Journal of Machine Learning Research 26 (2025) 1-46 Submitted 1/24; Revised 3/25; Published 4/25

Learning causal graphs via nonlinear sufficient dimension reduction

Eftychia Solea E.SOLEA@QMUL.AC.UK
School of Mathematical Sciences, Queen Mary University of London
Mile End, E1 4NS, London, UK

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

Kyongwon Kim KIMK@YONSEI.AC.KR
Department of Applied Statistics
Department of Statistics and Data Science
Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul, 03722, South Korea

Editor: Elias Bareinboim

Abstract

We introduce a new nonparametric methodology for estimating a directed acyclic graph (DAG) from observational data. Our method is nonparametric in nature: it does not impose any specific form on the joint distribution of the underlying DAG. Instead, it relies on a linear operator on reproducing kernel Hilbert spaces to evaluate conditional independence. However, a fully nonparametric approach would involve conditioning on a large number of random variables, subjecting it to the curse of dimensionality. To solve this problem, we apply nonlinear sufficient dimension reduction to reduce the number of variables before evaluating the conditional independence. We develop an estimator for the DAG, based on a linear operator that characterizes conditional independence, and establish the consistency and convergence rates of this estimator, as well as the uniform consistency of the estimated Markov equivalence class. We introduce a modified PC-algorithm to implement the estimating procedure efficiently such that the complexity depends on the sparseness of the underlying true DAG. We demonstrate the effectiveness of our methodology through simulations and a real data analysis.

Keywords: causality; conditional independence; directed graphs; pc algorithm; sufficient dimension reduction

1. Introduction

Inferring cause-effect relationships from observational data is fundamental in many scientific disciplines, such as epidemiology, neuroscience, genetics, sociology, and finance. See Spirtes et al. (2001). The underlying causal relationships among the data are often encoded in a directed acyclic graph (DAG). In this paper, we establish a new nonparametric framework for estimating the structure of a DAG from observational data.

Mathematically, let $X = (X^1, \ldots, X^p)$ be a p-dimensional random vector, and let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ be a DAG that consists of a set of vertices $\mathsf{V} = \{1, \ldots, p\}$, corresponding to the p random variables, $X^1, \ldots, X^p$ and a set of directed edges $\mathsf{E} \subset \{(i, j) \in \mathsf{V} \times \mathsf{V}, i \neq j\}$ such that there are no directed cycles. We denote every ordered pair $(i, j) \in \mathsf{E}$ by $i \to j$, which indicates that X^i has a direct

causal effect on X^j. We say that the joint distribution, P_X, of X is faithful with respect to G, if the conditional independence among the variables can be inferred from the structure of the DAG via the concept of d-separation in the graph G and vice-versa. See Section 2 which contains a brief review of the key concepts and definitions for DAGs that are relevant to this paper. In general, it is well known that there are typically a class of DAGs with respect to the joint distribution P_X is faithful. For this reason, we can only estimate the so-called Markov equivalence class of G, that is, the set of all DAGs with the same d-separation structure (Kalisch and Bühlman, 2007; Peters et al., 2014).

Many approaches have been proposed to estimate the Markov equivalence class of the true DAG. For example, in the Gaussian setting, Chickering (2003) introduced the greedy equivalence search (GES) which belongs to the category of the greedy score-based search approaches. Given a starting graph and a BIC-score, GES performs a greedy search on the space of Markov equivalence classes. van de Geer and Bühlmann (2013) proposed an ℓ_0-penalized likelihood approach for estimating the Markov equivalence class in the Gaussian setting. Another class of methods are the constraint-based methods like the PC-algorithm (Spirtes et al., 2001) that conducts a sequence of conditional independence tests to infer the Markov equivalence class. For example, Kalisch and Bühlman (2007) considered the PC-algorithm for estimating the equivalence class of a high-dimensional DAG under the Gaussian assumption, and proved its consistency in the high-dimensional sparse setting. According to Kalisch and Bühlman (2007), the PC-algorithm is computationally feasible and has the property that, if the underlying true DAG is sparse, its computational efficiency is determined by the level of sparseness rather than the size of the network.

Since the Gaussian assumption is restrictive in practice, many recent approaches have been developed to relax this assumption. A main methodological challenge is to relax the Gaussian assumption without resorting to high-dimensional kernels which can cause the curse of the dimensionality and reduce estimation accuracy. To this end, Harris and Drton (2013) propose the Gaussian copula DAG model using rank correlations. However, the Gaussian copula assumption can be violated under nonlinear interactions. To further relax the Gaussian copula assumption, one can resort to fully nonparametric approaches based on kernel-based conditional independence tests (Gretton et al., 2009; Sun, 2008; Zhang et al., 2011). However, such approaches involve the computation of high-dimensional kernels, and thus often suffer from the curse of dimensionality, a problem that is more severe for large networks. To strike a balance between model flexibility and dimensionality, Lee et al. (2020) proposed a fully nonparametric estimation approach based on a new statistical relation called additive conditional independence (ACI), originally proposed by Li et al. (2014). ACI is a three-way statistical relation that follows the spirit of conditional independence but its nonparametric characterization involves only one-dimensional kernels, thus avoiding the curse of dimensionality. See also Li and Solea (2018); Lee et al. (2016a).

1.1 Our proposal and contributions

In this paper, we propose an alternative nonparametric approach for estimating the Markov equivalence class with the PC-algorithm. Instead of relying on ACI to mitigate the curse of dimensionality, our proposal utilizes recent nonlinear sufficient dimension reduction (SDR) techniques (Lee et al., 2013) to construct the graph. In particular, the estimation procedure proceeds in two steps. First, we perform nonlinear sufficient dimension reduction on the high-dimensional conditioning vector to produce a low dimensional-conditioning vector. Second, we feed the low dimension-dimensional vector into a linear operator, called the *conjoined conditional cross-covariance*

operator (CCCO), to evaluate conditional independence. This operator is defined on reproducing kernel Hilbert spaces (RKHS) and does not rely on any parametric distribution assumption. The combined feature of dimension reduction and distribution-free makes our method attract in particularly attractive for handling high-dimensional, non-Gaussian causal graphs. To implement the methodology, we propose a modified PC-algorithm by integrating dimension reduction with the classical PC-algorithm such that the computational complexity does not depend on the size of the network but instead on its level of sparseness. We establish the consistency of the Markov equivalence class and the uniform consistency under a strong faithfulness condition (Uhler et al., 2013). We illustrate our methodology by simulation and a real data analysis.

1.2 Related work

Conditional independence testing is an important problem, especially in learning directed graphical models and causal discovery. Due to the curse of dimensionality, nonparametric testing for conditional independence of continuous variables is particularly challenging. Recently, some work has been proposed based on RKHS. For example, Fukumizu et al. (2004, 2007a) proposed the conditional cross-covariance operator to characterise conditional independence. Fukumizu et al. (2007b) introduced the normalized conditional cross-covariance operator that completely characterizes conditional independence, and has the additional advantage of removing the marginal variations from the covariance operators. Sun et al. (2007), Gretton et al. (2009) and Fukumizu et al. (2007b) proposed a causal learning algorithm that uses the Hilbert-Schmidt norm of the conditional cross-covariance operator to measure conditional dependence and combines the PC algorithm with a permutation test of independence. Zhang et al. (2011) developed the asymptotic distribution of the empirical cross-covariance operator (CCO) under the null hypothesis of conditional independence.

However, these kernel-based conditional independence tests rely on multi-dimensional kernels to characterize conditional independence, which can lead to substantial drop in accuracy known as curse of dimensionality. As a result, these methods often perform poorly in practice when dealing with large networks. Our solution addresses this limitation by applying nonlinear SDR before the evaluation of conditional independence. Nonlinear SDR generalizes linear SDR and seeks a set of nonlinear functions in the predictors that best predict the response. Lee et al. (2013) proposed the Generalized Sliced Inverse Regression (GSIR) and the Generalized Sliced Average Variance Estimator (GSAVE), which extend the Sliced Inverse Regression (Li, 1991) and the Sliced Average Variance Estimator (SAVE) (Cook and Weisberg, 1991), respectively. Both GSIR and GSAVE are not restricted to linear constraints, and can capture nonlinear relations that linear SDR methods can miss, thus achieving further dimension reduction. Employing of nonlinear sufficient dimension reduction before evaluating conditional independence to avoid curse of dimensionality, our method is significantly different from the above-mentioned fully nonparametric approaches. Corresponding to this difference in methodology, our asymptotic analysis is significantly different from the mentioned previous works as well. Specifically, we establish the consistency of CCCO taking into account the error introduced from the step for estimating the sufficient predictors, which requires a much more involved asymptotic analysis than in the previous works. We also establish the uniform consistency of the estimated DAG, not available in the previous works (Sun et al., 2007; Gretton et al., 2009; Fukumizu et al., 2007b; Zhang et al., 2011).

Our work is also substantially different from Lee et al. (2020), which employed additive conditional independence (ACI) as the criterion to construct the graph and proposed the normalised additive conditional covariance operator to characterize ACI. In contrast, we use conditional inde-

pendence, a more widely accepted criterion as the criterion to determine the DAG as required by Lee et al. (2020). Additionally, our method is not constrained by additive structures.

We should also mention that there is a class of hybrid algorithms that combine ideas from the constraint-based and the score-based approaches, typically by employing a greedy search over a restricted space determined through conditional independence tests. Examples of such algorithms include the max-min hill climbing (MMHC) algorithm (Tsamardinos et al., 2006) and the adaptively restricted GES (Nandy et al., 2018). Numerous works have also been proposed to identify the DAG under additional distributional assumptions. Examples include linear structural equation models (SEMs) with non-Gaussian noise, linear non-Gaussian acyclic models (LiNGAM, Shimizu et al. (2006)), nonlinear SEMs with additive noise (Peters et al., 2014; Bühlmann et al., 2014), and linear Gaussian SEMs with equal error variances (Peters and Bühlmann, 2014). Finally, these approaches have been further developed to accommodate latent variables (Richardson and Spirtes, 2002; Hoyer et al., 2008; Shimizu and Hyvärinen, 2007), time series data (Hyvärinen et al., 2010), interventional distributions (Hauser and Bühlmann, 2012; Tian and Pearl, 2001; Cooper and Yoo, 1999; Eaton and Murphy, 2007; He and Geng, 2008; Tong and Koller, 2001; Eberhardt, 2008; Peng et al., 2020), and directed cycles (Lacerda et al., 2008; Eberhardt et al., 2010) (for a detailed overview, see Shimizu (2014)). Our proposal significantly advances several key areas this line of research, including constraint-based learning methods for random variables, DAG estimation, and kernel-based conditional independence testing.

1.3 Organization of the paper

The rest of the paper is organized as follows. In Section 2, we briefly review the main concepts and theory of graphical models, and introduce some notations. In Section 3, we propose the DAG by incorporating sufficient dimension reduction with a DAG, and describe its estimation process at the population level. In Section 4, we develop the algorithm to implement the methodology. In Section 5, we study the asymptotic properties such as estimation consistency, uniform consistency and convergence rates. In Section 6, we demonstrate the performance of our methodology by simulation studies and a pathway analysis. We make some concluding remarks in Section 7, and put all proofs and some additional results in the Appendix.

2. Preliminaries and notation

In this section, we briefly review the theory of graphical models. A graph $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ consists of a set $\mathsf{V} = \{1, 2, \ldots, p\}$ of p vertices and a set of edges $\mathsf{E} \subseteq \{(i, j) \in \mathsf{V} \times \mathsf{V}, i \neq j\}$. A graph G is called a directed graph if for any $(i, j) \in \mathsf{V} \times \mathsf{V}$, with $i \neq j$, at most one of the ordered pairs (i, j) belongs to E. If $(i, j) \in \mathsf{E}$, then we write $i \to j$. An edge $(i, j) \in \mathsf{E}$ is called undirected if $(i, j) \in \mathsf{E}$ if and only if $(j, i) \in \mathsf{E}$, and is denoted by $i - j$. A graph G is called undirected if all its edges are undirected (Kalisch and Bühlman, 2007). In a graph $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ (directed or undirected), if there is an edge, (i, j) in E, then the nodes i and j are said to be adjacent. A path on G is a sequence of distinct vertices $i_1, \ldots, i_n$ such that successive nodes i_k and i_{k+1} are adjacent for all $k = 1, \ldots, n - 1$. If $i_k \to i_{k+1}$ for all k, then the path is called directed. A path of length at least three, for which the endpoint vertices i_1 and i_n are the same, but for which all other vertices are distinct from each other, is called a cycle. A DAG is a directed graph that contains no directed cycles.

Let $X = (X^1, \ldots, X^p)$ be a random vector associated with the DAG G such that each node $i \in \mathsf{V}$ corresponds to the random variable X^i. The joint distribution P_X of X is said to be *faithful*

with respect to G if for any $i, j \in \mathsf{V}$ with $i \neq j$ and any set $S \subseteq \mathsf{V} \setminus \{i, j\}$,

$$\text{node } i \text{ and node } j \text{ are d-separated by the set } S \text{ in } \mathsf{G} \Leftrightarrow X^i \perp\!\!\!\perp X^j \mid X^S,$$

where the notation $U \perp\!\!\!\perp V \mid W$ indicates that random variables U and V are conditionally independent given W, and d-separation means that every path in G between nodes i and j is blocked by S (see Koller and Friedman (2009)). The main reason for imposing this assumption is that non-faithful distributions form a set that has a Lebesgue measure of zero, as it can be viewed as a collection of hypersurfaces within a hypercube (Uhler et al., 2013). Examples of faithful distributions include the Gaussian distribution, the nonparanormal distribution (Liu et al., 2009) and distributions generated by linear structural equation models (SEM) with the additive errors drawn from a mixture of Cauchy distributions (see Harris and Drton (2013)). Sadeghi (2017) provided sufficient and necessary conditions for a distribution that factorizes with respect to a DAG to be faithful to a given graph. These are singleton-transitivity and ordered upward- and downward- stabilities (see Sadeghi (2017, Corollary 34)). The Gaussian distribution and binary distributions satisfy these properties. More generally, strictly positive densities also satisfy these conditions. We remark that examples of non-faithful distributions is given in Spirtes et al.(2000, Chapter 3.5.2), Peters et al. (2014) and Peters (2015).

The skeleton of a DAG, G, is the undirected graph constructed from G by dropping the orientation on the edges, and is denoted by $\mathrm{ske}(\mathsf{G})$. A v-structure in a DAG, G, is a triplet of nodes (i, j, k) such that $i \to j$ and $k \to j$, and i and k are not adjacent.

Two DAGs are called Markov equivalent if they possess the same d-separation relations and, hence, they cannot be distinguished from one another using conditional independence. The class of all DAGs sharing the same d-separation relations is called the Markov equivalence class of the true DAG, G. Markov equivalent DAGs are characterized by having the same skeleton and the same v-structures (Verma and Pearl, 2022; Drton and Maathuis, 2017; Kalisch and Bühlman, 2007).

It is common to visualize each Markov equivalence class by a completed partially directed acyclic graph (CPDAG) that may have directed and undirected edges (Andersson et al., 1997; Roverato, 2005). Every directed edge in a CPDAG, exists in all DAGs in the equivalence class of the true DAG, and if there exists a DAG with $i \to j$ and a DAG $j \to i$ in the equivalence class, then the undirected edge $i - j$ exists in the CPDAG. Hence, the goal is to estimate the Markov equivalence class of the true DAG, G, or equivalently, its CPDAG, from a simple random sample from X.

Assuming that P_X is faithful with respect to the true DAG, constraint-based methods like the PC algorithm (Spirtes et al., 2000) seek to estimate the CPDAG from data. The PC algorithm proceeds in two steps. The first step uses a sequence of conditional independence tests to estimate the skeleton of the true DAG. In particular, for any $i, j \in \mathsf{V}$, $i \neq j$,

$$i \text{ and } j \text{ are not connected in } \mathrm{ske}(\mathsf{G}) \;\Leftrightarrow\; X^i \perp\!\!\!\perp X^j \mid X^S \text{ for some } S \subset \mathsf{V} \setminus \{i, j\}. \tag{1}$$

The second step uses the results of the conditional independence tests to learn the partial orientation of the edges in the form of a CPDAG. For a compact description of the PC algorithm, see Drton and Maathuis (2017) and Kalisch and Buhlman (2007). In this process, the performance of the conditional independence test to infer $X^i \perp\!\!\!\perp X^j | X^S$ is of crucial importance, as Kalisch and Buhlman (2007) stated: "if this part is done correctly, the orientation of the edges in the CPDAG will be correct". The main point of the methodology proposed in this paper is to improve the performance of the conditional independence tests through dimension reduction.

We now introduce some notations that will be used throughout the article. For two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, let $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ denote the class of all bounded linear operators from $\mathcal{H}_1$ to $\mathcal{H}_2$. The space $\mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$ is a Banach space endowed with the operator norm which is denoted by $\|\cdot\|_{\text{op}}$. Let $\mathcal{B}_2(\mathcal{H}_1, \mathcal{H}_2)$ denote the class of all Hilbert-Schmidt operators from $\mathcal{H}_1$ to $\mathcal{H}_2$. The class $\mathcal{B}_2(\mathcal{H}_1, \mathcal{H}_2)$ is a Hilbert space with inner product $\langle\cdot,\cdot\rangle_{\text{HS}}$ and induced norm $\|\cdot\|_{\text{HS}}$. Moreover, it can be shown that $\mathcal{B}_2(\mathcal{H}_1, \mathcal{H}_2) \subset \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$. For convenience, when $\mathcal{H} = \mathcal{H}_1 = \mathcal{H}_2$, we use $\mathcal{B}(\mathcal{H})$ and $\mathcal{B}_2(\mathcal{H})$ to denote $\mathcal{B}(\mathcal{H}, \mathcal{H})$ and $\mathcal{B}_2(\mathcal{H}, \mathcal{H})$, respectively. For any operator $A \in \mathcal{B}(\mathcal{H}_1, \mathcal{H}_2)$, let A^* denote the adjoint operator of A, $\ker(A)$ the kernel space of A, $\text{ran}(A)$ the range of A, and $\overline{\text{ran}}(A)$ the closure of the range of A. For any operator $A \in \mathcal{B}(\mathcal{H})$, $\overline{\text{ran}}(A) = \ker(A^*)^\perp$.

3. Methodology

3.1 Nonlinear sufficient dimension reduction and directed acyclic graphs

For a univariate response Y, classical linear SDR seeks a $p \times d$ matrix β such that $Y \perp\!\!\!\perp X \,|\, \beta^\intercal X$, where $d \leq p$. The space spanned by the column vectors $\beta_1, \ldots, \beta_d$ of β is called the dimension reduction subspace. The greatest dimension reduction of the predictor vector is achieved by the smallest dimension reduction subspace, called the central subspace, denoted by $\mathcal{S}_{Y|X}$. See Li (1991); Cook and Weisberg (1991); Li (2018b). A commonly used method is the sliced inverse regression (SIR, Li (1991)). Linear SDR was generalized to the nonlinear case by Lee et al. (2013) and Li et al. (2011), which considered $Y \perp\!\!\!\perp X \,|\, \mathcal{G}$, where $\mathcal{G}$ is a sub-field of the σ-field, $\sigma(X)$, generated by X. The sub σ-field $\mathcal{G}$ is called the SDR σ-field for Y versus X. The intersection of all sub σ-field is called the central σ-field, and is denoted by $\mathcal{G}_{Y|X}$. By making analogy with classical SDR, the authors proposed the GSIR and the GSAVE, as extension to the SIR and the SAVE, respectively. This general framework was built upon several earlier kernel dimension reduction methods, such as the KCCA (Fukumizu et al., 2007a; Wu, 2008) and the kernel sliced inverse regression (Wang, 2008).

Following Kim (2022) and Li and Kim (2024), our idea is to connect nonlinear SDR with the DAG. For any $W \subseteq \mathsf{V}$, let $X^W = \{X^k : k \in W\}$ and $\sigma(X^W)$ be the σ-field generated by X^W.

Assumption 1 *For each $i, j \in \mathsf{V}$, $i \neq j$, and $S \subseteq \mathsf{V} \setminus \{i, j\}$, there is a sub σ-field $\mathcal{G}^S$ of $\sigma(X^S)$ such that*

$$(X^i, X^j) \perp\!\!\!\perp X^S \,|\, \mathcal{G}^S. \tag{2}$$

The smallest sub σ-field $\mathcal{G}^S$ of $\sigma(X^S)$ that satisfies (2) is called the central sub σ-field for X^S versus (X^i, X^j), and is denoted by $\mathcal{G}_{(X^i X^j)|X^S}$. Examples of joint distributions that satisfy Assumption 1 are given in Li and Kim (2024, Appendinx J), and include the Gaussian distribution, the nonparanormal distribution (Liu et al., 2009), and a multivariate distribution determined by a type of a regular graph in which each pair of vertices can have at most d neighbors. The following theorem is found in Li and Kim (2024, Theorem 1).

Theorem 1 *If Assumption 1 holds, then*

$$X^i \perp\!\!\!\perp X^j \,|\, X^S \;\Leftrightarrow\; X^i \perp\!\!\!\perp X^j \,|\, \mathcal{G}_{(X^i, X^j)|X^S}.$$

This theorem suggests that we can combine nonlinear SDR with d-separation to estimate the Markov equivalence class.

Definition 2 *Suppose that Assumption 1 holds. We say that the joint distribution P_X of the random vector X is faithful with respect to a DAG G if for any $i, j \in \mathsf{V}$ with $i \neq j$ and any set $S \subset \mathsf{V} \setminus \{i, j\}$,*

$$\textit{node } i \textit{ and node } j \textit{ are } d\textit{-separated by the set } S \textit{ in } \mathsf{G} \Leftrightarrow X^i \perp\!\!\!\perp X^j \,|\, \mathcal{G}_{(X^i, X^j)|X^S}. \tag{3}$$

Under faithfulness, we use $X^i \perp\!\!\!\perp X^j \,|\, \mathcal{G}_{(X^i X^j)|X^S}$ as a criterion to learn the skeleton of the DAG, G, after performing nonlinear SDR of (X^i, X^j) versus X^S for each $i, j \in \mathsf{V}, i \neq j$ and $S \subset V \setminus \{i, j\}$. Specifically, the skeleton of the DAG, denoted by $\mathrm{ske}(\mathsf{G})$, is defined by

$$\mathrm{ske}(\mathsf{G}) = \{(i, j) \in \mathsf{V} \times \mathsf{V} : i \neq j,\ X^i \perp\!\!\!\perp X^j \,|\, \mathcal{G}_{(X^i X^j)|X^S} \text{ for all } S \subset \mathsf{V} \setminus \{i, j\}\}.$$

3.2 Reproducing kernel Hilbert spaces and covariance operators

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $(\Omega_X, \mathcal{F}_X)$ a measurable space, and $X : \Omega \mapsto \Omega_X$ a p-dimensional random vector with distribution P_X. Suppose $\Omega_X = \Omega_{X^1} \times \cdots \times \Omega_{X^p}$, where Ω_{X^i} is the support of the ith component, X^i, of X. For each $i \in \mathsf{V}$, let $\mathcal{H}_{X^i}$ be an RKHS of functions on Ω_{X^i} to $\mathbb{R}$ defined by a positive definite kernel $\kappa_{X^i} : \Omega_{X^i} \times \Omega_{X^i} \mapsto \mathbb{R}$. Let $L_2(P_{X^i})$ be the class of all square-integrable functions in Ω_{X^i} such that $\int f^2 dP_{X^i} < \infty$, where P_{X^i} is the distribution of X^i. The following integrability assumption ensures that $\mathcal{H}_{X^i} \subset L_2(P_{X^i})$ (see Fukumizu et al. (2007b)).
Assumption 2 $\mathbb{E}(\kappa_{X^i}(X^i, X^i)) < \infty$, $i = 1, \ldots, p$.

