 (Li et al. (2014)) *Let G be a directed acyclic graph. Suppose X follows a multivariate Gaussian copula distribution with transforming functions $(f^1, \ldots, f^p)$; that is, there exist one-to-one transformations $f^1, \ldots, f^p$ such that $[f^1(X^1), \ldots, f^p(X^p)]$ follows a multivariate Gaussian distribution. Suppose $\mathscr{H}_{X^i} = L_2(P_{X^i})$. Then we have*

$$X^i \perp\!\!\!\perp_A X^j \mid X^{\mathsf{S}} \ \Rightarrow \ X^i \perp\!\!\!\perp X^j \mid X^{\mathsf{S}}.$$

In addition to the above Proposition, if X follows a copula Gaussian distribution, Li et al. (2014, Section 3) showed that $X^i \perp\!\!\!\perp X^j \mid X^{\mathsf{S}} \ \Rightarrow \ X^i \perp\!\!\!\perp_A X^j \mid X^{\mathsf{S}}$ holds approximately. We would also like to mention that, investigating the relation between ACI and CI beyond the Gaussian copula condition is an important question, which would need more complete studies. For the interest of space we leave this part of theoretical development to future research. Nonetheless, we have conducted additional numerical analysis to justify that ACI can approximate CI reasonably well, when the Gaussian copula condition does not satisfied. Specifically, in Sections 7 and 8, we carried out simulations based on a quadratic relation between nodes, whose resulting distribution is neither Gaussian nor copula Gaussian, and

also applied our method on a real world data set. In terms of recovering the true causal diagram (which is based on CI), we see our proposed AF-PC (which is based on ACI) outperforms competing methods, including HSIC-PC (Tillman et al., 2009) and KCI-PC (Zhang et al., 2011), both of which are based on multi-dimensional kernel and likely suffer from the curse of dimensionality.

In summary, we regard the ACI as a pragmatic criterion for the DAG that strikes a balance between a parametric Gaussian model and a fully fledged nonparametric model, and it performs well numerically. We acknowledge that it may involve untestable assumptions when applied to the causal inference context.

2.4. Skeleton and v-structure of DAG

Suppose $\mathscr{C}_{\mathcal{A}} = \{(i, j, \mathsf{S}) \in \mathscr{T} : \ X^i \perp\!\!\!\perp_{\mathcal{A}} X^j | X^{\mathsf{S}}\}$. Then Definition 2 imposes equivalence between ACI structure $\mathscr{C}_{\mathcal{A}}$ and the d-separation structure $\mathscr{D}$, which is uniquely determined by a DAG G. Hence, $\mathscr{D}$ is what we can learn about G based on ACI. The situation is parallel to the classical setting, where conditional independence leads to the knowledge of $\mathscr{D}$ but not more. How much can $\mathscr{D}$ tell us about G? Verma and Pearl (1991, Theorem 1) showed that two DAGs share the same $\mathscr{D}$ if and only if they share the same skeleton and same set of v-structures. Specifically, the skeleton E_{SKE} of a DAG is simply the directed edge set E with all arrowheads removed (that is, $(i, j) \in \mathsf{E}$ or $(j, i) \in \mathsf{E}$ iff $(i, j) \in \mathsf{E}_{\text{SKE}}$ and $(j, i) \in \mathsf{E}_{\text{SKE}}$). A v-structure is a sub-graph of three nodes i, k, j where $i \to k$ and $j \to k$ but i and j are not connected (Pearl, 2009). For convenience we denote a v-structure by $(i \to k \leftarrow j)$. Let $\mathscr{V}$ be the set of all v-structures; that is

$$\mathscr{V} = \{(i \to k \leftarrow j) : \ i, j \in \mathsf{V}, \ k \in \mathsf{V} \setminus \{i, j\}\}.$$

Then the result of Verma and Pearl (1991) can be summarized as follows: if G and G' are two DAGs, then

$$\mathscr{D}(\mathsf{G}) = \mathscr{D}(\mathsf{G}') \ \Leftrightarrow \ \begin{cases} \mathsf{E}_{\text{SKE}}(\mathsf{G}) = \mathsf{E}_{\text{SKE}}(\mathsf{G}') \\ \mathscr{V}(\mathsf{G}) = \mathscr{V}(\mathsf{G}'). \end{cases}$$

Thus, for a given DAG G, its d-separation structure $\mathscr{D}(\mathsf{G})$ can tell us the skeleton of G and those arrowheads that appear in the v-structures. The set of all DAGs having the same E_{SKE} and $\mathscr{V}$ form an equivalence class, which is the target of our estimation.

3. Normalized additive conditional covariance operator

In this section we introduce the *Normalized Additive Conditional Covariance Operator* and establish its relation with ACI. To do so we first redefine the additive conditional covariance operator introduced in Li, Chun, and Zhao (2014) in terms of the additive RKHS-geometry.

3.1. Additive conditional covariance operators

Recall that, if X follows a multivariate Gaussian distribution, then

$$\text{cov}(X^i, X^j | X^{\mathsf{S}}) = 0 \quad \Leftrightarrow \quad X^i \perp\!\!\!\perp X^j | X^{\mathsf{S}}. \tag{6}$$

Li, Chun, and Zhao (2014) extended this relation by introducing the *additive conditional covariance operator* (ACCO)

$$\Lambda_{X^iX^j|X^{\mathsf{S}}} \triangleq \Lambda_{X^iX^j} - \Lambda_{X^iX^{\mathsf{S}}}\Lambda_{X^{\mathsf{S}}X^j},$$

where, for any subvectors $A, B \subseteq \mathsf{V}$, $\Lambda_{X^AX^B}$ is the covariance operator defined using $L_2(P)$-geometry; that is, $\Lambda_{X^AX^B} : \mathscr{H}_{X^B} \to \mathscr{H}_{X^A}$ is induced by $\langle f, \Lambda_{X^AX^B} g\rangle_{L_2(P_X)} = \mathrm{cov}[f(X^A), g(X^B)]$, for any $f \in \mathscr{H}_{X^A}$ and $g \in \mathscr{H}_{X^B}$. Note that $\Lambda_{X^{\mathsf{S}}X^{\mathsf{S}}}$ is indeed the identity. Because of this, $\Lambda_{X^iX^j|X^{\mathsf{S}}}$ can also be represented as $\Lambda_{X^iX^j} - \Lambda_{X^iX^{\mathsf{S}}}\Lambda^{-1}_{X^{\mathsf{S}}X^{\mathsf{S}}}\Lambda_{X^{\mathsf{S}}X^j}$, a form more similar to the conditional covariance in the classical setting. Using ACCO they extended the equivalence (6) to

$$\Lambda_{X^iX^j|X^{\mathsf{S}}} = 0 \quad \Leftrightarrow \quad X^i \perp\!\!\!\perp_A X^j | X^{\mathsf{S}}. \tag{7}$$

We first re-establish this equivalence under a different geometry, which is the RKHS-geometry. We make the following assumption on the correlation operators.

Assumption 3 *For $i \neq j$, $R_{X^iX^j}$ is compact.*

To provide intuition for this condition, note that a sufficient condition of Assumption 3 is that the correlation operator $R_{X^iX^j}$ is Hilbert-Schmidt, which in fact imposes a smoothness condition on the dependency between X^i and X^j. Indeed, if $\{(\lambda_i^\alpha, \phi_i^\alpha)\}_{\alpha=1}^\infty$ is the eigen-decomposition of $\Sigma_{X^iX^i}$, then the squared Hilbert-Schmidt norm $\|R_{X^iX^j}\|^2_{\mathrm{HS}}$ is equal to

$$\|R_{X^iX^j}\|^2_{\mathrm{HS}} = \textstyle\sum_{\alpha,\beta=1}^\infty \mathrm{cor}^2\left[\phi_i^\alpha(X^i), \phi_j^\beta(X^j)\right].$$

Therefore, assuming $\|R_{X^iX^j}\|_{\mathrm{HS}}$ to be finite requires that the correlations between the tail eigenfunctions to vanishes quickly. Moreover, because the marginal variances of these eigenfunctions, i.e. λ_i^α, diminish to zero quickly, the dependency between X^i and X^j needs to adequately concentrate on the leading eigenfunctions.

We should mention that Fukumizu et al. (2007) also gives a sufficient condition for Assumption 3: if the *mean square contingency* between X^i and X^j is finite, then $\|R_{X^iX^j}\|_{\mathrm{HS}}$ is also finite. The mean square contingency between X^i and X^j is defined as

$$M(X^i, X^j) = E^{1/2}\left[\frac{f_{i,j}(X^i, X^j)}{f_i(X^i)f_j(X^j)} - 1\right],$$

where $f_{i,j}, f_i, f_j$ are the densities of $(X^i, X^j), X^i, X^j$. The above immediately implies that, $M(X^i, X^j) = 0$ if and only if $X^i \perp\!\!\!\perp X^j$. The properties of $M(X^i, X^j)$ again regulate the dependence between X^i and X^j. For example, if X^i and X^j are perfectly correlated, i.e. $X^i = c_1X^j + c_2$ for some $c_1 \neq 0$ and $c_2 \in \mathbb{R}$, then $R_{X^iX^j}$ is not compact because it is the identity mapping.

Suppose, for all $i \in \mathsf{V}$, $\mathrm{null}(\Sigma_{X^iX^i}) = 0$, which is satisfied by removing all constant functions from $\mathscr{H}_{X^iX^i}$. Since constant functions are irrelevant to the construction of ACI, we can let $\mathscr{H}_{X^i} = \overline{\mathrm{ran}}\,(\Sigma_{X^iX^i})$. This implies the joint correlation operator R_{XX} is invertible. Furthermore, R^{-1}_{XX} is also bounded by Assumption 3.

Definition 4 *Suppose Assumptions 1 and 3 hold. Then the following operator*

$$\Sigma_{X^iX^j|X^{\mathsf{S}}} = \Sigma_{X^iX^j} - \Sigma^{1/2}_{X^iX^i} R_{X^iX^{\mathsf{S}}} R^{-1}_{X^{\mathsf{S}}X^{\mathsf{S}}} R_{X^{\mathsf{S}}X^j} \Sigma^{1/2}_{X^jX^j} \tag{8}$$

is called the Additive Conditional Covariance Operator of (X^i, X^j) *given* X^{S}.

Fukumizu et al. (2009) introduced the *Conditional Covariance Operator* (CCO) as

$$\Sigma_{X^iX^j|X^{\mathsf{S}}} = \Sigma_{X^iX^j} - \Sigma^{1/2}_{X^iX^i} \mathfrak{R}_{X^iX^{\mathsf{S}}} \mathfrak{R}_{X^{\mathsf{S}}X^j} \Sigma^{1/2}_{X^jX^j},$$

where $\mathfrak{R}_{X^iX^{\mathsf{S}}}$ and $\mathfrak{R}_{X^{\mathsf{S}}X^j}$ are the non-additive correlation operators. Note that $\mathfrak{R}_{X^iX^{\mathsf{S}}}$ and $\mathfrak{R}_{X^{\mathsf{S}}X^j}$ are structurally different from their additive counterparts. This is because, $\mathfrak{R}_{X^{\mathsf{S}}X^{\mathsf{S}}}$ is the identity mapping while $R_{X^{\mathsf{S}}X^{\mathsf{S}}}$ is a block matrix whose diagonal elements are the identity operators and off-diagonal elements are compact operators.

The following theorem parallels Theorem 5 in Li, Chun, and Zhao (2014); that is, equation (7), but with some adjustment of details to reflect the change of geometry.

Theorem 5 *Under Assumptions 1, 2, and 3,*

$$\Sigma_{X^iX^j|X^{\mathsf{S}}} = 0 \quad \Leftrightarrow \quad X^i \perp\!\!\!\perp_A X^j | X^{\mathsf{S}}. \tag{9}$$

Note that the equivalence of $X^i \perp\!\!\!\perp_A X^j|X^{\mathsf{S}}$ and $\Sigma_{X^iX^j|X^{\mathsf{S}}} = 0$ requires no preconditions on the distribution of X. This is very appealing because it implies that one can use ACCO to identify ACI under any distribution of X—in the same way as we use partial correlation to identify conditional independence under the Gaussian distribution assumption on X.

3.2. Normalized additive conditional covariance operator

By construction, ACCO not only reflects the dependence between X^i and X^j, but also the variations of X^i and X^j themselves through the presence of $\Sigma^{1/2}_{X^iX^i}$ and $\Sigma^{1/2}_{X^jX^j}$. However, the "pure" dependence between X^i and X^j should be unrelated to the marginal variations of X^i and X^j themselves. For this reason it is reasonable to use a scale-free version of ACCO that filters out the variations of X^i and X^j. There is more than one way to achieve this, and we focus on a construction which we call the normalized additive conditional covariance operator.

Definition 6 *Suppose Assumptions 1, 2, and 3 hold. Then the following operator from* $\mathscr{H}_{X^j}$ *to* $\mathscr{H}_{X^i}$

$$R_{X^iX^j|X^{\mathsf{S}}} \triangleq R_{X^iX^j} - R_{X^iX^{\mathsf{S}}} R^{-1}_{X^{\mathsf{S}}X^{\mathsf{S}}} R_{X^{\mathsf{S}}X^j},$$

is called the Normalized Additive Conditional Covariance Operator (NACCO) of (X^i, X^j) *given* X^{S}.

By comparing Definition 4 and Definition 6, it is easy to see that

$$\Sigma_{X^iX^j|X^{\mathsf{S}}} = \Sigma^{1/2}_{X^iX^i} R_{X^iX^j|X^{\mathsf{S}}} \Sigma^{1/2}_{X^jX^j}. \tag{10}$$

Thus, we remove $\Sigma^{1/2}_{X^iX^i}$ and $\Sigma^{1/2}_{X^jX^j}$ from ACCO, which carries the information about the marginal distributions of X^i and X^j. Relation (10) immediately leads to the next corollary.

Corollary 7 *Under the same Assumptions in Theorem 5,*

$$R_{X^iX^j|X^{\mathsf{S}}} = 0 \quad \Leftrightarrow \quad X^i \perp\!\!\!\perp_A X^j | X^{\mathsf{S}}.$$

The above result allows us to use NACCO as a measure of additive conditional dependence. Interestingly, in a nonadditive setting, Fukumizu, Bach, and Jordan (2009) showed a similar result which links the conditional variance operator with conditional independence. The joint additive correlation operator $R_{X^{\mathsf{S}}X^{\mathsf{S}}}$ here is different from theirs due to the additive structure. In particular, in their setting, the operator corresponding to $R_{X^{\mathsf{S}}X^{\mathsf{S}}}$ is the identity mapping; whereas here the (i,j)th element of $R_{X^{\mathsf{S}}X^{\mathsf{S}}}$ cannot be zero unless X^i and X^j are independent.

4. Estimation of NACCO

As a critical step towards estimating E, in this section we first develop the estimator of NACCO and its norm.

