Journal of Machine Learning Research 25 (2024) 1-48 Submitted 9/22; Revised 11/23; Published 3/24

Functional Directed Acyclic Graphs

Kuang-Yao Lee KUANG-YAO.LEE@TEMPLE.EDU
Department of Statistics, Operations, and Data Science
Temple University
Philadelphia, PA 19122, USA

Lexin Li LEXINLI@BERKELEY.EDU
Department of Biostatistics and Epidemiology
University of California at Berkeley
Berkeley, CA 94720, USA

Bing Li BXL9@PSU.EDU
Department of Statistics
Pennsylvannia State University
University Park, PA 16802, USA

Editor: Peter Spirtes

Abstract

In this article, we introduce a new method to estimate a directed acyclic graph (DAG) from multivariate functional data. We build on the notion of faithfulness that relates a DAG with a set of conditional independences among the random functions. We develop two linear operators, the conditional covariance operator and the partial correlation operator, to characterize and evaluate the conditional independence. Based on these operators, we adapt and extend the PC-algorithm to estimate the functional directed graph, so that the computation time depends on the sparsity rather than the full size of the graph. We study the asymptotic properties of the two operators, derive their uniform convergence rates, and establish the uniform consistency of the estimated graph, all of which are obtained while allowing the graph size to diverge to infinity with the sample size. We demonstrate the efficacy of our method through both simulations and an application to a time-course proteomic dataset.

Keywords: Graphical model, faithfulness, functional regression, linear operator, reproducing kernel Hilbert space, uniform consistency.

1. Introduction

In this article, we introduce a new method to estimate a directed acyclic graph (DAG) based on multivariate functional data. Functional graphical modeling is becoming increasingly important, as multivariate functional data, where the observations are sampled from a vector of random functions, are fast emerging in a wide variety of scientific applications. Examples include molecular network modeling based on time-course gene or phosphoprotein data (Hill et al., 2016), and brain effective connectivity analysis based on electrocorticography or functional magnetic resonance imaging data (Friston, 2011). A crucial problem in these applications is to investigate directional relations among the random functions, which is a challenging task. The DAG model offers a tractable solution, and yet functional DAG modeling is a relatively underdeveloped topic.

Consider a graph $\mathsf{G} = (\mathsf{V}, \mathsf{E})$, where $\mathsf{V} = \{1, \ldots, p\}$ denotes a set of vertices associated with a vector of random variables or functions, $X_1, \ldots, X_p$, and $\mathsf{E} \subseteq \{(i,j) \in \mathsf{V} \times \mathsf{V} : i \neq j\}$ a set of edges. Let $(\mathsf{V} \times \mathsf{V})_0$ denote the set $\{(i,j) \in \mathsf{V} \times \mathsf{V} : i \neq j\}$ that excludes the diagonal pairs. A pair $(i,j) \in \mathsf{E}$ denotes an edge between vertices i and j, and is said to be directed from i to j if $(j,i) \notin \mathsf{E}$. In this relation, i is called a parent of j, and j a child of i. If both (i,j) and (j,i) belong to E, then the edge is said to be undirected. If there is a directed path $i \to \ldots \to j$ from i to j, then i is called an ancestor of j, and j a descendant of i. A DAG is a graph that contains only directed edges and no directed cycles. If two edges meet head-to-head at a vertex i on a path, say $j \to i \leftarrow k$, then i is called a collider on the path. For $\mathsf{S} \subseteq \mathsf{V} \backslash \{i,j\}$, the vertices i and j are said to be d-connected by S if and only if there exists a path connecting i and j that satisfies: (i) every collider in the path is either in S or has a descendant in S, and (ii) no non-collider in the path is in S. By convention, (i) includes the cases where the path has no collider at all. Also note that the descendant of a collider in the path does not belong to the path. We say i and j are d-separated by S if they are not d-connected by S.

In the classical random variable setting, directional relations among the variables can be depicted by *faithfulness*. Formally, $X = (X_1, \ldots, X_p)^\intercal \in \mathbb{R}^p$ is said to be faithful with respect to a DAG G, if and only if the collection of the triplets $\{(i,j,\mathsf{S}) \in \mathcal{T} :$ i and j are d-separated by $\mathsf{S}\}$ is the same as $\{(i,j,\mathsf{S}) \in \mathcal{T} : X_i \perp\!\!\!\perp X_j \mid X_\mathsf{S}\}$, where $\perp\!\!\!\perp$ denotes statistical independence, and $\mathcal{T} = \{(i,j,\mathsf{S}) : (i,j) \in (\mathsf{V} \times \mathsf{V})_0, \mathsf{S} \in \mathsf{V} \backslash \{i,j\}\}$. Moreover, when X follows a multivariate Gaussian distribution, we have the equivalence that $X_i \perp\!\!\!\perp X_j \mid X_\mathsf{S} \Leftrightarrow \mathrm{cov}\,(X_i, X_j | X_\mathsf{S}) = 0$.

In the functional data setting, the notion of faithfulness is essentially the same. Specifically, suppose $X = (X_1, \ldots, X_p)^\intercal$ is a p-dimensional Gaussian random element in a Hilbert space of functions defined on an interval T in $\mathbb{R}$. We say X is *faithful* with respect to a DAG G, if and only if the following equivalence holds:

$$i \text{ and } j \text{ are d-separated by } \mathsf{S} \quad \Leftrightarrow \quad X_i \perp\!\!\!\perp X_j \mid X_\mathsf{S}. \tag{1}$$

The conditional independence in (1) is in terms of Hilbertian random elements, and is formally defined in Section 2. Through the faithfulness in (1), a DAG is associated with a collection of conditional independence relations among p random functions.

To evaluate the conditional independence in (1), we develop two linear operators, the conditional covariance operator (CCO), and the partial correlation operator (PCO). Under the assumption that the p-variate random function X follows a Gaussian distribution, the conditional independence can be completely characterized by CCO or PCO, in the sense that CCO or PCO is zero if and only if the conditional independence between the random functions holds true. This agrees with the classical result that the conditional independence between two Gaussian variables is equivalent to their conditional covariance or partial correlation being zero. Henceforth, CCO and PCO can be viewed as the functional counterparts of conditional covariance and partial correlation. We next estimate the DAG by repeatedly evaluating the Hilbert-Schmidt norms of CCO and PCO, whose computation is straightforward and only involves eigen-decomposition of linear operators. We further embed CCO and PCO in the commonly used PC-algorithm (Spirtes et al., 2000), and turn the NP-hard problem of evaluating conditional independence for every pair (i, j) and every possible subset $\mathsf{S} \subseteq \mathsf{V} \backslash \{i, j\}$ to a computationally efficient algorithm of order p^m, where m is the maximum degree of the DAG. We also establish the error bounds and the uniform convergence for the estimated CCO and PCO, as well as the uniform consistency of the estimated DAG. We carry out all these theoretical investigations by allowing the graph size to diverge to infinity along with the sample size.

We note that our theoretical analysis faces a number of challenges. The first is, when characterizing the conditional independence between two random functions, we need to deal with diverging numbers of random variables from the Karhunen-Loève expansions of the functions, which requires a much more involved asymptotic analysis than in the classical setting. The second challenge is, most existing concentration inequalities as well as their sufficient conditions have been tailored for high-dimensional random variables, rather than high-dimensional random functions. To fill this gap, we develop suitable new asymptotic tools, including functional versions of the sub-Gaussianity and Bernstein's inequality for Hilbert space-valued random elements. The third challenge is, because the covariances are replaced by the covariance operators, we need to establish several concentration bounds and uniform convergences for the relevant linear operators in the high-dimensional setting. In fact, the theoretical framework we develop here is fairly general, and we expect it to be useful in other functional data analysis settings as well, especially when the number of functions involved is large compared to the sample size. It is also noteworthy that there is some novelty in our presentation of the classical DAG theory: We describe the CPDAG as a member of the quotient space of the collection of all DAGs, which makes its subtle definition more transparent and explicit.

Our work is a natural step forward in the current research on statistical graphical modeling. The majority of existing solutions focus on undirected or directed graphical models for random variables. Notable examples of the former include Yuan and Lin (2007); Friedman et al. (2008); Ravikumar et al. (2011); Cai et al. (2011); Guo et al. (2011); Ren et al. (2015); Fan and Lv (2016); Liu et al. (2021), among others, whereas examples of the latter include Chickering (2002); Kalisch and Bühlmann (2007); Harris and Drton (2013); van de Geer and Bühlmann (2013); Li et al. (2020), among others. More recently, Zhu et al. (2016), Li and Solea (2018), Qiao et al. (2019), Qiao et al. (2020), Solea and Li (2020), and Zhao et al. (2022) extended undirected graphical models to random functions. Even though the undirected and directed graphs are related, we discuss their differences in Section 7.3.

Despite the recent progress, the problem of estimating functional directed graphical models remains largely underdeveloped. Gómez et al. (2021) reformulated DAG estimation as sparse function-on-function regression (Fan et al., 2015; Luo and Qi, 2016), but required the causal order is known a prior, which can be unrealistic in practice. Lee and Li (2022) relaxed this requirement and proposed to estimate DAG through a functional structural equation model (SEM) with two man steps: order determination, then sparse functional regression. Our proposal is considerably different in several ways. First, unlike Gómez et al. (2021), our method does not assume the order is known, and thus the problem is more challenging; see also van de Geer and Bühlmann (2013) for the differences between DAG estimation with and without a known order in the random variable setting. Second, our method is built upon the proposed functional PC-algorithm, and belongs to the category of independence-based solutions, whereas Lee and Li (2022) is built upon function-on-function regression, and belongs to the category of SEM-based approaches; see also Peters et al. (2014) for the differences of these two categories of solutions under the random variable setting. Due to this methodological difference, the subsequent asymptotic analysis becomes utterly different too. In Section 7.1, we establish a one-to-one correspondence between the functional DAG and the functional linear SEM, which in turn reveals how the functional DAG factorizes the joint distribution, and how to interpret the identified edges. We further develop in Section 7.2 the notion of *do-intervention* for possible causal interpretation in the functional setting. None of these results are available in Lee and Li (2022).

We also remark that our proposal hinges on linear operators, which are being increasingly employed in graphical modeling (Li et al., 2014; Lee et al., 2016, 2020). Nevertheless, we use linear operators to study a completely different problem. For instance, Lee et al. (2021) studied conditional undirected graphs that vary along with the external variables, whereas we target DAG estimation for multivariate random functions. Correspondingly, the estimation method and asymptotic theory are substantially different.

The rest of the article is organized as follows. We introduce DAG and its equivalence class in Section 2, and develop the linear operators, CCO and PCO, in Section 3. We derive their sample estimation and the modified PC-algorithm in Section 4, and develop the asymptotic theory in Section 5. We conduct numerical studies in Section 6, and further discuss the model in Section 7. We relegate all technical proofs and additional numerical results to the Appendix.

2. DAG and its equivalence class for functions

Let $(\Omega, \mathcal{F}, P)$ denote a probability space. For an interval $T \subseteq \mathbb{R}$ and $t \in T$, let $X(t) = [X_1(t), \ldots, X_p(t)]^\intercal$ denote a vector of multivariate random functions of dimension p defined on Ω and taking values in $\Omega_X = \Omega_{X_1} \times \cdots \times \Omega_{X_p}$, where each Ω_{X_i} is a Hilbert space of $\mathbb{R}$-valued functions on T, $i \in \mathsf{V}$. Let $\langle \cdot, \cdot \rangle_{\Omega_{X_i}}$ denote the inner product in Ω_{X_i}, and $\| \cdot \|_{\Omega_{X_i}}$ the norm induced by this inner product. We allow p to increase with the sample size n; that is, $p = p(n)$, and $\lim_{n\to\infty} p(n) = \infty$. Henceforth, V also depends on n implicitly. We abbreviate $X(t)$ and $X_i(t)$ as X and X_i whenever there is no confusion. Next, we introduce two assumptions on X.

Assumption 1 *There exists $M_0 > 0$ such that* $\max\{E\|X_i\|^2_{\Omega_{X_i}} : i \in \mathsf{V}\} \leq M_0$.

Assumption 2 *The p-variate random function X is a zero-mean Gaussian random element in Ω_X and is faithful with respect to a DAG G; i.e., X satisfies (1).*

Assumption 1 is standard in high-dimensional functional data analysis, and ensures that the trace of the covariance operator of X_i is uniformly bounded. Assumption 2 is our main distributional assumption. The zero-mean condition is to simplify the development, and can be easily relaxed. It is also possible to relax the Gaussian condition, by employing the notions of functional additive conditional independence (Li and Solea, 2018) or copula graphical models (Liu et al., 2012; Solea and Li, 2020). Nevertheless, we feel the Gaussian case is important for its own sake and is worthy of a full investigation. We leave the non-Gaussian extension as future research.

Under Assumption 1, the bilinear form $(f, g) \mapsto E(\langle f, X_i\rangle_{\Omega_{X_i}} \langle g, X_j\rangle_{\Omega_{X_j}})$ from $\Omega_{X_i} \times \Omega_{X_j}$ to $\mathbb{R}$ is bounded, and induces a bounded linear operator $\Sigma_{X_i X_j} : \Omega_{X_j} \to \Omega_{X_i}$ for each $(i, j) \in \mathsf{V} \times \mathsf{V}$. We call it the covariance operator from Ω_{X_j} to Ω_{X_i}. We then define the joint covariance operator $\Sigma_{XX} : \Omega_X \to \Omega_X$ as the $p \times p$ matrix of operators whose (i, j)th entry is $\Sigma_{X_i X_j}$. This means, for any $f = (f_1, \ldots, f_p)^\mathsf{T} \in \Omega_X$, we have $\Sigma_{XX} f = \Big(\sum_{j=1}^p \Sigma_{X_1 X_j} f_j, \ldots, \sum_{j=1}^p \Sigma_{X_p X_j} f_j\Big)^\mathsf{T}$. Moreover, for any subsets $\mathsf{A}, \mathsf{B} \subseteq \mathsf{V}$, we define $\Sigma_{X_\mathsf{A} X_\mathsf{B}}$ to be the matrix of operators whose entries are $\{\Sigma_{X_i X_j} : i \in \mathsf{A}, j \in \mathsf{B}\}$, and its dimension is $|\mathsf{A}| \times |\mathsf{B}|$, where $|\mathsf{A}|$ denotes the cardinality of A. We note that, in Assumption 2, an Ω_X-valued random element X is Gaussian if and only if $\langle f, X\rangle_{\Omega_X}$ is a Gaussian random variable for every $f = (f_1, \ldots, f_p)^\mathsf{T} \in \Omega_X$, or equivalently, $E\big(\exp \sum_{i=1}^p \iota \langle f_i, X_i\rangle_{\Omega_{X_i}}\big) = \exp\big(-1/2 \sum_{i=1}^p \sum_{j=1}^p \langle f_i, \Sigma_{X_i X_j} f_j\rangle_{\Omega_{X_i}}\big)$, where $\iota = \sqrt{-1}$.

Next, we formally define the notion of conditional independence of random functions. For the probability space $(\Omega, \mathcal{F}, P)$, suppose $(\Omega_X, \mathcal{F}_X)$, $(\Omega_Y, \mathcal{F}_Y)$, $(\Omega_Z, \mathcal{F}_Z)$ are measurable spaces, and $X : \Omega \to \Omega_X$, $Y : \Omega \to \Omega_Y$, $Z : \Omega \to \Omega_Z$ are random elements in $(\Omega_X, \mathcal{F}_X)$, $(\Omega_Y, \mathcal{F}_Y)$, $(\Omega_Z, \mathcal{F}_Z)$, respectively. We say that X and Y are conditionally independent given Z, if and only if, for every $A \in \mathcal{F}_X$, $B \in \mathcal{F}_Y$,

$$P(X \in A, Y \in B \mid Z) = P(X \in A \mid Z)\, P(Y \in B \mid Z) \quad \text{almost surely } P.$$

In our case, Ω_X, Ω_Y, Ω_Z are separable Hilbert spaces, and $\mathcal{F}_X$, $\mathcal{F}_Y$, $\mathcal{F}_Z$, are the Borel σ-fields generated by the open sets in Ω_X, Ω_Y, Ω_Z, respectively. The specific forms of Ω_X, Ω_Y, Ω_Z are determined by contexts. This general definition satisfies all the relevant axioms of conditional independence such as those described in Lauritzen (1996, Chapter 3). In particular, as in the setting of undirected graphs, under the Gaussian assumption, the pairwise Markov property is equivalent to the global Markov property.

For a DAG G defined on V, let $\mathsf{H} = \{(i, j, \mathsf{S}) : (i, j) \in \mathsf{V} \times \mathsf{V}, \mathsf{S} \subseteq \mathsf{V} \backslash \{i, j\}\}$, and $\mathsf{H}_0 = \{(i, j, \mathsf{S}) \in \mathsf{H} : i \neq j\}$. Moreover, let

$$\begin{aligned} \mathsf{D}(\mathsf{G}) &= \{(i, j, \mathsf{S}) \in \mathsf{H}_0 : i \text{ and } j \text{ are d-separated by } \mathsf{S} \text{ under } \mathsf{G}\}, \\ \mathsf{F} &= \{(i, j, \mathsf{S}) \in \mathsf{H}_0 : X_i \perp\!\!\!\perp X_j \mid X_\mathsf{S}\}. \end{aligned} \tag{2}$$

By definition, X is faithful to G if $\mathsf{D}(\mathsf{G}) = \mathsf{F}$. It is possible for two different DAGs, say G_1 and G_2, to share the same D; i.e., $\mathsf{D}(\mathsf{G}_1) = \mathsf{D}(\mathsf{G}_2)$. Consequently, by conditional independence and faithfulness, we can only determine the class of DAGs with the same D. Let $\mathcal{G}$ be the collection of all DAGs defined on V. We say that $\mathsf{G}_1, \mathsf{G}_2 \in \mathcal{G}$ are equivalent, and write $\mathsf{G}_1 \sim \mathsf{G}_2$ if and only if $\mathsf{D}(\mathsf{G}_1) = \mathsf{D}(\mathsf{G}_2)$, where $\sim$ is an equivalence relation. This is called the Markov equivalence (Peters et al., 2014). Thus, each $\mathsf{G} \in \mathcal{G}$ induces an equivalence class $\{\mathsf{G}' \in \mathcal{G} : \mathsf{G}' \sim \mathsf{G}\}$. The collection of all Markov equivalence classes forms a partition of $\mathcal{G}$, which is a quotient space of $\mathcal{G}$, denoted by $\mathcal{G}/\!\sim$ (Kelley, 1955), and is referred to as the quotient space of Markov equivalence.