This is a mild assumption that guarantees the square integrability of the members in $\mathcal{H}_{X^i}$, and is satisfied by bounded kernels such as the Gaussian radial basis function (RBF)

$$\kappa_{X^i} : \mathcal{H}_{X^i} \times \mathcal{H}_{X^i} \mapsto \mathbb{R}, \quad (x_1^i, x_2^i) \mapsto \exp(-\gamma_{X^i}(x_1^i - x_2^i)^2), \tag{4}$$

where $\gamma_{X^i} > 0$, and the Laplacian kernel $\kappa_{X^i}(x_1^i, x_2^i) = \exp(-\gamma_{X^i}|x_1^i - x_2^i|)$. See, for example, Fukumizu et al. (2009). Under Assumption 2, the linear functional $\mathcal{H}_{X^i} \ni f \mapsto \mathbb{E}(f(X^i))$ is bounded. Hence, by Riesz's representation theorem, there exists a unique function μ_{X^i} in $\mathcal{H}_{X^i}$ such that $\mathbb{E}(f(X^i)) = \langle f, \mu_{X^i} \rangle, \forall f \in \mathcal{H}_{X^i}$. Also under Assumption 2, for any $i, j \in \mathsf{V}$, the bilinear form defined by

$$\mathcal{H}_{X^i} \times \mathcal{H}_{X^j} \to \mathbb{R}, \quad (f, g) \mapsto \mathrm{cov}(f(X^i), g(X^j))$$

is bounded, which implies that there is a unique linear operator $\Sigma_{X^i X^j} \in \mathcal{B}(\mathcal{H}_{X^i}, \mathcal{H}_{X^j})$ such that, for all $f \in \mathcal{H}_{X^i}$, $g \in \mathcal{H}_{X^j}$,

$$\langle f, \Sigma_{X^i X^j} g \rangle_{\mathcal{H}_{X^i}} = \mathrm{cov}(f(X^i), g(X^j)).$$

See, for example, Conway (2019, Chapter II, Theorem 2.2). This operator is called the cross-covariance operator between X^i and X^j (Fukumizu et al., 2007b), and is a Hilbert-Schmidt operator under Assumption 1 (Gretton et al. (2005a, Lemma 1)). This operator extends the covariance matrix on Euclidean spaces to nonlinear spaces. Obviously, $\Sigma^*_{X^i X^j} = \Sigma_{X^j X^i}$. If $X^i = X^j$, $\Sigma_{X^i X^i}$ is called the covariance operator, which is self-adjoint and positive definite Hilbert-Schmidt operator (Fukumizu et al., 2007b).

Note that any f belongs to $\ker(\Sigma_{X^iX^i})$ if and only if $\mathrm{var}(f(X^i)) = 0$. Hence $\ker(\Sigma_{X^iX^i})$ consists of functions in $\mathcal{H}_{X^i}$ that are constant almost surely. Since constants are unimportant in conditional independence, we can, without loss of generality, assume $\ker(\Sigma_{X^iX^i}) = \{0\}$ for all $i = 1, \ldots, p$. Since $\Sigma_{X^iX^i}$ is self-adjoint, $\overline{\mathrm{ran}}(\Sigma_{X^iX^i}) = \ker(\Sigma_{X^iX^i})^\perp$. Hence, we reset $\mathcal{H}_{X^i}$ to be $\overline{\mathrm{ran}}(\Sigma_{X^iX^i})$, and we refer to it as the centered RKHS. Li and Song (2017, Lemma 1) derived an explicit expression for the centered RKHS, $\mathcal{H}_{X^i}$ as

$$\mathcal{H}_{X^i} = \overline{\mathrm{span}}\{\kappa_{X^i}(\cdot, x^i) - \mu_{X^i} : x^i \in \Omega_{X^i}\}.$$

For the rest of the paper, we work with the centered RKHS, $\mathcal{H}_{X^i}$. Next, we define the inverse of an operator. Because $\ker(\Sigma_{X^iX^i}) = \{0\}$, the covariance operator $\Sigma_{X^iX^i}$ defined on $\mathcal{H}_{X^i} = \overline{\mathrm{ran}}(\Sigma_{X^iX^i})$ is an injective function. We use $\Sigma^{-1}_{X^iX^i}$ to denote the inverse of $\Sigma_{X^iX^i}$. Note that $\Sigma^{-1}_{X^iX^i}$ is an unbounded operator since $\Sigma_{X^iX^i}$ is a Hilbert-Schmidt operator.

3.3 Generalized sliced inverse regression and regression operator

We use the GSIR for nonlinear sufficient dimension reduction. For each $S \subseteq \mathsf{V}$, let Ω_{X^S} be the range of X^S, which is the cartesian product of Ω_{X^k} for $k \in S$. Let $\kappa_{X^S} : \Omega_{X^S} \times \Omega_{X^S} \mapsto \mathbb{R}$ be a positive definite kernel. Let $\mathcal{H}_{X^S} = \overline{\mathrm{span}}\{\kappa_{X^S}(\cdot, x^S) - \mu_{X^S} : x^S \in \Omega_{X^S}\}$ be the centered RKHS. Similarly, for each $i, j \in \mathsf{V}$, let $\Omega_{X^iX^j} = \Omega_{X^i} \times \Omega_{X^j}$ be the range of (X^i, X^j), and let $\mathcal{H}_{X^iX^j}$ be the RKHS generated by the kernel $\kappa_{X^iX^j} : \Omega_{X^iX^j} \times \Omega_{X^iX^j} \mapsto \mathbb{R}$. We make the following assumption.

Assumption 3 *For any $i, j \in \mathsf{V}$, $i \neq j$, $S \subseteq \mathsf{V} \setminus \{i, j\}$, we have*

1. $\mathbb{E}(\kappa_{X^S}(X^S, X^S)) < \infty$, *and* $\mathbb{E}(\kappa_{X^iX^j}((X^i, X^j), (X^i, X^j))) < \infty$.
2. $\mathrm{ran}(\Sigma_{X^S(X^iX^j)}) \subseteq \mathrm{ran}(\Sigma_{X^SX^S})$.

Part 1 of Assumption 3 ensures the covariance operators $\Sigma_{X^S(X^iX^j)} \in \mathcal{B}(\mathcal{H}_{X^iX^j}, \mathcal{H}_{X^S})$ and $\Sigma_{X^SX^S} \in \mathcal{B}(\mathcal{H}_{X^S})$ are well-defined. Part 2 assumes a type of collective smoothness assumption in the relation between (X^i, X^j) and X^S. It requires that the operator $\Sigma_{X^S(X^iX^j)}$ sends any incoming function $f \in \mathcal{H}_{X^iX^j}$ to the eigenspaces of $\Sigma_{X^SX^S}$ corresponding to the leading eigenvalues, or to the low-frequency components of $\Sigma_{X^SX^S}$. This assumption is satisfied when the range of the operator $\Sigma_{X^S(X^iX^j)}$ is a finite-dimensional reducing subspace of $\Sigma_{X^SX^S}$. This is true, for example, when the polynomial kernel of finite order is used (Lee et al., 2016a,b). Otherwise, for kernels inducing infinite-dimensional spaces, it holds if the dependency between X^i and X^j given X^S has to be sufficiently concentrated on the leading eigenfunctions of $\Sigma_{X^SX^S}$. In practice, this assumption is satisfied since $\mathcal{H}_{X^S}$ can be approximated by the spanning space of a few leading eigenfunctions of $\Sigma_{X^SX^S}$. Under Assumption 3, the operator

$$B_{X^S(X^iX^j)} = \Sigma^{-1}_{X^SX^S}\Sigma_{X^S(X^iX^j)}$$

is well-defined, and we call it the regression operator from $\mathcal{H}_{X^iX^j}$ to $\mathcal{H}_{X^S}$, since it resembles the regression coefficient matrix in multivariate linear regression. The regression operator plays an important role in nonlinear SDR.

In order for the class of functions, $\mathcal{H}_{X^S}$, to characterize the central σ-field, $\mathcal{G}_{(X^i,X^j)|X^S}$, we need to make the following assumption: that projections onto $L_2(P_{X^S})$-the L_2-space with respect to the distribution P_{X^S} of X^S-can be well approximated by elements in the RKHS, $\mathcal{H}_{X^S}$.

Assumption 4 *For any $S \subseteq \mathsf{V}$, $\mathcal{H}_{X^S}$ is a dense subset of $L_2(P_{X^S})$ modulo constants; that is for any $f \in L_2(P_{X^S})$ and any $\epsilon > 0$, there exists a $g \in \mathcal{H}_{X^S}$ such that* $\mathrm{var}[f(X^S) - g(X^S)] < \epsilon$.

Assumption 4 ensures the kernel function, κ_{X^S}, to be sufficiently rich so that it is a characteristic kernel with respect to $L_2(P_{X^S})$. This means that projections onto $L_2(P_{X^S})$ can be well approximated by elements in the RKHS, $\mathcal{H}_{X^S}$. This assumption is satisfied by the Gaussian RBF kernel, but not by the polynomial kernel. For a compact metric space, an RKHS constructed by a universal kernel (Steinwart, 2002) is characteristic. See Sriperumbudur et al. (2011) and Fukumizu et al. (2007b, Lemma 1). Under Assumption 4, the class of functions in $\mathcal{H}_{X^S}$ that are $\mathcal{G}_{(X^iX^j)|X^S}$-measurable form the central class for X^S versus (X^i, X^j), and is denoted by $\mathfrak{S}_{(X^iX^j)|X^S}$. The central class is the target of estimation in nonlinear SDR.

Let $L_2(P_{X^S})$ be the L_2-space with respect to the distribution P_{X^S} of X^S. Similar to Lee et al. (2013), we assume that the central class, $\mathfrak{S}_{(X^iX^j)|X^S}$, is a complete class which means that for any $\mathcal{G}_{(X^iX^j)|X^S}$-measurable $f \in L_2(P_{X^S})$ such that $\mathbb{E}(f(X^S)|(X^i, X^j)) = 0$ almost surely, $f(X^S) = 0$ almost surely.

Assumption 5 *The central class $\mathfrak{S}_{(X^iX^j)|X^S}$ is a complete dimension reduction class of (X^i, X^j) versus X^S.*

Assumption 5 is a mild condition which is satisfied by most nonparametric models with additive error (See Lee et al. (2013, Propositions 1 and 2)). As shown in Lee et al. (2013), under the Assumptions 2 through 5, the closure of the range of the regression operator is equal to the central class

$$\overline{\mathrm{ran}}(B_{X^S(X^iX^j)}) = \mathfrak{S}_{(X^iX^j)|X^S}. \tag{5}$$

In most of the practical nonparametric regression problems, the response variable Y depends on the predictor X only through a few functions of X. For example, in the commonly-used mean regression model

$$Y = f(X) + \epsilon$$

where $X \perp\!\!\!\perp \epsilon$, the central σ-field is generated by one function $f(X)$. In the mean regression with heteroscedasticity:

$$Y = f(X) + g(X)\epsilon,$$

the central σ-field is generated by 2 functions. Based on this consideration, it is reasonable to assume that the central σ-field is generated by a finite number of functions, which amounts to assuming the rank of the regression operator is finite. The next assumption formalizes this observation. This assumption was also made in Li and Song (2017) and Li and Kim (2024). In most of the practical nonparametric regression problems, the response variable Y depends on the predictor X only through a few functions of X. For example, in the commonly-used mean regression model

$$Y = f(X) + \epsilon$$

where $X \perp\!\!\!\perp \epsilon$, the central σ-field is generated by one function $f(X)$. In the mean regression with heteroscedasticity:

$$Y = f(X) + g(X)\epsilon,$$

the central σ-field is generated by 2 functions. Based on this consideration, it is reasonable to assume that the central σ-field is generated by a finite number of functions, which amounts to assuming the rank of the regression operator is finite. The next assumption formalizes this observation. This assumption was also made in Li and Song (2017) and Li and Kim (2024).

Assumption 6 *$B_{X^S(X^iX^j)}$ is a finite-rank operator with rank d_S^{ij}.*

This assumption allows to establish the estimation procedure and recover the range of $B_{X^S(X^iX^j)}$ via finite d_S^{ij} steps of optimization. It means that for any $f \in \mathcal{H}_{X^iX^j}$, $\Sigma_{X^S(X^iX^j)}f$ is relatively smooth. It would be violated, for example, if one can find $f \in \mathcal{H}_{X^iX^j}$ such that $\Sigma_{X^S(X^iX^j)}f$ is arbitrarily choppy (Li and Kim, 2024). Moreover, in the simulation settings, we approximate each $\mathcal{H}_{X^i}$ through a few leading eigenfunctions, and hence Assumption 6 holds since all operators involved are finite-rank operators.

Relation (5) suggests that we use $\overline{\mathrm{ran}}(B_{X^S(X^iX^j)})$ to estimate the $\mathfrak{S}_{(X^iX^j)|X^S}$. Since $\overline{\mathrm{ran}}(B_{X^S(X^iX^j)}) = \overline{\mathrm{ran}}(\Sigma_{X^SX^S}^{-1}\Sigma_{X^S(X^iX^j)}A\Sigma_{(X^iX^j)X^S}\Sigma_{X^SX^S}^{-1})$ for any nonsingular and self-adjoint operator A in $\mathcal{B}(H_{X^iX^j}, \mathcal{H}_{X^iX^j})$, we can use $\Sigma_{X^SX^S}^{-1}\Sigma_{X^S(X^iX^j)}A\Sigma_{(X^iX^j)X^S}\Sigma_{X^SX^S}^{-1}$ to recover the central class. Following Li and Kim (2024), we choose $A = \Sigma_{(X^iX^j)(X^iX^j)}^{-1}$ since it achieves better scaling. Thus, we have

$$\overline{\mathrm{ran}}(\Sigma_{X^SX^S}^{-1}\Sigma_{X^S(X^iX^j)}\Sigma_{(X^iX^j)(X^iX^j)}^{-1}\Sigma_{(X^iX^j)X^S}\Sigma_{X^SX^S}^{-1}) = \mathfrak{S}_{(X^iX^j)|X^S} \tag{6}$$

The space in (6) can be recovered by performing a generalized eigenvalue problem of $\Sigma_{X^{(i,j)}X^S}\Sigma_{X^SX^S}^{-1}\Sigma_{X^SX^{(i,j)}}$ with respect to $\Sigma_{X^SX^S}$, which can be restated in terms of the sequential maximization problem given below in (8). We first need the following assumption.

Assumption 7 *For each $i, j \in \mathsf{V}$, $i \neq j$ and $S \subset \mathsf{V} \setminus \{i, j\}$,*

1. $\mathrm{ran}(\Sigma_{X^SX^{(i,j)}}) \subseteq \mathrm{ran}(\Sigma_{X^SX^S})$, $\mathrm{ran}(\Sigma_{X^{(i,j)}X^S}) \subseteq \mathrm{ran}(\Sigma_{X^{(i,j)}X^{(i,j)}})$;
2. *$\Sigma_{X^SX^{(i,j)}}$ is a finite rank-operator with rank d_S^{ij};*
3. *all the nonzero eigenvalues of $\Sigma_{X^{(i,j)}X^S}\Sigma_{X^SX^S}^{-1}\Sigma_{X^SX^{(i,j)}}$ are distinct.*

The operator $\Sigma_{X^SX^S}^{-1}\Sigma_{X^S(X^iX^j)}\Sigma_{(X^iX^j)(X^iX^j)}^{-1}\Sigma_{(X^iX^j)X^S}\Sigma_{X^SX^S}^{-1}$ exists by part 1 of Assumption 7. Part 2 ensures that it is finite-rank. Part 3 assumes that its eigenvalues satisfy $\lambda_1^{ij} > \cdots > \lambda_{d_S^{ij}}^{ij}$. This is to guarantee a unique solution for the d_S^{ij} iterations and to simplify the asymptotic analysis. Under Assumption 7, a basis $\{f_1^{ij}, \ldots, f_{d_S^{ij}}^{ij}\}$ of the central class, $\mathfrak{S}_{(X^iX^j)|X^S}$, can be found by solving the following iterative maximization problem for each $i, j \in \mathsf{V}$ and $S \subset \mathsf{V} \setminus \{i, j\}$. For each $k = 1, \ldots, d_S^{ij}$

$$\begin{aligned} &\text{maximize} \quad \langle f, \Sigma_{X^S(X^iX^j)}\Sigma_{(X^iX^j)(X^iX^j)}^{-1}\Sigma_{(X^iX^j)X^S}f\rangle \\ &\text{subject to} \quad f \in \mathcal{H}_{X^S}, \langle f, \Sigma_{X^SX^S}f\rangle = 1, \quad \langle f, \Sigma_{X^SX^S}f_1\rangle = \cdots = \langle f, \Sigma_{X^SX^S}f_{k-1}\rangle = 0, \end{aligned} \tag{7}$$

This means, we first solve the standard eigenvalue problem. For each $k = 1, \ldots, d_S^{ij}$

$$\begin{aligned} &\text{maximize} \quad \langle g, \Sigma_{X^S X^S}^{-1/2} \Sigma_{X^S (X^i X^j)} \Sigma_{(X^i X^j)(X^i X^j)}^{-1} \Sigma_{(X^i X^j) X^S} \Sigma_{X^S X^S}^{-1/2} g \rangle \\ &\text{subject to} \quad g \in \mathcal{H}_{X^S}, \langle g, g \rangle = 1, \quad \langle g, g_1 \rangle = \cdots = \langle g, g_{k-1} \rangle = 0, \end{aligned} \tag{8}$$

The operator $\Sigma_{X^S X^S}^{-1/2} \Sigma_{X^S (X^i X^j)} \Sigma_{(X^i X^j)(X^i X^j)}^{-1} \Sigma_{(X^i X^j) X^S} \Sigma_{X^S X^S}^{-1/2}$ in (8) is nonnegative definite, self-adjoint and compact. By Assumption 7, it is finite-rank and its eigenvalues satisfy $\lambda_1^{ij} > \cdots > \lambda_{d_S^{ij}}^{ij}$. Thus, by the Theorem 4.2.5 in Hsing and Eubank (2015), the maximum in (8) is achieved at its kth eigenfunction g_k^{ij} corresponding to the kth largest eigenvalue. Furthermore, since $\mathfrak{S}_{(X^i X^j)|X^S}$ is a complete class, the transformed functions $f_k^{ij} = \Sigma_{X^S X^S}^{-1/2} g_k^{ij}$, $k = 1, \ldots, d_S^{ij}$ guarantees the central σ-field.

Any sample estimator targeting the central class, $\mathfrak{S}_{(X^i X^j)|X^S}$, is referred to as GSIR. To summarize, our goal consists of two steps: first, to use GSIR to estimate the central class $\mathfrak{S}_{(X^i X^j)|X^S}$ for any $i, j \in \mathsf{V}$, $S \subset \mathsf{V} \setminus \{i, j\}$, and second, to estimate the skeleton of the true DAG based on the reduced data $\{(X^i, X^j, \mathfrak{S}_{(X^i X^j)|X^S}) : i, j \in \mathsf{V}, S \subset \mathsf{V} \setminus \{i, j\}\}$. In particular, let $f_1^{ij}(X^S), \ldots, f_{d_S^{ij}}^{ij}(X^S)$ be the d_S^{ij} basis functions of $\mathfrak{S}_{(X^i X^j)|X^S}$, where $d_S^{ij} < |S|$ and $|S|$ denotes the cardinality of any set $S \subseteq \mathsf{V}$. We aim to evaluate

$$\begin{aligned} \text{node } i \text{ and node } j \text{ are not connected in } \mathrm{ske}(\mathsf{G}) \;\Leftrightarrow\; X^i \perp\!\!\!\perp X^j \,|\, f_1^{ij}(X^S), \ldots, f_{d_S^{ij}}^{ij}(X^S) \\ \text{for some } S \subseteq \mathsf{V} \setminus \{i, j\}. \end{aligned}$$

3.4 Conjoined conditional cross-covariance operator

In this section, we introduce the conjoined conditional cross-covariance operator (CCCO) to characterize $X^i \perp\!\!\!\perp X^j \,|\, f_1^{ij}(X^S) \ldots, f_{d_S^{ij}}^{ij}(X^S)$, and hence the skeleton of the DAG. This operator was originally proposed by Fukumizu et al. (2007b) for kernel-based conditional independence testing and was also used in Li and Kim (2024) for the construction of undirected graphical models. For each $i = 1, \ldots, p$, let $\Omega_{X^i X^S} = \Omega_{X^i} \times \Omega_{X^S}$, and let $\mathcal{H}_{X^i X^S}$ be the centered RKHS generated by the kernel $\kappa_{X^i X^S} : \Omega_{X^i X^S} \times \Omega_{X^i X^S} \mapsto \mathbb{R}$. We need the following assumption.

Assumption 8 *For any $i, j \in \mathsf{V}$, $i \neq j$, $S \subseteq \mathsf{V} \setminus \{i, j\}$, we assume*

1. $\mathbb{E}(\kappa_{X^S}(X^S, X^S)) < \infty$, $\quad \mathbb{E}(\kappa_{X^i X^S}((X^i, X^S), (X^i, X^S))) < \infty$, *and* $\mathbb{E}(\kappa_{X^j X^S}((X^j, X^S), (X^j, X^S))) < \infty$.
2. $\mathrm{ran}(\Sigma_{X^S (X^i X^S)}) \subseteq \mathrm{ran}(\Sigma_{X^S X^S})$ *and* $\mathrm{ran}(\Sigma_{X^S (X^j X^S)}) \subseteq \mathrm{ran}(\Sigma_{X^S X^S})$.