4.1. Empirical operators

Let $X_1, \ldots, X_n$ be an i.i.d. sample of X, and let X_k^i denote the ith component of X_k. Let $\hat{\Sigma}_{X^iX^j}$ be the *empirical covariance operator* defined through the relation

$$\langle f, \hat{\Sigma}_{X^iX^j} g \rangle_{\mathscr{H}_{X^i}} = E_n\left[f(X^i)g(X^j)\right] - E_n f(X^i) E_n g(X^j) \triangleq \mathrm{cov}_n(f,g),$$

for any $f \in \mathscr{H}_{X^i}$ and $g \in \mathscr{H}_{X^j}$. Here, E_n and cov_n are the sample moments with respect to the empirical distribution; for example, $E_n f(X^i) = n^{-1}\sum_{k=1}^n f(X_k^i)$. The operators $\hat{\Sigma}_{X^iX^j}$ are then used to build up the matrices of operators $\hat{\Sigma}_{X^iX^{\mathsf{S}}}$, $\hat{\Sigma}_{X^{\mathsf{S}}X^j}$, and $\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}}$, for any subset $\mathsf{S} \subseteq \mathsf{V}$. For example, $\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}}$ is the $\mathrm{card}(\mathsf{S}) \times \mathrm{card}(\mathsf{S})$ matrix, whose components are the operators $\{\hat{\Sigma}_{X^iX^j} : i,j \in \mathsf{S}\}$, where $\mathrm{card}(\mathsf{S})$ indicates the cardinality of S. We define the empirical version of ACCO to be

$$\hat{\Sigma}_{X^iX^j|X^{\mathsf{S}}} = \hat{\Sigma}_{X^iX^j} - \hat{\Sigma}_{X^iX^{\mathsf{S}}} \hat{\Sigma}^{\star}_{X^{\mathsf{S}}X^{\mathsf{S}}} \hat{\Sigma}_{X^{\mathsf{S}}X^j}, \tag{11}$$

where $\star$ represents one of the two forms of regularized inverses to be detailed in Sections 4.3 and 4.4. Correspondingly, the empirical version of NACCO is defined as

$$\hat{R}_{X^iX^j|X^{\mathsf{S}}} = \hat{\Sigma}^{\star 1/2}_{X^iX^i} \hat{\Sigma}_{X^iX^j|X^{\mathsf{S}}} \hat{\Sigma}^{\star 1/2}_{X^jX^j}, \tag{12}$$

where $\hat{\Sigma}^{\star 1/2}_{X^iX^i}$ stands for $(\hat{\Sigma}^{\star}_{X^iX^i})^{1/2}$.

4.2. Coordinate representation

The empirical operators in Section 4.1 are to be realized as matrices of numbers through their coordinate representations, adopted from Horn and Johnson (1985). Suppose $\mathscr{H}$ is an n-dimensional Hilbert space spanned by $\{h_1, \ldots, h_n\}$. Let B denote the vector-valued function $(h_1, \ldots, h_n)^{\mathsf{T}}$. Then, any $f \in \mathscr{H}$ can be represented as $f = \sum_{i=1}^n ([f]_B)_i h_i = B^{\mathsf{T}}[f]_B$, where $[f]_B \in \mathbb{R}^n$ is called the coordinate of f with respect to B. Let $\mathscr{H}'$ be another Hilbert space spanned by $B' = (h'_1, \ldots, h'_m)^{\mathsf{T}}$, and A be an operator from $\mathscr{H}$ to $\mathscr{H}'$. Then it can be

shown that $[Af]_{B'} = ({}_{B'}[A]_B)[f]_B$, where ${}_{B'}[A]_B \in \mathbb{R}^{m\times n}$ is the matrix $([Ah_1]_{B'}, \ldots, [Ah_n]_{B'})$, which is called the coordinate of the operator A relative to the bases B and B'. If $A' : \mathscr{H}' \to \mathscr{H}''$ where $\mathscr{H}''$ is a third Hilbert space with bases B'', then we also have that ${}_{B''}[A'A]_B = ({}_{B''}[A']_{B'})({}_{B'}[A]_B)$. For simplicity, $[A]$ and $[f]$ will be used for ${}_{B'}[A]_B$ and $[f]_B$ when the bases involved are clear from context. It is also true that $[A^\alpha] = [A]^\alpha$ for any $\alpha > 0$. Hereafter we will reserve $[\cdot]$ exclusively for expressing the coordinate representations.

For more details about these representations and the above coordinate relations, see Li, Chun, and Zhao (2012), Lee, Li, and Chiaromonte (2013) and Li, Chun, and Zhao (2014).

4.3. Estimation with ridge-regression inverse

We first develop the coordinate representation and the norm of $\hat{R}_{X^iX^j|X^{\mathsf{S}}}$ using the ridge-regression inverse (also known as Tychonoff regularization). Let $K_{X^i} = \{\kappa(X^i_s, X^i_t)\}^n_{s,t=1}$ be the kernel matrix for X^i based on the samples $X^i_1, \ldots, X^i_n$ and the reproducing kernel κ for $\mathscr{H}_{X^i}$. Let $\{\phi^i_k\}^n_{k=1}$ be the collection of functions in $\mathscr{H}_{X^i}$ defined by

$$\phi^i_k(\cdot) = \kappa(\cdot, X^i_k) - E_n(\kappa(\cdot, X^i)),\ k = 1, \ldots, n. \tag{13}$$

Note that, for any $(i, j) \in \mathsf{V} \times \mathsf{V}$, the ranges of $\hat{\Sigma}_{X^iX^i}$ and $\hat{\Sigma}_{X^jX^j}$ are contained in the spaces spanned by the functions in (13). That is,

$$\begin{aligned}
\operatorname{ran}(\hat{\Sigma}_{X^iX^i}) &= \operatorname{null}(\hat{\Sigma}_{X^iX^i})^\perp = \operatorname{span}\{\phi^i_1, \ldots, \phi^i_n\} \triangleq \mathscr{H}^{(n)}_{X^i};\\
\operatorname{ran}(\hat{\Sigma}_{X^jX^j}) &= \operatorname{null}(\hat{\Sigma}_{X^jX^j})^\perp = \operatorname{span}\{\phi^j_1, \ldots, \phi^j_n\} \triangleq \mathscr{H}^{(n)}_{X^j}.
\end{aligned}$$

Hence we can restrict our development to these finite-dimensional spaces. This observation is critical because it links the infinite-dimensional operators to their finite-dimensional representations.

Let G_{X^i} be the doubly centered kernel matrix $G_{X^i} = QK_{X^i}Q$ with $Q = I_n - 1_n1_n^{\mathsf{T}}/n$ and 1_n being the n-dimensional vector $(1, \ldots, 1)^{\mathsf{T}}$. Following similar arguments in Li, Chun, and Zhao (2012), it can be shown that $[\hat{\Sigma}_{X^iX^j}] = n^{-1}G_{X^j}$ for all $(i, j) \in \mathsf{V} \times \mathsf{V}$. We can then build up the coordinate representations for $\hat{\Sigma}_{X^iX^{\mathsf{S}}}$, $\hat{\Sigma}_{X^{\mathsf{S}}X^j}$, $\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}}$ using $[\hat{\Sigma}_{X^iX^j}]$ as matrix blocks. For example, $\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}}$ is a $\operatorname{card}(\mathsf{S}) \times \operatorname{card}(\mathsf{S})$ matrix of matrices, whose (i, j)th entry is $[\Sigma_{X^iX^j}] = n^{-1}G_{X^j}$. Furthermore, it can be shown that, for any $f_1, f_2 \in \mathscr{H}^{(n)}_{X^i}$, we have

$$\langle f_1, f_2\rangle = [f_1]^{\mathsf{T}}G_{X^i}[f_2] = n[f_1]^{\mathsf{T}}[\Sigma_{X^iX^i}][f_2].$$

Replacing the regularized inverse $\hat{\Sigma}^\star_{X^{\mathsf{S}}X^{\mathsf{S}}}$ in (11) and $\hat{\Sigma}^\star_{X^iX^i}$ and $\hat{\Sigma}^\star_{X^jX^j}$ in (12) by the ridge-regression inverses, we have

$$\begin{aligned}
\hat{\Sigma}^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}} &= \hat{\Sigma}_{X^iX^j} - \hat{\Sigma}_{X^iX^{\mathsf{S}}}(\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon_n)^{-1}\hat{\Sigma}_{X^{\mathsf{S}}X^j},\\
\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} &= (\hat{\Sigma}_{X^iX^i} + \delta_nI)^{-1/2}\hat{\Sigma}^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}}(\hat{\Sigma}_{X^jX^j} + \delta_nI)^{-1/2},
\end{aligned}$$

where $\epsilon_n, \delta_n \to 0$ are tuning constants. Hence the coordinate representation of $\hat{\Sigma}^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}}$ is

$$[\hat{\Sigma}^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}}] = [\hat{\Sigma}_{X^iX^j}] - [\hat{\Sigma}_{X^iX^{\mathsf{S}}}]([\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}}] + \epsilon_nI_{n\times\operatorname{card}(\mathsf{S})})^{-1}[\hat{\Sigma}_{X^{\mathsf{S}}X^j}],$$

and that of $\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}$ is

$$[\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}] = ([\hat{\Sigma}_{X^iX^i}] + \delta_n I_n)^{-1/2}[\hat{\Sigma}^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}}]([\hat{\Sigma}_{X^jX^j}] + \delta_n I_n)^{-1/2}.$$

To find the operator norm of $\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}$ amounts to solving the following generalized eigenvalue problem

$$\begin{aligned} &\text{maximize} \quad \langle f, \hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} g\rangle \\ &\text{subject to} \quad f \in \mathscr{H}^{(n)}_{X^i}, \quad g \in \mathscr{H}^{(n)}_{X^j}, \quad \langle f, f\rangle = \langle g, g\rangle = 1. \end{aligned}$$

Using coordinate representation, the above can be equivalently represented as the following singular value decomposition problem:

$$\begin{aligned} &\text{maximize} \quad n[f]^{\mathsf{T}}[\Sigma_{X^iX^i}][\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}][g] \\ &\text{subject to} \quad n[f]^{\mathsf{T}}[\Sigma_{X^iX^i}][f] = n[g]^{\mathsf{T}}[\Sigma_{X^jX^j}][g] = 1. \end{aligned} \tag{14}$$

To transform this into a standard eigenvalue problem, let

$$u = G^{1/2}_{X^i}[f] = n^{1/2}[\Sigma_{X^iX^i}]^{1/2}[f], \quad v = G^{1/2}_{X^j}[g] = n^{1/2}[\Sigma_{X^jX^j}]^{1/2}[g].$$

Solving these equations for $[f]$ and $[g]$ with ridge-regression regularizations, we have

$$[f] = n^{-1/2}([\Sigma_{X^iX^i}] + \delta_n I_n)^{-1/2}u, \quad [g] = n^{-1/2}([\Sigma_{X^jX^j}] + \delta_n I_n)^{-1/2}v. \tag{15}$$

Substituting (15) into (14), we have the following standard eigenvalue problem

$$\begin{aligned} &\text{maximize} \quad u^{\mathsf{T}}([\Sigma_{X^iX^i}] + \delta_n I_n)^{-1/2}[\hat{\Sigma}_{X^iX^i}][\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}]([\Sigma_{X^jX^j}] + \delta_n I_n)^{-1/2}v \\ &\text{subject to} \quad u^{\mathsf{T}}u = v^{\mathsf{T}}v = 1. \end{aligned}$$

Consequently, the operator norm $\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\|$ is simply the largest singular value of the $n \times n$ matrix

$$([\Sigma_{X^iX^i}] + \delta_n I_n)^{-1/2}[\Sigma_{X^iX^i}][\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}]([\Sigma_{X^jX^j}] + \delta_n I_n)^{-1/2}.$$

4.4. Estimation with principal-component inverse

The second approach to handle the inverse of $\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}}$ is to only invert its large eigenvalues, while setting all the small eigenvalues to 0. So it is akin to the principal-component or Moore-Penrose inverse. This approach is particularly attractive for handling large network because it can reduce the amount of computation substantially. Write the symmetric matrix G_{X^i} as

$$G_{X^i} = V_{X^i}\Lambda_{X^i}V^{\mathsf{T}}_{X^i},$$

where Λ_{X^i} is the diagonal matrix of the nonzero eigenvalues of G_{X^i}, and V_{X^i} is the matrix of eigenvectors corresponding to Λ_{X^i}. Usually, the rank of G_{X^i} is $n-1$, but this is not

always the case. So let m be the rank of G_{X^i} and, for simplicity, assume it to be the same for different i. It is then equivalent to work with $\mathscr{G}^{(n)}_{X^i} = \text{span}\{\psi^i_1, \ldots, \psi^i_m\}$, where

$$(\psi^i_1, \ldots, \psi^i_m)^\mathsf{T} = \Lambda^{-1/2}_{X^i} V^\mathsf{T}_{X^i} \phi^i \triangleq \psi^i.$$

It is easy to see that the coordinate representation of the inner product in $\mathscr{G}^{(n)}_{X^i}$ is: for any $f, g \in \mathscr{G}^{(n)}_{X^i}$,

$$\langle f, g \rangle = [f]^\mathsf{T}[g]$$

where $[\cdot]$ is the coordinate representation with respect to ψ^i.

The next proposition gives the coordinate representations of relevant operators using the new bases.

Proposition 8 *For any $\mathsf{S} = \{s_1, \ldots, s_d\} \subseteq V$, the operators*

$$\hat{\Sigma}_{X^iX^j} : \mathscr{G}^{(n)}_{X^j} \to \mathscr{G}^{(n)}_{X^i}, \quad \hat{\Sigma}_{X^iX^i} : \mathscr{G}^{(n)}_{X^i} \to \mathscr{G}^{(n)}_{X^i}, \quad \hat{\Sigma}_{X^\mathsf{S}X^\mathsf{S}} : \oplus_{i\in\mathsf{S}} \mathscr{G}^{(n)}_{X^i} \to \oplus_{i\in\mathsf{S}} \mathscr{G}^{(n)}_{X^i}$$

have the following coordinate representation:

$$[\hat{\Sigma}_{X^iX^j}] = n^{-1}\Lambda^{1/2}_{X^i} V^\mathsf{T}_{X^i} V_{X^j} \Lambda^{1/2}_{X^j}, \quad [\hat{\Sigma}_{X^iX^i}] = n^{-1}\Lambda_{X^i},$$
$$[\hat{\Sigma}_{X^iX^\mathsf{S}}] = n^{-1}\Lambda^{1/2}_{X^i} V^\mathsf{T}_{X^i} M_{X^\mathsf{S}}, \quad [\hat{\Sigma}_{X^\mathsf{S}X^j}] = n^{-1} M_{X^\mathsf{S}} V_{X^j} \Lambda^{1/2}_{X^j}, \quad [\hat{\Sigma}_{X^\mathsf{S}X^\mathsf{S}}] = n^{-1} M^\mathsf{T}_{X^\mathsf{S}} M_{X^\mathsf{S}},$$

where $M_{X^\mathsf{S}} = (V_{X^{s_1}}\Lambda^{1/2}_{X^{s_1}}, \ldots, V_{X^{s_d}}\Lambda^{1/2}_{X^{s_d}})$.