A partially directed graph is a graph in which some edges are undirected. A completed partially acyclic directed graph (CPDAG) is a partially directed acyclic graph L such that there exists a member $\mathcal{D}$ of $\mathcal{G}/\!\sim$ satisfying the following properties: (i) each directed edge in L is an edge in every DAG in $\mathcal{D}$; and (ii) for each undirected edge, say $i \leftrightarrow j$, in L, there is a DAG in $\mathcal{D}$ with $i \to j$ being one of its edges, and another DAG in $\mathcal{D}$ with $j \to i$

being one of its edges. Chickering (2002, Lemma 2) showed that two DAGs are Markov equivalent if and only if they have the same CPDAG. In other words, a member $\mathcal{D}$ of $\mathcal{G}/\sim$ corresponds to one and only one CPDAG. This also establishes a bijection between $\mathcal{G}/\sim$ and the collection of all CPDAGs.

3. Two linear operators

In this section, we formally develop the conditional covariance operator and the partial correlation operator to estimate the CPDAG, and derive their population properties.

We begin with some notation. Let Ω and Ω' be two Hilbert spaces, $\mathscr{B}(\Omega, \Omega')$ the collection of all bounded linear operators from Ω to Ω', $\mathscr{B}_2(\Omega, \Omega')$ the collection of all Hilbert-Schmidt operators from Ω to Ω', and $\mathscr{B}_1(\Omega, \Omega')$ the collection of all trace class operators from Ω to Ω'. When $\Omega' = \Omega$, we write $\mathscr{B}(\Omega, \Omega)$ as $\mathscr{B}(\Omega)$, $\mathscr{B}_2(\Omega, \Omega)$ as $\mathscr{B}_2(\Omega)$, and $\mathscr{B}_1(\Omega, \Omega)$ as $\mathscr{B}_1(\Omega)$. Let $\|\cdot\|$, $\|\cdot\|_{\mathrm{HS}}$, and $\|\cdot\|_{\mathrm{TR}}$ be the operator norm in $\mathscr{B}(\Omega)$, the Hilbert-Schmidt norm in $\mathscr{B}_2(\Omega)$, and the trace norm in $\mathscr{B}_1(\Omega)$ respectively. For a linear operator A, let $\ker(A)$, $\mathrm{ran}(A)$, and $\overline{\mathrm{ran}}(A)$ denote the kernel, range, and the closure of the range of A; i.e., $\ker(A) = \{f \in \Omega : Af = 0\}, \mathrm{ran}(A) = \{Af : f \in \Omega\}, \overline{\mathrm{ran}}(A) = \mathrm{cl}\,[\mathrm{ran}(A)]$, where $\mathrm{cl}(\cdot)$ stands for the closure of a set. For a bounded and self-adjoint linear operator A, its restriction on $\overline{\mathrm{ran}}(A)$ is an injective function from $\overline{\mathrm{ran}}(A)$ to $\mathrm{ran}(A)$. We call the inverse of this function $A|\overline{\mathrm{ran}}(A)$ as the Moore-Penrose inverse, and denote it by $A^\dagger$. That is, $A^\dagger$ is the mapping from $\mathrm{ran}(A)$ to $\overline{\mathrm{ran}}(A)$ such that, for any $x \in \mathrm{ran}(A)$, $A^\dagger x$ is the unique member $y \in \overline{\mathrm{ran}}(A)$ satisfying $Ay = x$. Let A^* denote the adjoint of A. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \preceq b_n$ or $b_n \succeq a_n$ if a_n/b_n is bounded, and write $a_n \prec b_n$ or $b_n \succ a_n$ if $a_n/b_n \to 0$. If two sequences a_n and b_n are ordered by $\preceq$, then we use $a_n \wedge b_n$ to denote the smaller sequence in terms of $\preceq$. Moreover, we write $a_n \asymp b_n$ if $a_n \preceq b_n$ and $b_n \preceq a_n$.

For each $i \in \mathsf{V}$, let $\{(\lambda_i^a, \eta_i^a)\}_{a \in \mathbb{N}}$ be the collection of eigenvalue-eigenfunction pairs of $\Sigma_{X_i X_i}$, with $\lambda_i^1 \geq \lambda_i^2 \geq \cdots \geq 0$ and $\mathbb{N}$ being the set of natural numbers $\{1, 2, \ldots, \}$. Then $X_i = \sum_{a \in \mathbb{N}} c_i^a \eta_i^a$ holds almost surely, where $c_i^a = \langle X_i, \eta_i^a \rangle$ are from i.i.d. $N(0, \lambda_i^a)$. This expansion is known as the Karhunen-Loève (K-L) expansion, and $c_i^1, c_i^2, \ldots$ are called the functional principal component scores (Bosq, 2000). We note that the K-L expansion has been widely used in functional data analysis (Yao et al., 2005; Yao and Müller, 2010; Li and Guan, 2014; Chen and Lei, 2015). Particularly, Qiao et al. (2019) also used the K-L expansion for undirected functional graphic modeling.

We next formally define the conditional covariance operator. Let

$$M_{X_\mathsf{A} X_\mathsf{B}} = \Sigma^\dagger_{X_\mathsf{A} X_\mathsf{A}} \Sigma_{X_\mathsf{A} X_\mathsf{B}}, \qquad \text{for } \mathsf{A}, \mathsf{B} \subseteq \mathsf{V}.$$

Given its resemblance to the regression coefficient matrix in the classical regression, we call $M_{X_\mathsf{A} X_\mathsf{B}}$ a *regression operator*, and the next assumption ensures it is well-defined.

Assumption 3 *Suppose* $\mathrm{ran}(\Sigma_{X_\mathsf{A} X_\mathsf{B}}) \subseteq \mathrm{ran}(\Sigma_{X_\mathsf{A} X_\mathsf{A}})$*, and* $M_{X_\mathsf{A} X_\mathsf{B}}$ *is Hilbert-Schmidt.*

The Hilbert-Schmidt assumption on $M_{X_\mathsf{A} X_\mathsf{B}}$ regulates the degree of smoothness in the dependence between X_A and X_B. That is, the output of $\Sigma_{X_\mathsf{A} X_\mathsf{B}}$ needs to sufficiently concentrate on the leading eigenfunctions of $\Sigma_{X_\mathsf{A} X_\mathsf{A}}$. Similar conditions are commonly imposed in the literature (see, e.g., Lee et al., 2016; Li, 2018).

For any $\mathsf{A} \subseteq \mathsf{V}$, let Ω_{X_A} be the direct sum of $\{\Omega_{X_i} : i \in \mathsf{A}\}$; i.e., Ω_{X_A} is the Cartesian product of Ω_{X_i}, $i \in \mathsf{A}$, and the inner product in Ω_{X_A} is the sum of the inner products in Ω_{X_i}, $i \in \mathsf{A}$. The next proposition shows that the regression operator $M_{X_\mathsf{A} X_\mathsf{B}}$ in fact uniquely determines the mapping $f \mapsto E(\langle f, X_\mathsf{B}\rangle_{\Omega_{X_\mathsf{B}}} \mid X_\mathsf{A})$, which characterizes $E(X_\mathsf{B} \mid X_\mathsf{A})$. In other words, $M_{X_\mathsf{A} X_\mathsf{B}}$ has a one-to-one correspondence with $E(X_\mathsf{B} \mid X_\mathsf{A})$.

Proposition 1 *Suppose Assumptions 1 to 3 hold. Then, for* $f \in \Omega_{X_\mathsf{B}}$, $\langle M_{X_\mathsf{A} X_\mathsf{B}} f, X_\mathsf{A}\rangle_{\Omega_{X_\mathsf{A}}} = E(\langle f, X_\mathsf{B}\rangle_{\Omega_{X_\mathsf{B}}} \mid X_\mathsf{A})$.

Definition 1 *Suppose Assumptions 1 to 3 hold. For* $(i, j, \mathsf{S}) \in \mathsf{H}$, *let* $\Sigma_{X_i X_j | X_\mathsf{S}} : \Omega_{X_j} \to \Omega_{X_i}$ *be the linear operator*

$$\Sigma_{X_i X_j} - M^*_{X_\mathsf{S} X_i} \Sigma_{X_\mathsf{S} X_\mathsf{S}} M_{X_\mathsf{S} X_j}.$$

We call $\Sigma_{X_i X_j | X_\mathsf{S}}$ *the conditional covariance operator (CCO) of* X_i *and* X_j *given* X_S.

We note that, an equivalent form of CCO is $\Sigma_{X_i X_j | X_\mathsf{S}} = \Sigma_{X_i X_j} - \Sigma_{X_i X_\mathsf{S}} \Sigma^\dagger_{X_\mathsf{S} X_\mathsf{S}} \Sigma_{X_\mathsf{S} X_j}$. However, the form in Definition 1 involves the regression operator, and is more conducive for subsequent proofs. The next theorem establishes the properties of $\Sigma_{X_i X_j | X_\mathsf{S}}$.

Theorem 1 *Suppose Assumptions 1 to 3 hold. Then, for every* $(i, j, \mathsf{S}) \in \mathsf{H}_0$,

(i) $\Sigma_{X_i X_j | X_\mathsf{S}} = 0$ *if and only if* $X_i \perp\!\!\!\perp X_j \mid X_\mathsf{S}$;

(ii) $\Sigma_{X_i X_j | X_\mathsf{S}} = \sum_{a,b \in \mathbb{N}} \mathrm{cov}(c_i^a, c_j^b \mid X_\mathsf{S})(\eta_i^a \otimes \eta_j^b)$, *where* c_i^a *is the ath K-L coefficient of* X_i, *and* η_i^a *is the eigenfunction of* $\Sigma_{X_i X_i}$ *associated with its ath largest eigenvalue.*

Theorem 1 (i) generalizes the classical result when X is a vector of Gaussian random variables to the functional setting. Consequently, we can use the CCO to characterize the conditional independence between X_i and X_j given X_S.

We next define the partial correlation operator, which extends partial correlation to the functional setting. This is motivated by the observation that partial correlation achieves better scaling in the classical random variable setting (Peng et al., 2009; Lee et al., 2016; Liu, 2017). The next theorem establishes the existence of PCO and its connection with the conditional independence. Its proof is similar to that of Lee, Li, and Zhao (2016, Theorem 1), and is thus omitted.

Theorem 2 *Suppose Assumptions 1 to 3 hold. Then, there exists a unique operator* $R_{X_i X_j | X_\mathsf{S}} \in \mathscr{B}(\Omega_{X_j}, \Omega_{X_i})$ *such that,*

(i) $\Sigma_{X_i X_j | X_\mathsf{S}} = \Sigma^{1/2}_{X_i X_i | X_\mathsf{S}} R_{X_i X_j | X_\mathsf{S}} \Sigma^{1/2}_{X_j X_j | X_\mathsf{S}}$;

(ii) $\|R_{X_i X_j | X_\mathsf{S}}\| \leq 1$, *for every* $(i, j, \mathsf{S}) \in \mathsf{H}_0$.

Moreover, $R_{X_i X_j | X_\mathsf{S}} = 0$ *if and only if* $X_i \perp\!\!\!\perp X_j \mid X_\mathsf{S}$.

Definition 2 *We call* $R_{X_i X_j | X_\mathsf{S}}$ *in Theorem 2 the partial correlation operator (PCO) of* X_i *and* X_j *given* X_S.

4. Estimation

Theorems 1 and 2 suggest two linear operators, $\Sigma_{X_i X_j | X_{\mathsf{S}}}$ and $R_{X_i X_j | X_{\mathsf{S}}}$, to characterize the conditional independence. In this section, we develop their sample estimators, first at the operator level, then at the coordinate level. We then develop a procedure for evaluating $X_i \perp\!\!\!\perp X_j \mid X_{\mathsf{S}}$ for a given triplet $(i, j, \mathsf{S}) \in \mathsf{H}_0$, and a modified PC-algorithm for estimating the entire DAG.

4.1 Operator-level estimation

We first derive the K-L expansion at the sample level. Suppose $X^1, \ldots, X^n$ are i.i.d. samples from X, where $X^k = (X_1^k, \ldots, X_p^k)^\intercal, k = 1, \ldots, n$. Let E_n represent the sample mean operator; i.e., $E_n(U) = \sum_{k=1}^n U^k / n$, for a set of samples $(U^1, \ldots, U^n)$. The covariance operator $\Sigma_{X_i X_j}$ is estimated by

$$\hat{\Sigma}_{X_i X_j} = E_n \left[(X_i - E_n X_i) \otimes (X_j - E_n X_j) \right],$$

for any $(i, j) \in \mathsf{V} \times \mathsf{V}$. For each $i \in \mathsf{V}$, let $\{(\hat{\lambda}_i^a, \hat{\eta}_i^a)\}_{a \in \mathbb{N}}$ be the sequence of eigenvalue-eigenfunction pairs of $\hat{\Sigma}_{X_i X_i}$. Then, the sample-level K-L expansion of $X_i^k - E_n(X_i)$ is $X_i^k - E_n(X_i) = \sum_{a \in \mathbb{N}} \hat{c}_i^{k,a} \hat{\eta}_i^a$, where $\hat{c}_i^{k,a} = \langle X_i^k - E_n(X_i), \hat{\eta}_i^a \rangle_{\Omega_{X_i}}$. To improve estimation efficiency, we further truncate this expansion at the dth term to obtain the approximation, $X_i^k - E_n(X_i) \approx \sum_{a=1}^d \hat{c}_i^{k,a} \hat{\eta}_i^a$, for all $k = 1, \ldots, n$, and $i \in \mathsf{V}$. Note that, we allow the truncation number d to depend on n. Correspondingly, the truncated estimate of $\Sigma_{X_i X_j}$ is $\hat{\Sigma}^d_{X_i X_j} = \sum_{a,b=1}^d E_n(\hat{c}_i^a \hat{c}_j^b)(\hat{\eta}_i^a \otimes \hat{\eta}_j^b)$.

We next estimate the regression operators $M_{X_{\mathsf{S}} X_i}$ by

$$\hat{M}_{X_{\mathsf{S}} X_i} = (\hat{\Sigma}^d_{X_{\mathsf{S}} X_{\mathsf{S}}} + \epsilon I)^{-1} \hat{\Sigma}^d_{X_{\mathsf{S}} X_i}, \tag{3}$$

for $i \in \mathsf{V}$, where $\epsilon > 0$ is a ridge-type tuning parameter, and I is the identity operator from $\Omega_{X_{\mathsf{S}}}$ to $\Omega_{X_{\mathsf{S}}}$. The inverse is done by taking the inverse of the eigenvalues; see Proposition 3 in Section 4.2. Following Theorem 1(ii), we estimate the CCO by

$$\hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_{\mathsf{S}}} = \sum_{a,b=1}^d \{ E_n(\hat{c}_i^a \hat{c}_j^b) - E_n[(\hat{M}_{X_{\mathsf{S}} X_i} \hat{\eta}_i^a)(X_{\mathsf{S}})(\hat{M}_{X_{\mathsf{S}} X_j} \hat{\eta}_j^b)(X_{\mathsf{S}})] \} (\hat{\eta}_i^a \otimes \hat{\eta}_j^b).$$

Similarly, following Theorem 2(i), we estimate the PCO by

$$\hat{R}^{d,\epsilon,\delta}_{X_i X_j | X_{\mathsf{S}}} = (\hat{\Sigma}^{d,\epsilon}_{X_i X_i | X_{\mathsf{S}}} + \delta I)^{-1/2} \hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_{\mathsf{S}}} (\hat{\Sigma}^{d,\epsilon}_{X_j X_j | X_{\mathsf{S}}} + \delta I)^{-1/2}, \tag{4}$$

where $\delta > 0$ is a ridge parameter that regularizes the inverses of the two matrices.

Note that in (3) and (4), we both truncate the K-L expansion and employ the ridge-type regularization. We choose to do so for the following reasons. First of all, we note that the rank of $\hat{\Sigma}^d_{X_{\mathsf{S}} X_{\mathsf{S}}}$ is dn_s, where n_s is the cardinality of S. This rank can be larger than the sample size n, and thus we need to introduce an extra ridge-type regularization in (3). Meanwhile, since the ranks of $\hat{\Sigma}^{d,\epsilon}_{X_i X_i | X_{\mathsf{S}}}$ and $\hat{\Sigma}^{d,\epsilon}_{X_j X_j | X_{\mathsf{S}}}$ are d, which is typically smaller than n, an alternative estimator of the PCO is not to employ the ridge regularization as in (3) and (4), which leads to

$$\tilde{R}^d_{X_i X_j | X_{\mathsf{S}}} = (\tilde{\Sigma}^d_{X_i X_i | X_{\mathsf{S}}})^{-1/2} \tilde{\Sigma}^d_{X_i X_j | X_{\mathsf{S}}} (\tilde{\Sigma}^d_{X_j X_j | X_{\mathsf{S}}})^{-1/2}, \tag{5}$$

where $\tilde{\Sigma}^d_{X_iX_j|X_\mathsf{S}} = \hat{\Sigma}^d_{X_iX_j} - \hat{\Sigma}^d_{X_iX_\mathsf{S}} \hat{\Sigma}^{\dagger d}_{X_\mathsf{S}X_\mathsf{S}} \hat{\Sigma}^d_{X_\mathsf{S}X_j}$, and $\hat{\Sigma}^{\dagger d}_{X_\mathsf{S}X_\mathsf{S}} = \sum_{a=1}^d (\lambda^a_\mathsf{S})^{-1}(\eta^a_\mathsf{S} \otimes \eta^a_\mathsf{S})$ is the Moore-Penrose inverse of $\hat{\Sigma}^d_{X_\mathsf{S}X_\mathsf{S}}$, for $(i, j, \mathsf{S}) \in \mathsf{H}_0$. Comparing (5) to (4), this alternative PCO estimator has two fewer tuning parameters, and thus is computationally easier. However, as we remark after Theorem 4 in Section 5.2, the asymptotic analysis of (4) is much easier than that of (5). Moreover, when $d = 1$, (5) is closely related to one of the competing methods, linear-PC, that we numerically compare in Section 6.1, and we show that (4) achieves a better empirical performance than (5). Therefore, we propose (4) as our PCO estimator, and build our DAG estimation based on (4).

4.2 Coordinate-level evaluation for conditional independence

We begin with constructing the spaces Ω_{X_i} and Ω_X using functional bases. Let $\mathcal{H}$ be a generic finite-dimensional Hilbert space spanned by a set of functions $\mathcal{B} = \{b_1, \ldots, b_m\}$ defined on T. For any $h \in \mathcal{H}$, let $[h]_\mathcal{B} = ([h]_{\mathcal{B},1}, \ldots, [h]_{\mathcal{B},m})^\intercal$ denote the coordinate of h with respect to $\mathcal{B}$; that is, $h = \sum_{i=1}^m [h]_{\mathcal{B},i}\, b_i = b_{1:m}^\intercal\, [h]_\mathcal{B}$, where $b_{1:m}$ denotes the vector of functions $(b_1, \ldots, b_m)^\intercal$. Let $K_\mathcal{B} = \{\langle b_s, b_t\rangle_\mathcal{H}\}_{s,t=1}^m$ be the Gram kernel matrix. Then the inner product $\langle h_1, h_2\rangle_\mathcal{H}$ can be expressed as $[h_1]_\mathcal{B}^\intercal\, K_\mathcal{B}\, [h_2]_\mathcal{B}$.