Assumption 8 ensures that the covariance operators $\Sigma_{X^S (X^i X^S)}$, $\Sigma_{X^S (X^j X^S)}$, $\Sigma_{(X^i X^S)(X^j X^S)}$, $\Sigma_{X^S X^S}^{-1} \Sigma_{X^S (X^i X^S)}$ and $\Sigma_{X^S X^S}^{-1} \Sigma_{X^S (X^j X^S)}$ are well-defined. Part 2 of Assumption 8 has similar interpretation as part 2 of Assumption 3. Under Assumption 8, the CCCO exists. The following definition is due to Fukumizu et al. (2007b).

Definition 3 *Suppose Assumption 8 holds. Then, the following operator*

$$\Sigma_{(X^i X^S)(X^j X^S)|X^S} = \Sigma_{(X^i X^S)(X^j X^S)} - \Sigma_{(X^i X^S) X^S} \Sigma_{X^S X^S}^{-1} \Sigma_{X^S (X^j X^S)} \tag{9}$$

is well-defined and is called the conjoined conditional cross-covariance operator (CCCO) of (X^i, X^j) given X^S.

The Definition 3 is needed to state the following proposition which is due to Fukumizu et al. (2007b), a proof of a special case of which is given in Fukumizu et al. (2004). For a random vector X, a RKHS, $\mathscr{H}_X$ based on a kernel κ_X is probability determining if and only if the mapping $P_X \mapsto E_{P_X}[\kappa_X(\cdot, X)]$ is injective. For example, the Gaussian RBF is probability determining, but the polynomial kernel is not.

Proposition 4 *Suppose that for any $i, j \in \mathsf{V}$ and $i \neq j$, any $S \subseteq \mathsf{V} \setminus \{i, j\}$*

1. *$\mathcal{H}_{X^i X^S} \otimes \mathcal{H}_{X^j X^S}$ is probability determining.*
2. *For each $f \in \mathcal{H}_{X^i X^S}$, $\mathbb{E}[f(X^i, X^S)|X^S = \cdot]$ belongs to $\mathcal{H}_{X^S}$.*
3. *For each $g \in \mathcal{H}_{X^j X^S}$, $\mathbb{E}[g(X^j, X^S)|X^S = \cdot]$ belongs to $\mathcal{H}_{X^S}$.*

where $\otimes$ is a tensor product. Then,

$$\Sigma_{(X^i X^S)(X^j X^S)|X^S} = 0 \Leftrightarrow X^i \perp\!\!\!\perp X^j \,|\, X^S.$$

Let $U^{ij,S} = (f_1^{ij}(X^S), \ldots, f_{d_S^{ij}}^{ij}(X^S))$ denote the population-level output from the nonlinear SDR step, obtained by solving the maximization problem (8). Let $\kappa_{X^i U^{ij,S}} : \Omega_{X^i U^{ij,S}} \times \Omega_{X^i U^{ij,S}} \mapsto \mathbb{R}$ be a positive definite kernel, where $\Omega_{X^i U^{ij,S}} = \Omega_{X^i} \times \Omega_{U^{ij,S}}$. Let $\mathcal{H}_{X^i U^{ij,S}}$ be the centered RKHS generated by $\kappa_{X^i U^{ij,S}}$. Similarly, let $\mathcal{H}_{U^{ij,S}}$ be the centered RKHS generated by the positive definite kernel $\kappa_{U^{ij,S}} : \Omega_{U^{ij,S}} \times \Omega_{U^{ij,S}} \mapsto \mathbb{R}$. We apply the CCCO to $(X^i, X^j, U^{ij,S})$ to determine whether $X^i \perp\!\!\!\perp X^j \,|\, U^{ij,S}$ is true for all $i, j \in \mathsf{V}$, $i \neq j$ and $S \subset \mathsf{V} \setminus \{i, j\}$. In order for the CCCO to be defined on $(X^i, X^j, U^{ij,S})$, we need the following assumption.

Assumption 9 *Assumption 8 and conditions 1 through 3 in Proposition 4 are satisfied with X^S replaced by $U^{ij,S}$ for all $i, j \in \mathsf{V}$, $i \neq j$ and $S \subset V \setminus \{i, j\}$.*

Under Assumption 9, the CCCO

$$\Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}} = \Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})} - \Sigma_{(X^i U^{ij,S})U^{ij,S}} \Sigma^{-1}_{U^{ij,S} U^{ij,S}} \Sigma_{X^i U^{ij,S}(X^j U^{ij,S})} \tag{10}$$

is well-defined and satisfies $\Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}} = 0 \Leftrightarrow X^i \perp\!\!\!\perp X^j \,|\, U^{ij,S}$. This result allows us to use $\Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}}$ to measure conditional independence and characterize the skeleton of the DAG, G.

Corollary 5 *Suppose Assumption 1 holds and X is faithful to a DAG, G. Then, under Assumption 9, for all $i, j \in \mathsf{V}$, $i \neq j$*

$$\textit{nodes } i \textit{ and } j \textit{ are not connected in } \mathrm{ske}(\mathsf{G}) \;\Leftrightarrow\; \Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}} = 0, \quad \textit{for some } S \subset \mathsf{V} \setminus \{i, j\}. \tag{11}$$

Hence, for a given DAG, G,

$$\mathrm{ske}(\mathsf{G}) = \{(i, j) \in \mathsf{V} \times \mathsf{V} : i \neq j, \|\Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}} \neq 0 \quad \text{for all} \quad S \subset \mathsf{V} \setminus \{i, j\}\}.$$

The above result allows to use $\Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}}$ as a criterion for conditional independence that is a based on a low-dimensional vector $U^{ij,S}$ obtained from the first SDR step. Furthermore, the equivalence (11) does not require any distributional assumptions, and avoids the computation of high-dimensional kernels for the estimation of the skeleton. It thus avoids the curse of dimensionality and is able to handle large-scale networks.

4. Estimation

In this section, we develop the algorithms to implement the methodology. We first develop algorithms for the sample versions of GSIR and CCCO, followed by a cross-validation scheme to select tuning parameters in the GSIR and CCCO algorithms and a strategy to accelerate computation. Then we develop a new version of the PC-algorithm to estimate the DAG that incorporates dimension reduction. To formulate the algorithms we need to represent the operators as matrices using a coordinate representation system (Horn and Johnson, 2012).

4.1 Coordinate representation

Let $\mathscr{G}_1$ be an n-dimensional Hilbert space spanned by the set of functions $\mathscr{B}_1 = \{b_1^1, \ldots, b_n^1\}$. Then, each member f of $\mathscr{G}_1$ can be written as a linear combination of $b_1^1, \ldots, b_n^1$. The vector of coefficients is called the coordinate of f with respect to $\mathscr{B}_1$, and is written as $[f]_{\mathscr{B}_1}$. Let $\mathscr{G}_2$ be another Hilbert space spanned by $\mathscr{B}_2 = \{b_1^2, \ldots, b_m^2\}$. For any linear operator $A : \mathscr{G}_1 \mapsto \mathscr{G}_2$, the matrix ${}_{\mathscr{B}_2}[A]_{\mathscr{B}_1} = ([Ab_1^1]_{\mathscr{B}_2}, \ldots, [Ab_n^1]_{\mathscr{B}_2}) \in \mathbb{R}^{m\times n}$ is called the coordinate of A relative to the bases $\mathscr{B}_1$ and $\mathscr{B}_2$. The mapping ${}_{\mathscr{B}_2}[\cdot]_{\mathscr{B}_1} : A \mapsto {}_{\mathscr{B}_2}[A]_{\mathscr{B}_1}$ is called the coordinate mapping. Let $\mathscr{G}_3$ be a third Hilbert space with base $\mathscr{B}_3$, and $A' : \mathscr{G}_2 \mapsto \mathscr{G}_3$ a linear operator, then the coordinate representation of the composite operator $A'A : \mathscr{G}_1 \mapsto \mathscr{G}_3$ is ${}_{\mathscr{B}_3}[A'A]_{\mathscr{B}_1} = {}_{\mathscr{B}_3}[A']_{\mathscr{B}_2\mathscr{B}_2}[A]_{\mathscr{B}_1}$. For more details about the coordinate representations, see Horn and Johnson (2012), page 30 and Solea and Li (2022), Theorem 5.

4.2 Step 1: implementation of GSIR

In this section, we derive the coordinate representation of the estimator of the GSIR problem (8). Let $X_1, \ldots, X_n$ be an i.i.d sample of X, where $X_i = (X_i^1, \ldots, X_i^p)$ for $i = 1, \ldots, n$. At the sample level, the true distribution P_X is replaced by the empirical distribution based on the sample. For each $i = 1, \ldots, p$, the centered RKHS $\mathcal{H}_{X^i}$ is spanned by

$$\phi_{X^i} = \{\kappa_{X^i}(\cdot, X_a^i) - \mathbb{E}_n(\kappa_{X^i}(\cdot, X^i)) : a = 1, \ldots, n\}, \tag{12}$$

where $\mathbb{E}_n$ is the mean with respect to the empirical distribution, that is, $\mathbb{E}_n\kappa_{X^i}(\cdot, X^i) = \frac{1}{n}\sum_{a=1}^n \kappa_{X^i}(\cdot, X_a^i)$. For any subset S of V, the centered RKHS $\mathcal{H}_{X^S}$ is spanned by $\phi_{X^S} = \{\kappa_{X^S}(\cdot, X_a^S) - \mathbb{E}_n(\kappa_{X^S}(\cdot, X^S)) : a = 1, \ldots, n\}$. Furthermore, for each $i = 1, \ldots, p$ let $K_{X^i} = \{\kappa_{X^i}(X_\ell^i, X_m^i)\}_{\ell,m=1}^n \in \mathbb{R}^{n\times n}$ be the Gram matrix for X^i, and $G_{X^i} = QK_{X^i}Q \in \mathbb{R}^{n\times n}$ be its centered version, where $Q = I_n - 1_n 1_n^\intercal/n$, 1_n is the n-dimension vector of 1's and I_n is the $n\times n$ identity matrix. Similarly, for every $(i, j) \in \mathsf{V}$, $i \neq j$, let $K_{X^iX^j}$ be the $n \times n$ Gram matrix whose (ℓ, m)th entry is $\kappa_{X^iX^j}((X_\ell^i, X_\ell^j), (X_m^i, X_m^j))$, and $G_{X^iX^j} = QK_{X^iX^j}Q \in \mathbb{R}^{n\times n}$ be its centered version. It can be shown that for any $f, g \in \mathcal{H}_{X^i}$ $\langle f, g\rangle = [f]^\intercal G_{X^i}[g]$. Let K_S be the $n \times n$ matrix whose (a, b)th entry is $\kappa_{X^S}(X_a^S, X_b^S)$ and let $G_{X^S} = QK_SQ$. We have the following coordinate representations for the operators (Li, 2018a):

$$\begin{gathered}[\Sigma_{X^iX^j}] \propto G_{X^j}, \quad [\Sigma_{X^S{}_{(X^iX^j)}}] \propto G_{X^iX^j}, \quad [\Sigma_{X^SX^S}] \propto G_{X^S} \\ [\Sigma_{(X^iX^j)(X^iX^j)}] \propto G_{X^iX^j}, \quad [\Sigma_{(X^iX^j)X^S}] \propto G_{X^S}.\end{gathered} \tag{13}$$

Using the coordinate expressions in (13), the maximization problem in (8) can be represented as the following standard eigenvalue problem with Tychonoff regularized inverse (see Li and Kim

(2024) for more details): for each $i, j \in \mathsf{V}$, $i \neq j$ and $S \subset \mathsf{V} \setminus \{i, j\}$ and for each $k = 1, \ldots, d_S^{ij}$,

$$\begin{aligned} &\text{maximize} \quad v^\mathsf{T}(G_{X^S} + \eta_n I_n)^{-1} G_{X^S} G_{X^i X^j} (G_{X^i X^j} + \epsilon_n I_n)^{-1} G_{X^S} (G_{X^S} + \eta_n I_n)^{-1} v \\ &\text{subject to} \quad v^\mathsf{T} v = 1,\, v^\mathsf{T} v_1 = \ldots = v^\mathsf{T} v_{k-1} = 0, \end{aligned}$$

where $\epsilon_n > 0$ and $\eta_n > 0$ are tuning constants such that $\epsilon_n \to 0$ and $\eta_n \to 0$ as $n \to \infty$. In other words, $v_1, \ldots, v_{d_S^{ij}}$ are the first d_S^{ij} eigenvectors of the matrix

$$(G_{X^S} + \eta_n I_n)^{-1} G_{X^S} G_{X^i X^j} (G_{X^i X^j} + \epsilon_n I_n)^{-1} G_{X^S} (G_{X^S} + \eta_n I_n)^{-1}. \tag{14}$$

Then, the GSIR sufficient predictors are

$$\hat{f}_r^{ij,S} = \hat{U}_r^{ij,S} = v_r^\mathsf{T} (G_{X^S} + \eta_n I)^{-1} K_{X^S} Q, \quad r = 1, \ldots, d_S^{ij}, \tag{15}$$

will be used for the implementation of the CCCO in tne next section.

4.3 Step 2: implementation of the conjoined conditional cross-covariance operator

Having derived the estimated sufficient predictors $\hat{U}^{ij,S}$ for all $i, j \in \mathsf{V}$, $i \neq j$ and for any $S \subset \mathsf{V} \setminus \{i, j\}$, we next evaluate conditional independence by thresholding the Hilbert-Schmidt norm of the CCCO defined in (10). That is, for any $i, j \in \mathsf{V}$, $i \neq j$ and for any $S \subset \mathsf{V} \setminus \{i, j\}$,

$$X^i \perp\!\!\!\perp X^j \,|\, \hat{U}^{ij,S} \quad \Leftrightarrow \quad \|\Sigma_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}}\|_{\mathrm{HS}} < \rho_n,$$

for some threshold constant $\rho_n > 0$ that depends on n. Thus we have the following estimate of the skeleton of the SDAG G,

$$\widehat{\mathrm{ske}(\mathsf{G})} = \{(i, j) \in \mathsf{V} \times \mathsf{V} : i \neq j, \|\Sigma_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}}\|_{\mathrm{HS}} > \rho_n \text{ for all } S \subset \mathsf{V} \setminus \{i, j\}\}. \tag{16}$$

Next, we turn to the coordinate representation of $\Sigma_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S}}$ and its Hilbert-Schmidt norm. For each $i, j \in \mathsf{V}$, $i \neq j$, and $S \subset \mathsf{V} \setminus \{i, j\}$, we define the centered RKHS

$$\begin{aligned} \mathcal{H}_{X^i \hat{U}^{ij,S}} &= \mathrm{span}\{\kappa_{X^i U^{ij,S}}(\cdot, (X_a^i, \hat{U}_a^{ij,S})) - \mathbb{E}_n[\kappa_{X^i \hat{U}^{ij,S}}(\cdot, (X^i, \hat{U}^{ij,S}))] : a = 1, \ldots, n\}, \\ \mathcal{H}_{X^j \hat{U}^{ij,S}} &= \mathrm{span}\{\kappa_{X^j \hat{U}^{ij,S}}(\cdot, (X_a^j, \hat{U}_a^{ij,S})) - \mathbb{E}_n[\kappa_{X^j \hat{U}^{ij,S}}(\cdot, (X^j, \hat{U}^{ij,S}))] : a = 1, \ldots, n\}, \\ \mathcal{H}_{\hat{U}^{ij,S}} &= \mathrm{span}\{\kappa_{\hat{U}^{ij,S}}(\cdot, \hat{U}_a^{ij,S}) - \mathbb{E}_n[\kappa_{\hat{U}^{ij,S}}(\cdot, \hat{U}^{ij,S})] : a = 1, \ldots, n\}. \end{aligned}$$

By Theorem 12.1 of Li (2018a), the coordinate representations of the covariance operators are

$$\begin{aligned} [\Sigma_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})}] \propto G_{X^j \hat{U}^{ij,S}}, \quad [\Sigma_{(X^i \hat{U}^{ij,S})\hat{U}^{ij,S}}] \propto G_{\hat{U}^{ij,S}}, \\ [\Sigma_{\hat{U}^{ij,S}\hat{U}^{ij,S}}] \propto G_{\hat{U}^{ij,S}}, \quad [\Sigma_{\hat{U}^{ij,S}(X^j \hat{U}^{ij,S})}] \propto G_{X^j \hat{U}^{ij,S}}. \end{aligned}$$

Consequently, the coordinate representation of the CCCO (Li and Kim, 2024) is

$$[\Sigma_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}}] = G_{X^j \hat{U}^{ij,S}} - G_{\hat{U}^{ij,S}} (G_{\hat{U}^{ij,S}} + \delta_n I_n)^{-1} G_{X^j \hat{U}^{ij,S}},$$

where $\delta_n \to 0$ is a positive tuning constant. By Li and Kim (2024), the Hilbert-Schmidt norm of CCCO has the following coordinate representation

$$\|\Sigma_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}}\|_{\mathrm{HS}} = \|G^{1/2}_{X^i \hat{U}^{ij,S}} G^{1/2}_{X^j \hat{U}^{ij,S}} - G^{1/2}_{X^i \hat{U}^{ij,S}} G_{\hat{U}^{ij,S}} (G_{\hat{U}^{ij,S}} + \delta_n I_n)^{-1} G^{1/2}_{X^j \hat{U}^{ij,S}}\|_{\mathrm{F}},$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenious matrix norm.

4.3.1 RANK-REDUCED KERNEL

To reduce computation time, so that our method can be readily applied to relatively large networks, we propose a simplified algorithm to estimate the CCCO. In practice, the Gram matrices G_{X^i} often have a few dominating eigenvalues, indicating that the RKHS $\mathcal{H}_{X^i}$ are of lower dimensions (Lee and Huang, 2007; Chen et al., 2010). This motivates us to use the leading eigenfunctions to construct $\mathcal{H}_{X^i}$, which reduces the amount of computation without incurring much loss of accuracy. Suppose that, for each $i = 1, \ldots, p$, the Gram matrices G_{X^i} have the following spectral-decomposition

$$G_{X^i} = V_{X^i} D_{X^i} V_{X^i}^{\mathsf{T}} + \tilde{V}_{X^i} \tilde{D}_{X^i} \tilde{V}_{X^i}^{\mathsf{T}},$$

where $V_{X^i} D_{X^i} V_{X^i}^{\mathsf{T}}$ corresponds to the first r_i eigenvalues of G_{X^i} and $\tilde{V}_{X^i} \tilde{D}_{X^i} \tilde{V}_{X^i}^{\mathsf{T}}$ corresponds to the last $n - r_i$ eigenvalues of G_{X^i}. Instead of the original bases $\phi_{X^i}^{\mathsf{T}} = (\phi_{1,X^i}, \ldots, \phi_{n,X^i})^{\mathsf{T}}$ in (12), we use

$$\psi_{X^i}^{\mathsf{T}} = (\psi_{1,X^i}, \ldots, \psi_{r_i,X^i})^{\mathsf{T}} = (\phi_{1,X^i}, \ldots, \phi_{n,X^i})^{\mathsf{T}} V_{X^i} D_{X^i}^{-1/2},$$

as the orthonormal basis for $\mathcal{H}_{X^i}$. One can show that the coordinate representation of the inner product with respect to the new bases is $\langle f, g \rangle = [f]^{\mathsf{T}}[g]$ for any $f, g \in \mathcal{H}_{X^i}$. The next proposition gives the coordinate representations of relevant operators using the new basis $\psi_{X^i}^{\mathsf{T}} = (\psi_{1,X^i}, \ldots, \psi_{r_i,X^i})^{\mathsf{T}}$. The proof uses essentially the same arguments in the proof of Proposition 8 in Lee et al. (2020), and is thus omitted.

Proposition 6 *For any $i, j \in \mathsf{V}$, $i \neq j$, and $S \subset \mathsf{V} \setminus \{i, j\}$ the operators $\Sigma_{X^i X^i}$, $\Sigma_{X^i X^j}$, $\Sigma_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})}$, $\Sigma_{(X^i \hat{U}^{ij,S})\hat{U}^{ij,S}}$, $\Sigma_{\hat{U}^{ij,S}\hat{U}^{ij,S}}$ and $\Sigma_{\hat{U}^{ij,S}(X^j \hat{U}^{ij,S})}$ have the following coordinate representation with respect to the new basis $\psi_{X^i}^{\mathsf{T}} = (\psi_{1,X^i}, \ldots, \psi_{r_i,X^i})^{\mathsf{T}}$*

$$\begin{aligned}
[\Sigma_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})}] &= n^{-1} \Lambda_{X^i \hat{U}^{ij,S}}^{1/2} V_{X^i \hat{U}^{ij,S}}^{\mathsf{T}} V_{X^j \hat{U}^{ij,S}} \Lambda_{X^j \hat{U}^{ij,S}}^{1/2}, \\
[\Sigma_{(X^i \hat{U}^{ij,S})\hat{U}^{ij,S}}] &= n^{-1} \Lambda_{X^i \hat{U}^{ij,S}}^{1/2} V_{X^i \hat{U}^{ij,S}}^{\mathsf{T}} V_{\hat{U}^{ij,S}} \Lambda_{\hat{U}^{ij,S}}^{1/2}, \\
[\Sigma_{\hat{U}^{ij,S}\hat{U}^{ij,S}}] &= n^{-1} \Lambda_{\hat{U}^{ij,S}}^{1/2} V_{\hat{U}^{ij,S}}^{\mathsf{T}} V_{\hat{U}^{ij,S}} \Lambda_{\hat{U}^{ij,S}}^{1/2}, \\
[\Sigma_{\hat{U}^{ij,S}(X^j \hat{U}^{ij,S})}] &= n^{-1} \Lambda_{\hat{U}^{ij,S}}^{1/2} V_{\hat{U}^{ij,S}}^{\mathsf{T}} V_{X^j \hat{U}^{ij,S}} \Lambda_{X^j \hat{U}^{ij,S}}^{1/2},
\end{aligned}$$

The coordinate representation of the CCCO is then

$$[\Sigma_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}}] = [\Sigma_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})}] - [\Sigma_{(X^i \hat{U}^{ij,S})\hat{U}^{ij,S}}][\Sigma_{\hat{U}^{ij,S}\hat{U}^{ij,S}}]^{\dagger}(\delta_n)[\Sigma_{\hat{U}^{ij,S}(X^j \hat{U}^{ij,S})}],$$

where, for any matrix $A \in \mathbb{R}^{n \times n}$ and $\epsilon > 0$, $A^{\dagger}(\epsilon) = \sum_{i=1}^{n} \lambda_i^{-1} I(\lambda_i > \epsilon) u_i u_i^{\mathsf{T}}$, and $(\lambda_i, u_i), i = 1, \ldots, n$, are the eigenpairs of A. Using the coordinate representations of proposition 6, we obtain the coordinate representation of the CCCO with respect to the new basis as

$$[\Sigma_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}}] = n^{-1} \Lambda_{X^i \hat{U}^{ij,S}}^{1/2} V_{X^i \hat{U}^{ij,S}}^{\mathsf{T}} \big(I_n - V_{\hat{U}^{ij,S}} \mathrm{diag}(I_{m_{\hat{U}^{ij,S}}}, 0) V_{\hat{U}^{ij,S}}^{\mathsf{T}}\big) V_{X^j \hat{U}^{ij,S}} \Lambda_{X^j \hat{U}^{ij,S}}^{1/2}, \tag{17}$$

where $m_{\hat{U}^{ij,S}}$ is the number of eigenvalues in $\Lambda_{\hat{U}^{ij,S}}$ greater than δ_n. Correspondingly, the Hilbert-Schmidt norm of the CCCO can be computed via

$$\|\Sigma_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}}\|_{\mathrm{HS}}^2 = \mathrm{tr}\big([\Sigma_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}}][\Sigma_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}}]^{\mathsf{T}}\big). \tag{18}$$

4.4 Tuning

In this subsection, we propose the tuning procedures involved in various steps in our method. Specifically, for step 1, the tuning parameters include the kernel parameters κ_{X^i}, $i = 1, \ldots, p$, the Tychonoff regularization tuning constants η_n and ϵ_n, and the dimension d_S^{ij} of the sufficient predictor $\hat{U}^{ij,S}$. For step 2, the tuning parameters include the number of leading eigenvalues of G_{X^i}, r_i, the Tychonoff regularization parameter for the CCCO, δ_n, and the thresholding constant ρ_n in the estimation of the skeleton in (16).