Expressions (11) and (12), and the rules for coordinate manipulation in Section 4.2 suggest the following coordinate representation for $\hat{R}_{X^iX^j|X^\mathsf{S}}$ as

$$[\hat{R}_{X^iX^j|X^\mathsf{S}}] = [\hat{\Sigma}_{X^iX^i}]^{-1/2} \left([\hat{\Sigma}_{X^iX^j}] - [\hat{\Sigma}_{X^iX^\mathsf{S}}][\hat{\Sigma}_{X^\mathsf{S}X^\mathsf{S}}]^{-1}[\hat{\Sigma}_{X^\mathsf{S}X^j}] \right) [\hat{\Sigma}_{X^jX^j}]^{-1/2}. \tag{16}$$

However, to enhance performance and reduce the amount of computation we propose to regularize the inversions of $[\hat{\Sigma}_{X^iX^i}]$, $[\hat{\Sigma}_{X^jX^j}]$, and $[\hat{\Sigma}_{X^\mathsf{S}X^\mathsf{S}}]$, as follows. For a generic symmetric matrix $A \in \mathbb{R}^{k\times k}$ and a number $\epsilon > 0$, let

$$A^\dagger(\epsilon) = \textstyle\sum_{i=1}^k \lambda_i^{-1} I(\lambda_i > \epsilon) v_i v_i^\mathsf{T},$$

where $(\lambda_1, v_1), \ldots, (\lambda_r, v_r)$ are the eigenvalue-eigenvector pairs of A. That is, we ignore all the eigenvalues of A that are smaller than ϵ and invert only the eigenvalues larger than ϵ. We replace the full inverses in (16) by their regularized versions:

$$[\hat{\Sigma}_{X^iX^i}]^\dagger(\delta_n), \quad [\hat{\Sigma}_{X^\mathsf{S}X^\mathsf{S}}]^\dagger(\epsilon_n), \quad [\hat{\Sigma}_{X^jX^j}]^\dagger(\delta_n),$$

where $0 < \delta_n \to 0$ and $0 < \epsilon_n \to 0$ are tuning constants. With these regularized inverses the NACCO estimator is defined through its coordinate representation

$$[\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^\mathsf{S}}] = [\hat{\Sigma}_{X^iX^i}]^{\dagger 1/2}(\delta_n) \left([\hat{\Sigma}_{X^iX^j}] - [\hat{\Sigma}_{X^iX^\mathsf{S}}][\hat{\Sigma}_{X^\mathsf{S}X^\mathsf{S}}]^\dagger(\epsilon_n)[\hat{\Sigma}_{X^\mathsf{S}X^j}] \right) [\hat{\Sigma}_{X^jX^j}]^{\dagger 1/2}(\delta_n). \tag{17}$$

The operator norm of $\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}$ can be found by solving the singular-value problem

$$\begin{aligned} &\text{maximize} \quad \langle f, \hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} g\rangle \\ &\text{subject to} \quad f \in \mathscr{G}^{(n)}_{X^i},\ g \in \mathscr{G}^{(n)}_{X^j}, \quad \langle f, f\rangle = \langle g, g\rangle = 1. \end{aligned}$$

Because

$$\langle f, \hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} g\rangle = [f]^{\mathsf{T}}[\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}][g], \quad \langle f, f\rangle = [f]^{\mathsf{T}}[f], \quad \langle g, g\rangle = [g]^{\mathsf{T}}[g],$$

$\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\|$ is the largest singular value of the matrix $[\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}]$. The next proposition gives the explicit form of this matrix.

Proposition 9 *The operator norm $\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\|$ is the largest singular value of the matrix*

$$\operatorname{diag}(I_{r_i}, 0) V_{X^i}^{\mathsf{T}} \left[I_n - M_{X^{\mathsf{S}}}(M_{X^{\mathsf{S}}}^{\mathsf{T}} M_{X^{\mathsf{S}}})^{\dagger}(\epsilon_n) M_{X^{\mathsf{S}}}^{\mathsf{T}}\right] V_{X^j} \operatorname{diag}(I_{r_j}, 0), \tag{18}$$

where $r_i = r_i(n)$ and $r_j = r_j(n)$ are the numbers of eigenvalues in Λ_{X^i} and Λ_{X^j} that are greater than δ_n. Moreover, $\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\| \le 1$.

In appearance, (18) involves spectral decomposition of $M_{X^{\mathsf{S}}}^{\mathsf{T}} M_{X^{\mathsf{S}}}$, which has dimension $[m \cdot \operatorname{card}(\mathsf{S})] \times [m \cdot \operatorname{card}(\mathsf{S})]$. This can be large when both the sample size n and the dimension of the network p are large. However, note that $M_{X^{\mathsf{S}}}(M_{X^{\mathsf{S}}}^{\mathsf{T}} M_{X^{\mathsf{S}}})^{\dagger}(\epsilon_n) M_{X^{\mathsf{S}}}^{\mathsf{T}}$ is simply the projection on to the space spanned by the eigenvectors of $M_{X^{\mathsf{S}}} M_{X^{\mathsf{S}}}^{\mathsf{T}}$ whose eigenvalues are greater than ϵ_n, and $M_{X^{\mathsf{S}}} M_{X^{\mathsf{S}}}^{\mathsf{T}}$ is an $n \times n$ matrix. Hence, in effect, we only need to compute the spectral decomposition of an $n \times n$ matrix in order to evaluate the norm of $\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}$. We record this fact as the following corollary.

Corollary 10 *The operator norm $\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\|$ is the largest singular value of the matrix*

$$\operatorname{diag}(I_{r_i}, 0) V_{X^i}^{\mathsf{T}} \left[I_n - \Pi_{X^{\mathsf{S}}}(\epsilon_n)\right] V_{X^j} \operatorname{diag}(I_{r_j}, 0),$$

where $\Pi_{X^{\mathsf{S}}}(\epsilon_n)$ is the projection on to the subspace spanned by those eigenvectors of $M_{X^{\mathsf{S}}} M_{X^{\mathsf{S}}}^{\mathsf{T}}$ whose eigenvalues are greater than ϵ_n.

4.5. Tuning constants δ_n and ϵ_n

Based on the cumulative percentages of total variation explained by leading eigenvalues (Jolliffe, 2002, Chapter 6), we recommend the following empirical rules for determining the tuning constants ϵ_n and δ_n:

$$\begin{aligned} \delta_n &= \max\{\delta : \textstyle\sum_{k=1}^m \lambda_k I(\lambda_k \ge \delta) / \sum_{k=1}^m \lambda_k \ge 1 - 10^{-6}\}, \\ \epsilon_n &= \max\{\epsilon : \textstyle\sum_{k=1}^n \tau_k I(\tau_k \ge \epsilon) / \sum_{k=1}^n \tau_k \ge 1 - 10^{-6}\}, \end{aligned} \tag{19}$$

where $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of G_{X^i}, and $\tau_1, \ldots, \tau_n$ are the eigenvalues of $M_{X^{\mathsf{S}}} M_{X^{\mathsf{S}}}^{\mathsf{T}}$. In other words, δ_n is selected so that the summation of the eigenvalues of G_{X^i} smaller than δ_n, is less than $10^{-6} \times \operatorname{trace}(G_{X^i})$; we choose ϵ_n similarly. Note that the λ's are allowed to depend on i, and the τ's are allowed to depend on S.

4.6. Summary

We summarize below the estimation procedure for the norm of NACCO for given $i, j \in \mathsf{V}$, $\mathsf{S} \subseteq \mathsf{V} \setminus \{i, j\}$, focusing on the principal-component-inverse version.

1. For $i = 1, \ldots, p$, marginally standardize $X_1^i, \ldots, X_n^i$ so that $E_n(X^i) = 0$, $\mathrm{var}_n(X^i) = 1$.

2. Choose a kernel function κ; for example, it can be the Gaussian radial basis function

$$\kappa(X_s^i, X_t^i) = \exp(-\gamma_i |X_s^i - X_t^i|^2),$$

   where γ_i is the bandwidth parameter. To compute γ_i we recommend the formula

$$1/\sqrt{\gamma_i} = \textstyle\sum_{s<t} |X_s^i - X_t^i| / \binom{n}{2},$$

   which was also used in Lee, Li, and Chiaromonte (2013); Lee, Li, and Zhao (2016).

3. Use the chosen κ and γ_i, compute the kernel matrices K_{X^i} and their centered versions G_{X^i} for $i = 1, \ldots, p$. Perform spectral decomposition on G_{X^i} to obtain V_{X^i} and Λ_{X^i}. Stack $V_{X^i} \Lambda_{X^i}^{1/2}$, $i \in \mathsf{S}$ to form M_{X^S}.

4. Determine the tuning constants ϵ_n, δ_n using (19).

5. Perform spectral decomposition on $M_{X^\mathsf{S}} M_{X^\mathsf{S}}^\mathsf{T}$ and compute $\Pi_{X^\mathsf{S}}(\epsilon_n)$ in Corollary 10.

6. Evaluate $\|\hat{R}^{\epsilon_n, \delta_n}_{X^i X^j | X^\mathsf{S}}\|$ according to Corollary 10.

5. Estimation of AFDAG

The estimators of NACCO in the last section suggests a way to determine ACI approximately based on the data: if the norm of the operator $\hat{R}^{\epsilon_n, \delta_n}_{X^i X^j | X^\mathsf{S}}$ is small, then we can expect $X^i \perp\!\!\!\perp_A X^j | X^\mathsf{S}$ to hold with large probability. Making this assessment for every possible triplet (i, j, S) for $i, j \in \mathsf{V}$, $\mathsf{S} \subseteq \mathsf{V} \setminus \{i, j\}$ then leads to an estimator of the set $\mathscr{C}_A = \mathscr{D}$ which, in turn, leads to estimators of $\mathsf{E}_{\mathrm{SKE}}$ and $\mathscr{V}$. Below we describe this process in detail. We begin with descriptions at the population level of how $\mathscr{D}$ (and hence also $\mathscr{C}_A$) determines $\mathsf{E}_{\mathrm{SKE}}$ and $\mathscr{V}$, and then mimic these relations at the sample level to create estimators for $\mathsf{E}_{\mathrm{SKE}}$ and $\mathscr{V}$. Finally, we develop a modified PC-algorithm (Spirtes et al., 2000) to implement the estimation procedure efficiently.

5.1. Characterizing $\mathsf{E}_{\mathrm{SKE}}$ and $\mathscr{V}$ by ACI at population level

We first describe how the ACI structure $\mathscr{C}_A$ determines the skeleton of an AFDAG model. The following result immediately follows from Verma and Pearl (1990, Lemma 1) and thus its proof is omitted.

Proposition 11 *For a given AFDAG model* $\mathsf{G} = (\mathsf{V}, \mathsf{E})$*,*

$$\mathsf{E}_{\mathrm{SKE}} = \{(i, j) : \ (i, j, \mathsf{S}) \notin \mathscr{C}_A \text{ for every } \mathsf{S} \subseteq \mathsf{V} \setminus \{i, j\}\}. \tag{20}$$

A semi-graphoid model is the graphical model that associates a semi-graphoid relation with graph separation. Lauritzen (Example 3.2, 1996) also provides a general relation of semi-graphoid models. Li, Chun, and Zhao (2014) used ACI to define an *additive semi-graphoid model* (ASG) for undirected graphs. In their setting, the undirected graph, denoted by E_{ASG}, is associated with ACI according to the following relation

$$(i,j) \notin \mathsf{E}_{\text{ASG}} \quad \Leftrightarrow \quad X^i \perp\!\!\!\perp_A X^j | X^{\mathsf{V}\setminus\{i,j\}}.$$

This and (20) together imply that $\mathsf{E}_{\text{SKE}} \subseteq \mathsf{E}_{\text{ASG}}$; that is, the skeleton of a DAG is a subset of the additive semi-graphoid. Therefore, one may also use E_{ASG} to estimate an upper bound of E_{SKE}, but in general it cannot fully specify E_{SKE}.

The next proposition describes how the ACI structure $\mathscr{C}_A$ determines the set of v-structures $\mathscr{V}$, a result that follows directly from Lemma 2 of Verma and Pearl (1990).

Proposition 12 *For a given AFDAG model* $\mathsf{G} = (\mathsf{V}, \mathsf{E})$*,*

$$\mathscr{V} = \{\{i,k,j\} : \ (i,j) \notin \mathsf{E}_{\text{SKE}}, \ (i,k) \in \mathsf{E}_{\text{SKE}}, \ (j,k) \in \mathsf{E}_{\text{SKE}}, \ k \notin \mathsf{C}_{i,j}\}, \tag{21}$$

where $\mathsf{C}_{i,j} = \{\mathsf{S} : \ (i,j,\mathsf{S}) \in \mathscr{C}_A\}$.

Since E_{SKE} and $\mathsf{C}_{i,j}$ are determined by $\mathscr{C}_A$, $\mathscr{V}$ itself is determined by $\mathscr{C}_A$.

5.2. Estimation of E_{SKE} and $\mathscr{V}$

Propositions 11 and 12 suggest methods to estimate E_{SKE} and $\mathscr{V}$ once we have an estimate of the ACI structure $\mathscr{C}_A$, and Corollary 7 suggests that we can estimate ACI by the smallness of $\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}$. Specifically, we propose to estimate $\mathscr{C}_A$ by

$$\hat{\mathscr{C}}_A = \{(i,j,\mathsf{S}) \in \mathscr{T} : \ \|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\| < \rho\}, \tag{22}$$

where ρ is a pre-determined threshold. We then mimic (20) to estimate E_{SKE}:

$$\hat{\mathsf{E}}_{\text{SKE}} = \{(i,j) : \ (i,j,\mathsf{S}) \notin \hat{\mathscr{C}}_A \text{ for each } \mathsf{S} \subseteq \mathsf{V} \setminus \{i,j\}\}.$$

Finally, we mimic (21) to estimate $\mathscr{V}$:

$$\hat{\mathscr{V}} = \{\{i,k,j\} : \ (i,j) \notin \hat{\mathsf{E}}_{\text{SKE}}, \ (i,k) \in \hat{\mathsf{E}}_{\text{SKE}}, \ (j,k) \in \hat{\mathsf{E}}_{\text{SKE}}, \ k \notin \hat{\mathsf{C}}_{i,j}\},$$

where $\hat{\mathsf{C}}_{i,j} = \{\mathsf{S} : \ (i,j,\mathsf{S}) \in \hat{\mathscr{C}}_A\}$.

5.3. A permutation test-based threshold

In Section 5.2 we used a constant threshold ρ to construct $\hat{\mathsf{E}}_{\text{SKE}}$ and $\hat{\mathscr{V}}$. In this subsection we introduce a permutation test for additive conditional independence, which leads to a more natural and data driven threshold for estimating the AFDAG.

We adopt a similar technique used in both Fukumizu et al. (2008) and Tillman et al. (2009), where the idea is to first divide the conditioned variables into groups, and then to permute the random elements within the same groups to break the dependency. More specifically, suppose we want to break the conditional dependence between X^i and X^j given

X^{S}. We first partition $\{X_1^{\mathsf{S}}, \ldots, X_n^{\mathsf{S}}\}$ into L subsets, say $C_1, \ldots, C_L$ using the distance in the X^S-space, which can be implemented via the K-means clustering (Lloyd, 1982). For each $\ell = 1, \ldots, L$, let D_ℓ denote the subset $\{k : X_k^{\mathsf{S}} \in C_\ell\} = \{k_{\ell,1}, \ldots, k_{\ell,d_\ell}\}$ and let X_{D_ℓ} denote the subvector $(X_{k_{\ell,1}}, \ldots, X_{k_{\ell,d_\ell}})$.