We next derive the coordinate representation of X^k, $k = 1, \ldots, n$. Suppose each X^k is measured on u_k time points $T_k = \{t_{k1}, \ldots, t_{ku_k}\}$, $k = 1, \ldots, n$. Let $T_{1:n} = \cup_{k=1}^n T_k$, $\tau_1 < \tau_2 < \cdots < \tau_\ell$ be all the time points in $T_{1:n}$, and $\ell = |T_{1:n}|$ be the total number of distinctive time points. Let Ω_{X_i} be the linear space spanned by the functions $\{\kappa_T(\cdot, \tau) : \tau \in T_{1:n}\}$. Let $K_T = \{\kappa_T(\tau_i, \tau_j)\}_{i,j=1}^\ell$. Suppose K_T is of rank r, and has the spectral decomposition $K_T = U_T D_T U_T^\intercal$, where $D_T \in \mathbb{R}^{r\times r}$ is the diagonal matrix whose diagonal elements are the sorted nonzero eigenvalues of K_T, and $U_T \in \mathbb{R}^{\ell\times r}$ is the matrix whose columns are the eigenvectors corresponding to the eigenvalues in D_T. Let $(b_1, \ldots, b_r)^\intercal = D_T^{-1/2} U_T^\intercal [\kappa_T(\cdot, \tau_1), \ldots, \kappa_T(\cdot, \tau_\ell)]^\intercal$. It is straightforward to verify that $\mathcal{B}_r = \{b_1, \ldots, b_r\}$ is an orthonormal basis of Ω_{X_i}. Using this basis, each X_i^k can be represented as $X_i^k(\cdot) = \sum_{t=1}^r [X_i^k]_{\mathcal{B}_r,t}\, b_t(\cdot) = b_{1:r}^\intercal(\cdot)\, [X_i^k]_{\mathcal{B}_r}$. Note that the kth individual X_i^k is observed only at u_k time points in T_k. Let $X_i^k(T_k) = \big(X_i^k(t_{k1}), \ldots, X_i^k(t_{ku_k})\big)^\intercal$, and $b_{1:r}(T_k) = \big(b_{1:r}(t_{k1}), \ldots, b_{1:r}(t_{ku_k})\big)$. Therefore, we have $X_i^k(T_k) = b_{1:r}(T_k)^\intercal\, [X_i^k]_{\mathcal{B}_r}$, on both side of which we then multiply $b_{1:r}(T_k)$ to get $b_{1:r}(T_k) X_i^k(T_k) = b_{1:r}(T_k) b_{1:r}(T_k)^\intercal\, [X_i^k]_{\mathcal{B}_r}$. Solving this linear equation with a ridge-type regularization, we obtain the following coordinate representation of $[X_i^k]_{\mathcal{B}_r}$, for $i \in \mathsf{V}$ and $k = 1, \ldots, n$,

$$[X_i^k]_{\mathcal{B}_r} = [b_{1:r}(T_k) b_{1:r}(T_k)^\intercal + \epsilon_T^k I_r]^{-1}\, b_{1:r}(T_k) X_i^k, \tag{6}$$

where I_r is the $r \times r$ identity matrix, and ϵ_T^k is a ridge tuning parameter.

We next derive the coordinate representations of the truncated sample covariance operators. Let $\mathcal{H}$ and $\mathcal{H}'$ be two finite-dimensional Hilbert spaces spanned by $\mathcal{B} = \{b_1, \ldots, b_m\}$ and $\mathcal{B}' = \{b'_1, \ldots, b'_{m'}\}$, respectively. Let $A : \mathcal{H} \to \mathcal{H}'$ be a linear operator. The coordinate representation of A with respect to $\mathcal{B}$ and $\mathcal{B}'$ is defined as $([Ab_1]_{\mathcal{B}'}, \ldots, [Ab_m]_{\mathcal{B}'}) \equiv {}_{\mathcal{B}'}[A]_\mathcal{B}$. If $\mathcal{H}''$ is a third Hilbert space with basis $\mathcal{B}''$, and A' is a linear operator from $\mathcal{H}'$ to $\mathcal{H}''$, then ${}_{\mathcal{B}''}[A'A]_\mathcal{B} = ({}_{\mathcal{B}''}[A']_{\mathcal{B}'})({}_{\mathcal{B}'}[A]_\mathcal{B})$. For simplicity, we abbreviate ${}_{\mathcal{B}'}[A]_\mathcal{B}$ by $[A]$ when the bases $\mathcal{B}, \mathcal{B}'$ are obvious from the context.

Proposition 2 *For each* $(i, j) \in \mathsf{V} \times \mathsf{V}$, $[\hat{\Sigma}_{X_iX_j}] = E_n\left([X_i - E_n X_i]\,[X_j - E_n X_j]^\intercal\right)$.

Therefore, by the definition of $\hat{M}_{X_\mathsf{S} X_i}$, the coordinate representation of $\hat{M}_{X_\mathsf{S} X_i} f$ is

$$[\hat{M}_{X_\mathsf{S} X_i} f] = ([\hat{\Sigma}_{X_\mathsf{S} X_\mathsf{S}}] + \epsilon I_{r n_s})^{-1} [\hat{\Sigma}_{X_\mathsf{S} X_i}][f],$$

for each $f \in \Omega_{X_i}$, $i \in \mathsf{V}$, and $\mathsf{S} \subseteq \mathsf{V}$, with n_s being the cardinality of S.

Let $\{(\hat{\lambda}_i^a, [\hat{\eta}_i^a])\}_{a=1}^r$ denote the collection of eigenvalue-eigenvector pairs of $[\hat{\Sigma}_{X_i X_i}]$, $i \in \mathsf{V}$. Then the sample-level K-L expansion of $X_i^k - E_n X_i$ is

$$X_i^k - E_n X_i = \textstyle\sum_{a=1}^d \langle X_i^k - E_n X_i, \hat{\eta}_i^a \rangle_{\Omega_{X_i}} \hat{\eta}_i^a = \sum_{a=1}^d \left[X_i^k - E_n X_i\right]^\mathsf{T} \left[\hat{\eta}_i^a\right] \hat{\eta}_i^a. \tag{7}$$

Finally, we derive the coordinate representations of the sample CCO and PCO. Recall that $\hat{c}_i^{k,a}$ denotes the inner product $\langle X_i^k - E_n X_i, \hat{\eta}_i^a \rangle_{\Omega_{X_i}}$. For $i \in \mathsf{V}$, $a = 1, \ldots, d$, let C_i^a be the n-dimensional vector $\{c_i^{k,a} : k = 1, \ldots, n\}$, and $C_i^{1:d}$ be the $n \times d$ matrix whose ath column is C_i^a. For $\mathsf{S} = \{a_1, \ldots, a_{n_s}\} \subseteq \mathsf{V} \backslash \{i, j\}$, let C_S be the $1 \times n_s$ block matrix $\left(C_{a_1}^{1:d}, \ldots, C_{a_{n_s}}^{1:d}\right)$. The sample estimates of CCO and PCO are given in Proposition 3. Its proof can be derived from Proposition 2 and (7), and is omitted.

Proposition 3 *The coordinate representations of* $\hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_\mathsf{S}}$ *and* $\hat{R}^{d,\epsilon,\delta}_{X_i X_j | X_\mathsf{S}}$ *with respect to* $\mathcal{B}_i^* = \{\hat{\eta}_i^1, \ldots, \hat{\eta}_i^d\}$ *and* $\mathcal{B}_j^* = \{\hat{\eta}_j^1, \ldots, \hat{\eta}_j^d\}$ *are:*

$$\begin{aligned} {}_{\mathcal{B}_i^*}[\hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_\mathsf{S}}]_{\mathcal{B}_j^*} &= n^{-1} \{(c_i^a)^\mathsf{T} [I_n - D(\mathsf{S})] c_j^b\}_{a,b=1}^d \equiv M_{i,j|\mathsf{S}}(\epsilon), & (8) \\ {}_{\mathcal{B}_i^*}[\hat{R}^{d,\epsilon,\delta}_{X_i X_j | X_\mathsf{S}}]_{\mathcal{B}_j^*} &= [M_{i,i|\mathsf{S}}(\epsilon) + \delta I_d]^{-1/2} M_{i,j|\mathsf{S}}(\epsilon) [M_{j,j|\mathsf{S}}(\epsilon) + \delta I_d]^{-1/2}, & (9) \end{aligned}$$

where $D(\mathsf{S}) = C_\mathsf{S} \left[(C_\mathsf{S}^\mathsf{T} C_\mathsf{S} + \epsilon I_{n_s d})^{-1} C_\mathsf{S}^\mathsf{T} C_\mathsf{S} (C_\mathsf{S}^\mathsf{T} C_\mathsf{S} + \epsilon I_{n_s d})^{-1}\right] C_\mathsf{S}^\mathsf{T}$.

Based on (8), we compute the squared Hilbert-Schmidt (H-S) norm of $\hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_\mathsf{S}}$ as

$$\begin{aligned} \|\hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_\mathsf{S}}\|^2_{\mathrm{HS}} &= \textstyle\sum_{a=1}^d \langle \hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_\mathsf{S}} \hat{\eta}_j^a, \hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_\mathsf{S}} \hat{\eta}_j^a \rangle_{\Omega_{X_i}} \\ &= \textstyle\sum_{a=1}^d \left[\hat{\eta}_j^a\right]^\mathsf{T}_{\mathcal{B}_j^*} ({}_{\mathcal{B}_j^*}[\hat{\Sigma}^{d,\epsilon}_{X_j X_i | X_\mathsf{S}}]_{\mathcal{B}_i^*}) ({}_{\mathcal{B}_i^*}[\hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_\mathsf{S}}]_{\mathcal{B}_j^*}) \left[\hat{\eta}_j^a\right]_{\mathcal{B}_j^*} = \|{}_{\mathcal{B}_i^*}[\hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_\mathsf{S}}]_{\mathcal{B}_j^*}\|^2_{\mathrm{F}}, \end{aligned}$$

where $\|\cdot\|_\mathrm{F}$ is the Frobenius norm. In other words, the H-S norm of $\hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_\mathsf{S}}$ is simply the Frobenius norm of the matrix ${}_{\mathcal{B}_j^*}[\hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_\mathsf{S}}]_{\mathcal{B}_i^*}$. Similarly, the H-S norm of $\hat{R}^{d,\epsilon,\delta}_{X_i X_j | X_\mathsf{S}}$ is the Frobenius norm of the matrix ${}_{\mathcal{B}_i^*}[\hat{R}^{d,\epsilon,\delta}_{X_i X_j | X_\mathsf{S}}]_{\mathcal{B}_j^*}$ in (9). We then threshold the H-S norms $\|\hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_\mathsf{S}}\|_{\mathrm{HS}}$ and $\|\hat{R}^{d,\epsilon,\delta}_{X_i X_j | X_\mathsf{S}}\|_{\mathrm{HS}}$ to evaluate the conditional independence; that is, we declare $X_i \perp\!\!\!\perp X_j \mid X_\mathsf{S}$ if

$$\|{}_{\mathcal{B}_j^*}[\hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_\mathsf{S}}]_{\mathcal{B}_i^*}\|_\mathrm{F} < \rho_{\mathrm{CCO}}, \quad \|{}_{\mathcal{B}_j^*}[\hat{R}^{d,\epsilon,\delta}_{X_i X_j | X_\mathsf{S}}]_{\mathcal{B}_i^*}\|_\mathrm{F} < \rho_{\mathrm{PCO}}, \tag{10}$$

where ρ_{CCO} and ρ_{PCO} are the threshold values determined adaptively given the data. Observing that for the random variable case, the partial correlation can be tested using its Fisher z-transformation, whose variance is approximated by $(n - |\mathsf{S}| - 3)^{-1}$ (Harris and Drton, 2013), we take ρ_{CCO} and ρ_{PCO} to be proportional to $(n - |\mathsf{S}| - 3)^{-1/2}$. Our numerical experiments have found that the results are not overly sensitive to the choice of these threshold values as long as they are within a reasonable range.

We summarize the above estimation procedure in Algorithm 1.

Algorithm 1 Evaluation of $X_i \perp\!\!\!\perp X_j \mid X_{\mathsf{S}}$ for a given triplet $(i,j,\mathsf{S}) \in \mathsf{H}_0$.

```
Choose a kernel $\kappa_T$; e.g., the Brownian motion function $\kappa_T(s,t) = \min(s,t)$, or the radial basis function $\kappa_T(s,t) = \exp\{-\gamma_T(s-t)^2\}$, for $(s,t) \in T \times T$, where $\gamma_T = \left\{\sum_{s<t} |\tau_s - \tau_t| / \binom{\ell}{2}\right\}^{-2}$. Compute the Gram matrix $K_T$, and perform spectral decomposition to obtain $U_T$ and $D_T$.
Compute the coordinate $[X_i^k]$ using (6), for $i \in \mathsf{V}$, $k = 1, \ldots, n$, where the ridge parameter is set as $\epsilon_T^k = 0.01 \times \sigma_{\max}[b_{1:r}(T_k) b_{1:r}(T_k)^\intercal]$, $k = 1, \ldots, n$, with $\sigma_{\max}(\cdot)$ denoting the largest eigenvalue.
Perform the spectral decomposition on $E_n\left([X_i - E_n X_i]\,[X_i - E_n X_i]^\intercal\right)$ to obtain its $a$th eigenvector $[\hat{\eta}_i^a]$, and K-L coefficient $\hat{c}_i^{k,a}$ for $X_i^k$ using (7), $i \in \mathsf{V}$ and $a = 1, \ldots, d$, where the parameter $d$ is set as $d = [n^{1/5}]$. Then obtain $C_i^a$, $C_i^{1:d}$, and $C_{\mathsf{S}}$.
Compute the coordinates of $\hat{\Sigma}^{d,\epsilon}_{X_i X_j | X_{\mathsf{S}}}$ and $\hat{R}^{d,\epsilon,\delta}_{X_i X_j | X_{\mathsf{S}}}$ using (8) and (9), where the tuning parameters are set at $\epsilon = 0.1 \times \sigma_{\max}(C_{\mathsf{S}} C_{\mathsf{S}}^\intercal)$, and $\delta = 0.5 \times \max\{\sigma_{\max}[M_{i,i|\mathsf{S}}(\epsilon)], \sigma_{\max}[M_{j,j|\mathsf{S}}(\epsilon)]\}$.
Evaluate $X_i \perp\!\!\!\perp X_j \mid X_{\mathsf{S}}$ using (10), where we take $\rho_{\mathrm{CCO}} = c^{-1} \times (n - |\mathsf{S}| - 3)^{-1/2}$, and $\rho_{\mathrm{PCO}} = \Phi^{-1}(1 - c/2) \times (n - |\mathsf{S}| - 3)^{-1/2}$, with $\Phi(\cdot)$ being the normal cumulative distribution function, $0 < c < 1$ being a constant and set as $c = 0.05$.
```

4.3 Functional PC-algorithm

Algorithm 1 allows us to evaluate (10) for all possible triplets $(i,j,\mathsf{S}) \in \mathsf{H}_0$, which leads to an estimate of the collection F defined in (2). By faithfulness, we then have an estimate of $\mathsf{D}(\mathsf{G})$ in (2), and hence an estimate for CPDAG. However, despite the simplicity of this idea, the amount of computation needed to achieve it can be prohibitively large, as one has to evaluate the conditional independence $X_i \perp\!\!\!\perp X_j \mid X_{\mathsf{S}}$ for every triplet $(i,j,\mathsf{S}) \in \mathsf{H}_0$, which is an NP-hard problem.

A solution to address this issue is the PC-algorithm, which involves two main steps. In the first step, it recursively deletes edges from an initial complete undirected graph based on conditional independence evaluations. This results in a skeleton. In the second step, it extends the skeleton to a CPDAG. This algorithm brings down the computation time considerably by avoiding an exhaustive search: the amount of search is determined by the sparsity, rather than the size, of the graph. In the worst scenario, its runtime grows exponentially with p, but when the true DAG is sparse, the runtime reduces to the polynomial time.

Next, we extend the PC-algorithm to our DAG setting, which, like the classical PC-algorithm, also consists of two steps. For any undirected graph $\mathsf{M} \subseteq (\mathsf{V} \times \mathsf{V})_0$ and any $i \in \mathsf{V}$, let $\mathsf{V}(i,\mathsf{M}) = \{k \in \mathsf{V} : (i,k) \in \mathsf{M}\}$ be the neighborhood of i in M. For any $j \in \mathsf{V}$, let $\mathsf{V}(i,-j,\mathsf{M}) = \mathsf{V}(i,\mathsf{M}) \backslash \{j\}$ be the neighborhood of i in M with j removed. For $\ell = 0,1,2,\ldots$, let $\Lambda(\mathsf{M},\ell) = \{(i,j) \in \mathsf{M} : |\mathsf{V}(i,-j,\mathsf{M})| \geq \ell\}$. For $(i,j) \in \Lambda(\mathsf{M},\ell)$, let $\Gamma(i,j,\mathsf{M},\ell) = \{\mathsf{S} \subseteq \mathsf{V}(i,-j,\mathsf{M}) : |\mathsf{S}| = \ell\}$. Moreover, define two test functions:

$$\phi_1(i,j,\mathsf{S}) = 1 \text{ if } X_i \perp\!\!\!\perp X_j \mid X_{\mathsf{S}} \text{ is accepted by CCO, and 0 otherwise;}$$
$$\phi_2(i,j,\mathsf{S}) = 1 \text{ if } X_i \perp\!\!\!\perp X_j \mid X_{\mathsf{S}} \text{ is accepted by PCO, and 0 otherwise.}$$

Algorithm 2 Step 1 of the PC-algorithm

```
initialize: set $\ell = -1$ and $\mathsf{M}$ = the complete undirected graph
repeat
  set $\ell = \ell + 1$ and $\mathsf{R}_\ell = \emptyset$, which denotes the edge set to be removed
  repeat
    choose $(i, j) \in \Lambda(\mathsf{M}, \ell)$
    repeat
      evaluate $\phi(i, j, \mathsf{S})$ for each $\mathsf{S} \in \Gamma(i, j, \mathsf{M}, \ell)$
    until $\phi(i, j, \mathsf{S}) = 1$, then reset $\mathsf{R}_\ell$ to $\mathsf{R}_\ell \cup \{(i, j)\}$ and record $\mathsf{S}_{i,j} = \mathsf{S}$ for later use;
    if $\phi(i, j, \mathsf{S}) = 0$ for all $\mathsf{S} \in \Gamma(i, j, \mathsf{M}, \ell)$, then keep $\mathsf{R}_\ell$ the same
  until all $(i, j) \in \Lambda(\mathsf{M}, \ell)$ are chosen, then reset $\mathsf{M}$ to $\mathsf{M} \backslash \mathsf{R}_\ell$
until $\Lambda(\mathsf{M}, \ell) = \emptyset$
```

The first step of our extended PC-algorithm is summarized in Algorithm 2, where the test function ϕ can be either ϕ_1 or ϕ_2. Also note that in Algorithm 2, we need to select a sequence of pairs (i, j) from $\Lambda(\mathsf{M}, \ell)$, but the output does not depend on the choice of the sequence. See Colombo and Maathuis (2014) for more discussion on the order-dependent issue of the PC-algorithm. The output of Step 1 is a skeleton $\hat{\mathsf{E}}_{\mathrm{SKE}}$, along with a collection of sets of vertices $\{\mathsf{S}_{i,j}\}$.