For each $i = 1, \ldots, p$, we take $\kappa_{X^i}(\cdot, \cdot)$ to be the Gaussian RBF. We choose γ_{X^i} by the criterion used in Lee et al. (2013):

$$1/\sqrt{\gamma_{X^i}} = \binom{n}{2}^{-1} \sum_{a<b} |X_a^i - X_b^i|. \tag{19}$$

For the choice of ϵ_n and η_n, we use a generalized cross validation (GCV) scheme, which follows from Li and Kim (2024),

$$\mathrm{GCV}(\epsilon) = \mathrm{argmin}_\epsilon \sum_{i<j} \frac{\|G_1 - G_2^\mathsf{T}(G_2 + \epsilon\lambda_{\max}(G_2))^{-1} G_1\|_\mathrm{F}}{n^{-1}\mathrm{tr}(I_n - G_2^\mathsf{T}(G_2 + \epsilon\lambda_{\max}(G_2))^{-1})},$$

where $G_1, G_2 \in \mathbb{R}^{n\times n}$ are positive definite matrices and $\lambda_{\max}(G_2)$ is the largest eigenvalue of G_2. The matrices are taken to be $G_1 = G_{X^S}$ and $G_2 = G_{X^iX^j}$ for ϵ_n and $G_1 = G_{X^iX^j}$ and $G_2 = G_{X^S}$ for η_n. We minimize $\mathrm{GCV}(\epsilon)$ over a grid of values to choose ϵ. For selecting the dimension d_S^{ij} of $U^{ij,S}$, to the best of our knowledge there is no general procedure for choosing the dimension of the central class for nonlinear SDR. Nevertheless, in practice a small dimension such as 1 or 2 is usually sufficient. For example, in the classical nonparametric regression models $Y = f(X) + \epsilon$, with $X \perp\!\!\!\perp \epsilon$, the dimension of the central class is by definition 1 (Li and Kim, 2024).

For step 2, we choose r_i such that the cumulative percentage of total variation of G_{X^i} exceeds 0.99, as in Lee et al. (2020). That is, we choose r_i according to

$$r_i = \min\Big\{r' : \frac{\sum_{i=1}^{r'} \lambda_i(G_{X^i})}{\sum_{i=1}^{n} \lambda_i(G_{X^i})} \geq 0.99\Big\}. \tag{20}$$

We choose δ_n similarly; that is,

$$\delta_n = \min\Big\{\epsilon : \frac{\sum_{i=1}^{n} \lambda_i(\Lambda_{\hat{U}^{ij,S}}) I\{\lambda_i(\Lambda_{\hat{U}^{ij,S}}) \geq \epsilon\}}{\sum_{i=1}^{n} \lambda_i(\Lambda_{\hat{U}^{ij,S}})} \geq 0.99\Big\}. \tag{21}$$

To determine the thresholding constant, ρ_n, we follow the permutation-test based approach as Lee et al. (2020); Fukumizu et al. (2007b); Sun (2008). Specifically, we first partition $\{\hat{U}_1^{ij,S}, \ldots, \hat{U}_n^{ij,S}\}$ into n_C clusters using, K-means. We then randomly shuffle ℓ times the observations of the vector $(X_j, \hat{U}^{ij,S})$ with respect to the n_C clusters. For each ℓth sample, we calculate the CCCO as in (17) and its Hilbert-Schmidt norm as in (18):

$$R_h = \|\Sigma^h_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})|\hat{U}^{ij,S}}\|^2_{\mathrm{HS}}, \quad h = 1, \ldots, \ell.$$

For a given significance level, α, we use the $1 - \alpha$ sample upper quantile of the $R_1, \ldots, R_\ell$ as the thresholding constant ρ_n.

4.5 A new PC-algorithm

We first summarize the implementation procedures for steps 1 and 2 developed in Section 4.4 for a given $i, j \in \mathsf{V}$, $i \neq j$, and $S \subset \mathsf{V} \setminus \{i, j\}$:

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

2. Choose the kernels κ_{X^i} for $\mathcal{H}_{X^i}$ to be the Gaussian RBF in (4), and the tuning parameter γ_{X^i} by (19).

3. Compute the matrices K_{X^i}, K_{X^j} and K_{X^S} and their centered versions G_{X^i}, G_{X^j} and G_{X^S}, respectively.

4. Solve the eigenvalue problem (14) with ϵ_n and η_n chosen by minimizing $\mathrm{GCV}(\epsilon)$ over a grid of values of ϵ. Let $\hat{v}_1, \ldots, \hat{v}_{d_S^{ij}}$ be the first d_S^{ij} eigenvectors, and form the sufficient predictor $\hat{U}^{ij,S} = (\hat{f}_1^{ij,S}, \ldots, \hat{f}_{d_S^{ij}}^{ij,S})$ according to (15).

5. Form the Gram matrices $G_{\hat{U}^{ij,S}}$, $G_{X^i\hat{U}^{ij,S}}$ and $G_{X^j\hat{U}^{ij,S}}$, and perform spectral decomposition to obtain the matrices $\Lambda_{\hat{U}^{ij,S}}$, $\Lambda_{X^i\hat{U}^{ij,S}}$ and $\Lambda_{X^j\hat{U}^{ij,S}}$ and $V_{\hat{U}^{ij,S}}$, $V_{X^i\hat{U}^{ij,S}}$ and $V_{X^j\hat{U}^{ij,S}}$.

6. Determine r_i and δ_n using (20) and (21), respectively.

7. Calculate $\Sigma_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})|\hat{U}^{ij,S}}$ and $\|\Sigma_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})|\hat{U}^{ij,S}}\|_{\mathrm{HS}}^2$ according to (17) and (18), respectively.

Next, we develop a modified version of the PC-algorithm (Spirtes et al., 2000) to identify the Markov equivalence class of the DAG at the sample level. The new PC algorithm incorporates dimension reduction to reduce the dimension of the conditional variable while exploiting the sparseness of the graph to reduce the amount of computation, as in Kalisch and Bühlman (2007). Like the original PC-algorithm, the complexity depends on the sparseness of the DAG instead of the size of the network. We display the new version of the PC-algorithm in the Algorithm 1 in the form of pseudo-codes. Algorithm 1 below describes only the first part of the DAG-PC algorithm that identifies the skeleton of the DAG. The output of this part is the estimated skeleton, $\widehat{\mathrm{ske}(\mathsf{G})}$, and the separation sets denoted, S. Then, the second part of the DAG-PC algorithm uses these separation sets to extend the skeleton to the equivalence class, by identifying the partial orientation of the edges. The output is the CPDAG (Meek, 1995). We omit the second part of the DAG-PC algorithm as it is essentially the same as that in Kalisch and Bühlman (2007). In algorithm 1 $\mathrm{adj}(\mathsf{G}, j)$ denotes the set of all vertices i that are adjacent to j in DAG G (connected by a directed or undirected edge).

5. Asymptotic theory

In this section, we develop the asymptotic theory of our method, including the consistency and the convergence rates of the CCCO, and the consistency and uniform consistency of the DAG-PC algorithm. We focus on the case where the network size, p and the dimension d_S^{ij} are fixed, while allowing the tuning parameters, η_n, δ_n and ϵ_n, and the thresholding constant, ρ_n to depend on the sample size n. For two sequences of positive numbers $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, we write $a_n \prec b_n$ if $a_n/b_n \to 0$ as $n \to \infty$, and $a_n \preceq b_n$ if a_n/b_n is a bounded sequence.

Algorithm 1 The DAG-PC algorithm

Set $\ell = -1$ and $\widehat{\mathrm{ske}(\mathsf{G})}$ to be the complete undirected graph $\tilde{C}$
repeat
  $\ell = \ell + 1$
  repeat
    Select a new ordered pair of nodes (i, j) that are adjacent in $\widehat{\mathrm{ske}(\mathsf{G})}$ such that $|\mathrm{adj}(\widehat{\mathrm{ske}(\mathsf{G})}, i) \setminus \{j\}| \geq \ell$
    repeat
      select new $S \subset \mathrm{adj}(\widehat{\mathrm{ske}(\mathsf{G})}, i) \setminus \{j\})$ with $|S| = \ell$
      Implement GSIR to obtain the d_S^{ij} sufficient predictors $\hat{U}^{ij,S} = (\hat{f}_1^{ij}, \ldots, \hat{f}_{d_S^{ij}}^{ij})$
      if $\|\Sigma_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})|\hat{U}^{ij,S}}\|_{\mathrm{HS}} \leq \rho$ **then**
        delete edge $i - j$ from $\widehat{\mathrm{ske}(\mathsf{G})}$
        Save $S_{i,j} = S$
      end if
    until
    edge $i - j$ is deleted or all $S \subset \mathrm{adj}(\widehat{\mathrm{ske}(\mathsf{G})}, i) \setminus \{j\})$ with $|S| = \ell$ have been chosen
  until
  all ordered pairs of adjacent variables i and j such that $|\mathrm{adj}(\widehat{\mathrm{ske}(\mathsf{G})}, i) \setminus \{j\}| \geq \ell$ and $S \subseteq \mathrm{adj}(\widehat{\mathrm{ske}(\mathsf{G})}, i) \setminus \{j\})$ with $|S|$ have been tested for conditional independence
until
$|\mathrm{adj}(\widehat{\mathrm{ske}(\mathsf{G})}, i) \setminus \{j\}| < \ell$ for each ordered pair of adjacent nodes (i, j).

5.1 Convergence rate of the GSIR estimator

In this subsection, we first present the convergence rate of the GSIR estimator $\hat{U}^{ij,S} = (\hat{f}_1^{ij,S}, \ldots, \hat{f}_{d_S^{ij}}^{ij,S})$, which was given in Li and Kim (2024), and will be important when deriving the consistency of the CCCO as we can replace $\hat{U}^{ij,S}$ with $U^{ij,S}$. We make the following assumption.

Theorem 7 *Suppose Assumptions 3-7 hold. If*

$$n^{-1/2} \prec \eta_n \prec 1, \quad n^{-1/2} \prec \epsilon_n \prec 1, \tag{22}$$

then $\|\hat{U}^{ij,S} - U^{ij,S}\| = O_p(\eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n + \epsilon_n)$.

5.2 Convergence rate of CCCO

Let $\hat{\Sigma}_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})|\hat{U}^{ij,S}}$ be the sample-level of CCCO. In this subsection, we derive the convergence rate of the Hilbert-Schmidt norm $\|\hat{\Sigma}_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})|\hat{U}^{ij,S}} - \Sigma_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}}$, when the estimated GSIR $\hat{U}^{ij,S}$ satisfies (22). The approach used in establishing this result is of independent interest, as it tells us how to deal with estimated quantities such as $\hat{U}^{ij,S}$ that are in valued in the sampled estimate of a linear operator. We make the following assumption for the reproducing kernels $\kappa_{X^iU^{ij,S}}$, $\kappa_{X^jU^{ij,S}}$ and $\kappa_{U^{ij,S}}$ which implies a type of Lipschitz continuity (see also Li and Kim (2024)).

Assumption 10 *Let $\kappa : \Omega \times \Omega \to \mathbb{R}$ be a positive kernel that satisfies*

1. *For each $t \in \Omega$, the function $s \mapsto \kappa(s,t)$ is twice differentiable.*
2. $\frac{\partial \kappa(s,t)}{\partial s}\big|_{s=t} = 0$.
3. *The minimum and maximum eigenvalues of the Hessian matrix $H(s,t) = \frac{\partial^2 \kappa(s,t)}{\partial s \partial s^\intercal}$, $\lambda_{\min}(H(s,t))$ and $\lambda_{\max}(H(s,t))$, respectively satisfy*

$$C_1 \le \lambda_{\min}(H(s,t)) \le \lambda_{\max}(H(s,t)) \le C_2,$$

for some constants $C_1 > 0$ and $C_2 > 0$.

Let $\mathcal{H}_1$ be an RKHS generated by the kernel $\kappa_1(s,t)$ that satisfies Assumption 10 (for example, the Gaussian RBF satisfies Assumption 10, but the Laplace kernel does not (Li and Kim, 2024)). Let $\mathcal{H}_0$ be an RKHS generated by a kernel κ_0. Then, according to Theorem 2 of Li and Kim (2024), there exists a constant $C > 0$ such that for any, $f, g \in \mathcal{H}_0$ and $a \in \Omega$,

$$\|\kappa_1(\cdot, f(a)) - \kappa_1(\cdot, g(a))\| \le C\sqrt{\kappa_0(a,a)}\|f - g\|_{\mathcal{H}_1}. \tag{23}$$

A consequence of (23) is the following Theorem (Theorem 2, Li and Kim, 2024).

Theorem 8 *Suppose the conditions of Theorem 7 are satisfied such that $\|\hat{U}^{ij,S} - U^{ij,S}\| = O_p(b_n)$, where $b_n = \eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n + \epsilon_n \prec 1$. Suppose, furthermore, that Assumptions 9 and 10 hold for the reproducing kernels $\kappa_{X^i U^{ij,S}}$, $\kappa_{X^j U^{ij,S}}$ and $\kappa_{U^{ij,S}}$. Then,*

1. $\|\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} - \hat{\Sigma}_{U^{ij,S}U^{ij,S}}\|_{\mathrm{HS}} = O_p(b_n)$,
2. $\|\hat{\Sigma}_{(X^i\hat{U}^{ij,S})\hat{U}^{ij,S}} - \hat{\Sigma}_{(X^iU^{ij,S})U^{ij,S}}\|_{\mathrm{HS}} = O_p(b_n)$,
3. $\|\hat{\Sigma}_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})} - \hat{\Sigma}_{(X^iU^{ij,S})(X^jU^{ij,S})}\|_{\mathrm{HS}} = O_p(b_n)$.

Using Theorem 8, we can derive the convergence rate of $\hat{\Sigma}_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})|\hat{U}^{ij,S}}$. We need the following assumption.

Assumption 11 *For each $i, j \in \mathsf{V}, i \neq j$, $S \setminus \{i,j\}$, $\Sigma^{-1}_{U^{ij,S}U^{ij,S}}\Sigma_{U^{ij,S}(X^jU^{ij,S})}$ is a Hilbert-Schmidt operator.*

This assumption ensures a degree of smoothness in the relation between X^i and $U^{ij,S}$. Since $\Sigma^{-1}_{U^{ij,S}U^{ij,S}}$ is an unbounded operator, in order for $\Sigma^{-1}_{U^{ij,S}U^{ij,S}}\Sigma_{U^{ij,S}(X^jU^{ij,S})}$ to be a Hilbert-Schmidt operator, the range space of $\Sigma_{U^{ij,S}(X^jU^{ij,S})}$ should be sufficiently concentrated on the eigenspaces of $\Sigma_{U^{ij,S}U^{ij,S}}$ corresponding to large eigenvalues, or the low-frequency components of $\Sigma_{U^{ij,S}U^{ij,S}}$.

Theorem 9 *Suppose the conditions of Theorem 7 are satisfied such that $\|\hat{U}^{ij,S} - U^{ij,S}\| = O_p(b_n)$, where $b_n = \eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n + \epsilon_n \prec 1$, and the Assumptions 9 and 10 hold for the reproducing kernels $\kappa_{X^iU^{ij,S}}$, $\kappa_{X^jU^{ij,S}}$ and $\kappa_{U^{ij,S}}$. Suppose, furthermore, that Assumptions 11 holds, and*

$$n^{-1/2} \prec \eta_n \prec 1, \quad n^{-1/2} \prec \epsilon_n \prec 1, \quad n^{-1/2} \prec \delta_n^{3/2}, \quad b_n \prec \delta_n^{3/2},$$

where $\delta_n \prec 1$, $b_n = \eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n \prec 1$. Then,

$$\|\hat{\Sigma}_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})|\hat{U}^{ij,S}} - \Sigma_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}} = O_p(\delta_n^{-1}b_n + \delta_n^{-1}n^{-1/2} + \delta_n^{1/2}).$$

5.3 Consistency of the estimator of the DAG

In this section, we prove the consistency of the estimated CPDAG of the true DAG. Let $\widehat{\mathrm{ske}(\mathsf{G})}$ and $\widehat{\mathrm{CPDAG}}(\mathsf{G})$ be the estimated skeleton and the estimated CPDAG of the true DAG, G, respectively. To highlight the dependence of $\widehat{\mathrm{ske}(\mathsf{G})}$ and $\widehat{\mathrm{CPDAG}}(\mathsf{G})$ on $\eta_n, \epsilon_n, \delta_n$ and ρ_n, we write them as $\widehat{\mathrm{ske}(\mathsf{G})}(\eta_n, \epsilon_n, \delta_n, \rho_n)$ and $\widehat{\mathrm{CPDAG}}(\mathsf{G})(\eta_n, \epsilon_n, \delta_n, \rho_n)$, respectively.

Theorem 10 *Suppose Assumption 1 is satisfied and X is faithful with respect to a DAG, G, according to Definition 3. Then, under the assumptions of Theorem 9, and if $\delta_n^{-1} b_n + \delta_n^{-1} n^{-1/2} + \delta_n^{1/2} \prec \rho_n \prec 1$,*

$$\mathbb{P}(\widehat{\mathrm{ske}(\mathsf{G})}(\eta_n, \epsilon_n, \delta_n, \rho_n) = \mathrm{ske}(\mathsf{G})) \to 1, \quad \textit{as} \quad n \to \infty,$$

where $\widehat{\mathrm{ske}(\mathsf{G})}$ is as defined in (16).

Theorem 11 *Suppose Assumption 1 is satisfied and X is faithful with respect to a DAG, G, according to Definition 3. Then, under the assumptions of Theorem 9, and if $\delta_n^{-1} b_n + \delta_n^{-1} n^{-1/2} + \delta_n^{1/2} \prec \rho_n \prec 1$,*

$$\mathbb{P}(\widehat{\mathrm{CPDAG}}(\mathsf{G})(\eta_n, \epsilon_n, \delta_n, \rho_n) = \mathrm{CPDAG}(\mathsf{G})) \to 1, \quad \textit{as} \quad n \to \infty.$$

5.4 Uniform consistency of the estimator of the DAG

In the last subsection, the consistency was established in terms of the true distribution P_X. Sometimes the uniform consistency might be preferable if we want to control type I and type II errors (Lee et al., 2020). For a DAG, G, a Gaussian distribution P is called τ-strongly faithful with respect to G for some $\tau > 0$, if and only if P is faithful with respect to G, and

$$\min\{|\mathrm{cor}(X^j, X^k|X^S)| : \mathrm{cor}(X^j, X^k|X^S) \neq 0, (j,k) \in \mathsf{V} \times \mathsf{V}, S \subseteq \mathsf{V} \setminus \{j,k\}\} > \tau.$$

Strong faithfulness can be viewed as a condition of signal strength in terms of non-zero partial correlations. Uhler et al. (2013) showed that strong faithfulness is a strong assumption for many DAGs, but it is essentially preferable for statistical inference. Let $\mathcal{P}$ be the class of distributions of X. The following definition extends the strong faithfulness condition to our setting.

Definition 12 *A family of distributions $\mathcal{P}$ of X is τ-strongly faithful with respect to G, if the faithfulness condition* (3) *holds, and there is a $\tau > 0$ such that $\forall P \in \mathcal{P}$*

$$\min\left\{\|\Sigma_{(X^{i}U^{ij,S})(X^{j}U^{ij,S})|U^{ij,S}}(P)\|_{\mathrm{HS}} \neq 0, i, j \in \mathsf{V}, S \subseteq \mathsf{V} \setminus \{i,j\}\right\} > \tau. \tag{24}$$

The next theorem establishes the uniform consistency of the $\widehat{\mathrm{CPDAG}}(\mathsf{G})(\eta_n, \delta_n, \epsilon_n, \rho_n)$.

Theorem 13 *Suppose*

(a) $\mathcal{P}$ is τ-strongly faithful with respect to G.

(b) For all $i, j \in \mathsf{V}$, $S \subseteq \mathsf{V} \setminus \{i,j\}$ and $\epsilon > 0$,

$$\limsup_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(\|\hat{\Sigma}_{(X^{i}\hat{U}^{ij,S})(X^{j}\hat{U}^{ij,S})|\hat{U}^{ij,S}} - \Sigma_{(X^{i}U^{ij,S})(X^{j}U^{ij,S})|U^{ij,S}}(P)\|_{\mathrm{HS}} > \epsilon) = 0. \tag{25}$$

Then, for any $0 < \rho_n < \tau$,

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}(\widehat{\mathrm{CPDAG}}(\mathsf{G})(\eta_n, \epsilon_n, \delta_n, \rho_n) = \mathrm{CPDAG}(\mathsf{G})) = 0.$$

Condition (b) can be viewed as the uniform extension of $\mathbb{P}(\|\hat{\Sigma}_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})|\hat{U}^{ij,S}} - \Sigma_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}}(P)\|_{\mathrm{HS}} > \epsilon)$ which was proved in Theorem 9. A similar condition was considered in Lee et al. (2020).