Next, we permute the data within D_ℓ. Given $\ell = 1, \ldots, L$, let $\pi_\ell : D_\ell \to D_\ell$ be a permutation (i.e. a bijection), and let $\mathfrak{S}_\ell$ be the set of all such permutations of this form. For each $i \in \{1, \ldots, p\}$ and $\pi_\ell \in \mathfrak{S}_\ell$, let

$$g_{\pi_\ell, i} : \underbrace{\mathbb{R}^p \times \cdots \times \mathbb{R}^p}_{d_\ell} \to \underbrace{\mathbb{R}^p \times \cdots \times \mathbb{R}^p}_{d_\ell}$$

be the function that permutes the ith component of a set of d_ℓ vectors by π_ℓ. That is, if $(a_1, \ldots, a_{d_\ell})$ is a set vectors in $\mathbb{R}^p$, then $g_{\pi_\ell,i}(a_1, \ldots, a_{d_\ell}) = (b_1, \ldots, b_{d_\ell})$, where for each $k \in \{1, \ldots, d_\ell\}$, b_k is a vector in $\mathbb{R}^p$ whose jth component is defined by

$$b_k^j = \begin{cases} a_k^j & j \neq i \\ a_{\pi_\ell(k)}^j & j = i. \end{cases}$$

Let $\hat{R}^{\epsilon_n, \delta_n}_{X^i X^j | X^{\mathsf{S}}}[g_{\pi_1,i}(X_{D_1}), \cdots, g_{\pi_L,i}(X_{D_L})]$ denote the estimate of $R_{X^i X^j | X^{\mathsf{S}}}$ constructed using the partially permuted sample $[g_{\pi_1,i}(X_{D_1}), \cdots, g_{\pi_L,i}(X_{D_L})]$. The empirical distribution of the norm of this operator based on the permutations is

$$\hat{F}_{\text{NACCO}}(x) = (\Pi_{\ell=1}^L d_\ell)^{-1} \sum_{\pi_1 \in \mathfrak{S}_1} \cdots \sum_{\pi_L \in \mathfrak{S}_L} \mathbf{1}\left\{ \|\hat{R}^{\epsilon_n, \delta_n}_{X^i X^j | X^{\mathsf{S}}}[g_{\pi_1,i}(X_{D_1}), \cdots, g_{\pi_L,i}(X_{D_L})]\| \leq x \right\}.$$

We can use $\rho = \hat{F}^{-1}_{\text{NACCO}}(1-\alpha)$ as the threshold in (22) for significance level α.

The cardinality of $\mathfrak{S}_1 \times \cdots \times \mathfrak{S}_L$ scales fast and the associated computation can easily become infeasible. In practice, we use the following approximation of $\hat{F}_{\text{NACCO}}(x)$. Suppose $\pi_1^1 \times \cdots \times \pi_L^1, \ldots, \pi_1^b \times \cdots \times \pi_L^b$ are b random draws from the discrete uniform distribution on $\{\pi_1 \times \cdots \times \pi_L : \pi_\ell \in \mathfrak{S}_\ell, \ell = 1, \ldots, L\}$. We then estimate $\hat{F}_{\text{NACCO}}(x)$ via

$$\hat{F}^b_{\text{NACCO}}(x) = \frac{1}{b} \sum_{j=1}^b \mathbf{1}\left\{ \|\hat{R}^{\epsilon_n, \delta_n}_{X^i X^j | X^{\mathsf{S}}}[g_{\pi_1^j,i}(X_{D_1}), \cdots, g_{\pi_L^j,i}(X_{D_L})]\| \leq x \right\},$$

where b is a sufficiently large integer.

5.4. PC-algorithm

In appearance, (22) involves evaluating the criterion $\|\hat{R}^{\epsilon_n, \delta_n}_{X^i X^j | X^{\mathsf{S}}}\| < \rho$ for all distinct pairs (i, j) and all distinct subsets $\mathsf{S} \subseteq \mathsf{V} \setminus \{i, j\}$, which would be an enormous task. However, we can make the process much more efficient by adapting the PC-algorithm (Spirtes et al., 2000) to this criterion. That is, we gradually prune a complete graph according to this criterion and, after each pruning action, focus only on those edges that have not been pruned. In this way the complexity of the algorithm is not determined by the dimension of X but by the level of sparseness of the DAG. We refer this modified algorithm as the AF-PC algorithm (AF referring to additively faithful).

For any (i, j), let $\text{conn}(i, -j, \hat{\mathsf{E}}_{\text{SKE}})$ denote the collection of all vertices connected to i in $\hat{\mathsf{E}}_{\text{SKE}}$ with the jth vertex removed; that is,

$$\text{conn}(i, -j, \hat{\mathsf{E}}_{\text{SKE}}) = \{j : (i, j) \in \hat{\mathsf{E}}_{\text{SKE}}\} \setminus \{j\}.$$

Pseudo codes: AF-PC skeleton-algorithm
initialize: set $l = -1$ and $\hat{\mathsf{E}}_{\text{SKE}}$ to be the complete graph
repeat
  set $l = l + 1$
  repeat
    select a new ordered pair $(i, j) \in \hat{\mathsf{E}}_{\text{SKE}}$ such that $\text{card}\{\text{conn}(i, -j, \hat{\mathsf{E}}_{\text{SKE}})\} \geq l$
    repeat
      select new $\mathsf{S} \subseteq \text{conn}(i, -j, \hat{\mathsf{E}}_{\text{SKE}})$ with $\text{card}\{\mathsf{S}\} = l$
      if $\|\hat{R}_{X^i X^j | X^{\mathsf{S}}}\| < \rho$ then remove (i, j) from $\hat{\mathsf{E}}_{\text{SKE}}$ and save $\mathsf{S}_{i,j} = \mathsf{S}$
    until $(i, j) \notin \hat{\mathsf{E}}_{\text{SKE}}$ or all $\mathsf{S} \subseteq \text{conn}(i, -j, \hat{\mathsf{E}}_{\text{SKE}})$ with $\text{card}\{\mathsf{S}\} = l$ have been chosen
  until all ordered pairs $(i, j) \in \hat{\mathsf{E}}_{\text{SKE}}$ such that $\text{card}\{\text{conn}(i, -j, \hat{\mathsf{E}}_{\text{SKE}})\} \geq l$ have been chosen
until $l = p - 2$ or there is no $(i, j) \in \hat{\mathsf{E}}_{\text{SKE}}$ such that $\text{card}\{\text{conn}(i, -j, \hat{\mathsf{E}}_{\text{SKE}})\} \geq l$.

The AF-PC algorithm is given below in the form of pseudo codes:

Note that the outputs of AF-PC algorithm is the estimated skeleton $\hat{\mathsf{E}}_{\text{SKE}}$ and the separating sets $\mathsf{S}_{i,j}$. We then use $\mathsf{S}_{i,j}$ to identify the v-structures and orient the edges in $\hat{\mathsf{E}}_{\text{SKE}}$ accordingly, using the rules described in the previous subsection. We can further orient as many edges as we can under the constraint that the terminal graph is a DAG. One can show that the output is a *completed partially directed acyclic graphs* (CPDAG, Kalisch and Bühlmann, 2007), which we denote by $\hat{\mathsf{E}}_{\text{CPDAG}}$. The algorithm of this portion is omitted here as it can be found in standard PC-algorithms such as Kalisch and Bühlmann (2007, Algorithm 2).

6. Consistency

We now establish the consistency of the estimator of the AFDAG skeleton E_{SKE} and its v-structures $\mathscr{V}$ described in Section 5. For simplicity, we will focus on the case where ρ is fixed. We carry this out in three steps: the consistency of the ACCO estimator, the consistency of the NACCO estimator, and then the consistency of $\hat{\mathscr{C}}_A$, which implies the consistency of $\hat{\mathsf{E}}_{\text{SKE}}$ and $\hat{\mathscr{V}}$. Since ridge-regression-inverse version (Section 4.3) is analytically simpler to handle, we focus on that version. However, we expect that consistency of the principal-component-inverse version (as described in Section 4.4) can be established similarly. To emphasize the dependence of $\hat{\mathscr{C}}_A$, $\hat{\mathsf{E}}_{\text{SKE}}$, and $\hat{\mathscr{V}}$ on ϵ_n, δ_n, and ρ, we write them as $\hat{\mathscr{C}}_A(\epsilon_n, \delta_n, \rho)$, $\hat{\mathsf{E}}_{\text{SKE}}(\epsilon_n, \delta_n, \rho)$, and $\hat{\mathscr{V}}(\epsilon_n, \delta_n, \rho)$ henceforth.

6.1. Consistency of ACCO

To prove the convergence of $\hat{\Sigma}^{\epsilon_n}_{X^i X^j | X^{\mathsf{S}}}$ to $\Sigma_{X^i X^j | X^{\mathsf{S}}}$, we need the following lemma.

Lemma 13 *Suppose Assumptions 2 and 3 hold. Then, for any* $S \subseteq \mathsf{V}$,

(a) $\|(\hat{\Sigma}_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon I)^{-1}\| = O_P(\epsilon^{-1})$, $\|(\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon I)^{-1}\| = O(\epsilon^{-1})$;

(b) $\|(\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon I)^{-1/2} D^{1/2}_{X^{\mathsf{S}} X^{\mathsf{S}}}\| = O_P(1)$.

We introduce the following intermediate operator

$$\Sigma^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}} \triangleq \Sigma_{X^iX^j} - \Sigma_{X^iX^{\mathsf{S}}}(\Sigma_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon_n I)^{-1}\Sigma_{X^{\mathsf{S}}X^j}. \tag{23}$$

The next lemma first shows that the norm of $\Sigma^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}}$ is bounded by a constant for all n.

Lemma 14 *Under the same assumptions in Lemma 13, given any sequence $\epsilon_n > 0$, $\|\Sigma^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}}\|$ is bounded by a constant.*

For two positive sequences a_n and b_n, we write $a_n \prec b_n$ when $a_n/b_n \to 0$, and $a_n \preceq b_n$ when a_n/b_n converges to 0 or is bounded. In particular, $a_n \prec 1$ means $a_n \to 0$. The next two lemmas show the closeness of the intermediate operator to $\hat{\Sigma}^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}}$ and $\Sigma_{X^iX^j|X^{\mathsf{S}}}$.

Lemma 15 *If Assumptions 1, 2, 3 hold, and $n^{-1/2} \prec \epsilon_n \prec 1$, then*

$$\|\hat{\Sigma}^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}} - \Sigma^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}}\| = O_P(\epsilon_n^{-1}n^{-1/2}).$$

Lemma 16 *If Assumptions 1, 2, 3 hold, then*

$$\|\Sigma_{X^iX^{\mathsf{S}}}(\Sigma_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon I)^{-1}\Sigma_{X^{\mathsf{S}}X^j} - \Sigma^{1/2}_{X^iX^i}R_{X^iX^{\mathsf{S}}}R^{-1}_{X^{\mathsf{S}}X^{\mathsf{S}}}R_{X^{\mathsf{S}}X^j}\Sigma^{1/2}_{X^jX^j}\| = O(\epsilon^{1/2}).$$

The consistency of ACCO follows immediately.

Theorem 17 *Suppose $\hat{\Sigma}^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}}$ is the empirical additive covariance operator defined in (11) with ϵ_n satisfying $n^{-1/2} \prec \epsilon_n \prec 1$. Then under the same assumptions in Lemma 16, we have*

$$\|\hat{\Sigma}^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}} - \Sigma_{X^iX^j|X^{\mathsf{S}}}\| = O_P(\epsilon_n^{1/2} + n^{-1/2}\epsilon_n^{-1}).$$

6.2. Consistency of NACCO

Similar to (23), we define the intermediate operator between $R_{X^iX^j|X^{\mathsf{S}}}$ and $\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}$ as

$$R^{\delta_n}_{X^iX^j|X^{\mathsf{S}}} = (\Sigma_{X^iX^i} + \delta_n I)^{-1/2}\Sigma_{X^iX^j|X^{\mathsf{S}}}(\Sigma_{X^jX^j} + \delta_n I)^{-1/2}.$$

By the triangular inequality, for the consistency of $\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}$ it suffices to show that, as $n \to \infty$,

$$\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R^{\delta_n}_{X^iX^j|X^{\mathsf{S}}}\| \xrightarrow{P} 0 \quad \text{and} \quad \|R^{\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R_{X^iX^j|X^{\mathsf{S}}}\| \to 0. \tag{24}$$

We first state a lemma (Fukumizu, Bach, and Gretton, 2007, Lemma 8).

Lemma 18 *Let $\{A_n : n = 1, 2, \ldots\}$ and A be self-adjoint random operators in $\mathscr{B}(\mathscr{H})$. If $\|A_n - A\| \xrightarrow{P} 0$, then $\|A_n^{3/2} - A^{3/2}\| = O(\|A_n - A\|)$.*

The next two lemmas show the two convergence results in (24).

Lemma 19 *If Assumptions 1, 2, 3 hold and*

$$n^{1/2} \prec \epsilon_n \prec 1, \quad n^{-1/2}\epsilon_n^{-1} + \epsilon_n^{1/2} \prec \delta_n^{3/2}, \tag{25}$$

then $\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R^{\delta_n}_{X^iX^j|X^{\mathsf{S}}}\| \xrightarrow{P} 0$.

Lemma 20 *Under the same Assumptions in Lemma 19, we have, as* $\delta_n \to 0$,

$$\|R^{\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R_{X^iX^j|X^{\mathsf{S}}}\| \to 0.$$

Consistency of $\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}$ now follows from the previous two lemmas.

Theorem 21 *Under the same assumptions in Lemma 19 we have*

$$\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R_{X^iX^j|X^{\mathsf{S}}}\| \xrightarrow{P} 0.$$

Fukumizu et al. (2008, Theorem 5) established the consistency of (non-additive) normalized conditional correlation operator. As previously discussed, our setting is additive and different from theirs; therefore, Theorem 21 can not be directly implied by their result.

6.3. Consistency of the estimator of AFDAG skeleton

We are now ready to prove the consistency of the estimated skeleton of AFDAG.

Theorem 22 *Suppose* X *is additively faithful with respect to a directed acyclic graph* G. *Then, under Assumptions 1, 2, and 3, we have*

$$P(\hat{\mathscr{C}}_{A}(\epsilon_n, \delta_n, \rho) = \mathscr{C}_{A}) \to 1,$$

for (ϵ_n, δ_n) *satisfying (25) and sufficiently small* ρ. *Consequently,*

$$P(\hat{\mathsf{E}}_{\mathrm{SKE}}(\epsilon_n, \delta_n, \rho) = \mathsf{E}_{\mathrm{SKE}}) \to 1, \quad P(\hat{\mathscr{V}}(\epsilon_n, \delta_n, \rho) = \mathscr{V}) \to 1.$$

7. Strong additive faithfulness and uniform consistency

7.1. Uniform consistency

The consistency established in the last section is in terms of the true distribution. Sometimes a stronger form of consistency—uniform consistency—is preferable so that we can control the worst-case type I and type II errors. Zhang and Spirtes (2002) proved the uniform consistency for the Gaussian DAG model, if X is *strongly faithful* (SF) with respect to G; that is,

$$i \text{ and } j \text{ are d-separated by } \mathsf{S} \text{ under } \mathsf{G} \quad \Leftrightarrow \quad |\mathrm{cor}(X^i, X^j|X^{\mathsf{S}})| \le \lambda, \tag{26}$$

for some $\lambda > 0$. The necessary part of (26) is equivalent to

$$\min\{|\mathrm{cor}(X^i, X^j|X^{\mathsf{S}})| : \quad (i, j, \mathsf{S}) \notin \mathscr{D}\} \ > \lambda. \tag{27}$$

We now extend uniform consistency to our setting. Let $\mathscr{P}$ be a class of distributions of X, and, for each $P \in \mathscr{P}$, let $\mathscr{C}_{A}(P)$ be the ACI structure corresponding distribution P; that is, $\mathscr{C}_{A}(P) = \{(i, j, \mathsf{S}) \in \mathscr{T} : X^i \perp\!\!\!\perp_{A} X^j|X^{\mathsf{S}} \text{ under } P\}$. We first extend the strong faithfulness assumption to our setting.