The second step of the PC-algorithm transforms the skeleton $\hat{\mathsf{E}}_{\mathrm{SKE}}$ to a CPDAG by applying several deterministic operations based on $\hat{\mathsf{E}}_{\mathrm{SKE}}$ and $\{\mathsf{S}_{i,j}\}$. This step is exactly the same as the classical PC-algorithm (see, e.g., Meek, 1995, Phase I-S2 and Phase II), and its presentation is omitted. We denote the resulting CPDAG as $\hat{\mathsf{E}}_{\text{CPDAG-fCCO}}$ or $\hat{\mathsf{E}}_{\text{CPDAG-fPCO}}$, depending on the thresholding criterion used in (10). We also denote the number of iterations in the functional-PC algorithm as $\hat{\ell}_{\mathrm{fCCO}}$ or $\hat{\ell}_{\mathrm{fPCO}}$.

The next proposition shows that, at the population-level, the output of the PC-algorithm indeed recovers the true CPDAG, and the number of iterations needed is no greater than the maximum degree of the true skeleton. Define a population-level PC-algorithm, denoted by functional-PC^0, where we replace the evaluation of functional conditional independence using (10) with the ground truth of $X_i \perp\!\!\!\perp X_j \mid X_{\mathsf{S}}$. Denote the true edge set of G by $\mathsf{E}_{\mathrm{DAG}}$, the CPDAG of G by $\mathsf{E}_{\mathrm{CPDAG}}$, and the skeleton of $\mathsf{E}_{\mathrm{DAG}}$ by $\mathsf{E}_{\mathrm{SKE}}$. Also, denote the CPDAG from functional-PC^0 by $\mathsf{E}^0_{\mathrm{CPDAG}}$, and the number of iterations in functional-PC^0 by ℓ^0.

Proposition 4 *Suppose Assumptions 1 to 3 hold. Then,*

(i) $\mathsf{E}^0_{\mathrm{CPDAG}} = \mathsf{E}_{\mathrm{CPDAG}}$*;*

(ii) $\ell^0 \leq m$, *where* $m = \max_{i \in \mathsf{V}} |\{j : (i, j) \in \mathsf{E}_{\mathrm{SKE}}\}|$ *is the maximum degree of* $\mathsf{E}_{\mathrm{SKE}}$.

5. Asymptotic theory

We derive the uniform convergence rates of the sample estimates of CCO and PCO, and establish the uniform consistency of the estimated CPDAG. Our asymptotic theory allows both the number of functions p and the number of leading K-L expansion d to diverge with the sample size n. Moreover, many of our theoretical results only require the sub-Gaussian distribution of the random functions, which is weaker than the Gaussian assumption. Nevertheless, the Gaussianity is still needed for CCO and PCO to characterize the conditional independence.

We first formally define a sub-Gaussian, Hilbertian random element.

Definition 3 *Suppose X is an $\mathcal{H}$-valued random element with a trace-class covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$, where $\mathcal{H}$ is a generic separable Hilbert space. We say X follows a sub-Gaussian distribution, if there exists $b > 0$, such that $E\left(\exp\langle f, X\rangle_{\mathcal{H}}\right) \leq \exp\left(b\left\langle \Sigma f, f\right\rangle_{\mathcal{H}}/2\right)$, for all $f \in \mathcal{H}$. We denote it as $X \sim \mathrm{subG}(\Sigma, b)$.*

The notion of sub-Gaussianity in Hilbert space was introduced by Antonini (1997, Definition 1.1). See also Chen and Yang (2021); Mirshani and Reimherr (2021); Zapata et al. (2021); Qin Fang and Qiao (2023); Waghmare et al. (2023). Our Definition is slightly different: Antonini (1997) allows $b = 0$, which leads to a degenerate distribution, whereas we require b to be strictly positive.

5.1 Uniform convergence of CCO estimation

We first study the CCO. We begin with an assumption on the smoothness level of X_i.

Assumption 4 *There exists $\gamma > 1$, such that $\lambda_i^a \asymp a^{-\gamma}$ and $\lambda_i^a - \lambda_i^{a+1} \succeq a^{-1-\gamma}$, as $a \to \infty$, for every $i \in \mathsf{V}$.*

As our estimators are built on the leading K-L coefficients, it is reasonable to assume the tail eigenvalues of $\Sigma_{X_iX_i}$ diminish sufficiently fast. The first part of Assumption 4 implies that $\max_{i\in\mathsf{V}}(\sum_{a=d+1}^{\infty}\lambda_i^a) \preceq d^{-\gamma}$. That is, the decaying rate of the tail eigenvalues of $\Sigma_{X_iX_i}$ is in a polynomial order of d. The second part of Assumption 4 requires the decaying rate of the gaps of the adjacent eigenvalues, $\lambda_i^a - \lambda_i^{a+1}$, to be greater than a polynomial order of a, for $a \in \mathbb{N}$. The parameter γ imposes a level of smoothness of the decaying rate; the larger the value of γ, the faster the decaying rate.

For any $m \in \mathbb{N} \cup \{0\}$, let $\mathsf{H}_0(m) = \{(i, j, \mathsf{S}) \in \mathsf{H}_0 : |\mathsf{S}| \leq m\}$, $\mathsf{H}(m) = \{(i, j, \mathsf{S}) \in \mathsf{H} : |\mathsf{S}| \leq m\}$, and $\mathsf{H}_1(m) = \mathsf{H}(m)\backslash\mathsf{H}_0(m)$. In the following development, we allow $p \to \infty, d \to \infty, m \to \infty, \epsilon \to 0$ as $n \to \infty$. Moreover, given $m \in \mathbb{N}$, let $t(m) = \min\{\|\Sigma_{X_iX_j|X_\mathsf{S}}\|_{\mathrm{HS}} : \Sigma_{X_iX_j|X_\mathsf{S}} \neq 0, (i, j, \mathsf{S}) \in \mathsf{H}_0(m)\}$, and $\zeta(m, d, p, \epsilon, n) = md^{3+\gamma}(\log p)^{1/2}/\ (n^{1/2}\epsilon) + m\epsilon^{-1}d^{-\gamma} + \epsilon^{1/2}s(m)$. Here $t(m)$ is the minimal H-S norm of the nonzero CCO, and thus $t(m) \preceq 1$. Next, we introduce an assumption on $t(m)$.

Assumption 5 *Suppose $t(m) \succ \zeta(m, d, p, \epsilon, n)$.*

Assumption 5 places a lower bound on the order of $t(m)$ to prevent it from diminishing too fast. Note that, in the random variable setting, the partial correlation-based PC-algorithm requires a strong faithfulness assumption to ensure the uniform consistency of the estimated graph (Uhler et al., 2013). Assumption 5 is similar and can be viewed as the functional version of strong faithfulness. Also note that, we require $t(m)$ to go to 0 at a slower rate than $\zeta(m, d, p, \epsilon, n)$, whose first term is $md^{3+\gamma}(\log p)^{1/2}/(n^{1/2}\epsilon)$. By comparison, for the partial correlation-based PC-algorithm, the order of the minimal nonzero partial correlation has to be greater than $(m \log p/n)^{1/2}$ (Uhler et al., 2013).

We next establish the uniform convergence rate of $\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_{\mathsf{S}}}$, and the uniform consistency of the estimated CPDAG, $\hat{\mathsf{E}}_{\text{CPDAG-fCCO}}$, based on CCO.

Theorem 3 *(i) Uniform convergence rate of $\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_{\mathsf{S}}}$: Suppose Assumptions 1, 3, 4, and 5 hold, $X_i \sim \text{subG}(\Sigma_{X_iX_i}, b_0)$ with $E(X_i) = 0$, $\epsilon \prec 1$, $md^{3+\gamma}(\log p)^{1/2}/(n^{1/2}\epsilon) \preceq 1$, and the threshold $\rho_{\text{CCO}} = t(m)/2$. Then,*

$$\max_{(i,j,\mathsf{S})\in\mathsf{H}(m)} \|\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_{\mathsf{S}}} - \Sigma_{X_iX_j|X_{\mathsf{S}}}\|_{\text{HS}} = O_P[md^{3+\gamma}(\log p)^{1/2}/(n^{1/2}\epsilon) + m\epsilon^{-1}d^{-\gamma} + \epsilon^{1/2}s(m)],$$

where $s(m) = \max_{(i,i,\mathsf{S})\in\mathsf{H}_1(m)} \|M_{X_{\mathsf{S}}X_i}\|_{\text{HS}}$. (ii) Uniform graph consistency based on CCO: If we further assume Assumption 2 holds, then,

$$P(\hat{\mathsf{E}}_{\text{CPDAG-fCCO}} = \mathsf{E}^0_{\text{CPDAG}}) \to 1 \quad \textit{and} \quad P(\hat{\ell}_{\text{fCCO}} = \ell^0) \to 1, \quad \textit{as } n \to \infty.$$

Theorem 3 requires the maximal degree m to satisfy $md^{3+\gamma}(\log p)^{1/2}/(n^{1/2}\epsilon) \preceq 1$, which implies that m can grow at most at a polynomial rate, and thus in turn imposes a level of sparsity on the graph. This rate for m is consistent with the classical settings. For instance, in the sparse regression, the order of magnitude of the sparsity parameter can only grow in a polynomial order of n. The classical linear PC-algorithm (Kalisch and Bühlmann, 2007, condition A3) also requires m to grow in a polynomial order of n.

5.2 Uniform convergence of PCO estimation

We next study the PCO. We show that the norm of $\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_{\mathsf{S}}}$ is no greater than 1, which resembles the property of the partial correlation. We then introduce two assumptions.

Proposition 5 *For each $(i,j,\mathsf{S}) \in \mathsf{H}_0$, we have $\|\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_{\mathsf{S}}}\| \leq 1$.*

Assumption 6 *There exists $c_0 > 0$, such that $\max\left\{\sum_{a\in\mathbb{M}_{i,\mathsf{S}}}\sum_{b\in\mathbb{M}_{j,\mathsf{S}}}(\rho^{a,b}_{i,j,\mathsf{S}})^2/(\mu^a_{i,\mathsf{S}}\mu^b_{j,\mathsf{S}}) : (i,j,\mathsf{S}) \in \mathsf{H}_0\right\} \leq c_0$, where $\rho^{a,b}_{i,j,\mathsf{S}} = \text{cor}(\langle\nu^a_{i,\mathsf{S}}, X_i\rangle_{\Omega_{X_i}}, \langle\nu^b_{j,\mathsf{S}}, X_j\rangle_{\Omega_{X_j}})$, $\mathbb{M}_{i,\mathsf{S}} = \{a \in \mathbb{N} : \mu^a_{i,\mathsf{S}} > 0\}$, $\mu^1_{i,\mathsf{S}} \geq \mu^2_{i,\mathsf{S}} \geq \cdots$ and $\nu^1_{i,\mathsf{S}}, \nu^2_{i,\mathsf{S}} \cdots$ are the eigenvalues and eigenfunctions of $\Sigma_{X_iX_i|X_{\mathsf{S}}}$.*

Assumption 6 places a level of smoothness on the relation between X_i and X_j given X_{S}. Under the Gaussian assumption, the correlation $\rho^{a,b}_{i,j,\mathsf{S}}$ measures the strength of dependency between X_i and X_j given X_{S}. Moreover, because $\Sigma_{X_iX_i|X_{\mathsf{S}}}$ is a trace-class operator, its eigenvalues decay to 0 sufficiently fast so that $\sum_{a\in\mathbb{N}}\mu^a_{i,\mathsf{S}} < \infty$. Assumption 6 implies that $\rho^{a,b}_{i,j,\mathsf{S}}$ needs to converge to 0 faster than the product $\mu^a_{i,\mathsf{S}}\mu^b_{j,\mathsf{S}}$, as $a \to \infty, b \to \infty$. Hence, intuitively, the dependency between X_i and X_j given X_{S} has to be sufficiently concentrated on the leading eigenfunctions.

Similar to Assumption 5 on CCO, we also require the minimum H-S norm of the nonzero PCO to be sufficiently large. For $m \in \mathbb{N}$, let $u(m) = \min\{\|R_{X_iX_j|X_{\mathsf{S}}}\|_{\text{HS}} : R_{X_iX_j|X_{\mathsf{S}}} \neq 0, (i,j,\mathsf{S}) \in \mathsf{H}_0(m)\}$. The next assumption is the functional version of strong faithfulness based on PCO, which prevents $u(m)$ to go to zero too fast.

Assumption 7 *Suppose $u(m) \succ \delta^{-3/2}\zeta(m,d,p,\epsilon,n) + \delta^{1/2}$.*

We next establish the uniform convergence rate of $\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_\mathsf{S}}$, and the uniform consistency of the estimated CPDAG, $\hat{\mathsf{E}}_{\text{CPDAG-fPCO}}$, based on PCO.

Theorem 4 *(i) Uniform convergence rate of PCO: Suppose Assumptions 1, 3, 4, 6, and 7 hold, $X_i \sim \text{subG}(\Sigma_{X_iX_i}, b_0)$ with $E(X_i) = 0$, $\zeta(m, d, p, \epsilon, n) \preceq 1$, $\delta \prec 1$, and the threshold $\rho_{\text{PCO}} = u(m)/2$. Then,*

$$\max_{(i,j,\mathsf{S})\in\mathsf{H}_0(m)} \|\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_\mathsf{S}} - R_{X_iX_j|X_\mathsf{S}}\|_{\text{HS}} = O_P[\delta^{-3/2}\zeta(m, d, p, \epsilon, n) + \delta^{1/2}].$$

(ii) Uniform graph consistency based on PCO: If we further assume Assumption 2 holds, then,

$$P(\hat{\mathsf{E}}_{\text{CPDAG-fPCO}} = \mathsf{E}_{\text{CPDAG}}) \to 1 \quad \textit{and} \quad P(\hat{\ell}_{\text{fPCO}} = \ell^0) \to 1, \quad \textit{as } n \to \infty.$$

Bühlmann and van de Geer (2011, Theorem 13.1) derived the consistency of the PC-algorithm of Kalisch and Bühlmann (2007) for the random variable setting. We next compare our Theorem 4 to theirs. Specifically, both results establish the uniform consistency of the parameter estimation and the uniform graph consistency: whereas Bühlmann and van de Geer (2011) was based on the partial correlation, our result is based on the proposed partial correlation operator. As such, Theorem 4 can be viewed as the functional extension of the above-mentioned Theorem 13.1, though such an extension is far from trivial. The conditions imposed by the two theorems are generally similar. Both allow the graph size p to diverge at an exponential order of the sample size n, and both require the maximum degree of the DAG m to diverge at a polynomial order of n. However, there are some minor differences. For example, Bühlmann and van de Geer (2011, condition A4) required the absolute value of the smallest non-zero partial correlation to go to zero at a rate slower than $(m \log p/n)^{1/2}$. By contrast, our Assumption 7 requires the minimal H-S norm of all non-zero partial correlation operators $u(m)$ to converge to zero at a rate slower than $\delta^{-3/2}md^{3+\gamma}(\log p)^{1/2}/(n^{1/2}\epsilon)$. Our rate is slower than that of Bühlmann and van de Geer (2011), but we feel this is reasonable, as our setting involves infinite-dimensional functions and is more complicated. Also, Bühlmann and van de Geer (2011, condition A4) imposed a regularization on the dependency among the random variables by upper-bounding the partial correlations. We impose a similar condition, by requiring the maximum of the H-S norms of the regression operator $M_{X_\mathsf{S}X_i}$, i.e., $\epsilon^{1/2}s(m) \preceq 1$, for any $i \in \mathsf{V}$ and any subset $\mathsf{S} \subseteq \mathsf{V}\backslash\{i\}$ with $|\mathsf{S}| \leq m$. Note that ϵ goes to zero in a polynomial rate of n. If $\epsilon \asymp n^{-b}$ with some $b \in (0, 1/2)$, then this implies that $s(m) \preceq n^b$. As such, our condition also regulates the dependency among the random functions.

Following up our discussion in the last paragraph of Section 4.1, using the estimator defined in (4), Theorem 4 establishes the uniform convergence rate of our proposed PCO estimator $\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_\mathsf{S}}$ in (4). By contrast, the theoretical analysis of the alternative estimator $\tilde{R}^d_{X_iX_j|X_\mathsf{S}}$ in (5) would have involved inversion of $\hat{\Sigma}^{d,\epsilon}_{X_iX_i|X_\mathsf{S}}$, where the norm of this inverse is identical to the smallest eigenvalue of $\hat{\Sigma}^{d,\epsilon}_{X_iX_i|X_\mathsf{S}}$. Let $\mu^d_{i,\mathsf{S}}$ denote the smallest eigenvalue of $\Sigma^d_{X_iX_i|X_\mathsf{S}}$, i.e., the population version of $\hat{\Sigma}^{d,\epsilon}_{X_iX_i|X_\mathsf{S}}$. Then, it is easy to see that, the rate of convergence of $\tilde{R}^d_{X_iX_j|X_\mathsf{S}}$ in (5) depends on the rate at which $\mu^d_{i,\mathsf{S}}$ approaches to zero. On the other hand, the norm of the ridge-type estimator $(\hat{\Sigma}^{d,\epsilon}_{X_iX_i|X_\mathsf{S}} + \delta I)^{-1}$ in $\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_\mathsf{S}}$ in (4) is upper bounded by δ^{-1}, regardless of the magnitude of $\mu^d_{i,\mathsf{S}}$. Consequently, the uniform rate of convergence of $\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_\mathsf{S}}$ in (4) is independent of $\mu^d_{i,\mathsf{S}}$. Therefore, the asymptotic analysis of our PCO estimator is simpler than that of the alternative estimator defined by (5).

6. Numerical studies

We first evaluate the empirical performance of the proposed method through two simulation examples, then illustrate with an analysis of a time-course proteomic dataset.