6. Numerical Study

In this section, we evaluate the performance of our DAG estimator, referred to as the DAG-PC algorithm, through simulation comparisons with other methods and a data application. We compare it with three existing PC algorithms: the Gaussian-PC algorithm, which is based on the sample partial correlation (Kalisch and Bühlman, 2007); the semiparametric rank-PC algorithm, which is based on rank correlation statistics (Harris and Drton, 2013), and the fully nonparametric kernel-PC algorithm of Gretton et al. (2009) without dimension reduction (Székely et al., 2007; Gretton et al., 2005b). For simplicity, these methods are labeled as

Method A :	Gaussian-PC algorithm
Method B :	Rank-PC algorithm
Method C :	Kernel-PC algorithm
Method D :	DAG-PC algorithm.

We evaluate theses methods by their accuracy in estimating both the skeleton and the CPDAG. We use the structural Hamming distance (SHD; Tsamardinos et al., 2006) to measure the efficiency of estimating the CPDAG, and use the area under the curve (AUC) of the receiver operating characteristic curve (ROC) to measure the estimation efficiency of the skeleton. As a fully nonparametric approach, we expect DAG to perform better in capturing dependence structures in non-Gaussian DAGs. A dimension reduction approach, which mitigates the curse of dimensionality, we expect our method to perform well for larger networks.

To demonstrate these features, we generate both Gaussian and non-Gaussian DAGs. Similar to Kalisch and Bühlman (2007), we generate a $p \times p$ dimensional adjacency matrix D as follows: after deciding on the topological ordering among nodes, we fill the lower-diagonal elements of D with 0s or 1s according to the Bernoulli distribution with parameter s, which denotes the probability of success. Here, s can be regarded as a sparsity parameter, and the expected number of neighbors of each node i, denoted by $\mathbb{E}[N_i]$, is $s(p-1)$. Given the adjacency matrix D, we generate a

p-dimensional random vector $X = (X^1, \ldots, X^p)$ sequentially via

$$\begin{aligned}
&\text{Step I}: \ \epsilon^1, \ldots, \epsilon^p \overset{\text{i.i.d}}{\sim} N(0,1),\\
&\text{Step II}: \ X^1 = \epsilon^1,\\
&\text{Step III}: \ \text{Gaussian case} \ X^i = \sum_{j=1}^{i-1} d_{ij} X^j + \epsilon^i,\\
&\qquad\qquad \text{non-Gaussian case} \ X^i = \sum_{j=1}^{i-1} d_{ij} \cos(X^j) + \epsilon^i,\\
&\qquad\qquad \text{Mixed non-Gaussian case}: \ X^i = \sum_{j=1}^{i-1} d_{ij} f_{ij}(X^j) + \sum_{(j,k)\in P(i)} e_{ijk} h_{jk}(X^j, X^k) + \epsilon^i,
\end{aligned}$$

where d_{ij} is the (i,j)-th element of the adjacency matrix D, f_{ij} and h_{jk} are randomly selected from a bank of functions, including $\cos(x)$, x^2, $4/(1+\exp(-x))$, to allow for diverse causal mechanisms. We select e_{ijk} randomly from Uniform$([0,1])$. In the Mixed non-Gaussian case, the error terms ϵ^i are sampled from $N(0, \sigma_i^2)$, where σ_i^2 is chosen randomly from Uniform$[0.5, 2]$. We also introduce random interactions among parent nodes. We determine the number of interactions randomly based on the network size p, using a random number generator. This ensures that the complexity of the network scales appropriately with its size, with the number of interactions increasing as p grows. This allows the model to represent more intricate causal dynamics. To secure a comprehensive comparison, we choose the sample size and the network size to be the same as Gaussian and non-Gaussian cases. To make a comprehensive comparison, we choose the sample size n to be 100, 150 and 200 and the network size p to be 5, 10 and 50. We consider two scenarios where the expected values $\mathbb{E}[N_i]$ are 2 and 4, respectively

We choose the tuning parameters, η_n and ϵ_n in step 1, r_i and δ in step 2, and the threshold constant ρ using the procedures described in Section 4.4. Finally, since there is no available procedure for determining the dimension of the sufficient predictor, in the simulations, we experimented with both dimensions, $d_S^{ij} = 1$ and $d_S^{ij} = 2$, which lead to very similar result. For this reason, we present only the results for $d_S^{ij} = 1$.

6.1 Estimation of CPDAG(G)

In this subsection, we compare the estimated CPDAG to the true CPDAG. We use SHD to quantify the directional information in the CPDAG. SHD measures the extent to which the estimated CPDAG captures the true CPDAG by penalizing addition, deletion, and re-orientation of the estimated CPDAG to match the true CPDAG. Lower value SHD indicates higher accuracy. Methods A and B were implemented using the pcalg package (Kalisch et al., 2012) in R, while for Method C, we used the kpcalg package (Verbyla et al., 2017) in R. We repeated each model 10 times and plotted the boxplots of SHD in Figures 1 to 6.

Figure 1 shows the results for the Gaussian and $\mathbb{E}[N_i] = 2$ scenario. As expected, both linear methods (A and B) outperform the nonparametric methods (C and D) in this Gaussian scenario. We also observe that our Method D outperforms the Kernel-PC algorithm method (Method C), which shows the benefit of dimension reduction. Furthermore, for the higher dimension $p = 50$, our method D demonstrates improved efficiency relative to the other three methods, which again demonstrates the benefit of dimension reduction.

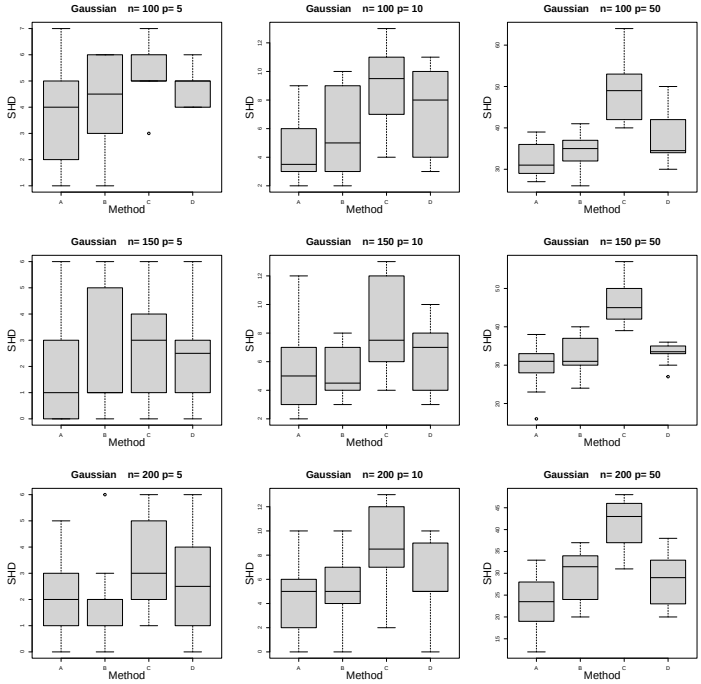

Figure 1: Comparisons of SHD values for Gaussian settings under $\mathbb{E}[N_i] = 2$.

Gaussian		$\mathbb{E}[N_i] = 2$			$\mathbb{E}[N_i] = 4$		
n	Method	$p = 5$	$p = 10$	$p = 50$	$p = 5$	$p = 10$	$p = 50$
100	A	0.73(0.12)	0.81(0.06)	0.79(0.02)	0.65(0.04)	0.66(0.06)	0.73(0.02)
	B	0.71(0.09)	0.80(0.09)	0.78(0.01)	0.67(0.06)	0.64(0.05)	0.72(0.02)
	C	0.59(0.06)	0.65(0.06)	0.67(0.02)	0.61(.0.3)	0.58(0.03)	0.61(0.01)
	D	0.64(0.04)	0.75(0.08)	0.74(0.02)	0.64(0.3)	0.63(0.05)	0.67(0.02)
150	A	0.88(0.12)	0.82(0.08)	0.80(0.04)	0.66(0.04)	0.68(0.06)	0.76(0.01)
	B	0.85(0.14)	0.83(0.05)	0.81(0.03)	0.66(0.04)	0.69(0.06)	0.75(0.01)
	C	0.77(0.13)	0.69(0.07)	0.69(0.02)	0.61(0.03)	0.62(0.03)	0.63(0.01)
	D	0.81(0.14)	0.75(0.06)	0.77(0.02)	0.65(0.05)	0.65(0.05)	0.70(0.01)
200	A	0.86(0.09)	0.85(0.08)	0.85(0.03)	0.64(0.02)	0.72(0.05)	0.76(0.02)
	B	0.85(0.07)	0.83(0.06)	0.84(0.03)	0.64(0.04)	0.72(0.06)	0.76(0.02)
	C	0.76(0.07)	0.70(0.10)	0.73(0.02)	0.61(0.03)	0.62(0.05)	0.64(0.02)
	D	0.80(0.08)	0.80(0.08)	0.81(0.03)	0.62(0.03)	0.69(0.04)	0.71(0.03)

Table 1: Comparison of AUC values between the true and estimated graph for Gaussian settings with $\mathbb{E}[N_i] = 2, 4$.

The results for the case with $\mathbb{E}[N_i] = 4$ are presented in Figure 2. We observe that Methods A, B, and C perform slightly better than the DAG (Method D) in this case.

The non-Gaussian and sparse scenarios are shown in Figure 3. We see that our method substantially outperforms the other three methods in all cases, particularly for higher dimensions. The

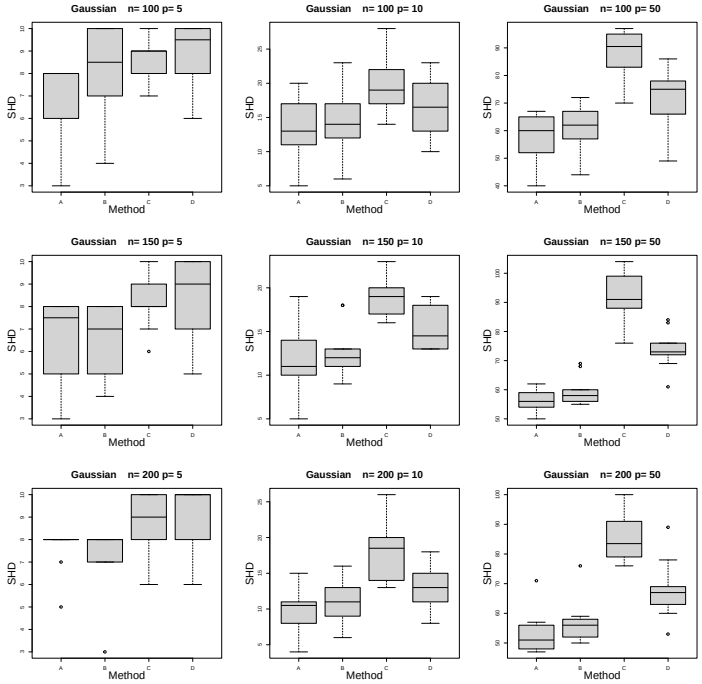

Figure 2: Comparisons of SHD values for Gaussian settings under $\mathbb{E}[N_i] = 4$.

Non-Gaussian		$\mathbb{E}[N_i] = 2$			$\mathbb{E}[N_i] = 4$		
n	Method	$p = 5$	$p = 10$	$p = 50$	$p = 5$	$p = 10$	$p = 50$
100	A	0.57(0.09)	0.54(0.03)	0.53(0.01)	0.54(0.03)	0.53(0.54)	0.54(0.01)
	B	0.55(0.09)	0.53(0.03)	0.53(0.01)	0.55(0.02)	0.54(0.02)	0.55(0.01)
	C	0.59(0.09)	0.57(0.02)	0.53(0.01)	0.56(0.02)	0.55(0.02)	0.54(0.01)
	D	0.62(0.08)	0.57(0.04)	0.56(0.01)	0.57(0.03)	0.56(0.03)	0.57(0.01)
150	A	0.52(0.03)	0.52(0.03)	0.55(0.02)	0.54(0.02)	0.52(0.03)	0.55(0.01)
	B	0.51(0.02)	0.53(0.05)	0.55(0.01)	0.55(0.02)	0.55(0.05)	0.57(0.02)
	C	0.59(0.08)	0.61(0.05)	0.58(0.02)	0.57(0.02)	0.56(0.02)	0.57(0.01)
	D	0.61(0.11)	0.61(0.07)	0.60(0.01)	0.60(0.03)	0.58(0.04)	0.61(0.01)
200	A	0.54(0.06)	0.53(0.03)	0.54(0.01)	0.55(0.03)	0.54(0.03)	0.56(0.01)
	B	0.54(0.06)	0.53(0.03)	0.54(0.01)	0.56(0.03)	0.56(0.04)	0.58(0.02)
	C	0.66(0.13)	0.59(0.03)	0.61(0.02)	0.59(0.03)	0.61(0.04)	0.61(0.02)
	D	0.70(0.11)	0.61(0.04)	0.63(0.02)	0.61(0.04)	0.63(0.03)	0.64(0.02)

Table 2: Comparison of AUC values between the true and estimated graph for non-Gaussian settings with $\mathbb{E}[N_i] = 2, 4$.

same pattern of comparison is also seen in Figure 4, which presents the non-Gaussian and the case of $\mathbb{E}[N_i] = 4$.

Figures 5 and 6 present the outcomes under the mixed non-Gaussian setting. For the complex and realistic causal mechanisms under a non-Gaussian scenario, we included random parent interactions, multiple nonlinearity, and randomly sampled error variances. As we can observe from

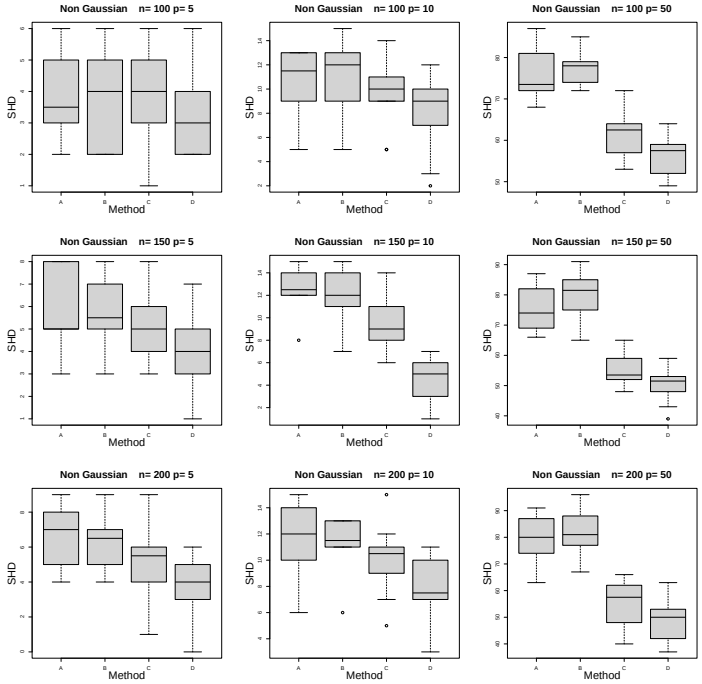

Figure 3: Comparisons of SHD values for non-Gaussian settings under $\mathbb{E}[N_i] = 2$.

mixed Model		$E[N_i] = 2$			$E[N_i] = 4$		
n	Method	$p = 5$	$p = 10$	$p = 50$	$p = 5$	$p = 10$	$p = 50$
100	A	0.51(0.05)	0.53(0.03)	0.54(0.01)	0.53(0.02)	0.50(0.02)	0.55(0.01)
	B	0.51(0.07)	0.53(0.03)	0.54(0.01)	0.55(0.02)	0.51(0.02)	0.56(0.01)
	C	0.56(0.04)	0.56(0.05)	0.53(0.01)	0.56(0.03)	0.54(0.03)	0.54(0.01)
	D	0.58(0.07)	0.55(0.04)	0.56(0.02)	0.56(0.02)	0.56(0.03)	0.57(0.01)
150	A	0.51(0.07)	0.54(0.03)	0.53(0.02)	0.55(0.02)	0.53(0.04)	0.57(0.01)
	B	0.50(0.06)	0.53(0.03)	0.54(0.02)	0.56(0.03)	0.55(0.05)	0.58(0.01)
	C	0.60(0.08)	0.57(0.05)	0.56(0.01)	0.59(0.03)	0.57(0.03)	0.58(0.01)
	D	0.65(0.15)	0.60(0.08)	0.59(0.01)	0.60(0.04)	0.58(0.04)	0.62(0.01)
200	A	0.56(0.07)	0.54(0.04)	0.53(0.02)	0.56(0.02)	0.54(0.05)	0.53(0.01)
	B	0.55(0.06)	0.57(0.03)	0.54(0.02)	0.56(0.01)	0.54(0.05)	0.54(0.01)
	C	0.60(0.06)	0.58(0.05)	0.61(0.03)	0.58(0.02)	0.58(0.04)	0.60(0.02)
	D	0.62(0.05)	0.62(0.05)	0.64(0.01)	0.62(0.04)	0.61(0.03)	0.63(0.03)

Table 3: Comparison of AUC values between the true and estimated graph for mixed Non-Gaussian settings with $\mathbb{E}[N_i] = 2, 4$.

Figures 5 and 6, the results demonstrate that our proposed method, Method D, consistently outperforms Methods A, B, and C, especially in high-dimensional settings. For both $\mathbb{E}[N_i] = 2$ and $\mathbb{E}[N_i] = 4$, Method D demonstrates strong performance across all dimensions. However, the advantage of Method D becomes especially noticeable in the high-dimensional cases, where the complexity of the causal graph makes accurate estimation more challenging. These result highlight the

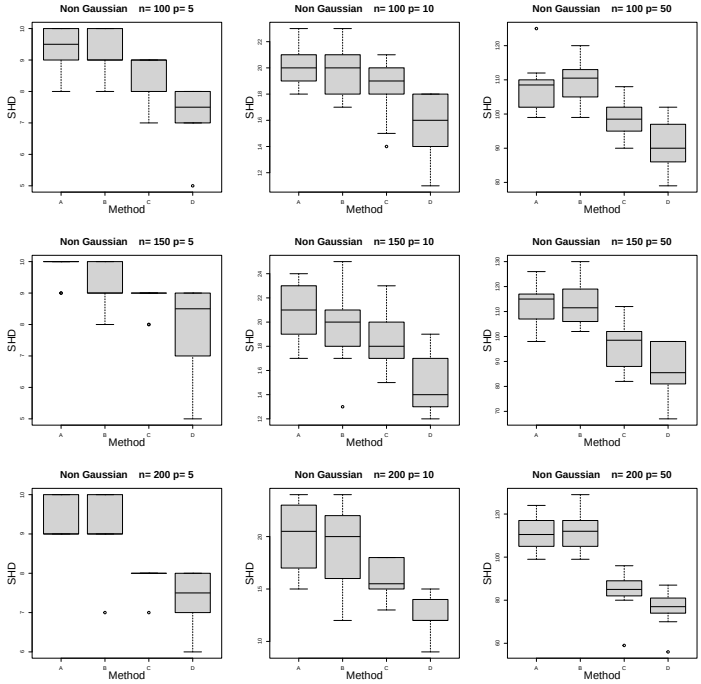

Figure 4: Comparisons of SHD values for non-Gaussian settings under $\mathbb{E}[N_i] = 4$.

benefit of dimension reduction of our method in maintaining high accuracy as the dimensionality grows. Methods A, B, and C show a decline in performance in high-dimensions.

In Tables 1, 2, and 3, we present the AUC values to evaluate the performance of the different methods in various settings while SHD is commonly used in causal structure learning, it is known to be less effective in sparse settings ($\mathbb{E}[N_i] = 2$) settings due to its inability to distinguish between different types of errors such as false positives and false negatives. This is particularly evident in high-dimensional cases where estimators may perform similarly in terms of SHD. AUC provides a more balanced evaluation of these trade-offs by considering both true positives and false positives.

In Table 1, where the distribution is Gaussian, while Method D shows only slightly lower AUC values compared than Methods A and B, it outperforms the nonparametric approach, Method C. This result demonstrates that Method D captures the causal structure more effectively than a non-parametric approach. Tables 2 and 3, which focus on non-Gaussian and mixed non-Gaussian settings, show that Method D consistently outperforms Methods A, B, and C. This performance in non-Gaussian environments reconfirms the advantage of our approach in scenarios where more complex, nonlinear relationships exist between variables. Finally, the lower SHD and higher AUC for Method D in these settings indicate that it is reliable in accurately recovering the underlying causal structure.

6.2 Estimation of ske(G)

In Tables 4, 5 and 6 we report the averaged AUC, to assess the performance of estimating true skeleton based on different methods when $\mathbb{E}[N_i] = 2$ and $\mathbb{E}[N_i] = 4$. Table 4 presents results in the Gaussian setting. Here, although Method D achieves slightly lower AUC values than Methods A

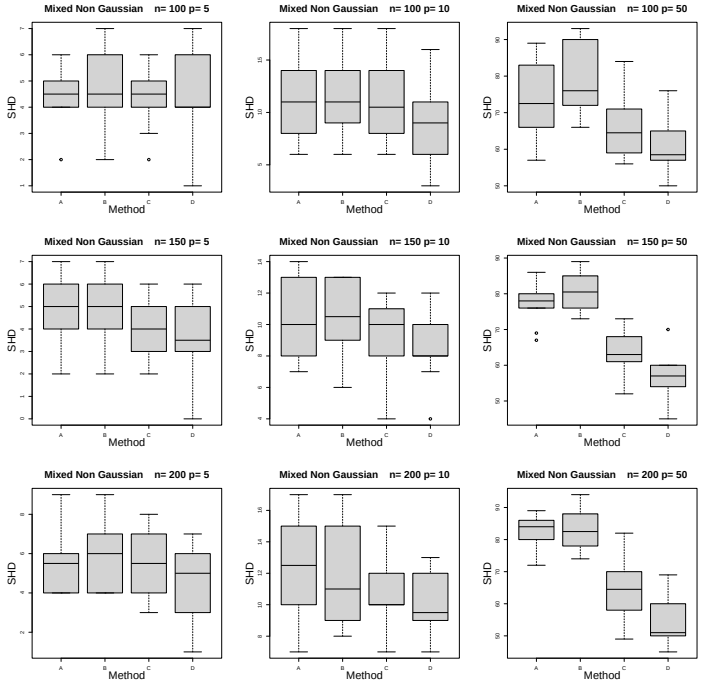

Figure 5: Comparisons of SHD values for mixed Non-Gaussian settings under $\mathbb{E}[N_i] = 2$.

and B, it still outperforms Method C. This suggests that Method D remains competitive even when assumptions are more favorable to Methods A and B. The results indicate that even in Gaussian settings, Method D performs better than fully nonparametric approaches, showing robustness as we move between $\mathbb{E}[N_i] = 2$ and $\mathbb{E}[N_i] = 4$.