Definition 23 *The family of distribution $\mathscr{P}$ is strongly additively faithful (SAF) with respect to G, if the Additive global Markov condition holds, and that there is a $\lambda > 0$ such that*

$$\min\{\|R_{X^i X^j|X^{\mathsf{S}}}(P)\| : \ (i,j,\mathsf{S}) \notin \mathscr{D}\} > \lambda \quad \forall P \in \mathscr{P}.$$

Let $\mathscr{P}_0$ be the subset of $\mathscr{P}$ whose members are faithful with respect to G; that is, $\mathscr{P}_0 = \{P \in \mathscr{P} : \ \mathscr{C}_A(P) = \mathscr{D}\}$. We then establish the uniform consistency of $\hat{\mathscr{C}}_A$.

Theorem 24 *Suppose*

a. $\mathscr{P}$ is strongly additively faithful with respect to G;

b. for each $(i,j,\mathsf{S}) \in \mathscr{T}$ and $\epsilon > 0$, we have

$$\lim_{n\to\infty} \sup_{P\in\mathscr{P}} P\left(\left\|\hat{R}^{\epsilon_n,\delta_n}_{X^i X^j|X^{\mathsf{S}}} - R_{X^i X^j|X^{\mathsf{S}}}(P)\right\| > \epsilon\right) = 0.$$

Then, for any $0 < \rho < \lambda$, we have

$$\lim_{n\to\infty} \sup_{P\in\mathscr{P}_0} P(\hat{\mathscr{C}}_A \neq \mathscr{D}) = 0, \quad \lim_{n\to\infty} \sup_{P\in\mathscr{P}_0^c} P(\hat{\mathscr{C}}_A = \mathscr{D}) = 0. \tag{28}$$

Condition b. is the uniform version of $P(\|\hat{R}^{\epsilon_n,\delta_n}_{X^i X^j|X^{\mathsf{S}}} - R_{X^i X^j|X^{\mathsf{S}}}(P)\| > \epsilon) \to 0$, which is proved in Section 6. It depends on the nature of the family $\mathscr{P}$. This type of condition is not regarded as very restrictive in the classical setting (see, for example Bickel et al. (1993, page 18).

Because there is a one-to-one correspondence between $\mathscr{D}$ and $(\mathsf{E}_{\text{SKE}}, \mathscr{V})$, Theorem 24 also implies the uniform convergence of $\hat{\mathsf{E}}_{\text{SKE}}$ and $\hat{\mathscr{V}}$, as recorded in the next corollary.

Corollary 25 *Under the conditions of Theorem 24, we have*

$$\lim_{n\to\infty} \sup_{P\in\mathscr{P}_0} P(\{\hat{\mathsf{E}}_{\text{SKE}} \neq \mathsf{E}_{\text{SKE}}\} \cup \{\hat{\mathscr{V}} \neq \mathscr{V}\}) = 0,$$
$$\lim_{n\to\infty} \sup_{P\in\mathscr{P}_0^c} P(\{\hat{\mathsf{E}}_{\text{SKE}} = \mathsf{E}_{\text{SKE}}\} \cap \{\hat{\mathscr{V}} = \mathscr{V}\}) = 0.$$

7.2. Comparison between SAF and SF

As concluded in Uhler et al. (2013), the SF condition in (27) can be very restrictive. In this subsection we show, through a numerical investigation, that SAF is a weaker condition than SF. For demonstration, we only consider a simple example of a complete DAG with three nodes. The distribution of the random vector $(X^1, X^2, X^3)^{\mathsf{T}}$ is generated via a *structural equation model* (SEM, see Pearl, 2009):

$$\begin{aligned} X^1 &= \varepsilon^1, \\ X^2 &= a_{21}X^1 + a_{22}X^1 \cdot X^1 + \varepsilon^2, \\ X^3 &= a_{31}X^1 + a_{32}X^1 \cdot X^1 + a_{33}X^2 + a_{34}X^2 \cdot X^2 + \varepsilon^3, \end{aligned} \tag{29}$$

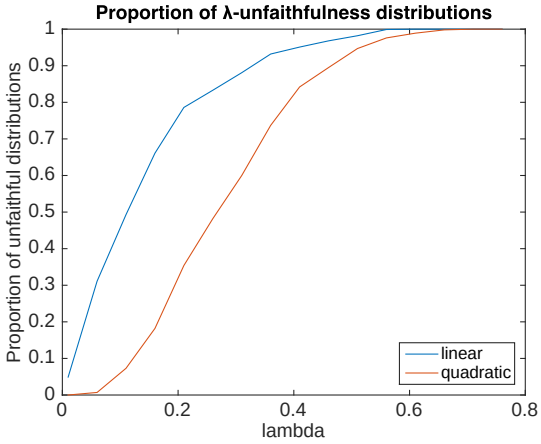

Figure 1: Proportion of unfaithfulness distributions.

where ε^i are *i.i.d.* standard Gaussian variables and $(a_{21}, \ldots, a_{34}) \in [-1, 1]^6$.

To characterize the level of *additive unfaithfulness* when SAF does not hold, we use the volume of the additively unfaithful distributions in the following sense. Let

$$\{P \in \mathscr{P} : \text{ there exists } (i, j, \mathsf{S}) \in \mathscr{D}^c \text{ such that } \|R_{X^i X^j | X^{\mathsf{S}}}(P)\| \leq \lambda\}$$

be the set of unfaithful distributions. In our example it has a one-to-one correspondence with the set

$$\{(a_{21}, \ldots, a_{34}) \in [-1, 1]^6 : \min_{i,j,\mathsf{S}} \|R_{X^i X^j | X^{\mathsf{S}}}(a_{21}, \ldots, a_{34})\| \leq \lambda\}, \tag{30}$$

where $R_{X^i X^j | X^{\mathsf{S}}}(a_{21}, \ldots, a_{34})$ is the NACCO corresponding to each $(a_{21}, \ldots, a_{34}) \in [-1, 1]^6$. We then use the volume of (30) to measure the level of unfaithfulness.

Because model (29) consists of both linear and quadratic functions, we use the polynomial kernel of order two; that is, $\kappa(a, b) = (1 + ab)^2$, $a, b \in \mathbb{R}$. The volume of (30) is empirically approximated as follows. Starting with an independent draw $(a_{21}, \ldots, a_{34})$ from Uniform$([-1, 1]^6)$, we compute $R_{X^i X^j | X^{\mathsf{S}}}(a_{21}, \ldots, a_{34})$ using the polynomial kernel. If there is some i, j, $\mathsf{S} \subseteq \mathsf{V} \backslash (i, j)$ such that $\|R_{X^i X^j | X^{\mathsf{S}}}(a_{21}, \ldots, a_{34})\| \leq \lambda$, then we count the distribution as additively unfaithful. We repeat the process 1,000 times to compute the proportion of additively unfaithful distributions. Uhler et al. (2013) considered a similar setting to estimate the proportion of unfaithful distributions based on linear models, where the coefficients of the quadratic terms a_{22}, a_{32}, a_{34} are set zeros. Figure 1 shows the proportions of both additive unfaithfulness and unfaithfulness with different values of λ. The curve of the additive unfaithfulness is consistently below that of the faithfulness. In particular, when $\lambda \leq 0.1$, it is very unlikely to have additive unfaithfulness. This indicates that the strong additive faithfulness is less restrictive than the strong faithfulness.

8. Numerical studies

8.1. Simulation settings

We compute the AFDAG estimator based on the principal-component-type procedure in Section 4.6 and AF-PC algorithm in Section 5.4. We also compare our estimator with two existing parametric and semi-parametric estimators: the linear-PC algorithm (Spirtes, Glymour, and Scheines, 2000; Kalisch and Bühlmann, 2007), i.e. PC algorithm combined with partial correlation test, and the rank-PC algorithm (Harris and Drton, 2013). A distinct feature of AFDAG is that it is able to capture interdependent structures beyond the Gaussian copula models. In order to reflect this feature, it is necessary to generate non-Gaussian random variables, which can be accomplished via SEM. We first describe how to produce non-Gaussian (or non-copula Gaussian) graphical models. Suppose an edge set E is given and the natural ordering of the nodes is $1 \to p$. Then the random vector $X = (X^1, \ldots, X^p)$ can be sequentially generated via

$$\begin{aligned} X^1 &= \varepsilon^1, \\ X^i &= f_i\left[\{X^j : (j,i) \in \mathsf{E}\}, \varepsilon^i\right], \quad i = 2, \ldots, p, \end{aligned}$$

where $f_i(\cdot)$ specifies the dependency of node i on its parent nodes, and $\varepsilon^1, \ldots, \varepsilon^p$ are *i.i.d.* standard Gaussian variables. The joint distribution of X from this model is non-Gaussian unless all f_i's are linear.

We then use the following two models for f_i, $i = 2, \ldots, p$,

$$\begin{aligned} \text{Linear}: &\quad f_i\left[\{X^j : (j,i) \in \mathsf{E}\}, \varepsilon^i\right] = \textstyle\sum_{(i,j)\in\mathsf{E}} \alpha_{i,j} X^j + \varepsilon^i, \\ \text{Quadratic}: &\quad f_i\left[\{X^j : (j,i) \in \mathsf{E}\}, \varepsilon^i\right] = \textstyle\sum_{(i,j)\in\mathsf{E}} \alpha_{i,j} X^j \cdot X^j + \varepsilon^i. \end{aligned}$$

To complete the simulation setting, we generate the graph E and the coefficients $\alpha_{i,j}$'s in the same way as Kalisch and Bühlmann (2007). Specifically, the graph is determined by the binary variable $I\left[(i,j) \in \mathsf{E}\right]$ and the value of $\alpha_{i,j} = I\left[(i,j) \in \mathsf{E}\right] A_{i,j}$, where

$$I\left[(i,j) \in \mathsf{E}\right] \sim \text{Bernoulli}(d), \quad A_{i,j} \sim \text{Uniform}(0.1, 1).$$

The parameter d controls the complexity of the graph: a larger d means a less sparse graph.

8.2. Comparison of $\hat{\mathsf{E}}_{\text{SKE}}$

We fix the network size at $p = 5$, the sparse parameter at $d = 0.1$, and vary the sample size n between 50, 100, and 300. We also fix the number of resamplings in the approximated permutation test $b = 5000$. For each model and each DAG, we compare the estimated skeleton $\hat{\mathsf{E}}_{\text{SKE}}$ and the truth E_{SKE}, and compute the corresponding ROC curves. The process is repeated 50 times and the averaged ROC curves are reported in Figure 2 for the linear model, and Figure 3 for the quadratic model.

We should mention that in our simulation experiments the unfaithfulness is less serious of an issue because the edge weight is generated from Uniform$(0.1, 1)$ with a support bounding away from 0. As noticed by (Uhler et al., 2013), when the causal parameters are large, the strong faithfulness becomes less problematic. Therefore, in our simulation we expect the performance of linear-PC (i.e. based on partial correlation) would perform the best when the

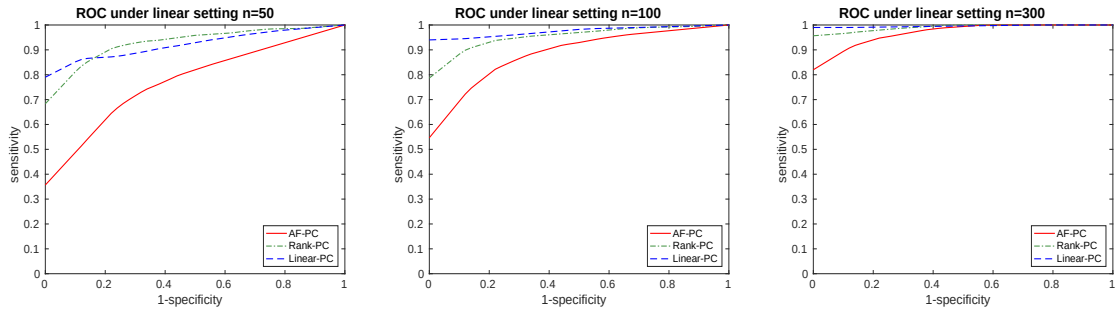

Figure 2: Comparisons of the ROC curves by AF-PC, rank-PC, and linear-PC on linear models.

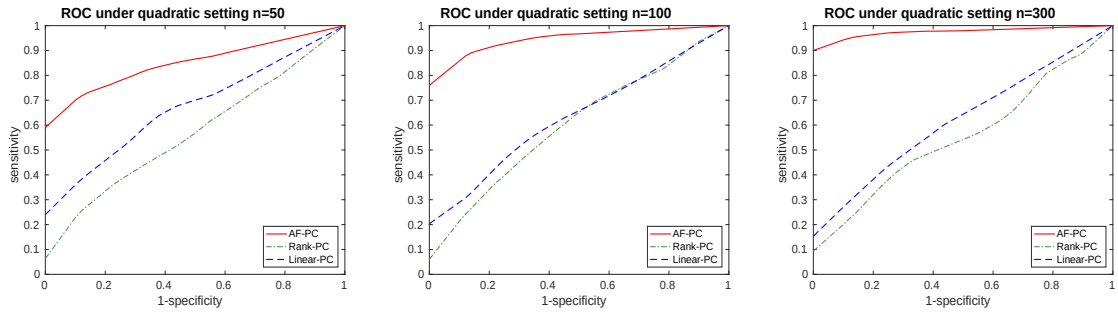

Figure 3: Comparisons of the ROC curves by AF-PC, rank-PC, and linear-PC on quadratic models.