6.1 Simulations

Without loss of generality, we assume that $\{1, 2, \ldots, p\}$ is the order of the true DAG. Moreover, let $\mathrm{pa}(i) = \{j : (j, i) \in \mathsf{E}_{\mathrm{DAG}}\}$ denote the parents of i, and $\{u_1, \ldots u_k\}$ a k-grid in $[0, 1]$ with $u_1 = 1/k, \ldots, u_k = 1$. We generate the p-dimensional vector of random functions $X(t) = [X_1(t), \ldots, X_p(t)]^\intercal$ in a sequential manner as,

$$\text{Model I} : X_1(t) = \epsilon_1(t),\;\; X_i(t) = \textstyle\sum_{(j,i)\in\mathsf{E}_{\mathrm{DAG}}} X_j(t) + \epsilon_i(t),\; i = 2, \ldots, p,$$
$$\text{Model II} : X_i(t) = U_{i1}\eta_1(t) + U_{i2}\eta_2(t),\; U_{i2} = (|\mathrm{pa}(i)| + 1)^{-1}\big(\textstyle\sum_{j\in\mathrm{pa}(i)} U_{j2} + \epsilon_i\big),\; i = 1, \ldots, p.$$

For Model I, the error functions $\epsilon_i(t) = \sum_{j=1}^{v} \xi_j \kappa(t, s_j), i = 1, \ldots, p,$ are i.i.d. Gaussian random process with the Brownian motion covariance function, where $\kappa(t, s) = \min(t, s)$, $\{s_1, \ldots, s_v\}$ is a v-grid in $[0, 1]$, $\xi_1, \ldots, \xi_v$ are i.i.d. normal variables with mean zero and standard deviation 5, and $v = 10$. For Model II, $U_{11}, \ldots, U_{p1}, \epsilon_1, \ldots, \epsilon_p$ are i.i.d. standard normal variables, $\eta_k(t) = a_k\chi_k(t)$, $\chi_k(t) = \sqrt{2}\sin((k - 0.5)\pi t)$ is the kth eigenfunction of the Brownian motion kernel, $k = 1, 2$, $a_1 = 2.5$, and $a_2 = 1$. We assume all subjects are observed at the same set of time points, $T_k = \{t_1, \ldots, t_u\}$, which is taken to be an u-grid in $[0, 1]$ with $u = 50$. We generate the edge set $\mathsf{E}_{\mathrm{DAG}}$ via the independent Bernoulli variable $I[(i, j) \in \mathsf{E}_{\mathrm{DAG}}]$, with $P\{I[(i, j) \in \mathsf{E}_{\mathrm{DAG}}]\} = 2q/(p - 1)$. The expected number of edges in $\mathsf{E}_{\mathrm{DAG}}$ is qp, with q controlling the sparsity of the graph, and $q = 1.05$.

We apply the proposed CCO and PCO estimators to the simulated data. We employ the Brownian motion kernel to construct Ω_{X_i} as the span of $\{\kappa_T(\cdot, t_s) : s = 1, \ldots, u\}$. Besides, we choose all the tuning parameters following the rules outlined in Algorithm 1. Our preliminary results have found that PCO outperforms CCO consistently, due to the benefit of proper scaling. As such, we only report the PCO results subsequently.

We also compare with some alternative methods. Specifically, we consider the PC-algorithm based on the partial correlation test (linear-PC, Spirtes et al., 2000), the PC-algorithm based on the rank correlation test (rank-PC, Harris and Drton, 2013), and the causal additive models based on high-dimensional penalized regressions (CAM, Bühlmann et al., 2014). To adapt them to the functional setting, for each subject k and node i, we first extract from the observed function X^k_i the first K-L expansion coefficient $\hat{c}^{k,1}_i = \langle X^k_i - E_nX_i, \hat{\eta}^1_i\rangle_{\Omega_{X_i}}$ using (7), which is the first functional PCA score. We then apply the three competing methods to the sample of the p-dimensional vectors, $\{(\hat{c}^{k,1}_1, \ldots, \hat{c}^{k,1}_p)^\intercal : k = 1, \ldots, n\}$, to estimate the CPDAG. We comment that such an adaption is intuitive, but there is no theoretical guarantee. We also note that the linear-PC method corresponds to the alternative estimator $\tilde{R}^d_{X_iX_j|X_\mathsf{S}}$ in (5) with $d = 1$.

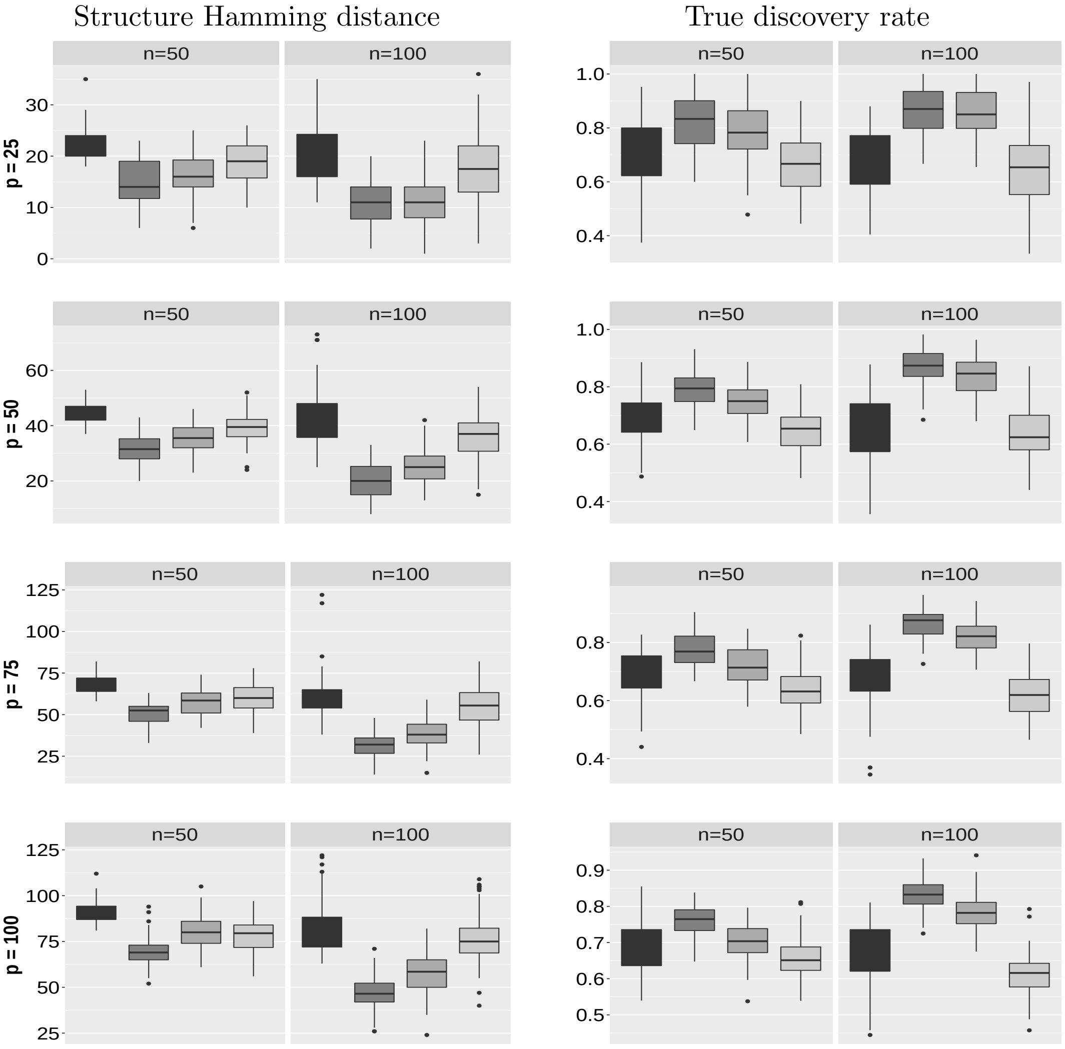


Figure 1: Empirical performance under Model I. Four methods are compared, from left to right, the modified PC-algorithm based on PCO, linear-PC, rank-PC, and CAM.

We evaluate the performance by two criteria, the structure Hamming distance (SHD, Tsamardinos et al., 2006), and the true discovery rate (TDR), which are defined as,

$$\begin{aligned}
\mathrm{SHD}(\hat{\mathsf{E}}_{\textsc{cpdag}}, \mathsf{E}_{\textsc{cpdag}}) = {} & |\hat{\mathsf{E}}_{\textsc{cpdag}} \cup \hat{\mathsf{E}}^{\mathsf{T}}_{\textsc{cpdag}} - \mathsf{E}_{\textsc{cpdag}} \cup \mathsf{E}^{\mathsf{T}}_{\textsc{cpdag}}|/2 \\
& + |\mathsf{E}_{\textsc{cpdag}} \cup \mathsf{E}^{\mathsf{T}}_{\textsc{cpdag}} - \hat{\mathsf{E}}_{\textsc{cpdag}} \cup \hat{\mathsf{E}}^{\mathsf{T}}_{\textsc{cpdag}}|/2 \\
& + |\hat{\mathsf{E}}_{\textsc{cpdag}} - (\mathsf{E}_{\textsc{cpdag}} \cup \hat{\mathsf{E}}^{\mathsf{T}}_{\textsc{cpdag}} - \mathsf{E}_{\textsc{cpdag}} \cup \mathsf{E}^{\mathsf{T}}_{\textsc{cpdag}}) - \mathsf{E}_{\textsc{cpdag}}|, \\
\mathrm{TDR}(\hat{\mathsf{E}}_{\textsc{cpdag}}, \mathsf{E}_{\textsc{cpdag}}) = {} & |\{(i,j) \in \hat{\mathsf{E}}_{\textsc{cpdag}} : (i,j) \in \mathsf{E}_{\textsc{cpdag}}\}|/|\hat{\mathsf{E}}_{\textsc{cpdag}}|,
\end{aligned}$$

where, for an edge set E, E^{T} stands for $\{(j,i) : (i,j) \in \mathsf{E}\}$, $\mathsf{E}_{\textsc{cpdag}}$ is the true CPDAG, and $\hat{\mathsf{E}}_{\textsc{cpdag}}$ is its estimate. We note that the three terms in $\mathrm{SHD}(\hat{\mathsf{E}}_{\textsc{cpdag}}, \mathsf{E}_{\textsc{cpdag}})$ represent the numbers of deletions, insertions, and reorientations needed to transform $\hat{\mathsf{E}}_{\textsc{cpdag}}$ to $\mathsf{E}_{\textsc{cpdag}}$,

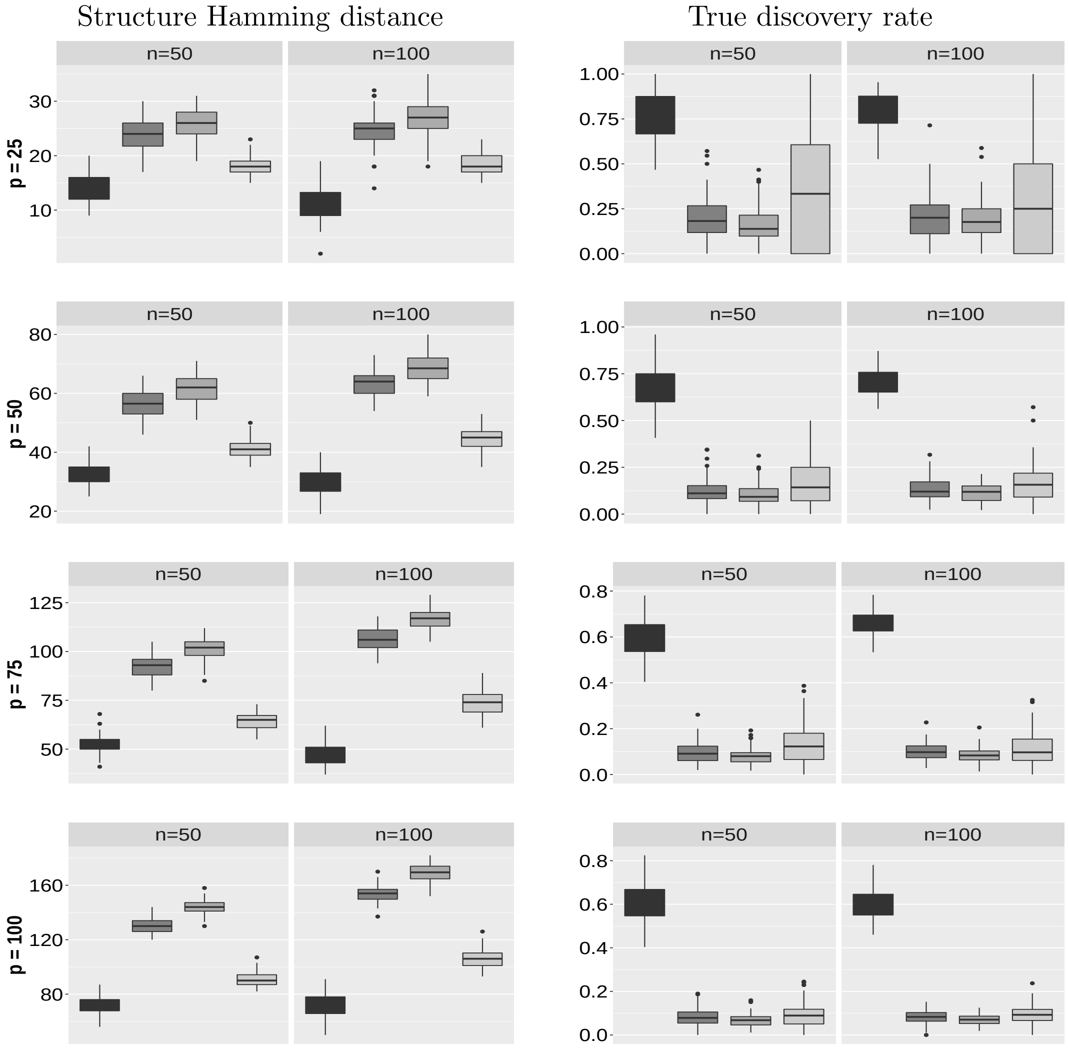


Figure 2: Empirical performance under Model II. Four methods are compared, from left to right, the modified PC-algorithm based on PCO, linear-PC, rank-PC, and CAM.

respectively. A smaller SHD or a higher TDR indicates more accurate estimation. We choose to report the true discovery rate, instead of the true positive rate or false positive rate, because the underlying DAG is sparse, and the proportions of the true positives and negatives are highly imbalanced.

We vary the sample size n in $\{50, 100\}$, and the graph size p in $\{25, 50, 75, 100\}$, resulting in 8 different scenarios. Figures 1 and 2 report the box plots for SHD and TDR based on 80 data replications, for Model I and Model II, respectively. We observe that the proposed PCO-based method has a comparable performance as the alternative methods in Model I, but clearly outperforms the alternatives in Model II. This is because the first PC captures most of variation in Model I, but not so in Model II. Moreover, the performance of PCO improves as the sample size increases, which agrees with our asymptotic theory.

We have also conducted additional simulations with more combinations of (n, p, q), different initializations, different kernel functions, and comparison with the SEM method of Lee and Li (2022). We report those results in Section A.5 of the Appendix. Overall our proposed PCO-based method achieves a competitive empirical performance.

6.2 Proteomic application

We illustrate our method with a DREAM breast cancer proteomic dataset (`https://www.synapse.org/#!Synapse:syn1720047/wiki/56213`). The goal of the study is to estimate the directed relations among different proteins given the time-course proteomic measurements. We consider the in silico data generated using a nonlinear dynamic model whose characteristic satisfies the Reverse Phase Protein Array (RPPA) quantitative proteomics technology (Hill et al., 2016). Based on various combinations of stimuli and inhibitors, the true network, as shown in the first panel in Figure 3, is used to generate the time-courses of phosphoprotein abundance levels. There are a total of 20 different conditions, and for each condition, 3 independent copies of 20 time-course protein levels were collected at time $t = 0, 1, 2, 4, 6, 10, 15, 30, 45, 60, 120$ minutes. After the removal of 4 conditions whose protein levels had unusual distributions, the data consists of $n = 48$ subjects, each with $p = 20$ protein levels measured at $m = 11$ time points. Figure 11 in the Appendix plots the time-course data for all 20 protein levels.

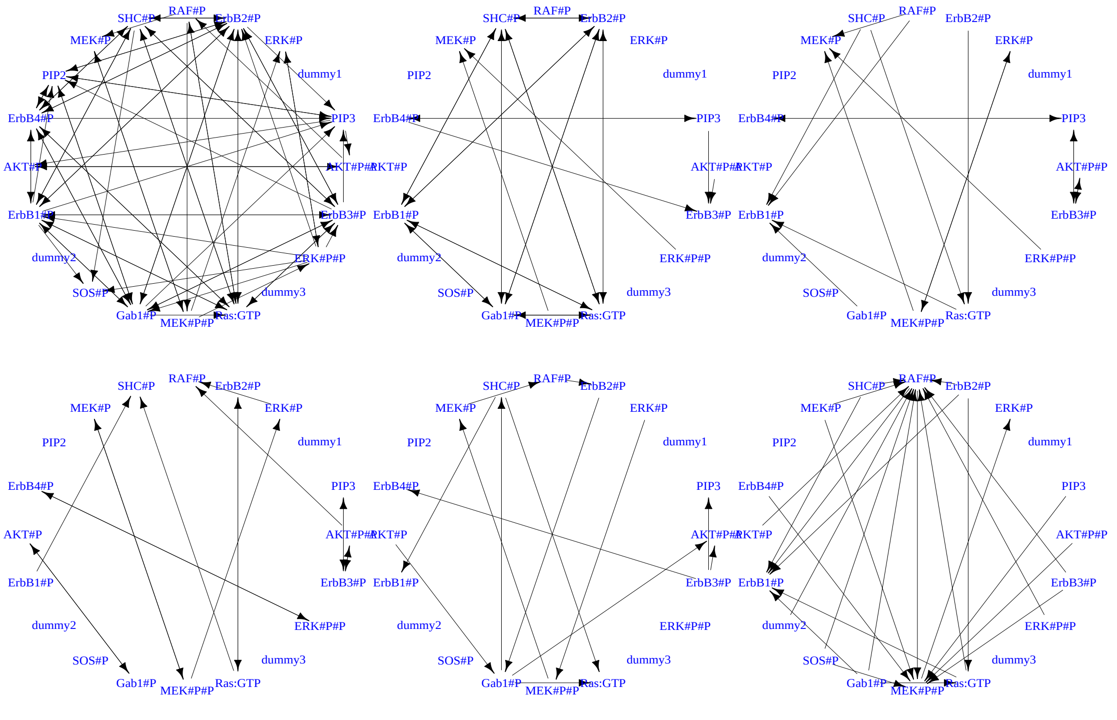


Figure 3: True and estimated graphs for the time-course proteomic data. From left to right, top to bottom: the truth, the modified PC-algorithm based on PCO, linear-PC, rank-PC, CAM, and SEM.

Table 1: The performance and comparison for the time-course proteomic data.

Method	PCO	linear-PC	rank-PC	CAM	SEM
Structure Hamming distance	**44**	50	48	51	50
True discovery rate	**0.70**	0.59	0.53	0.48	0.40

Figure 3 reports the estimated graphs by our method, linear-PC, rank-PC, CAM, and the SEM method of Lee and Li (2022), while Table 1 reports the corresponding structure Hamming distance and true discovery rate. We see that our PCO-based method performs the best, by achieving the smallest SHD and the highest TDR.