In Table 5, under the non-Gaussian scenario, Method D outperforms Methods A, B, and C, achieving higher AUC values for both $\mathbb{E}[N_i] = 2$ and $\mathbb{E}[N_i] = 4$. This indicates that Method D can capture nonlinear relationships more effectively than the other methods. Method C, while performing better here than in the Gaussian case, still falls short of Method D, demonstrating that our sufficient dimension reduction based method is competitive at complex settings. Finally, in Table 6, which is a mixed non-Gaussian model, Method D again outperforms the other methods.

6.3 Data application

We apply our method to the flow cytometry dataset (Sachs et al., 2005), which includes simultaneously measured $p = 11$ phosphorylated phosphoproteins and phospholipids in the single cell. This dataset can be downloaded from

$$https://github.com/fernandoPalluzzi/SEMgraph.$$

The goal of this application was to demonstrate that our method can accurately discover causal relationships in the latent signaling network. To demonstrate the effectiveness of our approach, we applied Methods A, B, C, and D to $n = 90$ observations, and compared the estimated CPDAG with the true CPDAG. The true DAG can be found in (Sachs et al., 2005). We repeated this subsampling procedure 10 times and reported the mean and standard deviation of the SHD values in Table 7.

Gaussian		$\mathbb{E}[N_i] = 2$			$\mathbb{E}[N_i] = 4$		
n	Method	$p = 5$	$p = 10$	$p = 50$	$p = 5$	$p = 10$	$p = 50$
100	A	0.85(0.07)	0.86(0.05)	0.86(0.03)	0.75(0.06)	0.72(0.05)	0.75(0.02)
	B	0.86(0.07)	0.86(0.07)	0.85(0.03)	0.75(0.05)	0.70(0.04)	0.75(0.03)
	C	0.75(0.14)	0.71(0.07)	0.72(0.02)	0.62(0.05)	0.64(0.04)	0.64(0.02)
	D	0.81(0.10)	0.79(0.07)	0.81(0.03)	0.70(0.04)	0.67(0.05)	0.70(0.02)
150	A	0.87(0.06)	0.89(0.06)	0.89(0.02)	0.78(0.05)	0.75(0.05)	0.80(0.03)
	B	0.88(0.06)	0.89(0.06)	0.89(0.02)	0.77(0.04)	0.75(0.05)	0.80(0.03)
	C	0.77(0.08)	0.78(0.08)	0.78(0.03)	0.63(0.05)	0.67(0.05)	0.65(0.04)
	D	0.84(0.08)	0.82(0.14)	0.85(0.03)	0.74(0.06)	0.73(0.06)	0.63(0.05)
200	A	0.95(0.07)	0.91(0.07)	0.90(0.02)	0.80(0.05)	0.80(0.06)	0.80(0.01)
	B	0.93(0.09)	0.91(0.07)	0.90(0.02)	0.80(0.05)	0.80(0.06)	0.81(0.01)
	C	0.90(0.08)	0.78(0.05)	0.81(0.03)	0.70(0.04)	0.68(0.05)	0.67(0.07)
	D	0.92(0.08)	0.84(0.05)	0.87(0.03)	0.78(0.05)	0.75(0.06)	0.76(0.01)

Table 4: Averaged AUC values of estimating true skeleton for Gaussian setting under $\mathbb{E}[N_i] = 2, 4$.

Non-Gaussian		$\mathbb{E}[N_i] = 2$			$\mathbb{E}[N_i] = 4$		
n	Method	$p = 5$	$p = 10$	$p = 50$	$p = 5$	$p = 10$	$p = 50$
100	A	0.53(0.07)	0.52(0.04)	0.54(0.02)	0.55(0.04)	0.56(0.03)	0.56(0.01)
	B	0.51(0.08)	0.53(0.06)	0.55(0.02)	0.55(0.03)	0.57(0.04)	0.56(0.01)
	C	0.63(0.10)	0.63(0.07)	0.58(0.04)	0.57(0.03)	0.60(0.04)	0.58(0.02)
	D	0.59(0.12)	0.62(0.08)	0.61(0.02)	0.59(0.06)	0.62(0.05)	0.60(0.02)
150	A	0.55(0.08)	0.55(0.09)	0.57(0.03)	0.57(0.05)	0.55(0.05)	0.59(0.02)
	B	0.55(0.07)	0.55(0.09)	0.57(0.02)	0.58(0.05)	0.56(0.05)	0.60(0.02)
	C	0.68(0.08)	0.69(0.08)	0.64(0.02)	0.65(0.04)	0.62(0.05)	0.63(0.02)
	D	0.71(0.08)	0.70(0.05)	0.67(0.03)	0.63(0.05)	0.63(0.05)	0.67(0.02)
200	A	0.54(0.08)	0.60(0.06)	0.55(0.03)	0.54(0.08)	0.58(0.05)	0.58(0.02)
	B	0.57(0.08)	0.60(0.07)	0.57(0.03)	0.60(0.07)	0.59(0.06)	0.59(0.03)
	C	0.79(0.10)	0.75(0.10)	0.65(0.04)	0.67(0.07)	0.67(0.05)	0.65(0.02)
	D	0.78(0.07)	0.78(0.10)	0.70(0.03)	0.68(0.07)	0.69(0.07)	0.68(0.02)

Table 5: Averaged AUC values of estimating true skeleton for non-Gaussian setting under $\mathbb{E}[N_i] = 2, 4$

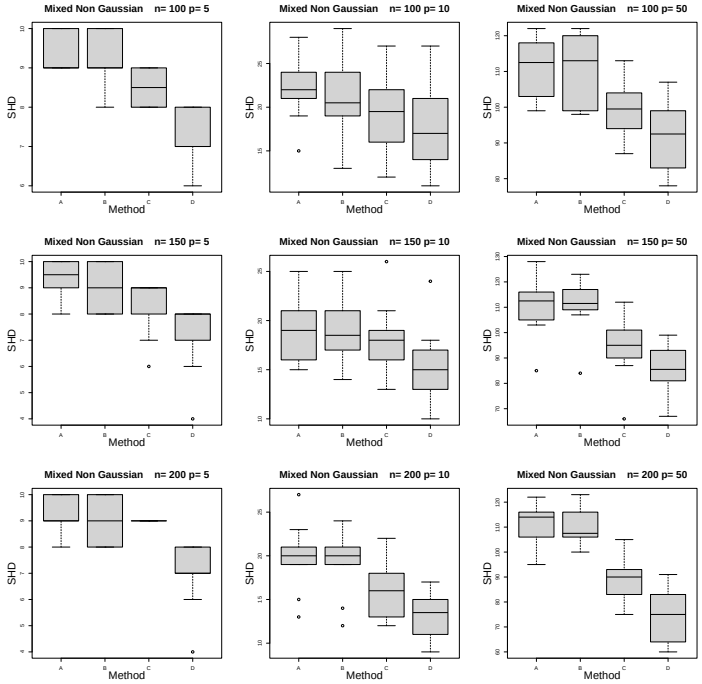

Figure 6: Comparisons of SHD values for mixed Non-Gaussian settings under $\mathbb{E}[N_i] = 4$.

Mixed Model		$\mathbb{E}[N_i] = 2$			$\mathbb{E}[N_i] = 4$		
n	Method	$p = 5$	$p = 10$	$p = 50$	$p = 5$	$p = 10$	$p = 50$
100	A	0.65(0.15)	0.63(0.06)	0.60(0.03)	0.54(0.04)	0.55(0.04)	0.56(0.01)
	B	0.59(0.11)	0.62(0.06)	0.59(0.04)	0.53(0.04)	0.56(0.04)	0.57(0.01)
	C	0.67(0.12)	0.68(0.04)	0.62(0.03)	0.58(0.04)	0.60(0.03)	0.57(0.01)
	D	0.70(0.13)	0.70(0.05)	0.66(0.03)	0.60(0.06)	0.62(0.02)	0.60(0.01)
150	A	0.69(0.09)	0.66(0.10)	0.63(0.04)	0.57(0.04)	0.54(0.04)	0.59(0.01)
	B	0.71(0.07)	0.66(0.12)	0.62(0.04)	0.58(0.04)	0.56(0.04)	0.59(0.02)
	C	0.76(0.11)	0.74(0.08)	0.68(0.04)	0.62(0.03)	0.63(0.03)	0.61(0.01)
	D	0.78(0.07)	0.75(0.08)	0.73(0.04)	0.63(0.04)	0.65(0.04)	0.65(0.01)
200	A	0.62(0.10)	0.65(0.10)	0.62(0.04)	0.55(0.04)	0.54(0.03)	0.59(0.02)
	B	0.58(0.10)	0.64(0.11)	0.62(0.04)	0.59(0.05)	0.54(0.03)	0.60(0.01)
	C	0.77(0.12)	0.78(0.06)	0.72(0.03)	0.66(0.05)	0.64(0.03)	0.64(0.01)
	D	0.81(0.11)	0.81(0.05)	0.75(0.03)	0.67(0.09)	0.67(0.04)	0.69(0.01)

Table 6: Averaged AUC values of estimating true skeleton for mixed non Gaussian setting under $\mathbb{E}[N_i] = 2, 4$.

Our proposed method (Method D) has the lowest SHD values, indicating its competitiveness for investigating causal relations among cells.

Method	A	B	C	D
SHD	21.2(1.31)	23.6(1.07)	21.1(1.91))	20.6(0.79)

Table 7: Comparison of the mean and standard deviation of SHD values among four methods, with the standard deviation presented in parentheses.

7. Discussion

In this paper, we introduce a novel nonparametric methodology, called DAG, for estimating a DAG using the PC-algorithm. This methodology applies nonlinear sufficient dimension reduction to replace the high-dimensional conditioning random variable with a low-dimensional variable in the evaluation of conditional independence when estimating the skeleton of the DAG with the PC-algorithm. The proposed approach achieves a substantial gain in estimation accuracy, particularly in high-dimensional settings, where the number of variables is large compared to the sample size. Our method also can capture nonlinear relations without requiring any distributional or linear structural assumptions. Under a reasonably mild set of conditions on the kernel, we established the consistency and the convergence rates of the estimators, taking into account the error in the estimated sufficient predictors $\widehat{U}^{ij,S}$. We also established the uniform consistency of the DAG under a strong faithfulness assumption. We developed a new PC-algorithm that integrates dimension reduction, whose complexity does not depend on the number of nodes but instead on the level of sparseness of the DAG. We illustrate the methodology by a variety of simulation studies and real data analysis using a flow cytometry dataset.

The idea advanced in this paper opens up a number of possibilities for further development. First, we have used the Hilbert-Schmidt norm of the conjoined conditional cross-covariance operator as the dependence measure of conditional independence. However, one could construct the normalized conjoined conditional cross-covariance operator similar to the NCCO of Fukumizu et al. (2007b), which removes the effect of marginal variation when evaluating interdependence. Second, our methodology assumes that the central class is complete and sufficient, so that GSIR is exhaustive and able to recover it. However, when the central class is not complete, GSIR is no longer exhaustive and there is no guarantee that it will recover the central class fully. In this case, we can employ GSAVE to recover a larger portion of the central class than GSIR (Lee et al., 2013; Li, 2018b). Third, one could use more sophisticated sparse penalty techniques than threshold, such as the LASSO (Tibshirani, 1996), and the adaptive LASSO (Zou, 2006). Finally, another possible extension of the current method is the situations where the observations on each vertex are random functions. This type of network structure is commonly encountered in medical applications such as electroencephalography and functional magnetic resonance imaging, where the data collected represents functional measurements. Recently, various undirected functional graphical models have been proposed in the literature Zhu et al. (2016); Qiao et al. (2019); Li and Solea (2018); Solea and Li (2022). However, learning causal relationships is fundamental in many disciplines such as genetics, epidemiology and finance. Therefore, in our future research, we will consider constructing a nonparametric functional DAG whose observations on the vertices are random functions. To address the curse of the dimensionality in the evaluation of conditional independence, a possible approach would be to utilize the functional generalized sliced inverse regression technique developed by Li and Song (2017).

Acknowledgments

We are grateful to the Editor, Action Editor, and two referees for their insightful comments and valuable suggestions, which have significantly improved our manuscript. Kyongwon Kim, the corresponding author, was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.RS-2023-00219212, RS-2025-00513476).

References

Steen A Andersson, David Madigan, and Michael D Perlman. A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526 – 2556, 2014. doi: 10.1214/14-AOS1260. URL https://doi.org/10.1214/14-AOS1260.

Pei-Chun Chen, Kuang-Yao Lee, Tsung-Ju Lee, Yuh-Jye Lee, and Su-Yun Huang. Multiclass support vector classification via coding and regression. *Neurocomputing*, 73(7-9):1501–1512, 2010.

David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null):507–554, mar 2003. ISSN 1532-4435. doi: 10.1162/153244303321897717. URL https://doi.org/10.1162/153244303321897717.

John B Conway. *A Course in Functional Analysis*, volume 96. Springer, 2019.

R Dennis Cook and Sanford Weisberg. Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

Gregory F. Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 116–125, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606149.

Mathias Drton and Marloes H Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

Daniel Eaton and Kevin Murphy. Exact bayesian structure learning from uncertain interventions. In *Artificial intelligence and statistics*, pages 107–114. PMLR, 2007.

Frederick Eberhardt. Almost optimal intervention sets for causal discovery. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, page 161–168, Arlington, Virginia, USA, 2008. AUAI Press. ISBN 0974903949.

Frederick Eberhardt, Patrik Hoyer, and Richard Scheines. Combining experiments to discover linear cyclic models with latent variables. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 185–192. JMLR Workshop and Conference Proceedings, 2010.

Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *J. Mach. Learn. Res.*, 5:73–99, dec 2004. ISSN 1532-4435.

Kenji Fukumizu, Francis R. Bach, and Arthur Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(14):361–383, 2007a. URL http://jmlr.org/papers/v8/fukumizu07a.html.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007b. URL https://proceedings.neurips.cc/paper/2007/file/3a0772443a0739141292a5429b952fe6-Paper.pdf.

Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871 – 1905, 2009. doi: 10.1214/08-AOS637. URL https://doi.org/10.1214/08-AOS637.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Algorithmic Learning Theory*, pages 63–77, Berlin, Heidelberg, 2005a. Springer Berlin Heidelberg. ISBN 978-3-540-31696-1.

Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, Bernhard Schölkopf, et al. Kernel methods for measuring independence. 2005b.

Arthur Gretton, Peter Spirtes, and Robert Tillman. Nonlinear directed acyclic structure learning with weakly additive noise models. *Advances in neural information processing systems*, 22, 2009.

Naftali Harris and Mathias Drton. Pc algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(11), 2013.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13 (1):2409–2464, 2012.

Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9(Nov):2523–2547, 2008.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

Tailen Hsing and Randall Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons, 2015.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11 (5), 2010.

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H Maathuis, and Peter Bühlmann. Causal inference using graphical models with the r package pcalg. *Journal of statistical software*, 47:1–26, 2012.

Kyongwon Kim. On principal graphical models with application to gene network. *Computational Statistics and Data Analysis*, 166:107344, 2022. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2021.107344. URL https://www.sciencedirect.com/science/article/pii/S016794732100178X.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Gustavo Lacerda, Peter Spirtes, Joseph Ramsey, and Patrik O. Hoyer. Discovering cyclic causal models by independent components analysis. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, page 366–374, Arlington, Virginia, USA, 2008. AUAI Press. ISBN 0974903949.

Kuang-Yao Lee, Bing Li, and Francesca Chiaromonte. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics*, 41(1):221–249, 2013.

Kuang-Yao Lee, Bing Li, and Hongyu Zhao. On an additive partial correlation operator and nonparametric estimation of graphical models. *Biometrika*, 103(3):513–530, 2016a.

Kuang-Yao Lee, Bing Li, and Hongyu Zhao. Variable selection via additive conditional independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1037–1055, 2016b.

Kuang-Yao Lee, Tianqi Liu, Bing Li, and Hongyu Zhao. Learning causal networks via additive faithfulness. *Journal of Machine Learning Research*, 21(51):1–38, 2020. URL http://jmlr.org/papers/v21/16-252.html.

Yuh-Jye Lee and Su-Yun Huang. Reduced support vector machines: A statistical theory. *IEEE Transactions on neural networks*, 18(1):1–13, 2007.

Bing Li. Linear operator-based statistical analysis: A useful paradigm for big data. *Canadian Journal of Statistics*, 46(1):79–103, 2018a.

Bing Li. *Sufficient dimension reduction: Methods and applications with R*. Chapman and Hall/CRC, 2018b.

Bing Li and Kyongwon Kim. On sufficient graphical models. *Journal of Machine Learning Research*, 25(17):1–64, 2024.

Bing Li and Eftychia Solea. A nonparametric graphical model for functional data with application to brain networks based on fmri. *Journal of the American Statistical Association*, 113(524):1637–1655, 2018.

Bing Li and Jun Song. Nonlinear sufficient dimension reduction for functional data. *The Annals of Statistics*, 45(3):1059–1095, 2017.

Bing Li, Andreas Artemiou, and Lexin Li. Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, 39(6):3182–3210, 2011.

Bing Li, Hyonho Chun, and Hongyu Zhao. On an additive semigraphoid model for statistical networks with application to pathway analysis. *Journal of the American Statistical Association*, 109(507):1188–1204, 2014.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(10), 2009.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 403–410, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859.

Preetam Nandy, Alain Hauser, and Marloes H Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018.

Si Peng, Xiaotong Shen, and Wei Pan. Reconstruction of a directed acyclic graph with intervention. *Electronic journal of statistics*, 14(2):4133, 2020.

Jonas Peters. On the intersection property of conditional independence and its application to causal discovery. *Journal of Causal Inference*, 3(1):97–108, 2015.

Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014. URL `http://jmlr.org/papers/v15/peters14a.html`.

Xinghao Qiao, Shaojun Guo, and Gareth M. James. Functional graphical models. *Journal of the American Statistical Association*, 114:211–222, 2019.

Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

Alberto Roverato. A unified approach to the characterization of equivalence classes of dags, chain graphs with no flags and chain graphs. *Scandinavian Journal of Statistics*, 32(2):295–312, 2005.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Kayvan Sadeghi. Faithfulness of probability distributions and graphs. *J. Mach. Learn. Res.*, 18(1): 5429–5457, jan 2017. ISSN 1532-4435.

Shohei Shimizu. Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

Shohei Shimizu and Aapo Hyvärinen. Discovery of linear non-gaussian acyclic models in the presence of latent classes. In *International Conference on Neural Information Processing*, pages 752–761. Springer, 2007.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Eftychia Solea and Bing Li. Copula gaussian graphical models for functional data. *Journal of the American Statistical Association*, 117(538):781–793, 2022. doi: 10.1080/01621459.2020.1817750. URL https://doi.org/10.1080/01621459.2020.1817750.

Peter Spirtes, Clark Glymour, and Richard Scheines. Causation, prediction, and search. adaptive computation and machine learning series. *The MIT Press*, 49:77–78, 2000.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, 2:67–93, mar 2002. ISSN 1532-4435. doi: 10.1162/153244302760185252. URL https://doi.org/10.1162/153244302760185252.

Xiaohai Sun. *Causal inference from statistical data*. PhD thesis, Karlsruhe, Univ., Diss., 2008, 2008.

Xiaohai Sun, Dominik Janzing, Bernhard Schölkopf, and Kenji Fukumizu. A kernel-based causal learning algorithm. In *Proceedings of the 24th international conference on Machine learning*, pages 855–862, 2007.

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. 2007.

Jin Tian and Judea Pearl. Causal discovery from changes. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, page 512–521, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Simon Tong and Daphne Koller. Active learning for structure in bayesian networks. In *International joint conference on artificial intelligence*, volume 17, pages 863–869. Citeseer, 2001.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65:31–78, 2006.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436 – 463, 2013. doi: 10.1214/12-AOS1080. URL https://doi.org/10.1214/12-AOS1080.

Sara van de Geer and Peter Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536 – 567, 2013. doi: 10.1214/13-AOS1085. URL https://doi.org/10.1214/13-AOS1085.

P Verbyla, NIB Desgranges, and L Wernisch. kpcalg: Kernel pc algorithm for causal structure detection. *URL https://CRAN. R-project. org/package= kpcalg, r package version*, 1(1), 2017.

Thomas S Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 221–236. 2022.

Yu Wang. *Nonlinear dimension reduction in feature space*. The Pennsylvania State University, 2008.

Joachim Weidmann. *Linear operators in Hilbert spaces*, volume 68. Springer Science & Business Media, 2012.

Han-Ming Wu. Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17(3):590–610, 2008.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard 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, page 804–813, Arlington, Virginia, USA, 2011. AUAI Press. ISBN 9780974903972.

Hongxiao Zhu, Nate Strawn, and David B. Dunson. Bayesian graphical models for multivariate functional data. *Journal of Machine Learning Research*, 14:1–27, 2016.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Appendix

Basics from bounded linear operators in Hilbert spaces

First, we recall some useful relations for bounded linear operators. For concrete references, see Weidmann (2012) and Conway (2019). The following result is given in Proposition 2.7 in Conway (2019).

Lemma 14 *Let $\mathcal{H}$ be a Hilbert space. If $A \in \mathcal{B}(\mathcal{H})$, then its adjoint operator, A^*, is also bounded, and*

(a) $\|A^*\|_{\text{op}} = \|A\|_{\text{op}}$.

(b) $\|A^*A\|_{\text{op}} = \|A\|^2_{\text{op}}$.

Proof

(a) Let $x \in \mathcal{H}$ such that $\|x\| = 1$. Suppose that $A^*(x) \neq 0$. Then, by the definition of the adjoint operator and the Cauchy-Schwarz inequality,

$$\|A^*(x)\|^2 = \langle A^*(x), A^*(x)\rangle = \langle x, AA^*(x)\rangle \leq \|A\|_{\text{op}}\|A^*(x)\|.$$

Dividing by $\|A^*(x)\|$ on both sides, and taking the supremum over $\|x\| = 1$ yields $\|A^*\|_{\text{op}} \leq \|A\|_{\text{op}}$. Next, using the fact that $(A^*)^* = A$, we have by the application of the above arguments $\|A\|_{\text{op}} = \|(A^*)^*\|_{\text{op}} \leq \|A^*\|_{\text{op}}$. Combining these two results yields $\|A^*\|_{\text{op}} = \|A\|_{\text{op}}$.

(b) First, for every $x \in \mathcal{H}$ such that $\|x\| = 1$,

$$\|A^*A(x)\| \leq \|A^*\|_{\text{op}}\|A(x)\| \leq \|A^*\|_{\text{op}}\|A\|_{\text{op}}.$$

Thus, A^*A is bounded and $\|A^*A\|_{\text{op}} \leq \|A^*\|_{\text{op}}\|A\|_{\text{op}} = \|A\|_{\text{op}}\|A\|_{\text{op}} = \|A\|^2_{\text{op}}$ because of the result in (a). Now, by the sub-multiplicative property of the operator norm, and using the result in (a),

$$\|A\|^2_{\text{op}} = \|A\|_{\text{op}}\|A\|_{\text{op}} = \|A^*\|_{\text{op}}\|A\|_{\text{op}} \geq \|A^*A\|_{\text{op}}.$$

Therefore, $\|A^*A\|_{\text{op}} = \|A\|^2_{\text{op}}$.