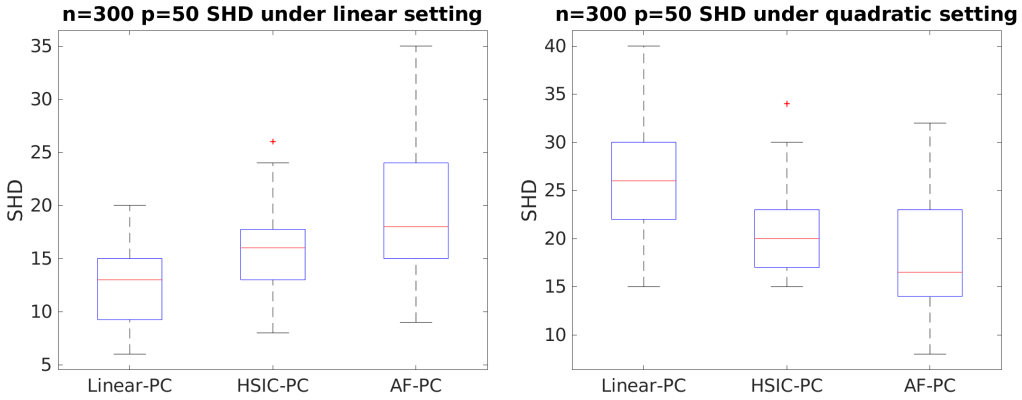

Figure 4: Boxplots of SHD($\hat{\mathsf{E}}_{\text{CPDAG}}$) from AF-PC, HSIC-PC, and linear-PC.

underlying distribution is precisely Gaussian. Nonetheless, when the underlying distribution deviates from normality, the improvement by our method over the other approaches is significant.

8.3. Comparison of $\hat{\mathsf{E}}_{\text{CPDAG}}$

We compare the performances of the estimators in recovering $\mathsf{E}_{\text{CPDAG}}$. Let $\hat{\mathsf{E}}_{\text{CPDAG}}$ denote an estimator of $\mathsf{E}_{\text{CPDAG}}$ by one of the methods. Because $\mathsf{E}_{\text{CPDAG}}$ is partially directed and provides more information about a DAG than the skeleton, we use a different criterion than the ROC to evaluate the performance of $\hat{\mathsf{E}}_{\text{CPDAG}}$, namely, the *structure Hamming distance* (Tsamardinos, Brown, and Aliferis, 2006). Given an estimate $\hat{\mathsf{E}}_{\text{CPDAG}}$, its *structure Hamming distance*, or SHD($\hat{\mathsf{E}}_{\text{CPDAG}}$), is the minimum number of single operations, including deletions, insertions, and re-orientations that are required to go from $\hat{\mathsf{E}}_{\text{CPDAG}}$ to $\mathsf{E}_{\text{CPDAG}}$. A smaller value of SHD($\hat{\mathsf{E}}_{\text{CPDAG}}$) indicates greater similarity between $\hat{\mathsf{E}}_{\text{CPDAG}}$ and $\mathsf{E}_{\text{CPDAG}}$.

Since the ROC curves in Figures 2 and 3 show that the linear-PC and the rank-PC produce very similar results, we only compare AF-PC with the first of the two. In addition, we also compare with the KCI-PC proposed by (Zhang et al., 2011). Furthermore, we also extend the comparison settings to include different network sizes ranging from 3 to 15. The averaged SHD($\hat{\mathsf{E}}_{\text{CPDAG}}$) from 50 replicates are reported in Tables 1 and 2. Again, AF-PC outperforms both linear-PC and KCI-PC when there exists nonlinearity between the nodes (in this case the Gaussian assumption no longer holds).

8.4. Large networks

Another feature of AFDAG is that it is capable of dealing with high-dimensional networks due to its computational simplicity, despite its nonparametric nature. To show this, we conduct a similar analysis with relatively large number of vertices: $p = 50$ and $d = 0.02$. Figure 4 shows the boxplots of 50 SHDs from both AF-PC and linear-PC. Moreover, we have added the comparison to HSIC-PC, the PC algorithm combined with CCO tests (Tillman et al., 2009). For the nonlinear model, AF-PC performs better than both linear-PC and HSIC-PC.

n	Method	$p=3$	$p=4$	$p=5$	$p=6$	$p=7$	$p=8$	$p=9$	$p=10$	$p=11$	$p=12$	$p=13$	$p=14$	$p=15$
100	Linear-PC	0.20	0.48	0.80	1.24	1.82	2.16	3.20	3.80	4.50	4.80	5.90	6.70	6.88
	KCI-PC	2.76	5.42	8.96	13.12	17.70	22.28	25.78	30.40	34.40	37.26	40.50	43.40	46.86
	AF-PC	0.30	0.72	0.98	1.64	2.28	2.78	3.66	4.24	5.50	6.66	7.58	8.68	10.32
300	Linear-PC	0.18	0.54	0.76	1.02	1.62	2.28	2.66	3.36	4.08	4.32	4.86	5.22	6.14
	KCI-PC	2.92	5.52	8.88	13.06	17.52	21.86	25.74	29.08	33.14	35.70	38.20	40.94	42.62
	AF-PC	0.30	0.60	0.84	1.22	1.88	2.82	3.30	4.12	4.76	5.34	6.44	7.22	8.56
500	Linear-PC	0.20	0.54	0.80	1.02	1.42	1.92	2.48	3.10	4.10	4.22	5.00	5.18	5.88
	KCI-PC	2.78	5.58	9.18	12.98	17.64	22.54	27.12	30.74	35.78	38.80	41.78	45.00	48.28
	AF-PC	0.24	0.54	0.74	1.20	1.86	2.26	3.12	3.94	4.58	5.02	6.02	7.14	8.08

Table 1: Comparisons of averaged SHD by AF-PC and linear-PC with network sizes $p = 3$ to 15 on linear models.

n	Method	$p=3$	$p=4$	$p=5$	$p=6$	$p=7$	$p=8$	$p=9$	$p=10$	$p=11$	$p=12$	$p=13$	$p=14$	$p=15$
100	Linear-PC	0.44	0.76	1.44	2.24	2.92	3.56	4.98	6.04	7.38	9.14	10.42	11.60	13.74
	KCI-PC	2.82	5.44	9.12	13.34	18.08	22.46	26.22	30.12	34.66	38.06	41.20	44.58	46.60
	AF-PC	0.30	0.56	0.98	1.50	1.94	2.46	3.46	4.02	4.74	6.10	6.94	8.48	9.74
300	Linear-PC	0.30	0.80	1.44	2.06	3.00	4.06	5.22	6.48.36	8.08	9.20	10.80	12.20	14.74
	KCI-PC	2.92	5.52	9.04	13.08	17.66	21.80	26.36	29.62	33.60	36.26	39.26	42.62	45.16
	AF-PC	0.26	0.50	0.90	1.24	1.74	2.28	3.06	3.98	4.32	5.32	6.36	6.78	8.48
500	Linear-PC	0.46	0.88	1.56	2.22	3.18	3.98	5.06	6.46	8.26	9.56	11.60	12.98	15.32
	KCI-PC	2.80	5.52	9.20	12.94	17.54	22.66	27.36	31.72	36.46	40.20	43.00	46.88	49.94
	AF-PC	0.22	0.46	0.80	1.20	1.94	2.48	3.28	3.80	4.50	5.08	6.00	7.16	7.84

Table 2: Comparisons of averaged SHD by AF-PC and linear-PC with network sizes $p = 3$ to 15 on quadratic models.

	AF-PC	HSIC-PC	KCI-PC	linear-PC
mean (std)	**16.80** (1.80)	17.67 (1.40)	24.45 (1.23)	19.20 (1.53)

Table 3: Mean (s.d.) of the $\text{SHD}(\hat{\mathsf{E}}_{\text{CPDAG}})$ by AF-PC, HSIC-PC, KCI-PC, and linear-PC.

8.5. The Mitogen-Activated Protein Kinase (MAPK) pathways

We next apply our method, the linear-PC, HSIC-PC, and KCI-PC to a flow cytometry data set from Sachs et al. (2005), in which $p = 11$ protein activities levels were measured on $n = 7466$ cells. The purpose of this study is to recover the causal networks of the protein signaling pathways within a human immune system. This data set has also been investigated by Friedman, Hastie, and Tibshirani (2008), Ellis and Wong (2008), and Luo and Zhao (2011). The underlying true DAG can be found in Friedman, Hastie, and Tibshirani (Figure 2, 2008). We conduct a stability analysis by first drawing a subsample of 2,000 cells. For each subsample, we then compute the $\text{SHD}(\hat{\mathsf{E}}_{\text{CPDAG}})$ from all competing methods. This process is repeated 20 times and the averaged SHDs and standard deviations are reported in Table 3.

AF-PC performs the best among all competing methods. This indicates that the CPDAG estimated by AF-PC is closer to the truth.

9. Concluding remarks

In this paper we introduced a statistical model, called the additively faithful directed acyclic graph, for causal network analysis, as well as a procedure to estimate the graph based on a normalized additive conditional covariance operator. Additive faithfulness is derived from additive conditional independence, which provides a balance between a parametric model and a fully nonparametric model: it enjoys the flexibility of a nonparametric method but at the same time avoids using multi-dimensional kernels. In particular, it can capture the type of nonlinear interactions that elude a Gaussian graphical model or a copula-Gaussian graphical model (see Li, Chun, and Zhao, 2014).

We also introduced an efficient PC-type algorithm to implement the estimator of the AFDAG, so that the complexity of the algorithm does not depend on the dimension of the network but instead on the level of sparseness of AFDAG. This feature makes the proposed method feasible for high-dimensional networks because the skeleton of a DAG is typically sparse (containing far fewer edges than the corresponding undirected graph). Along with theory of AFDAG and the methods for its estimation, we also established the consistency of the estimators, as well as their convergence rates. Furthermore, we established the uniform consistency of the proposed procedure under a strong additive faithfulness; we showed that this condition is weaker than the strong faithfulness in the linear setting, and is therefore more reasonable in modeling causality networks.

The methods developed in Sections 4 and 5 are but the first attempt to implement the general theory of AFDAG laid out in Sections 2 and 3, and there is much room for refinement and further development. For example, in this paper we only considered the thresholding or CI testing as a mechanism for deciding the absence of edges, but more sophisticated sparse penalized optimization, such as the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), and the adaptive LASSO (Zou, 2006) may be employed at the operator

level to further improve the accuracy of the estimation of AFDAG. These will be left for further research.

Acknowledgments

This research includes calculations carried out on HPC resources supported in part by the National Science Foundation (NSF) through major research instrumentation grant number 1625061 and by the US Army Research Laboratory under contract number W911NF-16-2-0189. The research was also supported in part by NSF DMS-1713078 grant awareded to Bing Li, and NSF DMS-1902903 and NIH R01 GM122078 grants awarded to Hongyu Zhao. We would alos like to thank two referees for their many constructive comments and suggestions.

Appendix: Proofs

Proof of Theorem 5: We first show that, for any $f \in \mathscr{H}_{X^i}$ and $g \in \mathscr{H}_{X^j}$,

$$\langle f, \Sigma_{X^iX^j|X^{\mathsf{S}}} g\rangle = \mathrm{cov}[f(X^i) - (T_{X^{\mathsf{S}}X^i} f)(X^{\mathsf{S}}), g(X^j) - (T_{X^{\mathsf{S}}X^j} g)(X^{\mathsf{S}})]. \tag{31}$$

By Lee et al. (2016, Proposition 2), for any $f \in \mathscr{H}_{X^i}$ and $g \in \mathscr{H}_{X^j}$,

$$\mathrm{cov}[(T_{X^{\mathsf{S}}X^i} f)(X^{\mathsf{S}}), g(X^j) - (T_{X^{\mathsf{S}}X^j} g)(X^j)] = 0.$$

Hence the right-hand side of (31) reduces to

$$\mathrm{cov}[f(X^i), g(X^j) - (T_{X^{\mathsf{S}}X^j} g)(X^{\mathsf{S}})] = \mathrm{cov}[f(X^i), g(X^j)] - \mathrm{cov}[f(X^i), (T_{X^{\mathsf{S}}X^j} g)(X^{\mathsf{S}})]. \tag{32}$$

By Assumption 2, $\langle f, \Sigma^{1/2}_{X^iX^i} R_{X^iX^{\mathsf{S}}} R^{-1}_{X^{\mathsf{S}}X^{\mathsf{S}}} R_{X^{\mathsf{S}}X^j} \Sigma^{1/2}_{X^jX^j} g\rangle = \langle f, \Sigma^{1/2}_{X^iX^i} R_{X^iX^{\mathsf{S}}} D^{1/2}_{X^{\mathsf{S}}X^{\mathsf{S}}} T_{X^{\mathsf{S}}X^j} g\rangle$, which is equal to $\langle f, \Sigma_{X^iX^{\mathsf{S}}} T_{X^{\mathsf{S}}X^j} g\rangle = \mathrm{cov}[f(X^i), (T_{X^{\mathsf{S}}X^j} g)(X^j)]$ by the definition of $\Sigma_{X^iX^{\mathsf{S}}}$. In the meantime, $\langle f, \Sigma_{X^iX^j} g\rangle = \mathrm{cov}[f(X^i), g(X^j)]$. Thus the right-hand side of (32) is simply $\langle f, \Sigma_{X^iX^j|X^{\mathsf{S}}} g\rangle$, proving (31).

By Definition 1, the right-hand side of (9) holds if and only if the right-hand side of (31) is 0 for all $f \in \mathscr{H}_{X^i}$ and $g \in \mathscr{H}_{X^j}$; the left-hand side of (9) holds if and only if the left-hand side of (31) is 0 for all $f \in \mathscr{H}_{X^i}$ and $g \in \mathscr{H}_{X^j}$. Thus the equality (31) implies the equivalence in (9). □

Proof of Proposition 8: For any $f \in \mathscr{G}^{(n)}_{X^i}$, $g \in \mathscr{G}^{(n)}_{X^j}$,

$$\begin{aligned}\langle f, \hat{\Sigma}_{X^iX^j} g\rangle &= \langle [f]^{\mathsf{T}}\psi^i, (\psi^i)^{\mathsf{T}}[\hat{\Sigma}_{X^iX^j}][g]\rangle \\ &= [f]^{\mathsf{T}}\Lambda^{-1/2}_{X^i} V^{\mathsf{T}}_{X^i} G_{X^i} V_{X^i} \Lambda^{-1/2}_{X^i}[\hat{\Sigma}_{X^iX^j}][g] = [f]^{\mathsf{T}}[\hat{\Sigma}_{X^iX^j}][g].\end{aligned} \tag{33}$$

By the definition of $\hat{\Sigma}_{X^iX^j}$, the left hand side also equals

$$\mathrm{cov}_n[f(X^i), g(X^j)] = n^{-1}[f]^{\mathsf{T}}\Lambda^{-1/2}_{X^i} V^{\mathsf{T}}_{X^i} G_{X^i} G_{X^j} V_{X^j} \Lambda^{-1/2}_{X^j}[g]. \tag{34}$$

Equate the right hand sides of (33) and (34) to obtain $[\hat{\Sigma}_{X^iX^j}]$. We can use the same argument to derive $[\hat{\Sigma}_{X^iX^i}]$. The last three coordinate representations are simply block matrices

with $\hat{\Sigma}_{X^iX^j}$ as blocks. For example, $[\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}}]$ is the $\mathrm{card}(\mathsf{S}) \times \mathrm{card}(\mathsf{S})$ matrix of submatrices whose (i,j)th block is $[\hat{\Sigma}_{X^iX^j}]$. Hence it has the asserted form. □

Proof of Proposition 9: By (17),

$$\begin{aligned}[\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}] = &[\hat{\Sigma}_{X^iX^i}]^{\dagger 1/2}(\delta_n)[\hat{\Sigma}_{X^iX^j}][\hat{\Sigma}_{X^jX^j}]^{\dagger 1/2}(\delta_n) \\ &- [\hat{\Sigma}_{X^iX^i}]^{\dagger 1/2}(\delta_n)[\hat{\Sigma}_{X^iX^{\mathsf{S}}}][\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}}]^{\dagger}(\epsilon_n)[\hat{\Sigma}_{X^{\mathsf{S}}X^j}][\hat{\Sigma}_{X^jX^j}]^{\dagger 1/2}(\delta_n)\end{aligned} \tag{35}$$