7. Discussions

In this section, we discuss the interpretation and implication of the functional DAG model, including its relation to the linear structural equation, the factorization of joint distribution, the causal interpretation, and the comparison to an undirected graph.

7.1 Relation to the functional linear structural equation model

We show that there is a one-to-one correspondence between the functional DAG and the functional linear structural equation model. Such a relation reveals how the functional DAG factorizes the joint distribution, and allows us to better interpret and understand the identified edges of the DAG.

We first formally define the linear structural equation model (SEM) for Hilbert space-valued random functions. For node $i \in \mathsf{V}$, subset $\mathsf{S} \subseteq \mathsf{V}\backslash\{i\}$, and linear operator $B(i, \mathsf{S}) \in \mathscr{B}(\Omega_{X_i}, \Omega_{X_\mathsf{S}})$, let $B_j(i, \mathsf{S})$ denote the jth suboperator of $B(i, \mathsf{S})$.

Definition 4 *We say that $X = (X_1, \ldots, X_p)^\intercal$ follows a zero-mean, linear structural equation model with respect to a DAG G if, for each $i \in \mathsf{V}$, there exists a $B(i, \text{pa}(i)) \in \mathscr{B}(\Omega_{X_i}, \Omega_{X_{\text{pa}(i)}})$, such that*

$$X_i = \textstyle\sum_{j \in \text{pa}(i)} B_j^*(i, \text{pa}(i)) X_j + \epsilon_i,$$

where ϵ_i is a zero-mean random element in Ω_{X_i}, and $\epsilon_1, \ldots, \epsilon_p$ are independent.

We next recall the notion of the global Markov property from (1). That is, $X = (X_1, \ldots, X_p)^\intercal$ satisfies the global Markov property with respect to G, if

$$i \text{ and } j \text{ are d-separated by } \mathsf{S} \text{ in } \mathsf{G} \quad \Rightarrow \quad X_i \perp\!\!\!\perp X_j \mid X_\mathsf{S}. \tag{11}$$

The next theorem establishes the equivalence between the functional linear SEM in Definition 4 and the global Markov property in (11) under the Gaussian assumption.

Theorem 5 *Suppose Assumptions 1 and 3 are satisfied, and $X = (X_1, \ldots, X_p)^\intercal$ is a zero-mean, Gaussian random element in Ω_{X_i}. Then the following two statements are equivalent: (i) X satisfies the global Markov property with respect to G, and (ii) X follows a linear structural equation model with respect to G.*

The structural equation factorizes the joint distribution of $X_1, \ldots, X_p$ into the product of the set of conditional distributions of $X_i \mid X_{\mathrm{pa}(i)}$, $i = 1, \ldots, p$, under G, and provides an interpretation of the edge directions. Consequently, we can interpret the edges of the functional DAG following the functional linear SEM. For instance, in the proteomic application, let $X_{\texttt{SOS}}, X_{\texttt{ERK}}, X_{\texttt{SHC}}$ and $X_{\texttt{ERBB1}}$ denote the random functions of the protein levels of $\texttt{SOS}, \texttt{ERK}, \texttt{SHC}$, and $\texttt{ERBB1}$, respectively, and suppose they all have zero-means. Because $\texttt{ERK}, \texttt{SHC}$, and $\texttt{ERBB1}$ are the parent nodes of $\texttt{SOS}$ in the ground truth, we have,

$$X_{\texttt{SOS}} = B_1^* X_{\texttt{ERK}} + B_2^* X_{\texttt{SHC}} + B_3^* X_{\texttt{ERBB1}} + \epsilon,$$

where ϵ is a zero-mean, Gaussian random error function that is independent of $X_{\texttt{ERK}}, X_{\texttt{SHC}}$ and $X_{\texttt{ERBB1}}$, and B_1^*, B_2^*, B_3^* are the linear operators from the ranges of $X_{\texttt{ERK}}, X_{\texttt{SHC}}$ and $X_{\texttt{ERBB1}}$, to the range of $X_{\texttt{SOS}}$, respectively.

7.2 Potential causal interpretation

Next, we introduce the *do-intervention* under the functional setting, and discuss potential causal interpretation of the functional DAG model.

For any $j \in \mathsf{V}$, and any $x = (x_1, \ldots, x_p) \in \Omega_X$, let $x_{\mathrm{pa}(j)} = \{x_k : k \in \mathrm{pa}(j)\}$, and let $\mathrm{LSE}_{j,\mathsf{G}} : \Omega_{X_{\mathrm{pa}(j)}} \times \Omega_{X_j} \to \Omega_{X_j}$ denote the mapping

$$\mathrm{LSE}_{j,\mathsf{G}}(x_{\mathrm{pa}(j)}, x_j) = \textstyle\sum_{k \in \mathrm{pa}(j)} B_k^*(j, \mathrm{pa}(j)) x_k + x_j.$$

For $y \in \Omega_X$, $\mathsf{A} \subseteq \mathsf{V}$, $i \in \mathsf{A}$, and $x_i \in \Omega_{X_i}$, let $y_{\mathsf{A}}(y_i \to x_i)$ be the vector $y_{\mathsf{A}} = \{y_k : k \in \mathsf{A}\}$ with its member y_i replaced by x_i. In other words, $y_{\mathsf{A}}(y_i \to x_i) = (y_{\mathsf{A}} \backslash \{y_i\}) \cup \{x_i\}$. Following Pearl (2009, Definition 3.2.1), we obtain the following definition.

Definition 5 *Suppose $X = (X_1, \ldots, X_p)^\mathsf{T}$ follows a linear structural equation model with respect to a DAG G, a set of regression operators $B(j, \mathrm{pa}(j)) \in \mathscr{B}(\Omega_{X_j}, \Omega_{X_{\mathrm{pa}(j)}})$, $j = 2, \ldots, p$, and the error random functions $\{\epsilon_j : j \in \mathsf{V}\}$. For a fixed $i \in \mathsf{V}$, and $x_i \in \Omega_{X_i}$, the causal effect of $X_i = x_i$ on $X_{\mathsf{V}\backslash\{i\}}$ is the joint distribution of $X_{\mathsf{V}\backslash\{i\}}$ induced by the following $p-1$ equations:*

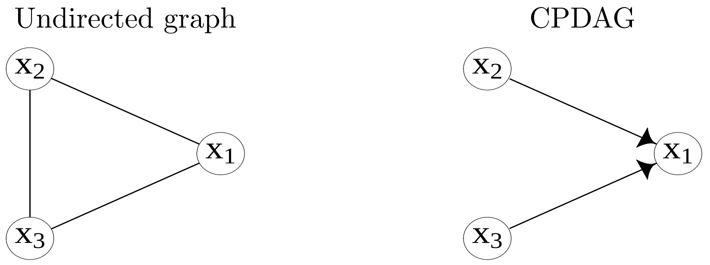
$$X_j = \begin{cases} \mathrm{LSE}_{j,\mathsf{G}}(X_{\mathrm{pa}(j)}(X_i \to x_i), \epsilon_j), & \text{if } \mathrm{pa}(j) \ni i, \\ \mathrm{LSE}_{j,\mathsf{G}}(X_{\mathrm{pa}(j)}, \epsilon_j), & \text{if } \mathrm{pa}(j) \not\ni i, \end{cases}$$

for all $j \in \mathsf{V}\backslash\{i\}$. We denote such a joint distribution as $P_{X_{\mathsf{V}\backslash\{i\}}|\mathrm{do}(x_i)}$, and call this distribution the interventional distribution at $X_i = x_i$.

For any $j \in \mathsf{V}\backslash\{i\}$, let $P_{X_j|\mathrm{do}(x_i)}$ denote the marginal distribution of X_j in the interventional distribution at $X_i = x_i$, P_{X_j} the marginal distribution of X_j, and $P_{X_j|x_i}$ the conditional distribution of $X_j \mid X_i = x_i$. Then by Definition 5, we have,

$$P_{X_j|\mathrm{do}(x_i)} = \begin{cases} P_{X_j}, & \text{if } j \text{ is not a descendant of } i, \\ P_{X_j|x_i}, & \text{if } j \text{ is a descendant of } i. \end{cases}$$

Definition 5 offers one way to define the causal effect in functional data, and is general and potentially useful for time-course interventional studies (Luo and Zhao, 2011). Based


Undirected graph

CPDAG

Figure 4: The induced undirected graph structure and the directed graph structure, based on the relation in model (12).

on this definition, it is possible to develop a full methodology and theory for modeling functional interventional data, following the lines of Maathuis et al. (2009); Hauser and Bühlmann (2015), who studied the random variable-based linear SEM for interventional data. However, it requires a substantial amount of additional efforts, and to avoid too much digestion, we leave it for future research.

7.3 Comparison with undirected graph

Finally, we illustrate the difference between the undirected functional graphical model and our functional DAG model by a specific example.

Example 1 Suppose $X = (X_1, X_2, X_3)^\intercal$ is a random element in $\mathcal{H} \times \mathcal{H} \times \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space, and $\epsilon_1, \epsilon_2, \epsilon_3$ are i.i.d. random elements in $\mathcal{H}$ with zero-mean and the covariance operator Λ. Furthermore, suppose

$$X_1 = \epsilon_1, \quad X_2 = \epsilon_2, \quad X_3 = X_1 + X_2 + \epsilon_3. \tag{12}$$

Note that the conditional covariance operator between X_1 and X_2 given X_3, by definition, is $\Sigma_{X_1 X_2 | X_3} = \Sigma_{X_1 X_2} - \Sigma_{X_1 X_3} \Sigma^\dagger_{X_3 X_3} \Sigma_{X_3 X_2} = -\Lambda \Lambda^\dagger \Lambda = -\Lambda$. Similarly, the conditional covariance operators $\Sigma_{X_1 X_3 | X_2} = \Sigma_{X_2 X_3 | X_1} = \Lambda$. Therefore, by Theorem 5, we have,

$$X_1 \not\perp\!\!\!\perp X_2 \mid X_3, \quad X_1 \not\perp\!\!\!\perp X_3 \mid X_2, \quad X_2 \not\perp\!\!\!\perp X_3 \mid X_1.$$

Figure 4 shows the undirected graph structure and the CPDAG, both induced by the relation in (12). We see that the the two graphs are very different. The undirected graphical model studied in Qiao et al. (2019); Li and Solea (2018) does not offer any structural simplification of the joint distribution of X_1, X_2, X_3 in this example, because there is no zero entry in the precision operator. On the other hand, the directed graphical model targeted by the functional DAG does provide a simplification of the joint distribution.

Appendix A. Appendix

In this Appendix, we first present some supporting theoretical results in Section A.1. We then prove the two main theorems, Theorems 3 and 4, in Sections A.2 and A.3. We collect the proofs of the rest of the theoretical results in Section A.4. We present additional numerical results in A.5.

A.1 Supporting theoretical results

We first derive a useful property of a zero-mean, Hilbert space-valued, Gaussian random element. Let $\pi : \mathsf{V} \to \mathsf{V}$ denote a permutation, i.e., an injective mapping from V to V. Let $[a]$ denote the vector $(1, \dots, a)^\intercal$ for integer $a \geq 1$.

Lemma 1 *Suppose Assumptions 1 and 3 hold, and the p-variate random function $X = (X_1, \dots, X_p)^\intercal$ is a zero-mean Gaussian random element in Ω_X. Then, for $i = 2, \dots, p$, there exists a linear operator $B(i, [i-i]) \in B_2(\Omega_{X_i}, \Omega_{X_{[i-1]}})$, and an Ω_{X_i}-valued random element ϵ_i, such that*

$$X_i = \textstyle\sum_{j=1}^{i-1} B_j^*(i, [i-1]) X_j + \epsilon_i. \tag{13}$$

Moreover, for $i = 2, \dots, p$, $B(i, [i-i]) = M_{X_{[i-1]}X_i}$, and $X_1, \epsilon_2, \dots, \epsilon_p$ are independent, zero-mean Gaussian random element in Ω_{X_i}, with $E(\epsilon_i \otimes \epsilon_i) = \Sigma_{X_i X_i | X_{[i-1]}} \in \mathscr{B}_1(\Omega_{X_i})$. This statement remains true if we replace $1, \dots, p$ by $\pi(1), \dots, \pi(p)$.

Proof of Lemma 1: We first show that, for any $i \in \mathsf{V}$ and any $\mathsf{S} \subseteq \mathsf{V} \backslash \{i\}$, $E(X_i \mid X_\mathsf{S}) \in \Omega_{X_i}$. This is because, under Assumptions 1 and 3, $M_{X_\mathsf{S} X_i}$ is defined and by Proposition 1, $E(X_i \mid X_\mathsf{S}) = M^*_{X_\mathsf{S} X_i} X_\mathsf{S}$, which is a member of Ω_{X_i}.

For $i = 2, \dots, p$, let $\epsilon_i = X_i - E(X_i \mid X_{[i-1]})$. Then $X_i = E(X_i \mid X_{[i-1]}) + \epsilon_i$. By Proposition 1, we have $X_i = M^*_{X_{[i-1]}X_i} X_{[i-1]} + \epsilon_i = \sum_{j=1}^{i-1} (M^*_{X_{[i-1]}X_i})_j X_j + \epsilon_i$. Therefore, (13) holds with $B(i, [i-1]) = M_{X_{[i-1]}X_i}$.

Next, we show that $X_1, \epsilon_2, \dots, \epsilon_p$ are independent. For convenience, denote X_1 by ϵ_1. Since $\epsilon_1, \dots, \epsilon_p$ are jointly Gaussian, we only need to check

$$\mathrm{cov}[\langle f, \epsilon_i \rangle_{\Omega_{X_i}}, \langle g, \epsilon_j \rangle_{\Omega_{X_j}}] = \langle f, \Sigma_{\epsilon_i \epsilon_j} g \rangle_{\Omega_{X_i}} = 0,$$

for every $f \in \Omega_{X_i}, g \in \Omega_{X_j}$. Suppose $i < j$. We have that,

$$\begin{aligned} \Sigma_{\epsilon_i \epsilon_j} = E(\epsilon_i \otimes \epsilon_j) &= E\{E[(X_i - E(X_i \mid X_{[i-1]})) \otimes (X_j - E(X_j \mid X_{[j-1]})) \mid X_{[i]}]\} \\ &= E\{(X_i - E(X_i \mid X_{[i-1]})) \otimes E(X_j - E(X_j \mid X_{[j-1]}) \mid X_{[i]})\} \\ &= E\{(X_i - E(X_i \mid X_{[i-1]})) \otimes [E(X_j \mid X_{[i]}) - E(X_j \mid X_{[i]})]\} = 0. \end{aligned}$$

It remains to show $\Sigma_{\epsilon_i \epsilon_i} = \Sigma_{X_i X_i | X_{[i-1]}}$ and $\Sigma_{\epsilon_i \epsilon_i}$ is a member of $\mathscr{B}_1(\Omega_{X_i})$. By definition,

$$\begin{aligned} \Sigma_{\epsilon_i \epsilon_i} &= E(X_i \otimes X_i) + E[E(X_i \mid X_{[i-1]}) \otimes E(X_i \mid X_{[i-1]})] - 2E[X_i \otimes E(X_i \mid X_{[i-1]})] \\ &= E(X_i \otimes X_i) - E[E(X_i \mid X_{[i-1]}) \otimes E(X_i \mid X_{[i-1]})], \end{aligned} \tag{14}$$

where the second equality holds because $E[X_i \otimes E(X_i \mid X_{[i-1]})] = E[E(X_i \mid X_{[i-1]}) \otimes E(X_i \mid X_{[i-1]})]$. Since the second term on the right-hand-side of (14) is equal to

$$\begin{aligned} E[E(X_i \mid X_{[i-1]}) \otimes E(X_i \mid X_{[i-1]})] &= E[(M^*_{X_{[i-1]}X_i} X_{[i-1]}) \otimes (M^*_{X_{[i-1]}X_i} X_{[i-1]})] \\ &= M^*_{X_{[i-1]}X_i} \Sigma_{X_{[i-1]}X_{[i-1]}} M_{X_{[i-1]}X_i}, \end{aligned}$$

we have that $E(\epsilon_i \otimes \epsilon_i) = \Sigma_{X_i X_i | X_{[i-1]}}$.

To show that $\Sigma_{\epsilon_i \epsilon_i}$ is a member of $\mathscr{B}_1(\Omega_{X_i})$, note that, for any $f \in \Omega_{X_i}$,

$$\langle f, (\Sigma_{X_i X_i} - \Sigma_{X_i X_i | X_{[i-1]}}) f \rangle_{\Omega_{X_i}} \geq 0 \quad \Rightarrow \quad \langle f, \Sigma_{X_i X_i} f \rangle_{\Omega_{X_i}} \geq \langle f, \Sigma_{X_i X_i | X_{[i-1]}} f \rangle_{\Omega_{X_i}},$$

which further implies that $\| \Sigma_{X_i X_i | X_{[i-1]}} \|_{\mathrm{TR}} \leq \| \Sigma_{X_i X_i} \|_{\mathrm{TR}} < \infty$ by Assumption 1. This completes the proof of Lemma 1. □

Let $\mathcal{H}$ be a generic separable Hilbert space. The next lemma extends the classical Bernstein's inequality (Boucheron et al., 2013, Chapter 2) to the functional setting.

Lemma 2 (Bernstein's inequality in Hilbert space) *Suppose X is a random element in $\mathcal{H}$ with $E(X) = 0$, and $X^1, \ldots, X^n$ are i.i.d. samples of X. If*

$$E(\|X\|_{\mathcal{H}}^{\ell}) \leq b^{\ell}\, \ell!, \quad \text{for some } b > 0, \text{ and each } \ell \in \mathbb{N}, \tag{15}$$

then, for any $t > 0$, $P(\|E_n X\|_{\mathcal{H}} > t) \leq 2 \exp\{-n\, [t/(4b) \wedge t^2/(8b^2)]\}$.

Proof of Lemma 2: By (15), we have $\sum_{i=1}^{n} E(\|X\|_{\mathcal{H}}^{\ell}) \leq n b^{\ell}\, \ell!$. Therefore, by Bosq (2000, Theorem 2.5),

$$P(\|E_n X\|_{\mathcal{H}} > t) \leq 2 \exp\left(-\frac{n t^2}{4b^2 + 2bt} \right) \equiv 2 \exp[-n f(t)].$$

Moreover, note that $f(t) > t/(4b)$ if $t > 2b$, and $f(t) \leq t^2/(8b^2)$ if $t \leq 2b$. This completes the proof of Lemma 2. □

Let $\{(\lambda_i^a, \eta_i^a)\}_{a \in \mathbb{N}}$ denote the eigenvalue-eigenfunction pairs of $\Sigma_{X_i X_i}$, with $\lambda_i^1 \geq \lambda_i^2 \geq \cdots \geq 0$. Let $c_i^a = \langle X_i, \eta_i^a \rangle$. The next lemma shows that, if X follows a sub-Gaussian distribution, then we can bound the moments $E(c_i^a)^{2\ell}$, for each $a \in \mathbb{N}$ and $\ell \in \mathbb{N}$.