The following result is given in Theorem 6.9 in Weidmann (2012).

Lemma 15 *If $A \in \mathcal{B}_2(\mathcal{H})$, then*

$$\|A\|_{\text{op}} \leq \|A\|_{\text{HS}} = \|A^*\|_{\text{HS}}.$$

Proof By Parseval's identity,

$$\begin{aligned}\|A^*(x)\|^2 = \textstyle\sum_{j=1}^{\infty}\langle A^*(x), \epsilon_j\rangle^2 &= \textstyle\sum_{j=1}^{\infty}\langle (x), A(\epsilon_j)\rangle^2 \\ &\leq \|x\|^2\textstyle\sum_{j=1}^{\infty}\|A(\epsilon_j)\|^2 = \|x\|^2\|A\|^2_{\text{HS}},\end{aligned}$$

where the inequality is due to the Cauchy-Schwarz inequality. Therefore, $\|A^*\|_{\text{op}} \leq \|A\|_{\text{HS}}$. Now use the fact that $\|A^*\|_{\text{op}} = \|A\|_{\text{op}}$ from Lemma 14(a) to complete the proof. The following inequality is well known-see (Weidmann, 2012).

Lemma 16 *Let $\mathcal{H}_1$, $\mathcal{H}_2$ and $\mathcal{H}_3$ be Hilbert spaces. If $B \in \mathcal{B}(\mathcal{H}_2, \mathcal{H}_1)$ and $A \in \mathcal{B}_2(\mathcal{H}_2, \mathcal{H}_3)$, then $BA \in \mathcal{B}_2(\mathcal{H}_3, \mathcal{H}_1)$, and*

$$\|BA\|_{\text{HS}} \leq \|B\|_{\text{op}}\|A\|_{\text{HS}}.$$

Proof Let $\{\epsilon_n\}_{n\in\natural}$ be an orthonormal basis in $\mathcal{H}$. By definition of the Hilbert-Schmidt norm,

$$\|BA\|^2_{\text{HS}} = \textstyle\sum_{n=1}^{\infty}\|BA(\epsilon_n)\|^2 \leq \|B\|^2_{\text{op}}\sum_{n=1}^{\infty}\|A(\epsilon_n)\|^2 = \|B\|^2_{\text{op}}\|A\|^2_{\text{HS}},$$

yielding the desired result. The following property is given in Fukumizu et al. (2007a) (Lemma 6).

Lemma 17 *If A and B are self-adjoint and invertible operators, then*

$$A^{-1/2} - B^{-1/2} = \{A^{-1/2}(B^{3/2} - A^{3/2}) + A - B\}B^{-3/2}. \tag{26}$$

The following Lemma is taken from Fukumizu et al. (2007a).

Lemma 18 *Suppose A and B are self-adjoint operators in a Hilbert space such that $0 \leq A \leq cI$ and $0 \leq B \leq cI$ for some positive constant c. Then,*

$$\|A^{3/2} - B^{3/2}\|_{\text{op}} \leq 3c^{1/2}\|A - B\|_{\text{op}}. \tag{27}$$

Basics of cross-covariance operators in RKHS

The following lemmas are critical to our subsequent developments. The following Lemma is due to Fukumizu et al. (2007a).

Lemma 19 *Suppose Assumption 2 holds and Assumption 8 is satisfied with X^S replaced by $U^{ij,S}$ for all $i, j \in \mathsf{V}$, $i \neq j$ and $S \subset V \setminus \{i, j\}$. Then,*

(a) $\|\hat{\Sigma}_{X^iX^j} - \Sigma_{X^iX^j}\|_{\mathrm{HS}} = O_p(n^{-1/2}), \|\hat{\Sigma}_{U^{ij,S}U^{ij,S}} - \Sigma_{U^{ij,S}U^{ij,S}}\|_{\mathrm{HS}} = O_p(n^{-1/2})$,

(b) $\|\hat{\Sigma}_{(X^iU^{ij,S})(X^jU^{ij,S})} - \Sigma_{(X^iU^{ij,S})(X^jU^{ij,S})}\|_{\mathrm{HS}} = O_p(n^{-1/2})$,

(c) $\|\hat{\Sigma}_{(X^iU^{ij,S})U^{ij,S}} - \Sigma_{(X^iU^{ij,S})U^{ij,S}}\|_{\mathrm{HS}} = O_p(n^{-1/2}), \|\hat{\Sigma}_{U^{ij,S}(X^jU^{ij,S})} - \Sigma_{U^{ij,S}(X^jU^{ij,S})}\|_{\mathrm{HS}} = O_p(n^{-1/2})$.

Lemma 20 *Suppose Assumption 2 holds and $\mathbb{E}\kappa_{U^{ij,S}}(U^{ij,S}, U^{ij,S}) < \infty$. Then, for any $n^{-1/2} \prec \delta_n \prec 1$,*

(a) $\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\|_{\mathrm{op}} = \|(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1}\|_{\mathrm{op}} = \|(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\|_{\mathrm{op}} = O_p(\delta_n^{-1})$,

(b) $\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\Sigma^{1/2}_{U^{ij,S}U^{ij,S}}\|_{\mathrm{op}} = O_p(\delta_n^{-1/2})$,

(c) $\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\Sigma_{U^{ij,S}(X^jU^{ij,S})}\|_{\mathrm{op}} = O_p(\delta_n^{-1/2})$,

(d) $\|\Sigma_{(X^iU^{ij,S})U^{ij,S}}(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1/2}\|_{\mathrm{op}} = O(1)$.

Proof

(a) This is because $\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\|_{\mathrm{op}} \leq \|(\delta_n I)^{-1}\| \leq \delta_n^{-1}$. This relation also holds if $\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I$ is replaced by $\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I$ or $\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I$.

(b) It is equivalent to show $\|\Sigma^{1/2}_{U^{ij,S}U^{ij,S}}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1})\|_{\mathrm{op}} = O_p(\delta_n^{-1/2})$. By using the fact that $\Sigma^{1/2}_{U^{ij,S}U^{ij,S}}$ is a self-adoint operator in $\mathcal{B}_2\mathcal{H}_{U^{ij,S}}$ and the Cauchy-Schwarz inequality, for any $f \in \mathcal{H}_{U^{ij,S}}$ such that $\|f\| \leq 1$,

$$
\begin{aligned}
&\|\Sigma^{1/2}_{U^{ij,S}U^{ij,S}}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1} f\|^2 \\
&= \langle(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1} f, \Sigma_{U^{ij,S}U^{ij,S}}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1} f\rangle \\
&\leq \|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\|_{\mathrm{op}}\|\Sigma_{U^{ij,S}U^{ij,S}}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\|_{\mathrm{op}} \\
&\leq \|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\|_{\mathrm{op}}[\|(\Sigma_{U^{ij,S}U^{ij,S}} - \hat{\Sigma}_{U^{ij,S}U^{ij,S}})(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\|_{\mathrm{op}} \\
&\quad + \|\hat{\Sigma}_{U^{ij,S}U^{ij,S}}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\|_{\mathrm{op}}.
\end{aligned}
$$

By (a), the first term on the right is $O_p(\delta_n^{-1})$. By Lemma 19 and (a) again, the second term is $O_p(n^{-1/2}\delta_n^{-1})$ and the third term is $O_p(1)$. Together with the assumption $n^{-1/2} \prec \delta_n \prec 1$, we get the desired bound.

(c) By definition

$$\begin{aligned}&\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1}\Sigma_{U^{ij,S}(X^jU^{ij,S})}\|_{\mathrm{op}}\\&\quad\le\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1}\Sigma^{1/2}_{U^{ij,S}U^{ij,S}}\|_{\mathrm{op}}\|R_{U^{ij,S}(X^jU^{ij,S})}\Sigma^{1/2}_{(X^jU^{ij,S})(X^jU^{ij,S})}\|_{\mathrm{op}},\end{aligned}$$

where the first term on the right is $O_p(\delta_n^{-1/2})$ by (b) and the second term is $O(1)$.

(d) Similarly, dy definition $\Sigma_{(X^iU^{ij,S})U^{ij,S}}=\Sigma^{1/2}_{(X^iU^{ij,S})(X^iU^{ij,S})}R_{(X^iU^{ij,S})U^{ij,S}}\Sigma^{1/2}_{U^{ij,S}U^{ij,S}}$, one obtains

$$\begin{aligned}&\|\Sigma_{(X^iU^{ij,S})U^{ij,S}}(\Sigma_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}\|_{\mathrm{op}}\\&\quad\le\|\Sigma^{1/2}_{(X^iU^{ij,S})(X^iU^{ij,S})}R_{(X^iU^{ij,S})U^{ij,S}}\|_{\mathrm{op}}\|\Sigma^{1/2}_{U^{ij,S}U^{ij,S}}(\Sigma_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}\|_{\mathrm{op}},\end{aligned}$$

which has order $O(1)$.

Proof of Theorem 9

First, by the triangle inequality we make the following decomposition

$$\begin{aligned}&\|\hat{\Sigma}_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})|\hat{U}^{ij,S}}-\Sigma_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}}\\&\quad\le\|\hat{\Sigma}_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})|\hat{U}^{ij,S}}-\hat{\Sigma}_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}}\\&\qquad+\|\hat{\Sigma}_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}}-\Sigma_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}}.\end{aligned}$$

Next, we derive the convergence rates of each term on the right-hand side.

Lemma 21 *Suppose the conditions of Theorem 7 are satisfied such that* $\|\hat{U}^{ij,S}-U^{ij,S}\|=O_p(b_n)$, *where* $b_n=\eta_n^{-3/2}\epsilon_n^{-1}n^{-1}+\eta_n^{-1}n^{-1/2}+\eta_n+\epsilon_n\prec 1$. *Suppose, furthermore, that Assumptions 9 and 10 hold for the reproducing kernels* $\kappa_{X^iU^{ij,S}}$, $\kappa_{X^jU^{ij,S}}$ *and* $\kappa_{U^{ij,S}}$. *Then,*

$$\|\hat{\Sigma}_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})|\hat{U}^{ij,S}}-\hat{\Sigma}_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}}=O_p(\delta_n^{-1}b_n).$$

Proof By definition and the triangular identity, the norm $\|\hat{\Sigma}_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})|\hat{U}^{ij,S}}-\hat{\Sigma}_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}}$ is decomposed by $\Theta_{1,n}+\Theta_{2,n}$, where

$$\begin{aligned}\Theta_{1,n}=&\|\hat{\Sigma}_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})}-\hat{\Sigma}_{(X^i\hat{U}^{ij,S})(X^j\hat{U}^{ij,S})}\|_{\mathrm{HS}}\\\Theta_{2,n}=&\|\hat{\Sigma}_{(X^iU^{ij,S})U^{ij,S}}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1}\hat{\Sigma}_{U^{ij,S}(X^jU^{ij,S})}\\&-\hat{\Sigma}_{(X^i\hat{U}^{ij,S})\hat{U}^{ij,S}}(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}}+\delta_n I)^{-1}\hat{\Sigma}_{\hat{U}^{ij,S}(X^j\hat{U}^{ij,S})}\|_{\mathrm{HS}}.\end{aligned}$$

Next, we derive the convergence rates of $\Theta_{1,n}$ and $\Theta_{2,n}$, respectively. First, by Theorem 8, relation (iii),

$$\Theta_{1,n}=O_p(b_n).\tag{28}$$

Next, we consider $\Theta_{2,n}$, which is upper bounded by $\Theta_{21,n} + \Theta_{22,n} + \Theta_{23,n}$, where

$$\begin{aligned}
\Theta_{21,n} &= \|(\hat{\Sigma}_{(X^i\hat{U}^{ij,S})\hat{U}^{ij,S}} - \hat{\Sigma}_{(X^iU^{ij,S})U^{ij,S}})(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1}\hat{\Sigma}_{\hat{U}^{ij,S}(X^j\hat{U}^{ij,S})}\|_{\mathrm{HS}}, \\
\Theta_{22,n} &= \|\hat{\Sigma}_{(X^iU^{ij,S})U^{ij,S}}[(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1} - (\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}]\Sigma_{\hat{U}^{ij,S}(X^j\hat{U}^{ij,S})}\|_{\mathrm{HS}}, \quad (29) \\
\Theta_{23,n} &= \|\hat{\Sigma}_{(X^iU^{ij,S})U^{ij,S}}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}(\Sigma_{\hat{U}^{ij,S}(X^j\hat{U}^{ij,S})} - \hat{\Sigma}_{U^{ij,S}(X^jU^{ij,S})})\|_{\mathrm{HS}}.
\end{aligned}$$

Next, we derive the convergence rates of $\Theta_{21,n}$, $\Theta_{22,n}$ and $\Theta_{23,n}$, respectively.

First, $\Theta_{21,n}$ is upper bounded by

$$\begin{aligned}
&\|\hat{\Sigma}_{(X^i\hat{U}^{ij,S})\hat{U}^{ij,S}} - \hat{\Sigma}_{(X^iU^{ij,S})U^{ij,S}}\|_{\mathrm{HS}} \\
&\times \big[\|(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1}(\hat{\Sigma}_{\hat{U}^{ij,S}(X^j\hat{U}^{ij,S})} - \hat{\Sigma}_{U^{ij,S}(X^jU^{ij,S})})\|_{\mathrm{op}} \\
&+ \|(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1}\hat{\Sigma}_{U^{ij,S}(X^jU^{ij,S})})\|_{\mathrm{op}}\big]
\end{aligned}$$

By Theorem 8, the first norm on the right-hand side is of order $O_p(b_n)$. By Theorem 8 again and Lemma 20, relation (a), the second norm is of order $O_p(\delta_n^{-1}b_n)$. The third term $\|(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1}\hat{\Sigma}_{U^{ij,S}(X^jU^{ij,S})})\|_{\mathrm{op}}$ is further bounded by

$$\begin{aligned}
&\|(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1}(\hat{\Sigma}_{U^{ij,S}(X^jU^{ij,S})} - \Sigma_{U^{ij,S}(X^jU^{ij,S})})\|_{\mathrm{op}} \\
&\quad + \|(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1}\Sigma_{U^{ij,S}(X^jU^{ij,S})}\|_{\mathrm{op}},
\end{aligned} \quad (30)$$

in which the first term is of order $O_p(\delta_n n^{-1/2})$ by Lemma 19 and Lemma 20, relation (c). For the second term $\|(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1}\Sigma_{U^{ij,S}(X^jU^{ij,S})}\|_{\mathrm{op}}$ at the right-hand side of (30) we have

$$\begin{aligned}
&\|(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1}\Sigma_{U^{ij,S}(X^jU^{ij,S})}\|_{\mathrm{op}} \\
&\quad \le \|[(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1} - (\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}]\Sigma_{U^{ij,S}(X^jU^{ij,S})}\|_{\mathrm{op}} \\
&\qquad + \|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\Sigma_{U^{ij,S}(X^jU^{ij,S})}\|_{\mathrm{op}}
\end{aligned}$$

By using the identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, the first term at the right-hand side is

$$\begin{aligned}
&\|(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1}(\hat{\Sigma}_{U^{ij,S}\hat{U}^{ij,S}} - \hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}})(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\Sigma_{U^{ij,S}(X^jU^{ij,S})}\|_{\mathrm{op}} \\
&\le \|(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1}\|\|\hat{\Sigma}_{U^{ij,S}\hat{U}^{ij,S}} - \hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}}\|_{\mathrm{HS}}\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\Sigma_{U^{ij,S}(X^jU^{ij,S})}\|_{\mathrm{op}}
\end{aligned}$$

which is of order $O_p(\delta^{-3/2}b_n)$ by the Theorem 8 again and Lemma 20, relations (a) and (c). Hence,

$$\|(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1}\Sigma_{U^{ij,S}(X^jU^{ij,S})}\|_{\mathrm{op}} = O_p(\delta^{-3/2}b_n) + O_p(\delta^{-1/2}). \quad (31)$$

Therefore,

$$\Theta_{21,n} \le O_p(b_n)[O_p(\delta_n^{-1}b_n) + O_p(\delta_n n^{-1/2}) + O_p(\delta_n^{-3/2}b_n) + O_p(\delta_n^{-1/2})] = O_p(\delta_n^{-1/2}b_n) \quad (32)$$

by the conditions $b_n \prec \delta_n^{3/2}$, $b_n \prec 1$ and $\delta_n \prec 1$. For $\Theta_{22,n}$, first note that

$$\begin{aligned}
&\Theta_{22,n} \le \|\hat{\Sigma}_{(X^iU^{ij,S})U^{ij,S}}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1/2}\|_{\mathrm{op}} \\
&\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{1/2}[(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1} - (\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}](\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{1/2}\|_{\mathrm{HS}} \\
&\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1/2}\hat{\Sigma}_{\hat{U}^{ij,S}(X^j\hat{U}^{ij,S})}\|_{\mathrm{op}}.
\end{aligned} \quad (33)$$

Now, the first norm $\|\hat{\Sigma}_{(X^iU^{ij,S})U^{ij,S}}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}\|_{\text{op}}$ at the right-hand side of (33) is upper bounded by

$$\begin{aligned}&\|(\hat{\Sigma}_{(X^iU^{ij,S})U^{ij,S}}-\Sigma_{(X^iU^{ij,S})U^{ij,S}})(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}\|_{\text{op}}\\&\quad+\|\Sigma_{(X^iU^{ij,S})U^{ij,S}}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}\|_{\text{op}},\end{aligned}\tag{34}$$

where the first norm is of order $O_p(\delta_n^{-1/2}n^{-1/2})$ by Lemma 19 and Lemma 20, relation (a). Now, by the definition $\Sigma_{(X^iU^{ij,S})U^{ij,S}}=\Sigma^{1/2}_{(X^iU^{ij,S})(X^iU^{ij,S})}R_{(X^iU^{ij,S})U^{ij,S}}\Sigma^{1/2}_{U^{ij,S}U^{ij,S}}$, the second term at the right-hand side of (34) is

$$\begin{aligned}&\|\Sigma_{(X^iU^{ij,S})U^{ij,S}}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}\|_{\text{op}}\\&\le\|\Sigma^{1/2}_{(X^iU^{ij,S})(X^iU^{ij,S})}R_{(X^iU^{ij,S})U^{ij,S}}\|_{\text{op}}\|\Sigma^{1/2}_{U^{ij,S}U^{ij,S}}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}\|_{\text{op}}\end{aligned}\tag{35}$$

where the first norm at the right-hand side is $O(1)$. Now, for the second term at the right-hand side of (35) we have

$$\begin{aligned}&\|\Sigma^{1/2}_{U^{ij,S}U^{ij,S}}[(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}-\Sigma_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}]\|_{\text{op}}\\&\quad+\|\Sigma^{1/2}_{U^{ij,S}U^{ij,S}}(\Sigma_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}\|_{\text{op}},\end{aligned}\tag{36}$$

where the second term is $O(1)$ since $\Sigma_{U^{ij,S}U^{ij,S}}\le\Sigma_{U^{ij,S}U^{ij,S}}+\delta_n I$. By Lemma 17 and Lemma 18, for the first term at the right-hand side of (36), we obtain

$$\begin{aligned}&\|\hat{\Sigma}^{-1/2}_{U^{ij,S}U^{ij,S}}[(\Sigma_{U^{ij,S}U^{ij,S}}+\delta_I)^{3/2}-(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_I)^{3/2}](\Sigma_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1}\|_{\text{op}}\\&\quad+\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}-\Sigma_{U^{ij,S}U^{ij,S}})(\Sigma_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1}\|_{\text{op}}\\&\le 3c^{1/2}\delta_n^{-3/2}\|\hat{\Sigma}_{U^{ij,S}U^{ij,S}}-\Sigma_{U^{ij,S}U^{ij,S}}\|_{\text{HS}}+3c^{1/2}\delta_n^{-1}\|\hat{\Sigma}_{U^{ij,S}U^{ij,S}}-\Sigma_{U^{ij,S}U^{ij,S}}\|_{\text{HS}}\\&=O_p(\delta_n^{-3/2}n^{-1/2})=O_p(1),\end{aligned}$$

by condition $n^{-1/2}\prec\delta_n^{3/2}$. Hence, using this fact and together with (35) we have shown

$$\|\hat{\Sigma}_{(X^iU^{ij,S})U^{ij,S}}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}\|_{\text{op}}=O_p(1).\tag{37}$$

Similarly, the third norm $\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}\hat{\Sigma}_{\hat{U}^{ij,S}(X^j\hat{U}^{ij,S})}\|_{\text{op}}$ at the right-hand side of (33) is bounded by

$$\begin{aligned}&\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}(\hat{\Sigma}_{\hat{U}^{ij,S}(X^j\hat{U}^{ij,S})}-\hat{\Sigma}_{U^{ij,S}(X^jU^{ij,S})})\|_{\text{op}}\\&\quad+\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}\hat{\Sigma}_{U^{ij,S}(X^jU^{ij,S})}\|_{\text{op}},\end{aligned}\tag{38}$$

in which the first norm is $O_p(\delta_n^{-1/2}b_n)$ by Theorem 8 and Lemma 20, relation (a). The second norm at the right-hand side of (38) is further bounded by

$$\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}(\hat{\Sigma}_{U^{ij,S}(X^jU^{ij,S})}-\Sigma_{U^{ij,S}(X^jU^{ij,S})})\|_{\text{op}}+\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}\Sigma_{U^{ij,S}(X^jU^{ij,S})}\|_{\text{op}}$$

which is of order $O_p(\delta_n^{-1/2}n^{-1/2})+O_p(1)=O_p(1)$, by (37). Combining the results, yields, $\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1/2}\hat{\Sigma}_{\hat{U}^{ij,S}(X^j\hat{U}^{ij,S})}\|_{\text{op}}=O_p(1)$. Next we derive the convergence rate of the second norm at the right-hand side of (33). By calculations,

$$\begin{aligned}&\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{1/2}[(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}}+\delta_n I)^{-1}-(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1}](\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{1/2}\|_{\text{op}}\\&=\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{1/2}(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}}+\delta_n I)^{-1}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{1/2}-I\|_{\text{HS}}.\end{aligned}\tag{39}$$

Now, because $AA^* - I$ and $A^*A - I$ have the same spectrum, the norm on the right of (39) can be re-written as

$$\begin{aligned}&\|(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1/2}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1/2} - I\|_{\mathrm{HS}} \\ &= \|(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1/2}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} - \hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}})(\hat{\Sigma}_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1/2}\|_{\mathrm{HS}},\end{aligned}$$

which by Theorem 8 and Lemma 20, relation (a) is of order $O_p(\delta_n^{-1}b_n)$.