After substituting the coordinate representations in Proposition 8, the first term on the right-hand side becomes

$$(n^{-1}\Lambda_{X^i})^{\dagger 1/2}(\delta_n)(n^{-1}\Lambda^{1/2}_{X^i}V^{\mathsf{T}}_{X^i}V_{X^j}\Lambda^{1/2}_{X^j})(n^{-1}\Lambda_{X^j})^{\dagger 1/2}(\delta_n) = \mathrm{diag}(I_{r_i},0)V^{\mathsf{T}}_{X^i}V_{X^j}\mathrm{diag}(I_{r_j},0).$$

Similarly, the second term (without the negative sign) on the right-hand side of (35) is

$$\begin{aligned}&(n^{-1}\Lambda_{X^i})^{\dagger 1/2}(\delta_n)(n^{-1}\Lambda^{1/2}_{X^i}V^{\mathsf{T}}_{X^i}M_{X^{\mathsf{S}}})(n^{-1}M^{\mathsf{T}}_{X^{\mathsf{S}}}M_{X^{\mathsf{S}}})^{\dagger}(\epsilon_n)(n^{-1}M^{\mathsf{T}}_{S}V_{X^j}\Lambda^{1/2}_{X^j})(n^{-1}\Lambda_{X^j})^{\dagger 1/2}(\delta_n) \\ &\quad = \mathrm{diag}(I_{r_i},0)V^{\mathsf{T}}_{X^i}M_{X^{\mathsf{S}}}(M^{\mathsf{T}}_{X^{\mathsf{S}}}M_{X^{\mathsf{S}}})^{\dagger}(\epsilon_n)M^{\mathsf{T}}_{S}V_{X^j}\mathrm{diag}(I_{r_j},0).\end{aligned}$$

Substitute these two terms into (35) to obtain the desired matrix. Finally, the norm is bounded by 1 because $I_n - M_{X^{\mathsf{S}}}(M^{\mathsf{T}}_{X^{\mathsf{S}}}M_{X^{\mathsf{S}}})^{\dagger}(\epsilon_n)M^{\mathsf{T}}_{X^{\mathsf{S}}}$ is a projection matrix. □

Proof of Lemma 13: *(a)* By the definition of empirical covariance operator, we have $\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}} \geq 0$. Therefore, $\|(\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon I)^{-1}\| \leq \epsilon^{-1}$. This is also true if we replace $\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}}$ by $\Sigma_{X^{\mathsf{S}}X^{\mathsf{S}}}$.
(b) First note that $\|(\Sigma_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon I)^{-1/2}\Sigma^{1/2}_{X^{\mathsf{S}}X^{\mathsf{S}}}\| \leq 1$. Since $R_{X^iX^j}$ is compact for all $i < j$, by Lemma A3 in Lee, Li, and Zhao (2016), there exists $C \in \mathscr{B}(\mathscr{H}_{X^{\mathsf{S}}})$ such that $D^{1/2}_{X^{\mathsf{S}}X^{\mathsf{S}}} = \Sigma^{1/2}_{X^{\mathsf{S}}X^{\mathsf{S}}}C$. Hence $\|(\Sigma_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon I)^{-1/2}D^{1/2}_{X^{\mathsf{S}}X^{\mathsf{S}}}\| \leq \|(\Sigma_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon I)^{-1/2}\Sigma^{1/2}_{X^{\mathsf{S}}X^{\mathsf{S}}}\|\|C\| \leq \|C\|$, which implies the asserted result. □

Proof of Lemma 14: It suffices to show that $\Sigma_{X^iX^{\mathsf{S}}}(\Sigma_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon_n I)^{-1}\Sigma_{X^{\mathsf{S}}X^j}$ is bounded. This term can be decomposed as

$$\Sigma_{X^iX^{\mathsf{S}}}(\Sigma_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon_n I)^{-1}\Sigma_{X^{\mathsf{S}}X^j} = \Sigma^{1/2}_{X^iX^i}R_{X^iX^{\mathsf{S}}}D^{1/2}_{X^{\mathsf{S}}X^{\mathsf{S}}}(\Sigma_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon_n I)^{-1}D^{1/2}_{X^{\mathsf{S}}X^{\mathsf{S}}}R_{X^{\mathsf{S}}X^j}\Sigma^{1/2}_{X^jX^j}.$$

By Lemma 13, $\|D^{1/2}_{X^{\mathsf{S}}X^{\mathsf{S}}}(\Sigma_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon_n I)^{-1}D^{1/2}_{X^{\mathsf{S}}X^{\mathsf{S}}}\|$ is bounded by a constant independent of n. Therefore, the norm of the above displayed operator is also bounded by a constant independent of n. □

Proof of Lemma 15: By the triangular inequality,

$$\begin{aligned}\|\hat{\Sigma}^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}} - \Sigma^{\epsilon_n}_{X^iX^j|X^{\mathsf{S}}}\| \leq &\|\hat{\Sigma}_{X^iX^j} - \Sigma_{X^iX^j}\| \\ &+ \|\hat{\Sigma}_{X^iX^{\mathsf{S}}}(\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon_n I)^{-1}\hat{\Sigma}_{X^{\mathsf{S}}X^j} - \Sigma_{X^iX^{\mathsf{S}}}(\Sigma_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon_n I)^{-1}\Sigma_{X^{\mathsf{S}}X^j}\|.\end{aligned}$$

The second term on the right can be decomposed as $\Theta_1 + \Theta_2 + \Theta_3$, where

$$\begin{aligned}\Theta_1 &= \|\hat{\Sigma}_{X^iX^{\mathsf{S}}}(\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon_n I)^{-1}\hat{\Sigma}_{X^{\mathsf{S}}X^j} - \Sigma_{X^iX^{\mathsf{S}}}(\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon_n I)^{-1}\hat{\Sigma}_{X^{\mathsf{S}}X^j}\|, \\ \Theta_2 &= \|\Sigma_{X^iX^{\mathsf{S}}}(\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon_n I)^{-1}\hat{\Sigma}_{X^{\mathsf{S}}X^j} - \Sigma_{X^iX^{\mathsf{S}}}(\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon_n I)^{-1}\Sigma_{X^{\mathsf{S}}X^j}\|, \\ \Theta_3 &= \|\Sigma_{X^iX^{\mathsf{S}}}(\hat{\Sigma}_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon_n I)^{-1}\Sigma_{X^{\mathsf{S}}X^j} - \Sigma_{X^iX^{\mathsf{S}}}(\Sigma_{X^{\mathsf{S}}X^{\mathsf{S}}} + \epsilon_n I)^{-1}\Sigma_{X^{\mathsf{S}}X^j}\|.\end{aligned}$$

By Lemma 13,

$$\Theta_1 \leq \epsilon_n^{-1} \|\hat{\Sigma}_{X^i X^{\mathsf{S}}} - \Sigma_{X^i X^{\mathsf{S}}}\| \, \|\hat{\Sigma}_{X^{\mathsf{S}} X^j}\|,$$

which is of order $O_P(\epsilon_n^{-1} n^{-1/2})$ by Lemma 4 in Fukumizu, Bach, and Gretton (2007). Similarly, $\Theta_2 = O_P(\epsilon_n^{-1} n^{-1/2})$. Regarding Θ_3 we have

$$\begin{aligned} \Theta_3 \leq & \|\Sigma_{X^i X^{\mathsf{S}}} (\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon_n I)^{-1/2}\| \, \|(\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon_n I)^{-1/2} \Sigma_{X^{\mathsf{S}} X^i}\| \\ & \|(\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon_n I)^{-1/2} (\hat{\Sigma}_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon_n I)(\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon_n I)^{-1/2} - I\|, \end{aligned} \tag{36}$$

where the last term in the above inequality is due to the following relation

$$\begin{aligned} & \|(\hat{\Sigma}_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon_n I)^{1/2} (\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon_n I)^{-1} (\hat{\Sigma}_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon_n I)^{1/2} - I\| \\ & \qquad = \|(\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon_n I)^{-1/2} (\hat{\Sigma}_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon_n I)(\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon_n I)^{-1/2} - I\|, \end{aligned}$$

which is identical to $\|(\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon_n I)^{-1/2} (\hat{\Sigma}_{X^{\mathsf{S}} X^{\mathsf{S}}} - \Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}})(\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon_n I)^{-1/2}\|$. Substitute this into (36) to obtain $\Theta_3 = O_P(\epsilon_n^{-1} n^{-1/2})$. □

Proof of Lemma 16: By Assumption 2, we have

$$\begin{aligned} & \Sigma_{X^i X^{\mathsf{S}}} (\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon I)^{-1} \Sigma_{X^{\mathsf{S}} X^j} - \Sigma^{1/2}_{X^i X^i} R_{X^i X^{\mathsf{S}}} R^{-1}_{X^{\mathsf{S}} X^{\mathsf{S}}} R_{X^{\mathsf{S}} X^j} \Sigma^{1/2}_{X^j X^j} \\ & = \Sigma_{X^i X^{\mathsf{S}}} ((\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon I)^{-1} \Sigma_{X^{\mathsf{S}} X^j} - T_{X^{\mathsf{S}} X^j}) = -\epsilon_n \Sigma_{X^i X^{\mathsf{S}}} (\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon I)^{-1} T_{X^{\mathsf{S}} X^j}, \end{aligned}$$

whose norm is bounded by

$$\epsilon \|\Sigma_{X^i X^{\mathsf{S}}} (\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon I)^{-1}\| \|T_{X^{\mathsf{S}} X^j}\| \leq \epsilon \|D^{1/2}_{X^{\mathsf{S}} X^{\mathsf{S}}} (\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon I)^{-1/2}\| \|(\Sigma_{X^{\mathsf{S}} X^{\mathsf{S}}} + \epsilon I)^{-1/2}\|.$$

The lemma now follows from Lemma 13. □

Proof of Lemma 19: Similar to the proof in Lemma 15, we have

$$\|\hat{R}^{\epsilon_n, \delta_n}_{X^i X^j | X^{\mathsf{S}}} - (\Sigma_{X^i X^i} + \delta_n I)^{-1/2} \Sigma_{X^i X^j | X^{\mathsf{S}}} (\Sigma_{X^j X^j} + \delta_n I)^{-1/2}\| \leq \Gamma_1 + \Gamma_2 + \Gamma_3,$$

where

$$\begin{aligned} \Gamma_1 &= \|((\hat{\Sigma}_{X^i X^i} + \delta_n I)^{-1/2} - (\Sigma_{X^i X^i} + \delta_n I)^{-1/2}) \hat{\Sigma}^{\epsilon_n}_{X^i X^j | X^{\mathsf{S}}} (\hat{\Sigma}_{X^j X^j} + \delta_n I)^{-1/2}\|, \\ \Gamma_2 &= \|(\Sigma_{X^i X^i} + \delta_n I)^{-1/2} (\hat{\Sigma}^{\epsilon_n}_{X^i X^j | X^{\mathsf{S}}} - \Sigma_{X^i X^j | X^{\mathsf{S}}}) (\hat{\Sigma}_{X^j X^j} + \delta_n I)^{-1/2}\|, \\ \Gamma_3 &= \|(\Sigma_{X^i X^i} + \delta_n I)^{-1/2} \Sigma_{X^i X^j | X^{\mathsf{S}}} ((\hat{\Sigma}_{X^j X^j} + \delta_n I)^{-1/2} - (\Sigma_{X^j X^j} + \delta_n I)^{-1/2})\|. \end{aligned}$$

To show Γ_1 is sufficiently small, we first observe that it is bounded from above by $\Gamma_{1,1} + \Gamma_{1,2}$, where

$$\begin{aligned} \Gamma_{1,1} = & \|(\Sigma_{X^i X^i} + \delta_n I)^{-1/2} ((\Sigma_{X^i X^i} + \delta_n I)^{3/2} - (\hat{\Sigma}_{X^i X^i} + \delta_n I)^{3/2}) \\ & (\hat{\Sigma}_{X^i X^i} + \delta_n I)^{-3/2} \hat{\Sigma}^{\epsilon_n}_{X^i X^j | X^{\mathsf{S}}} (\hat{\Sigma}_{X^j X^j} + \delta_n I)^{-1/2}\|, \\ \Gamma_{1,2} = & \|(\hat{\Sigma}_{X^i X^i} - \Sigma_{X^i X^i}) (\hat{\Sigma}_{X^i X^i} + \delta_n I)^{-3/2} \hat{\Sigma}^{\epsilon_n}_{X^i X^j | X^{\mathsf{S}}} (\hat{\Sigma}_{X^j X^j} + \delta_n I)^{-1/2}\|. \end{aligned}$$

Note that $\Gamma_{1,1}$ is no greater than $\delta_n^{-3/2}\|((\Sigma_{X^iX^i}+\delta_n I)^{3/2}-(\hat{\Sigma}_{X^iX^i}+\delta_n I)^{3/2})\|\,\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\|$. By Proposition 9 and Lemma 18, this term is of the order $\delta_n^{-3/2}O_P(n^{-1/2}\epsilon_n^{-1}+\epsilon_n^{1/2})$. The term $\Gamma_{1,2}$ is upper bounded by $\delta_n^{-1}\|(\hat{\Sigma}_{X^iX^i}-\Sigma_{X^iX^i})\| = \delta_n^{-1}O_P(n^{-1/2})$. Therefore, we have $\Gamma_1 = o_P(1)$. Similarly, one can show Γ_3 is bounded by $\delta_n^{-3/2}O_P(n^{-1/2}\epsilon_n^{-1}+\epsilon_n^{1/2})\,\|R^{\delta_n}_{X^iX^j|X^{\mathsf{S}}}\|$, which is $o_P(1)$ because $\|R^{\delta_n}_{X^iX^j|X^{\mathsf{S}}}\|$ is bounded by Lemma 20.