Lemma 3 *If $X_i \sim \mathrm{subG}(\Sigma, b)$, then $E(c_i^a)^{2\ell} \leq (4b\, \lambda_i^a)^{\ell}\, \ell!$, for $a \in \mathbb{N}$ and $\ell \in \mathbb{N}$.*

Proof of Lemma 3: By the definition of sub-Gaussianity of X_i, for any $s, t > 0$,

$$P(c_i^a > t) = P\left[\exp(s c_i^a) > \exp(st)\right] \leq \exp(b \lambda_i^a s^2/2 - st), \tag{16}$$

where the inequality follows from the Markov's inequality. Note that $\inf_{s>0} \exp(b\lambda_i^a s^2/2 - st) \leq \exp\left[-t^2/(2b\lambda_i^a)\right]$, which, together with (16), further implies that $P(c_i^a > t) \leq \exp\left[-t^2/(2b\lambda_i^a)\right]$ for each $t > 0$. Using a similar argument, we can show that $P(c_i^a < -t) \leq \exp\left[-t^2/(2b\lambda_i^a)\right]$, for $t > 0$. Therefore, we have

$$P(|c_i^a| > t) \leq 2 \exp\left[-t^2/(2b\lambda_i^a)\right]. \tag{17}$$

Moreover, we note that, for $\ell \in \mathbb{N}$,

$$E(c_i^a)^{2\ell} = \int_0^{\infty} P\left(|c_i^a| > t^{1/(2\ell)}\right) dt \leq 2 \int_0^{\infty} \exp\left[-t^{1/\ell}/(2b\lambda_i^a)\right] dt,$$

$$2 \int_0^{\infty} \exp\left[-t^{1/\ell}/(2b\lambda_i^a)\right] dt = 2(2b\lambda_i^a)^{\ell} \ell \int_0^{\infty} \exp(-x) x^{\ell-1} dx \leq (4b\lambda_i^a)^{\ell}\, \ell!.$$

Combining with (17), we obtain the desired bound for $E(c_i^a)^{2\ell}$. This completes the proof of Lemma 3. $\square$

The next lemma provides some properties regarding the Tychonoff regression. Its proof is similar to Lee et al. (2020, Lemma A4), and is omitted.

Lemma 4 *Let Σ and Γ be self-adjoint operators in $\mathscr{B}(\mathcal{H})$, and let I be the identity mapping. Then, for any $\epsilon > 0$,*

(i) $\|(\Sigma + \epsilon I)^{-1}\| \leq \epsilon^{-1}$*;*

(ii) $\|(\Sigma + \epsilon I)^{-1}\Gamma\| \leq 1 + \epsilon^{-1}\|\Sigma - \Gamma\|$*;*

(iii) $\|(\Sigma + \epsilon I)^{-1}\Gamma^{1/2}\| \leq \epsilon^{-1/2}(1 + \epsilon^{-1}\|\Sigma - \Gamma\|)^{1/2}$.

The next lemma is about the perturbation of the covariance operators.

Lemma 5 *For a given $i \in \mathsf{V}$, let $\{(\lambda_i^a, \eta_i^a)\}_{a=1}^{m+1}$ denote the leading $m+1$ eigenvalue-eigenfunction pairs of $\Sigma_{X_iX_i}$, with $\lambda_i^1 > \lambda_i^2 > \cdots > \lambda_i^{m+1}$. Then, $\max_{a=1,\ldots,m} \|\hat{\eta}_i^a - \eta_i^a\| \leq 4\kappa_m^{-1}\|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\|$, where $\kappa_m = \min\{\lambda_i^a - \lambda_i^{a+1} : a = 1, \ldots, m, i \in \mathsf{V}\}$.*

Proof of Lemma 5: For a given $i \in \mathsf{V}$, let $\tilde{\lambda}_i^a$ be the member of $\{\hat{\lambda}_i^1, \ldots, \hat{\lambda}_i^n\}$ that is closest to λ_i^a. Then by Kato (1980, Theorem 4.10), $\max\{|\tilde{\lambda}_i^a - \lambda_i^a| : a = 1, \ldots, n\} \leq \|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\|$. This implies, for all $a = 1, \ldots, n$, $\lambda_i^a - \|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\| \leq \tilde{\lambda}_i^a \leq \lambda_i^a + \|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\|$. Therefore, for all $a = m+1, \ldots, n$,

$$\tilde{\lambda}_i^a \leq \lambda_i^a + \|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\| \leq \lambda_i^{m+1} + \|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\|. \tag{18}$$

Similarly, for all $a = 1, \ldots, m$,

$$\tilde{\lambda}_i^a \geq \lambda_i^m - \|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\|. \tag{19}$$

If $\kappa_m \leq 2\|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\|$, then the asserted inequality holds automatically.

If $\kappa_m > 2\|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\|$, then $\lambda_i^m - \|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\| > \lambda_i^{m+1} + \|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\|$, which, together with (18) and (19), implies that $\min\{\tilde{\lambda}_i^1, \ldots, \tilde{\lambda}_i^m\} > \max\{\tilde{\lambda}_i^{m+1}, \ldots, \tilde{\lambda}_i^n\}$. Therefore, we have $\{\tilde{\lambda}_i^1, \ldots, \tilde{\lambda}_i^m\} = \{\hat{\lambda}_i^1, \ldots, \hat{\lambda}_i^m\}$. Moreover, for any $a = 1, \ldots, m-1$,

$$\tilde{\lambda}_i^{a+1} \leq \lambda_i^{a+1} + \|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\| < \lambda_i^a - \|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\| \leq \tilde{\lambda}_i^a,$$

which implies that $\tilde{\lambda}_i^a = \hat{\lambda}_i^a$, for all $a = 1, \ldots, m$ and $i \in \mathsf{V}$. Therefore, $\max\{|\hat{\lambda}_i^a - \lambda_i^a| : a = 1, \ldots, m\} \leq \|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\|$, which, by Kazdan (1971, Lemma 2), leads to the asserted inequality. This completes the proof of Lemma 5. $\square$

The next theorem establishes the concentration bound and the uniform convergence rate of $\|\hat{\Sigma}_{X_iX_j} - \Sigma_{X_iX_j}\|_{\mathrm{HS}}$.

Theorem 6 *Suppose Assumption 1 holds, and* $X_i \sim \mathrm{subG}(\Sigma_{X_i X_i}, b_0)$ *with* $E(X_i) = 0$ *for* $i \in \mathsf{V}$*. Then, for any* $t \geq 0$ *and* $(i, j) \in \mathsf{V} \times \mathsf{V}$*,*

$$P(\|\hat{\Sigma}_{X_i X_j} - \Sigma_{X_i X_j}\|_{\mathrm{HS}} > t) \leq 2 \exp\left[-n\left(\frac{t}{4C_0} \wedge \frac{t^2}{8C_0^2}\right)\right],$$

where $C_0 = \max(2M_0, 8M_0 b_0)$*, and* M_0 *is as defined in Assumption 1. Moreover, if* $\log p/n \to 0$*, as* $n \to \infty$*, then,*

$$\max_{i,j \in \mathsf{V}} \|\hat{\Sigma}_{X_i X_j} - \Sigma_{X_i X_j}\|_{\mathrm{HS}} = O_P[(\log p/n)^{1/2}].$$

Proof of Theorem 6: For convenience, for two sets I, J, we use $\sum_{i,j}^{I,J}$ to abbreviate the double sum $\sum_{i \in I}\sum_{j \in J}$. Similarly, for two integers r and s, we use $\sum_{i,j}^{r,s}$ to abbreviate the double sum $\sum_{i=1}^{r}\sum_{j=1}^{s}$.

We first note that, by the triangular and Jensen's inequalities,

$$\begin{aligned} E\|X_i \otimes X_j - E(X_i \otimes X_j)\|_{\mathrm{HS}}^{\ell} &\leq 2^{\ell-1}[E\|X_i \otimes X_j\|_{\mathrm{HS}}^{\ell} + \|E(X_i \otimes X_j)\|_{\mathrm{HS}}^{\ell}] \\ &\equiv 2^{\ell-1}[M_1(\ell) + M_2(\ell)]. \end{aligned}$$

We next bound $M_1(\ell)$ and $M_2(\ell)$, respectively.

For, $M_1(\ell)$, let $\mathbb{N}_i = \{a \in \mathbb{N} : \lambda_i^a \neq 0\}$. For $\ell = 1$,

$$M_1(\ell) \leq E^{1/2}[\textstyle\sum_{a,b}^{\mathbb{N},\mathbb{N}}(c_i^a)^2(c_j^b)^2] \leq \left\{\textstyle\sum_{a,b}^{\mathbb{N},\mathbb{N}} E[(c_i^a)^2(c_j^b)^2]\right\}^{1/2} \leq M_0 = M_0^1 1!.$$

For any $\ell \in \mathbb{N}_i, \ell \geq 2$, we have

$$\begin{aligned} M_1(\ell) &\leq E|\textstyle\sum_{a,b}^{\mathbb{N}_i,\mathbb{N}_i}\langle X_i \otimes X_j, \eta_i^a \otimes \eta_j^b\rangle_{\mathrm{HS}}^2|^{\ell/2} \\ &= E[\textstyle\sum_{a,b}^{\mathbb{N}_i,\mathbb{N}_i}(c_i^a)^2(c_j^b)^2]^{\ell/2} \leq M_0^{\ell}\left\{\textstyle\sum_{a,b}^{\mathbb{N}_i,\mathbb{N}_i}\lambda_i^a\lambda_j^b M_0^{-2}E[(c_i^a)^2(c_j^b)^2/(\lambda_i^a\lambda_j^b)]^{\ell/2}\right\}, \end{aligned}$$

where the last inequality follows from Jensen's inequality as applied to the convex function $f(u) = u^{\ell/2}$. By Lemma 3, we have $E(c_i^a)^{2\ell} \leq (4b_0\lambda_i^a)^{\ell}\,\ell!$. Substituting this into the right-hand-side above, we obtain that $M_1(\ell) \leq (4b_0 M_0)^{\ell}\,\ell!$ for $\ell \geq 2$. Therefore, for any $\ell \in \mathbb{N}$, we have $M_1(\ell) \leq [M_0 \max(1, 4b_0)]^{\ell}\,\ell!$.

For $M_2(\ell)$, by the Cauchy-Schwarz inequality,

$$M_2(\ell) = [\textstyle\sum_{a,b}^{\mathbb{N},\mathbb{N}}E^2(c_i^a c_j^b)]^{\ell/2} \leq \left(\textstyle\sum_{a,b}^{\mathbb{N},\mathbb{N}}\lambda_i^a\lambda_j^b\right)^{\ell/2} \leq M_0^{\ell}. \tag{20}$$

Putting the bounds for $M_1(\ell)$ and $M_2(\ell)$ together, we have $E\|X_i \otimes X_j - E(X_i \otimes X_j)\|_{\mathrm{HS}}^{\ell} \leq [\max(2M_0, 8M_0 b_0)]^{\ell}\,\ell! \equiv C_0^{\ell}\,\ell!$, which, by Lemma 2, implies the first statement of this theorem.

Next, note that, for any $t > 0$,

$$\begin{aligned} P\left(\max_{i,j \in \mathsf{V}}\|\hat{\Sigma}_{X_i X_j} - \Sigma_{X_i X_j}\|_{\mathrm{HS}} > t\right) &\leq \textstyle\sum_{i,j \in \mathsf{V}} P\left(\|\hat{\Sigma}_{X_i X_j} - \Sigma_{X_i X_j}\|_{\mathrm{HS}} > t\right) \\ &\leq 2p^2 \exp\left[-n\left(\frac{t}{4C_0} \wedge \frac{t^2}{8C_0^2}\right)\right], \end{aligned}$$

which implies that the second statement of this theorem holds when $\log p/n \to 0$. This completes the proof of Theorem 6. □

Let $\Sigma_{X_i X_j}^d = \sum_{a=1}^{d}\sum_{b=1}^{d}E(c_i^a c_j^b)(\eta_i^a \otimes \eta_j^b)$ be the truncated version of $\Sigma_{X_i X_j}$, $(i, j) \in \mathsf{V} \times \mathsf{V}$. The next theorem establishes the uniform convergence of $\|\hat{\Sigma}_{X_i X_j}^d - \Sigma_{X_i X_j}^d\|_{\mathrm{HS}}$.

Theorem 7 *Suppose Assumption 1 holds, $X_i \sim \mathrm{subG}(\Sigma_{X_iX_i}, b_0)$ with $E(X_i) = 0$ for $i \in \mathsf{V}$, and $d^{\gamma+1}(\log p)^{1/2}/n^{1/2} \preceq 1$. Then, we have*

$$\max_{i,j\in\mathsf{V}} \|\hat{\Sigma}^d_{X_iX_j} - \Sigma^d_{X_iX_j}\|_{\mathrm{HS}} = O_P\left[d^{3+\gamma}(\log p)^{1/2}/n^{1/2}\right].$$

Proof of Theorem 7: Note that $\max_{i,j\in\mathsf{V}} \|\hat{\Sigma}^d_{X_iX_j} - \Sigma^d_{X_iX_j}\|_{\mathrm{HS}}$ is upper-bounded by

$$\begin{aligned}&\max_{i,j\in\mathsf{V}} \|\textstyle\sum_{a,b}^{d,d}[E_n(\hat{c}_i^a\hat{c}_j^b) - E(c_i^ac_j^b)]\,\hat{\eta}_i^a \otimes \hat{\eta}_j^b\|_{\mathrm{HS}}\\ &+ \max_{i,j\in\mathsf{V}} \|\textstyle\sum_{a,b}^{d,d}E(c_i^ac_j^b)\,[(\hat{\eta}_i^a - \eta_i^a)\otimes\hat{\eta}_j^b + \eta_i^a\otimes(\hat{\eta}_j^b - \eta_j^b)]\|_{\mathrm{HS}} \equiv M_3(n) + M_4(n).\end{aligned}$$

Next, we derive the orders of magnitude for $M_3(n)$ and $M_4(n)$.

For $M_3(n)$, we have

$$\begin{aligned}M_3(n) \leq &\max_{i,j\in\mathsf{V}} \|\textstyle\sum_{a,b}^{d,d}E_n[(c_i^ac_j^b) - E(c_i^ac_j^b)]\,\hat{\eta}_i^a \otimes \hat{\eta}_j^b\|_{\mathrm{HS}}\\ &+ \max_{i,j\in\mathsf{V}} \|\textstyle\sum_{a,b}^{d,d}E_n[(\hat{c}_i^a - c_i^a)(\hat{c}_j^b - c_j^b)]\,\hat{\eta}_i^a \otimes \hat{\eta}_j^b\|_{\mathrm{HS}}\\ &+ \max_{i,j\in\mathsf{V}} \|\textstyle\sum_{a,b}^{d,d}E_n[c_i^a(\hat{c}_j^b - c_j^b)]\,\hat{\eta}_i^a \otimes \hat{\eta}_j^b\|_{\mathrm{HS}}\\ &+ \max_{i,j\in\mathsf{V}} \|\textstyle\sum_{a,b}^{d,d}E_n[(\hat{c}_i^a - c_i^a)c_j^b]\,\hat{\eta}_i^a \otimes \hat{\eta}_j^b\|_{\mathrm{HS}} \equiv M_{3,1}(n) + \cdots + M_{3,4}(n).\end{aligned}$$

For $M_{3,1}(n)$, it can be bounded as,

$$\begin{aligned}M_{3,1}(n) =&\max_{i,j\in\mathsf{V}} \|\textstyle\sum_{a,b}^{d,d}E_n\langle X_i\otimes X_j, -E(X_i\otimes X_j), \eta_i^a\otimes\eta_j^b\rangle_{\mathrm{HS}}\,\hat{\eta}_i^a \otimes \hat{\eta}_j^b\|_{\mathrm{HS}},\\ \leq&\max_{i,j\in\mathsf{V}} \textstyle\sum_{a,b}^{d,d}\|E_n\langle X_i\otimes X_j - E(X_i\otimes X_j), \eta_i^a\otimes\eta_j^b\rangle_{\mathrm{HS}}\,\hat{\eta}_i^a \otimes \hat{\eta}_j^b\|_{\mathrm{HS}}\\ \leq&\max_{i,j\in\mathsf{V}} \|E_n[X_i\otimes X_j - E(X_i\otimes X_j)]\|_{\mathrm{HS}}\,\left(\textstyle\sum_{a,b}^{d,d}\|\hat{\eta}_i^a\otimes\hat{\eta}_j^b\|_{\mathrm{HS}}\right)\\ =&d^2\max_{i,j\in\mathsf{V}} \|E_n[X_i\otimes X_j - E(X_i\otimes X_j)]\|_{\mathrm{HS}}.\end{aligned}$$

Therefore, by Theorem 6, $M_{3,1}(n) = O_P[d^2(\log p/n)^{1/2}]$.

For $M_{3,2}(n)$, it can be bounded as,

$$\begin{aligned}M_{3,2}(n) \leq&\max_{i,j\in\mathsf{V}} \|\textstyle\sum_{a,b}^{d,d}E_n\langle X_i\otimes X_j - E(X_i\otimes X_j), (\hat{\eta}_i^a - \eta_i^a)\otimes(\hat{\eta}_j^b - \eta_j^b)\rangle_{\mathrm{HS}}\,\hat{\eta}_i^a \otimes \hat{\eta}_j^b\|_{\mathrm{HS}}\\ &+ \max_{i,j\in\mathsf{V}} \|\textstyle\sum_{a,b}^{d,d}\langle E(X_i\otimes X_j), (\hat{\eta}_i^a - \eta_i^a)\otimes(\hat{\eta}_j^b - \eta_j^b)\rangle_{\mathrm{HS}}\,\hat{\eta}_i^a \otimes \hat{\eta}_j^b\|_{\mathrm{HS}}\\ \leq&16d^2\kappa_d^{-2}\max_{i,j\in\mathsf{V}}[\|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\|\,\|\hat{\Sigma}_{X_jX_j} - \Sigma_{X_jX_j}\|\\ &\times(\|E_n[X_i\otimes X_j - E(X_i\otimes X_j)]\|_{\mathrm{HS}} + \|E(X_i\otimes X_j)\|_{\mathrm{HS}})],\end{aligned}$$

where last inequality holds because, by Lemma 5, $\sum_{a,b}^{d,d}\|(\hat{\eta}_i^a - \eta_i^a)\otimes(\hat{\eta}_j^b - \eta_j^b)\|_{\mathrm{HS}} \leq 16d^2\kappa_d^{-2}\|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\|\,\|\hat{\Sigma}_{X_jX_j} - \Sigma_{X_jX_j}\|$. By (20), we have $\|E(X_i\otimes X_j)\|_{\mathrm{HS}} \leq M_0$. Therefore, $M_{3,2}(n) = O_P[d^2(\log p)/(n\kappa_d^2)]$.