Hence, $\Theta_{22,n} \leq O_p(\delta_n^{-1}b_n)$. Regarding, $\Theta_{23,n}$, we can obtain,

$$\begin{aligned}\Theta_{23,n} \leq O_p(b_n)\big[&\|(\hat{\Sigma}_{(X^iU^{ij,S})U^{ij,S}} - \Sigma_{(X^iU^{ij,S})U^{ij,S}})(\hat{\Sigma}_{U^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1}\|_{\mathrm{op}} \\ &+ \|\Sigma_{(X^iU^{ij,S})U^{ij,S}}(\Sigma_{\hat{U}^{ij,S}\hat{U}^{ij,S}} + \delta_n I)^{-1}\|_{\mathrm{op}}\big].\end{aligned}$$

By Theorem 8, $\Theta_{23,n}$ is of order $O_p(b_n)[O_p(\delta_n^{-1}n^{-1/2}) + O_p(\delta_n^{-1/2})] = O_p(b_n\delta_n^{-1/2})$ by Lemmas 19, 20 and conditions $b_n \prec 1$ and $n^{-1/2} \prec \delta_n \prec 1$.

Hence, by combining the results for $\Theta_{21,n}$, $\Theta_{22,n}$ and $\Theta_{23,n}$, we obtain that $\Theta_{2,n}$ is of order $O_p(\delta_n^{-1}b_n)$. This, together with (28) and the condition $b_n \prec 1$ lead to the desired result.

Lemma 22 *Suppose the conditions of Theorem 7 are satisfied such that $\|\hat{U}^{ij,S} - U^{ij,S}\| = O_p(b_n)$, where $b_n = \eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n + \epsilon_n \prec 1$, and the Assumptions 9 and 10 hold for the reproducing kernels $\kappa_{X^iU^{ij,S}}$, $\kappa_{X^jU^{ij,S}}$ and $\kappa_{U^{ij,S}}$. Suppose, furthermore, that Assumptions 11 holds, and*

$$n^{-1/2} \prec \eta_n \prec 1, \quad n^{-1/2} \prec \epsilon_n \prec 1, \quad n^{-1/2} \prec \delta_n^{3/2}, \quad b_n \prec \delta_n^{3/2},$$

where $\delta_n \prec 1$, $b_n = \eta_n^{-3/2}\epsilon_n^{-1}n^{-1} + \eta_n^{-1}n^{-1/2} + \eta_n \prec 1$. Then,

$$\|\hat{\Sigma}_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}} - \Sigma_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}} = O_p(\delta_n^{-1}n^{-1/2} + \delta_n^{1/2}). \tag{40}$$

Proof First, we introduce the intermediate operator

$$\Sigma^{\delta_n}_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}} = \Sigma_{(X^iU^{ij,S})(X^jU^{ij,S})} - \Sigma_{(X^iU^{ij,S})U^{ij,S}}(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\Sigma_{U^{ij,S}(X^jU^{ij,S})}.$$

Then, by the triangle inequality

$$\begin{aligned}&\|\hat{\Sigma}_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}} - \Sigma_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}} \\ &\quad\leq \|\hat{\Sigma}_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}} - \Sigma^{\delta_n}_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}} \\ &\qquad+ \|\Sigma^{\delta_n}_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}} - \Sigma_{(X^iU^{ij,S})(X^jU^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}} \\ &\quad= \Delta_{1,n} + \Delta_{2,n}.\end{aligned}$$

Next, we derive the convergence rates of $\Delta_{1,n}$ and $\Delta_{2,n}$, respectively. By the triangle inequality,

$$\Delta_{1,n} \leq \Delta_{11,n} + \Delta_{12,n},$$

where,

$$\begin{aligned}\Delta_{11,n} =&\|\hat{\Sigma}_{(X^iU^{ij,S})(X^jU^{ij,S})} - \Sigma_{(X^iU^{ij,S})(X^iU^{ij,S})}\|_{\mathrm{HS}}, \\ \Delta_{12,n} =&\|\Sigma_{(X^iU^{ij,S})U^{ij,S}}(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\Sigma_{U^{ij,S}(X^jU^{ij,S})} \\ &- \hat{\Sigma}_{(X^iU^{ij,S})U^{ij,S}}(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\hat{\Sigma}_{U^{ij,S}(X^jU^{ij,S})}\|_{\mathrm{HS}}.\end{aligned}$$

By Lemma 19, we immediately obtain $\Delta_{11,n} = O_p(n^{-1/2})$. The second term $\Delta_{12,n}$ is upper bounded by $\Delta_{121,n} + \Delta_{122,n} + \Delta_{123,n}$, where

$$\Delta_{121,n} = \|\hat{\Sigma}_{(X^i U^{ij,S})U^{ij,S}} - \Sigma_{(X^i U^{ij,S})U^{ij,S}})(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\hat{\Sigma}_{U^{ij,S}(X^j U^{ij,S})}\|_{\text{HS}}$$
$$\Delta_{122,n} = \|\Sigma_{(X^i U^{ij,S})U^{ij,S}}[(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1} - (\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}]\hat{\Sigma}_{U^{ij,S}(X^j U^{ij,S})}\|_{\text{HS}}$$
$$\Delta_{123,n} = \|\Sigma_{(X^i U^{ij,S})U^{ij,S}}(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}[\hat{\Sigma}_{U^{ij,S}(X^j U^{ij,S})} - \Sigma_{U^{ij,S}(X^j U^{ij,S})}]\|_{\text{HS}}.$$

We now derive the convergence rates of $\Delta_{121,n}, \Delta_{122,n}$ and $\Delta_{123,n}$, respectively. The term $\Delta_{121,n}$ is upper bounded by

$$\Delta_{121,n} \leq O_p(n^{-1/2})[\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}(\hat{\Sigma}_{U^{ij,S}(X^j U^{ij,S})} - \Sigma_{U^{ij,S}(X^j U^{ij,S})})\|_{op} + \|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}\Sigma_{U^{ij,S}(X^j U^{ij,S})}\|_{\text{op}}].$$

By Lemma 19 and Lemma 20(a), the norm $\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1}(\hat{\Sigma}_{U^{ij,S}(X^j U^{ij,S})}-\Sigma_{U^{ij,S}(X^j U^{ij,S})})\|_{\text{op}}$ is of order $O_p(\delta_n^{-1}n^{-1/2})$. By Lemma 20, again, relation (c), the term $\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}}+\delta_n I)^{-1}\Sigma_{U^{ij,S}(X^j U^{ij,S})}\|_{\text{op}}$ is of order $O_p(\delta_n^{-1/2})$. Therefore, by the fact that $n^{-1/2} \prec \delta_n \prec 1$,

$$\Delta_{121,n} = O_p(\delta_n^{-1}n^{-1}) + O_p(\delta_n^{-1/2}n^{-1/2}) = O_p(\delta_n^{-1/2}n^{-1/2}). \tag{41}$$

Regarding the term $\Delta_{122,n}$, we obtain the following decomposition

$$\begin{aligned}\Delta_{122,n} &\leq \|\Sigma_{(X^i U^{ij,S})U^{ij,S}}(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1/2}\|_{\text{op}} \\ &\times \|(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)^{1/2}[(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1} - (\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1}](\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)^{1/2}\|_{\text{HS}} \\ &\times \|(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1/2}\hat{\Sigma}_{U^{ij,S}(X^j U^{ij,S})}\|_{\text{op}}.\end{aligned} \tag{42}$$

By Lemma 20, relation (e), the first norm on the right hand side is of order $O(1)$. The third term on the right is

$$\begin{aligned}\|(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1/2}\hat{\Sigma}_{U^{ij,S}(X^j U^{ij,S})}\|_{\text{op}} &\leq \|(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1/2}(\hat{\Sigma}_{U^{ij,S}(X^j U^{ij,S})} - \Sigma_{U^{ij,S}(X^j U^{ij,S})})\|_{\text{op}} \\ &+ \|(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1/2}\Sigma_{U^{ij,S}(X^j U^{ij,S})}\|_{\text{op}} \\ &= O_p(\delta_n^{-1/2}n^{-1/2}) + O(1) = O_p(1)\end{aligned}$$

due to the fact that $n^{-1/2} \prec \delta_n \prec 1$. Using an argument similar to the term $\Theta_{2,n}$, the second norm on the right hand side of (42) is

$$\begin{aligned}&\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1/2}(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I)(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1/2} - I\|_{\text{HS}} \\ &\leq \|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1/2}\|\|\Sigma_{U^{ij,S}U^{ij,S}} - \hat{\Sigma}_{U^{ij,S}U^{ij,S}}\|_{\text{HS}}\|(\hat{\Sigma}_{U^{ij,S}U^{ij,S}} + \delta_n I)^{-1/2}\|\end{aligned}$$

which is of order $O_p(\delta_n^{-1}n^{-1/2})$ by Lemma 19 and Lemma 20, relation (a). Hence, $\Delta_{122,n} = O_p(\delta_n^{-1}n^{-1/2})$. Using similar arguments, we can obtain that the term $\Delta_{123,n}$ is $O_p(1)$. Thus, combining the results for $\Delta_{121,n}$, $\Delta_{122,n}$ and $\Delta_{123,n}$ we have that $\Delta_{12,n} = O_p(\delta_n^{-1}n^{-1/2})$. As a result,

$$\Delta_{1,n} = O_p(\delta_n^{-1}n^{-1/2}). \tag{43}$$

Next, we consider the term $\Delta_{2,n}$. By definition and Assumption 11

$$\begin{aligned}
&\Sigma_{(X^{i}U^{ij,S})(X^{j}U^{ij,S})|U^{ij,S}} - \Sigma^{\delta_n}_{(X^{i}U^{ij,S})(X^{j}U^{ij,S})|U^{ij,S}} \\
&= \Sigma_{(X^{i}U^{ij,S})U^{ij,S}}\big(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I\big)^{-1}\Sigma_{U^{ij,S}(X^{j}U^{ij,S})} - \Sigma_{(X^{i}U^{ij,S})U^{ij,S}}\Sigma^{-1}_{U^{ij,S}U^{ij,S}}\Sigma_{U^{ij,S}(X^{j}U^{ij,S})} \\
&= \Sigma_{(X^{i}U^{ij,S})U^{ij,S}}\big(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I\big)^{-1}\Sigma_{U^{ij,S}U^{ij,S}}T_{U^{ij,S}(X^{j}U^{ij,S})} - \Sigma_{(X^{i}U^{ij,S})U^{ij,S}}T_{U^{ij,S}(X^{j}U^{ij,S})} \\
&= -\delta_n\Sigma_{(X^{i}U^{ij,S})U^{ij,S}}\big(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I\big)^{-1}T_{U^{ij,S}(X^{j}U^{ij,S})}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\Delta_{2,n} &= \|\Sigma_{(X^{i}U^{ij,S})(X^{j}U^{ij,S})|U^{ij,S}} - \Sigma^{\delta_n}_{(X^{i}U^{ij,S})(X^{j}U^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}} \\
&= \delta_n\|\Sigma_{(X^{i}U^{ij,S})U^{ij,S}}\big(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I\big)^{-1}T_{U^{ij,S}(X^{j}U^{ij,S})}\|_{\mathrm{HS}} \\
&\le \delta_n\|\Sigma^{1/2}_{(X^{i}U^{ij,S})(X^{j}U^{ij,S})}R_{(X^{i}U^{ij,S})U^{ij,S}}\|_{\mathrm{HS}}\|\Sigma^{1/2}_{U^{ij,S}U^{ij,S}}\big(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I\big)^{-1/2}\|_{\mathrm{op}} \\
&\quad \times \|\big(\Sigma_{U^{ij,S}U^{ij,S}} + \delta_n I\big)^{-1/2}T_{U^{ij,S}(X^{j}U^{ij,S})}\|_{\mathrm{op}} \\
&= O(\delta_n)O(1)O(1)O(\delta_n^{-1/2}) = O(\delta_n^{1/2}).
\end{aligned} \tag{44}$$

The assertion follows by combining the results for $\Delta_{1,n}$ in (43) and $\Delta_{2,n}$ in (44).

Proof of Theorem 10

For convenience, let $\mathcal{C} = \{(i,j,S) \in \mathcal{Q} : X^i \perp\!\!\!\perp X^j \mid X^S\}$, where

$$\mathcal{Q} = \{(i,j,S) : i,j \in \mathsf{V}, i \neq j, S \subseteq \mathsf{V} \setminus \{i,j\}\}. \tag{45}$$

By Assumption 1, for every $(i,j,S) \in \mathcal{C}$, which also means $(i,j) \notin \mathrm{ske}(\mathsf{G})$ or $(i,j) \in \mathrm{ske}(\mathsf{G})^{\mathsf{C}}$, we have $\|\Sigma_{(X^{i}U^{ij,S})(X^{j}U^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}} = 0$. By Theorem 9 and the condition $\delta_n^{-1}b_n + \delta_n^{-1}n^{-1/2} + \delta_n^{1/2} \prec \rho_n \prec 1$, for any $(i,j) \notin \mathrm{ske}(\mathsf{G})$, $\mathbb{P}(\|\hat{\Sigma}_{(X^{i}\hat{U}^{ij,S})(X^{j}\hat{U}^{ij,S})|\hat{U}^{ij,S}}\|_{\mathrm{HS}} < \rho_n) \to 1$ as $n \to \infty$. Thus, by the definition of $\widehat{\mathrm{ske}(\mathsf{G})}(\eta_n, \epsilon_n, \delta_n, \rho_n)$, $\mathbb{P}((i,j) \notin \widehat{\mathrm{ske}(\mathsf{G})}(\eta_n, \epsilon_n, \delta_n, \rho_n)) \to 1$ as $n \to \infty$, which further implies that $\mathbb{P}(\mathrm{ske}(\mathsf{G})^{\mathsf{C}} \subseteq \widehat{\mathrm{ske}(\mathsf{G})}(\eta_n, \epsilon_n, \delta_n, \rho_n))^{\mathsf{C}}) \to 1$ as $n \to \infty$, or $\mathbb{P}(\widehat{\mathrm{ske}(\mathsf{G})}(\eta_n, \epsilon_n, \delta_n, \rho_n)) \subseteq \mathrm{ske}(\mathsf{G})) \to 1$ as $n \to \infty$.

We then show that $\mathrm{ske}(\mathsf{G}) \subseteq \widehat{\mathrm{ske}(\mathsf{G})}(\eta_n, \epsilon_n, \delta_n, \rho_n)$ with probability tending to 1. For any $(i,j) \in \mathrm{ske}(\mathsf{G})$, we have $\|\Sigma_{(X^{i}U^{ij,S})(X^{j}U^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}} > 2\rho_n > 0$ for all $S \subseteq \mathsf{V} \setminus \{i,j\}$. By Theorem 9 and the condition $\delta_n^{-1}b_n + \delta_n^{-1}n^{-1/2} + \delta_n^{1/2} \prec \rho_n \prec 1$, for any $(i,j) \in \mathrm{ske}(\mathsf{G})$, $\mathbb{P}(\|\hat{\Sigma}_{(X^{i}\hat{U}^{ij,S})(X^{j}\hat{U}^{ij,S})|U^{ij,S}}\|_{\mathrm{HS}} > \rho_n) \to 1$ as $n \to \infty$. Hence, by the definition of $\widehat{\mathrm{ske}(\mathsf{G})}(\eta_n, \epsilon_n, \delta_n, \rho_n)$, $\mathbb{P}((i,j) \in \widehat{\mathrm{ske}(\mathsf{G})}(\eta_n, \epsilon_n, \delta_n, \rho_n)) \to 1$ as $n \to \infty$, which further implies that $\mathbb{P}(\mathrm{ske}(\mathsf{G}) \subseteq \widehat{\mathrm{ske}(\mathsf{G})}(\eta_n, \epsilon_n, \delta_n, \rho_n)) \to 1$ as $n \to \infty$. Combining with the previous result, $\mathbb{P}(\widehat{\mathrm{ske}(\mathsf{G})}(\eta_n, \epsilon_n, \delta_n, \rho_n) \subseteq \mathrm{ske}(\mathsf{G})) \to 1$ as $n \to \infty$, we obtain the desired result. □

Proof of the Theorem 11

As mentioned in Section 2, to estimate the graphical structure of the CPDAG of the true graph, G, it is sufficient to estimate the correct skeleton and the separation sets, which are determined

by conditional independence. Hence, by Theorem 10 and the Continuous Mapping Theorem, $\mathbb{P}(\hat{C}_{ij} = C_{ij}) \to 1$ as $n \to \infty$, where $\hat{C}_{ij} = \{S \subset V \setminus \{i,j\} : (i,j,S) \notin \widehat{\text{ske}(\mathsf{G})}(\eta_n, \epsilon_n, \delta_n, \rho_n))\}$ and $C_{ij} = \{S \subset V \setminus \{i,j\} : (i,j,S) \notin \text{ske}(\mathsf{G}))\}$. In words, the probability of selecting the correct separation sets S tend to 1 as $n \to \infty$. Hence, the probability of estimating the true CPDAG tends to 1 as $n \to \infty$. □

Proof of the Theorem 13

We follow same arguments as in the proof of Theorem 24 in Lee et al. (2020). For convenience, for each $P \in \mathcal{P}$, let $\mathcal{I}(P) = \{(i,j,S) \in \mathcal{Q} : X^i \perp\!\!\!\perp X^j \,|\, X^S\}$ and $\widehat{I(P)} = \{(i,j,S) \in \mathcal{Q} : \|\hat{\Sigma}_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|U^{ij,S}}\|_{\text{HS}} \leq \rho_n\}$, where $\mathcal{Q}$ is defined in (45). Let also $\mathcal{D} = \{(i,j,S) \in \mathcal{Q} :$ i and j are d-separated by S in $\mathsf{G}\}$. It is sufficient to show that $\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_0} \mathbb{P}(\widehat{I(P)} \neq \mathcal{D}) = 0$. We define two events A_1 and A_2 as follows: $A_1 = \{\|\hat{\Sigma}_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|U^{ij,S}}\|_{\text{HS}} \geq \rho_n \ \forall (i,j,S) \notin \mathcal{D}\}$, and $A_2 = \{\|\hat{\Sigma}_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|U^{ij,S}}\|_{\text{HS}} < \rho_n \ \forall (i,j,S) \in \mathcal{D}\}$. Then, the event $\widehat{I(P)} \neq \mathcal{D}$ happens when the events A_1 and A_2 do not hold simultaneously. By the τ-strongly sufficient faithfulness condition (a), for each $P \in \mathcal{P}_0$, $\|\Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}}(P)\|_{\text{HS}} = 0$ for all $(i,j,S) \in \mathcal{D}$, and $\|\Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}}(P)\|_{\text{HS}} > \tau$ for all $(i,j,S) \notin \mathcal{D}$, for some $\tau > 0$. Hence, we have,

$$\begin{aligned}\mathbb{P}(\widehat{I(P)} \neq \mathcal{D}) = \mathbb{P}\big(&A_1 \text{ and } A_2 \text{ do not hold simultaneously },\\ &\|\Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}}(P)\|_{\text{HS}} = 0, \forall (i,j,S) \in \mathcal{D},\\ &\|\Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}}(P)\|_{\text{HS}} > \tau, \forall (i,j,S) \notin \mathcal{D}\big).\end{aligned}$$

Let S_n represent the event inside the parentheses on the right hand side. Then, for every $\epsilon > 0$ sufficiently small,

$$\begin{aligned}&\mathbb{P}(\widehat{I(P)} \neq \mathcal{D}) = \mathbb{P}(S_n)\\ &\leq \mathbb{P}\big(S_n, \|\hat{\Sigma}_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}} - \Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}}(P)\|_{\text{HS}} < \epsilon, \ \forall (i,j,S) \in \mathcal{Q}\big)\\ &\quad + \mathbb{P}\big(\|\hat{\Sigma}_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}} - \Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}}(P)\|_{\text{HS}} \geq \epsilon, \ \text{for some} (i,j,S) \in \mathcal{Q}\big)\\ &\leq \mathbb{P}\big(S_n, \|\hat{\Sigma}_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}} - \Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}}(P)\|_{\text{HS}} < \epsilon, \ \forall (i,j,S) \in \mathcal{Q}\big)\\ &\quad + \textstyle\sum_{(i,j,S)\in\mathcal{Q}} \mathbb{P}\big(\|\hat{\Sigma}_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}}(P) - \Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}}(P)\|_{\text{HS}} \geq \epsilon\big),\end{aligned}$$

where the last inequality follows by Boole's inequality. Taking the supremum over P and the limit superior, $\limsup_{n\to\infty}$, yields,

$$\begin{aligned}&\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}(\widehat{I(P)} \neq \mathcal{D})\\ &\leq \limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}(S_n, \|\hat{\Sigma}_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}} - \Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}}(P)\|_{\text{HS}} < \epsilon, \forall (i,j,S) \in \mathcal{Q})\\ &\quad + \limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \textstyle\sum_{(i,j,S)\in\mathcal{Q}} \mathbb{P}(\|\hat{\Sigma}_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}} - \Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}}(P)\|_{\text{HS}} \geq \epsilon),\end{aligned}$$

where the last limit on the right hand side is 0 due to condition (b) and the fact that the set $\mathcal{Q}$ is finite. This way, we then only need to show that the first limit on the right hand side is 0. Note

that the event $S_n \cap \{\|\hat{\Sigma}_{(X^i \hat{U}^{ij,S})(X^j \hat{U}^{ij,S})|\hat{U}^{ij,S}} - \Sigma_{(X^i U^{ij,S})(X^j U^{ij,S})|U^{ij,S}}(P)\|_{\mathrm{HS}} < \epsilon,\ \forall (i,j,S) \in \mathcal{Q}\}$ cannot happen for sufficiently small ϵ, and hence has probability equal to 0. Therefore, it follows immediately that

$$\limsup_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}(\widehat{I(P)} \neq \mathcal{D}) = 0.$$

Under the sufficient faithfulness condition 3, the skeleton $\mathrm{ske}(\mathsf{G})$ and its estimator $\widehat{\mathrm{ske}(\mathsf{G})}(\eta_n, \epsilon_n, \delta_n, \rho_n)$ are determined by $I(P)$ and $\widehat{I(P)}$, respectively, which implies

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}(\widehat{\mathrm{ske}(\mathsf{G})}(\eta_n, \epsilon_n, \delta_n, \rho_n) \neq \mathrm{ske}(\mathsf{G})) = 0.$$

Using a similar argument as with the proof of Theorem 11, we can obtain

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}(\widehat{\mathrm{CPDAG}}(\mathsf{G})(\eta_n, \epsilon_n, \delta_n, \rho_n) \neq \mathrm{CPDAG}(\mathsf{G})) = 0,$$

as desired.