Meanwhile, by Lemmas 15 and 16, $\Gamma_2 = \delta_n^{-1}\left[O_P(n^{-1/2}+\epsilon_n^{-1}n^{-1/2})+o(\delta_n)\right] = o_P(1)$, which completes the proof. □

Proof of Lemma 20: Note that $\|R^{\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R_{X^iX^j|X^{\mathsf{S}}}\| \le \Xi_1+\Xi_2+\Xi_3$, where

$$\begin{aligned}
\Xi_1 &= \|(\Sigma_{X^iX^i}+\delta_n I)^{-1/2}\Sigma_{X^iX^j}(\Sigma_{X^jX^j}+\delta_n I)^{-1/2} - R_{X^iX^j}\|,\\
\Xi_2 &= \|\left[(\Sigma_{X^iX^i}+\delta_n I)^{-1/2}\Sigma^{1/2}_{X^iX^i} - I\right] R_{X^iX^{\mathsf{S}}}R^{-1}_{X^{\mathsf{S}}X^{\mathsf{S}}}R_{X^{\mathsf{S}}X^j}\Sigma^{1/2}_{X^jX^j}(\Sigma_{X^jX^j}+\delta_n I)^{-1/2}\|,\\
\Xi_3 &= \|R_{X^iX^{\mathsf{S}}}R^{-1}_{X^{\mathsf{S}}X^{\mathsf{S}}}R_{X^{\mathsf{S}}X^j}\left[\Sigma^{1/2}_{X^jX^j}(\Sigma_{X^jX^j}+\delta_n I)^{-1/2} - I\right]\|.
\end{aligned}$$

By Fukumizu, Bach, and Gretton (2007, Lemma 7), we have $\Xi_1 = o(1)$. To show Ξ_2 converges to 0, we note that

$$\Xi_2 \le \|\left[(\Sigma_{X^iX^i}+\delta_n I)^{-1/2}\Sigma^{1/2}_{X^iX^i} - I\right] R_{X^iX^{\mathsf{S}}}R^{-1}_{X^{\mathsf{S}}X^{\mathsf{S}}}\|\,\|R_{X^{\mathsf{S}}X^j}\|\,\|\Sigma^{1/2}_{X^jX^j}(\Sigma_{X^jX^j}+\delta_n I)^{-1/2}\|,$$

where the last norm on the right is $O(1)$ by Lemma 13. Because $R_{X^iX^{\mathsf{S}}}R^{-1}_{X^{\mathsf{S}}X^{\mathsf{S}}}$ is compact, we have

$$\mathrm{range}(R_{X^iX^{\mathsf{S}}}R^{-1}_{X^{\mathsf{S}}X^{\mathsf{S}}}) \subseteq \overline{\mathrm{ran}}\,(\Sigma_{X^iX^i}) = \overline{\mathrm{ran}}\,(\Sigma^{1/2}_{X^iX^i}).$$

Also note that, for any $h' \in \mathscr{H}_{X^i}$,

$$\begin{aligned}
&\|((\Sigma_{X^iX^i}+\delta_n I)^{-1/2}\Sigma^{1/2}_{X^iX^i} - I)\Sigma^{1/2}_{X^iX^i}h'\|\\
&\quad = \|(\Sigma_{X^iX^i}+\delta_n I)^{-1/2}\Sigma^{1/2}_{X^iX^i}\left[\Sigma^{1/2}_{X^iX^i} - (\Sigma_{X^iX^i}+\delta_n I)^{1/2}\right]h'\|,
\end{aligned}$$

which has order $\|\Sigma^{1/2}_{X^iX^i} - (\Sigma_{X^iX^i}+\delta_n I)^{1/2}\| = o(1)$ as $\delta_n \to 0$.

Therefore, by Lemma A8 in Lee, Li, and Zhao (2016),

$$\|\left[(\Sigma_{X^iX^i}+\delta_n I)^{-1/2}\Sigma^{1/2}_{X^iX^i} - I\right] R_{X^iX^{\mathsf{S}}}R^{-1}_{X^{\mathsf{S}}X^{\mathsf{S}}}\| \to 0, \text{ as } \delta_n \to 0,$$

which implies $\Xi_2 = o(1)$. By a similar statement one can show $\Xi_3 = o(1)$. □

Proof of Theorem 22: Let the thresholding value be a sufficiently small positive number such that

$$\rho < \min\{\|R_{X^iX^j|X^{\mathsf{S}}}\| : (i,j,\mathsf{S}) \notin \mathscr{C}_A\}.$$

By Theorems 17 and 21, for any $(i,j) \in \mathscr{C}_A$, $\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\|$ converges to 0 in probability. Hence it is smaller than ρ with probability tending to 1. Therefore, the event $(i,j) \in \hat{\mathscr{C}}_A(\epsilon_n,\delta_n,\rho)$ has probability tending to 1. Similarly, for any $(i,j) \in \mathsf{E}_{\mathrm{SKE}}$, $\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\|$ converges to

$\|R_{X^iX^j|X^{\mathsf{S}}}\| > \rho$, implying that the event $(i,j,\mathsf{S}) \notin \hat{\mathscr{C}}_A(\epsilon_n,\delta_n,\rho)$ has probability tending to 1. The consistency of $\hat{\mathsf{E}}_{\mathrm{SKE}}(\epsilon_n,\delta_n,\rho)$ and that of $\hat{\mathscr{V}}(\epsilon_n,\delta_n,\rho)$ follow because they are continuous functions of $\hat{\mathscr{C}}_A(\epsilon_n,\delta_n,\rho)$ (with respect to the discrete topology). □

Proof of Theorem 24. To prove the first convergence in (28), we note that

$$\begin{aligned} P(\hat{\mathscr{C}}_A \neq \mathscr{D}) = P(&\text{ the statements } \|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\| < \rho \ \ \forall (i,j,\mathsf{S}) \in \mathscr{D} \\ &\text{and } \|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\| \geq \rho \ \ \forall (i,j,\mathsf{S}) \in \mathscr{D}^c \text{ are not both true}). \end{aligned} \tag{37}$$

By a., for each $P \in \mathscr{P}_0$, $\|R_{X^iX^j|X^{\mathsf{S}}}(P)\| = 0$ for all $(i,j,\mathsf{S}) \in \mathscr{D}$ and $\|R_{X^iX^j|X^{\mathsf{S}}}(P)\| > \lambda$ for all $(i,j,\mathsf{S}) \notin \mathscr{D}$. Hence the right hand side of (37) is

$$\begin{aligned} P(&\text{ the statements } \|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\| < \rho \ \ \forall (i,j,\mathsf{S}) \in \mathscr{D} \\ &\text{and } \|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\| \geq \rho \ \ \forall (i,j,\mathsf{S}) \in \mathscr{D}^c \text{ are not both true}, \\ &\|R_{X^iX^j|X^{\mathsf{S}}}(P)\| = 0 \ \forall\ (i,j,\mathsf{S}) \in \mathscr{D}, \ \ \|R_{X^iX^j|X^{\mathsf{S}}}(P)\| > \lambda \ \forall\ (i,j,\mathsf{S}) \notin \mathscr{D}). \end{aligned}$$

Let S_n represent the event inside the parentheses. Then

$$\begin{aligned} P(S_n) = &P(S_n, \ \|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R_{X^iX^j|X^{\mathsf{S}}}(P)\| < \epsilon \ \forall (i,j,\mathsf{S}) \in \mathscr{T}) \\ &+ P(S_n, \ \|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R_{X^iX^j|X^{\mathsf{S}}}(P)\| \geq \epsilon \text{ for some } (i,j,\mathsf{S}) \in \mathscr{T}) \\ \leq &P(S_n, \ \|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R_{X^iX^j|X^{\mathsf{S}}}(P)\| < \epsilon \ \forall (i,j,\mathsf{S}) \in \mathscr{T}) \\ &+ P(\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R_{X^iX^j|X^{\mathsf{S}}}(P)\| \geq \epsilon \text{ for some } (i,j,\mathsf{S}) \in \mathscr{T}). \end{aligned}$$

The second term on the right-hand side is

$$\begin{aligned} P(\cup_{(i,j,\mathsf{S})\in\mathscr{T}}\{\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R_{X^iX^j|X^{\mathsf{S}}}(P)\| \geq \epsilon\}) \\ \leq \textstyle\sum_{(i,j,\mathsf{S})\in\mathscr{T}} P(\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R_{X^iX^j|X^{\mathsf{S}}}(P)\| \geq \epsilon). \end{aligned}$$

It follows that

$$\begin{aligned} \sup_{P\in\mathscr{P}_0} P(S_n) \leq &\sup_{P\in\mathscr{P}_0} P(S_n, \ \|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R_{X^iX^j|X^{\mathsf{S}}}(P)\| < \epsilon \ \forall (i,j,\mathsf{S}) \in \mathscr{T}) \\ &+ \textstyle\sum_{(i,j,\mathsf{S})\in\mathscr{T}} \sup_{P\in\mathscr{P}_0} P(\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R_{X^iX^j|X^{\mathsf{S}}}(P)\| \geq \epsilon). \end{aligned}$$

Since the set $\mathscr{T}$ is finite, by condition b., the second term on the right-hand side tends to 0 as $n \to \infty$. Hence

$$\limsup_{n\to\infty} \sup_{P\in\mathscr{P}_0} P(S_n) \leq \limsup_{n\to\infty} \sup_{P\in\mathscr{P}_0} P(S_n, \ \|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R_{X^iX^j|X^{\mathsf{S}}}(P)\| < \epsilon \ \forall (i,j,\mathsf{S}) \in \mathscr{T}).$$

However, the statement inside $P(\cdots)$ on the right-hand side consists of, when spelled out fully, the following statements:

$$\begin{aligned} &\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\| < \rho \ \ \forall (i,j,\mathsf{S}) \in \mathscr{D} \text{ and } \|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}}\| \geq \rho \ \ \forall (i,j,\mathsf{S}) \in \mathscr{D}^c \text{ are not both true}, \\ &\|R_{X^iX^j|X^{\mathsf{S}}}(P)\| = 0 \ \forall\ (i,j,\mathsf{S}) \in \mathscr{D}, \text{ and } \|R_{X^iX^j|X^{\mathsf{S}}}(P)\| > \lambda \ \forall\ (i,j,\mathsf{S}) \notin \mathscr{D} \text{ are both true}, \\ &\|\hat{R}^{\epsilon_n,\delta_n}_{X^iX^j|X^{\mathsf{S}}} - R_{X^iX^j|X^{\mathsf{S}}}(P)\| < \epsilon \ \forall (i,j,\mathsf{S}) \in \mathscr{T}, \end{aligned}$$

which cannot happen for sufficiently small ϵ. Hence

$$\lim_{n\to\infty} \sup_{P\in\mathscr{P}_0} P(\hat{\mathscr{C}}_A \neq \mathscr{D}) = \lim_{n\to\infty} \sup_{P\in\mathscr{P}_0} P(S_n) = 0.$$

This proves the first convergence in (28). The second convergence can be proved similarly. □

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

F. R. Bach. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research*, 9:1179–1225, 2008.

C. R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press. Baltimore and London, 1993.

D. M. Chickering. Optimal structure identification with greedy search. *The Journal of Machine Learning Research*, 3(3):507–554, 2002.

B. Ellis and W. Wong. Learning causal Bayesian network structures from experimental data. *Journal of the American Statistical Association*, (103):778–789, 2008.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–41, 2008.

K. Fukumizu, F. R. Bach, and M. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

K. Fukumizu, F. R. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *The Journal of Machine Learning Research*, 8:361–383, 2007.

K. Fukumizu, a. Gretton, X. Sun, and B. Schölkopf. Kernel Measures of Conditional Dependence. *Advances in Neural Information Processing Systems*, 20:489–496, 2008.

K. Fukumizu, F. R. Bach, and M. Jordan. Kernel dimension reduction in regression. *Annals of Statistics*, 37(5):1871–1905, 2009.

A. Gretton, K. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13:723–773, 2012.

N. Harris and M. Drton. Pc algorithm for nonparanormal graphical models. *The Journal of Machine Learning Research*, 14(1):3365–3383, 2013.

Y.-B. He and Z. Geng. Active Learning of Causal Networks with Intervention Experiments and Optimal Designs . *Journal of Machine Learning Research*, (9):2523–2547, 2008.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems 21 (NIPS)*, 2009.

T. Huang. Testing conditional independence using maximal nonlinear conditional correlation. *Ann. Statist.*, 38:2047–2091, 2010.

I. T. Jolliffe. *Principal component analysis*. Springer, 2nd edition, 2002.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *The Journal of Machine Learning Research*, 8:613–636, 2007.

S. L. Lauritzen. *Graphical models*. Oxford University Press, 1996.

K.-Y. Lee, B. Li, and F. Chiaromonte. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics*, 41(1):221–249, 2013.

K.-Y. Lee, B. Li, and H. Zhao. Variable selection via additive conditional independence. *Journal of the Royal Statistical Society. Series B.*, 78(5):1037–1055, 2016.

B. Li, H. Chun, and H. Zhao. Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association*, 107:152–167, 2012.

B. Li, H. Chun, and H. Zhao. On an additive semi-graphoid model for statistical networks with application to pathway analysis. *Journal of the American Statistical Association*, 109:1188–1204, 2014.

O. Linton and P. Gozalo. Conditional Independence Restrictions: Testing and Estimation. Cowles Foundation Discussion Papers 1140, Cowles Foundation for Research in Economics, Yale University, November 1996. URL `https://ideas.repec.org/p/cwl/cwldpp/1140.html`.

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10: 2295–2328, 2009.

H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

S. P. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

R. Luo and H. Zhao. Bayesian Hierarchical Modeling for Signaling Pathway Inference From Single Cell Interventional Data. *The Annals of Applied Statistics*, 5(2A):725–745, 2011.

D. Margaritis. Distribution-free learning of Bayesian network structure in continuous domains. *In Proc. AAAI*, pages 825–830, 2005.

J. Mooij, D. Janzing, J. Peters, and B. Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 745–752, 2009.

K. Muandet, K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf. Kernel mean estimation and Stein effect. *ICML*, pages 10–18, 2014.

K. Muandet, K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf. Kernel Mean Shrinkage Estimators. *Journal of Machine Learning Research*, 17:1–41, 2016.

J. Pearl. *Causality: models, reasoning and inference*. Cambridge Univ Press, 2nd ed., 2009.

J. Pearl and T. Verma. The logic of representing dependencies by directed graphs. *Proceedings of the Sixth National Conference on Artificial Intelligence*, 1:374–379, 1987.

J. Pearl, D. Geiger, and T. Verma. Conditional independence and its representations. *Kybernetika*, 25:33–44, 1989.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning research*, 15:2009–2053, 2014.

K. Sachs, O. Perez, D. Pe'er, D. a Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–9, 2005.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

S. Shimizu, P. Hoyer, A. nad Hyvärinen, and A. J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, (7):2003–2030, 2006.

K. Song. Testing conditional independence via Rosenblatt transforms. *The Annals of Statistics*, 37(6B):4011–4045, December 2009.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, 2000.

L. Su and H. White. A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141(2):807–834, December 2007.

L. Su and H. White. A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24(04):233–36, April 2008.

X. Sun, D. Janzing, B. Schölkopf, and K. Fukumizu. *A kernel-based causal learning algorithm*. ICML, New York, New York, USA, June 2007.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

R. Tillman, A. Gretton, and P. Spirtes. Nonlinear directed acyclic structure learning with weakly additive noise models. *Proceedings of Advances in Neural Processing Information Systems*, 22:1847–1855, 2009.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463, April 2013.

S. van de Geer and P. Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Ann. Statist.*, 41(2):536–567, 2013.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. *Proc. Sixth Conference on Uncertainty in AI. Association for Uncertainty in AI, Inc.*, pages 220–227, 1990.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pages 255–270, New York, NY, USA, 1991. Elsevier Science Inc. ISBN 0-444-89264-8. URL `http://dl.acm.org/citation.cfm?id=647233.719736`.

J. Weidmann. *Linear Operators in Hilbert Spaces*. Springer, New York, 1980.

L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.*, 40(5):2541–2571, 2012.

J. Zhang and P. Spirtes. Strong faithfulness and uniform consistency in causal inference. In *In Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, 2002.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. *In Proceedings of the 25th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 804–813, Arlington, Virginia, United States, 2011. AUAI Press. ISBN 978-0-9749039-7-2. URL `http://dl.acm.org/citation.cfm?id=3020548.3020641`.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.