For $M_{3,3}(n)$, by Lemma 5 again, it can be bounded by

$$\begin{aligned}M_{3,3}(n) \leq 4d^2\kappa_d^{-1}\max_{i,j\in\mathsf{V}}\{&\|\hat{\Sigma}_{X_jX_j} - \Sigma_{X_jX_j}\|\\ &\times[\|E_n[X_i\otimes X_j - E(X_i\otimes X_j)]\|_{\mathrm{HS}} + \|E(X_i\otimes X_j)\|_{\mathrm{HS}}]\}.\end{aligned}$$

By Theorem 6 again, $M_{3,3}(n) = O_P[d^2(\log p)^{1/2}/(n^{1/2}\kappa_d)]$.

Similarly, we can show that $M_{3,4}(n)$ has the same order of magnitude as $M_{3,3}(n)$.

For $M_4(n)$, we have

$$\begin{aligned} M_4(n) &\leq \max_{i,j \in \mathsf{V}} \left[\|E(X_i \otimes X_j)\|_{\text{HS}} \textstyle\sum_{a,b}^{d,d} (\|\hat{\eta}_i^a - \eta_i^a\| + \|\hat{\eta}_j^b - \eta_j^b\|) \right] \\ &\leq 4d^2\kappa_d^{-1}M_0 \times \max_{i,j \in \mathsf{V}} (\|\hat{\Sigma}_{X_iX_i} - \Sigma_{X_iX_i}\| + \|\hat{\Sigma}_{X_jX_j} - \Sigma_{X_jX_j}\|), \end{aligned}$$

which leads to $M_4(n) = O_P[d^2(\log p)^{1/2}/(n^{1/2}\kappa_d)]$.

Combining the orders of magnitude of $M_{3,1}(n), \ldots, M_{3,4}(n)$, and $M_4(n)$, we obtain,

$$\begin{aligned} &\max_{i,j \in \mathsf{V}} \|\hat{\Sigma}^d_{X_iX_j} - \Sigma^d_{X_iX_j}\|_{\text{HS}} \\ &\quad = O_P[d^2(\log p/n)^2] + O_P[d^2 \log p/(n\kappa_d)] + O_p[d^2(\log p)^{1/2}/(n^{1/2}\kappa_d)]. \end{aligned}$$

Because $\kappa_d \to 0$ and $d^2(\log p)^{1/2}/(n^{1/2}\kappa_d) \preceq 1$, the third term on the right-hand-side is the dominating term, which is equal to the desired rate because $\kappa_d^{-1} \preceq d^{1+\gamma}$ by Assumption 4. This completes the proof of Theorem 7. □

Theorem 7 generalizes the convergence result for the high-dimensional covariance matrix (Bickel and Levina, 2008) to the high-dimensional covariance operator. We remark that, under the Gaussian assumption, Qiao et al. (2019) also established the concentration bound for the sample covariance of the leading K-L expansion coefficients. However, Theorem 7 differs from the result of Qiao et al. (2019), in that it provides the uniform convergence at the operator level, which can not be derived directly from their result. Moreover, Theorems 6 and 7 do not require the random process to be Gaussian, but only require the distribution to be sub-Gaussian.

A.2 Proof of Theorem 3

To prove this theorem, we first introduce an intermediate operator $\Sigma^{d,\epsilon}_{X_iX_j|X_\mathsf{S}}$. We next derive the order of magnitude for the differences between $\Sigma^{d,\epsilon}_{X_iX_j|X_\mathsf{S}}$ and $\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_\mathsf{S}}$ in Lemma 6, and between $\Sigma^{d,\epsilon}_{X_iX_j|X_\mathsf{S}}$ and $\Sigma_{X_iX_j|X_\mathsf{S}}$ in Lemma 7, respectively. These two Lemmas together lead to the first assertion, i.e. the uniform convergence of $\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_\mathsf{S}} - \Sigma_{X_iX_j|X_\mathsf{S}}$. Lastly, we derive the uniform convergence of the estimated graph.

For any $\mathsf{A}, \mathsf{B} \subseteq \mathsf{V}$, let $\Sigma^d_{X_\mathsf{A}X_\mathsf{B}}$ be the matrix of operators $\{\Sigma^d_{X_iX_j}\}_{i\in\mathsf{A}, j\in\mathsf{B}}$, and let

$$\Sigma^{d,\epsilon}_{X_iX_j|X_\mathsf{S}} = \Sigma^d_{X_iX_j} - \Sigma^d_{X_iX_\mathsf{S}}[\Sigma^d_{X_\mathsf{S}X_\mathsf{S}}(\epsilon)]^{\ddagger}\Sigma^d_{X_\mathsf{S}X_j},$$

where $\epsilon > 0$ is a tuning constant that decreases to 0 as $n \to \infty$, and $[A(\epsilon)]^{\ddagger}$ represents $(A + \epsilon I)^{-1}A(A + \epsilon I)^{-1}$. This term $\Sigma^{d,\epsilon}_{X_iX_j|X_\mathsf{S}}$ plays the role of an intermediate operator between $\Sigma_{X_iX_j|X_\mathsf{S}}$ and $\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_\mathsf{S}}$.

Lemma 6 *Suppose Assumptions 1 and 3 hold, $X_i \sim \text{subG}(\Sigma_{X_iX_i}, b_0)$ with $E(X_i) = 0$, for $i \in \mathsf{V}$, $m \succ 1$, $\epsilon \prec 1$, and $md^{3+\gamma}(\log p)^{1/2}/(n^{1/2}\epsilon) \preceq 1$. Then,*

$$\max_{\mathsf{H}(m)} \|\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_\mathsf{S}} - \Sigma^{d,\epsilon}_{X_iX_j|X_\mathsf{S}}\|_{\text{HS}} = O_P\{md^{3+\gamma}(\log p)^{1/2}/(n^{1/2}\epsilon)\}.$$

Proof of Lemma 6: Because $\hat{M}_{X_{\mathsf{S}}X_i} = (\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}} + \epsilon I)^{-1}\hat{\Sigma}^d_{X_{\mathsf{S}}X_i}$, we have $\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_{\mathsf{S}}} = \hat{\Sigma}^d_{X_iX_j} - \hat{\Sigma}^d_{X_iX_{\mathsf{S}}}[\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\ddagger}\hat{\Sigma}^d_{X_{\mathsf{S}}X_j}$. Therefore,

$$\begin{aligned}
&\max_{(i,j,\mathsf{S})\in\mathsf{H}(m)} \|\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_{\mathsf{S}}} - \Sigma^{d,\epsilon}_{X_iX_j|X_{\mathsf{S}}}\|_{\mathrm{HS}} \\
&\le \max_{i,j\in\mathsf{V}} \|\hat{\Sigma}^d_{X_iX_j} - \Sigma^d_{X_iX_j}\|_{\mathrm{HS}} + \max_{(i,j,\mathsf{S})\in\mathsf{H}(m)} \|(\hat{\Sigma}^d_{X_iX_{\mathsf{S}}} - \Sigma^d_{X_iX_{\mathsf{S}}})[\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\ddagger}\hat{\Sigma}^d_{X_{\mathsf{S}}X_j}\|_{\mathrm{HS}} \\
&+ \max_{(i,j,\mathsf{S})\in\mathsf{H}(m)} \|\Sigma^d_{X_iX_{\mathsf{S}}}\{[\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\ddagger} - [\Sigma^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\ddagger}\}\hat{\Sigma}^d_{X_{\mathsf{S}}X_j}\|_{\mathrm{HS}} \\
&+ \max_{(i,j,\mathsf{S})\in\mathsf{H}(m)} \|\Sigma^d_{X_iX_{\mathsf{S}}}[\Sigma^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\ddagger}(\hat{\Sigma}^d_{X_{\mathsf{S}}X_j} - \Sigma^d_{X_{\mathsf{S}}X_j})\|_{\mathrm{HS}} \equiv M_5(n) + M_6(n) + M_7(n) + M_8(n).
\end{aligned}$$

The order of magnitude of $M_5(n)$ is given in Theorem 7. We next derive the orders of magnitude of $M_6(n), M_7(n)$ and $M_8(n)$, respectively.

For $M_6(n)$, we have

$$\begin{aligned}
M_6(n) \le &\max_{(i,j,\mathsf{S})\in\mathsf{H}(m)} \|(\hat{\Sigma}^d_{X_iX_{\mathsf{S}}} - \Sigma^d_{X_iX_{\mathsf{S}}})(\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}} + \epsilon I)^{-1}\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}} + \epsilon I)^{-1}\Sigma^d_{X_{\mathsf{S}}X_j}\|_{\mathrm{HS}} \\
&+ \max_{(i,j,\mathsf{S})\in\mathsf{H}(m)} \|(\hat{\Sigma}^d_{X_iX_{\mathsf{S}}} - \Sigma^d_{X_iX_{\mathsf{S}}})(\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}} + \epsilon I)^{-1}\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}} \\
&\qquad \times (\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}} + \epsilon I)^{-1}(\hat{\Sigma}^d_{X_{\mathsf{S}}X_j} - \Sigma^d_{X_{\mathsf{S}}X_j})\|_{\mathrm{HS}}.
\end{aligned} \tag{21}$$

The first term on the right, by Lemma 4 and $\|\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}} + \epsilon I)^{-1}\| \le 1$, is upper-bounded by

$$\begin{aligned}
\epsilon^{-1} &\max_{(i,j,\mathsf{S})\in\mathsf{H}(m)} [\|\hat{\Sigma}^d_{X_iX_{\mathsf{S}}} - \Sigma^d_{X_iX_{\mathsf{S}}}\|_{\mathrm{HS}}\, \|\Sigma^d_{X_{\mathsf{S}}X_j}\|_{\mathrm{HS}}] \\
&\le \epsilon^{-1} \max_{(i,i,\mathsf{S})\in\mathsf{H}_1(m)} \|\hat{\Sigma}^d_{X_iX_{\mathsf{S}}} - \Sigma^d_{X_iX_{\mathsf{S}}}\|_{\mathrm{HS}} \times \max_{(j,j,\mathsf{S})\in\mathsf{H}_1(m)} \|\Sigma^d_{X_{\mathsf{S}}X_j}\|_{\mathrm{HS}}.
\end{aligned} \tag{22}$$

By a special case of (20), $\|\Sigma_{X_jX_i}\|_{\mathrm{HS}} \le M_0$. Henceforth,

$$\max_{(j,j,\mathsf{S})\in\mathsf{H}_1(m)} \|\Sigma^d_{X_{\mathsf{S}}X_j}\|_{\mathrm{HS}} = \max_{(j,j,\mathsf{S})\in\mathsf{H}_1(m)} \sqrt{\textstyle\sum_{i\in\mathsf{S}}\|\Sigma^d_{X_iX_j}\|^2_{\mathrm{HS}}} \le \max_{(j,j,\mathsf{S})\in\mathsf{H}_1(m)} \sqrt{\textstyle\sum_{i\in\mathsf{S}}\|\Sigma_{X_iX_j}\|^2_{\mathrm{HS}}} \le m^{1/2}M_0.$$

So the right-hand-side of (22) is further bounded by $\epsilon^{-1}m^{1/2}M_0(\max_{(i,i,\mathsf{S})\in\mathsf{H}_1(m)} \|\hat{\Sigma}^d_{X_iX_{\mathsf{S}}} - \Sigma^d_{X_iX_{\mathsf{S}}}\|_{\mathrm{HS}})$. Similarly, we can bound the second term on the right of (21) by

$$\epsilon^{-1} \max_{(i,j,\mathsf{S})\in\mathsf{H}(m)} (\|\hat{\Sigma}^d_{X_iX_{\mathsf{S}}} - \Sigma^d_{X_iX_{\mathsf{S}}}\|_{\mathrm{HS}}\, \|\hat{\Sigma}^d_{X_{\mathsf{S}}X_j} - \Sigma^d_{X_{\mathsf{S}}X_j}\|_{\mathrm{HS}}).$$

Therefore, we have,

$$\begin{aligned}
M_6(n) \le \epsilon^{-1}[&m^{1/2}\, M_0(\max_{(i,i,\mathsf{S})\in\mathsf{H}_1(m)} \|\hat{\Sigma}^d_{X_iX_{\mathsf{S}}} - \Sigma^d_{X_iX_{\mathsf{S}}}\|_{\mathrm{HS}}) \\
&+ \max_{(i,j,\mathsf{S})\in\mathsf{H}(m)} (\|\hat{\Sigma}^d_{X_iX_{\mathsf{S}}} - \Sigma^d_{X_iX_{\mathsf{S}}}\|_{\mathrm{HS}}\, \|\hat{\Sigma}^d_{X_{\mathsf{S}}X_j} - \Sigma^d_{X_{\mathsf{S}}X_j}\|_{\mathrm{HS}})].
\end{aligned}$$

For $M_7(n)$, let $[A(\epsilon)]^{\dagger}$ denote $(A + \epsilon I)^{-1}$, so that $[A(\epsilon)]^{\ddagger} = [A(\epsilon)]^{\dagger}A[A(\epsilon)]^{\dagger}$. In these notations, the difference $[\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\ddagger} - [\Sigma^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\ddagger}$ can be decomposed as

$$\begin{aligned}
&\{[\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\dagger} - [\Sigma^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\dagger}\}\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}}[\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\dagger} + [\Sigma^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\dagger}(\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}} - \Sigma^d_{X_{\mathsf{S}}X_{\mathsf{S}}})[\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\dagger} \\
&\qquad + [\Sigma^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\dagger}\Sigma^d_{X_{\mathsf{S}}X_{\mathsf{S}}}\{[\hat{\Sigma}^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\dagger} - [\Sigma^d_{X_{\mathsf{S}}X_{\mathsf{S}}}(\epsilon)]^{\dagger}\} \equiv \Gamma_1(n) + \Gamma_2(n) + \Gamma_3(n).
\end{aligned}$$

Moreover, $\|\Sigma^d_{X_i X_\mathsf{S}} \Gamma_1(n) \hat{\Sigma}^d_{X_\mathsf{S} X_j}\|_{\text{HS}}$ is bounded by

$$\|\Sigma^d_{X_i X_\mathsf{S}} [\Sigma^d_{X_\mathsf{S} X_\mathsf{S}}(\epsilon)]^\dagger\| \; \|\hat{\Sigma}^d_{X_\mathsf{S} X_\mathsf{S}} - \Sigma^d_{X_\mathsf{S} X_\mathsf{S}}\|_{\text{HS}} \; \|[\hat{\Sigma}^d_{X_\mathsf{S} X_\mathsf{S}}(\epsilon)]^\dagger \hat{\Sigma}^d_{X_\mathsf{S} X_\mathsf{S}} [\hat{\Sigma}^d_{X_\mathsf{S} X_\mathsf{S}}(\epsilon)]^\dagger \hat{\Sigma}^d_{X_\mathsf{S} X_j}\|. \tag{23}$$

The first norm in (23), by Baker (1973, Theorem 1) and Lemma 4, is upper-bounded by $\epsilon^{-1/2} \|\Sigma^d_{X_i X_i}\|^{1/2} \leq [\text{tr}(\Sigma_{X_i X_i})/\epsilon]^{1/2} = (M_0/\epsilon)^{1/2}$. The third norm in (23) is upper-bounded by $\epsilon^{-1/2}[\|(\hat{\Sigma}^d_{X_\mathsf{S} X_\mathsf{S}} + \epsilon I)^{-1/2} \Sigma^d_{X_\mathsf{S} X_j}\| + \|(\hat{\Sigma}^d_{X_\mathsf{S} X_\mathsf{S}} + \epsilon I)^{-1/2} (\hat{\Sigma}^d_{X_\mathsf{S} X_j} - \Sigma^d_{X_\mathsf{S} X_j})\|]$. By a similar argument as used to bound the first norm, we can be further bound the above by $\epsilon^{-1/2}[(M_0 + \epsilon^{-1} M_0 \|\hat{\Sigma}^d_{X_\mathsf{S} X_\mathsf{S}} - \Sigma^d_{X_\mathsf{S} X_\mathsf{S}}\|)^{1/2} + \epsilon^{-1/2} \|\hat{\Sigma}^d_{X_\mathsf{S} X_j} - \Sigma^d_{X_\mathsf{S} X_j}\|]$. Therefore,

$$\begin{aligned} \max_{(i,j,\mathsf{S}) \in \mathsf{H}(m)} \|\Sigma^d_{X_i X_\mathsf{S}} \Gamma_1(n) \hat{\Sigma}^d_{X_\mathsf{S} X_j}\|_{\text{HS}} &\leq M_0^{1/2} \epsilon^{-1} (\max_{(i,j,\mathsf{S}) \in \mathsf{H}(m)} \|\hat{\Sigma}^d_{X_\mathsf{S} X_\mathsf{S}} - \Sigma^d_{X_\mathsf{S} X_\mathsf{S}}\|_{\text{HS}}) \\ \times \max_{(i,j,\mathsf{S}) \in \mathsf{H}(m)} [(M_0 + \epsilon^{-1} M_0 \|\hat{\Sigma}^d_{X_\mathsf{S} X_\mathsf{S}} &- \Sigma^d_{X_\mathsf{S} X_\mathsf{S}}\|)^{1/2} + \epsilon^{-1/2} \|\hat{\Sigma}^d_{X_\mathsf{S} X_j} - \Sigma^d_{X_\mathsf{S} X_j}\|]. \end{aligned} \tag{24}$$

Moreover, $\max_{(i,j,\mathsf{S}) \in \mathsf{H}(m)} \|\Sigma^d_{X_i X_\mathsf{S}} \Gamma_2(n) \hat{\Sigma}^d_{X_\mathsf{S} X_j}\|_{\text{HS}}$ and $\max_{(i,j,\mathsf{S}) \in \mathsf{H}(m)} \|\Sigma^d_{X_i X_\mathsf{S}} \Gamma_3(n) \hat{\Sigma}^d_{X_\mathsf{S} X_j}\|_{\text{HS}}$ can be bounded by the right-hand-side of (24). Therefore, $M_7(n)$ is bounded by three times of the quantity on the right-hand-side of (24).

For $M_8(n)$, similar to the derivation of the bound for $M_6(n)$, we can show that

$$M_8(n) \leq \epsilon^{-1} m^{1/2} \, M_0 (\max_{(i,j,\mathsf{S}) \in \mathsf{H}(m)} \|\hat{\Sigma}^d_{X_\mathsf{S} X_j} - \Sigma^d_{X_\mathsf{S} X_j}\|_{\text{HS}}).$$

On the other hand, note that

$$\max_{(i,j,\mathsf{S}) \in \mathsf{H}(m)} \|\hat{\Sigma}^d_{X_\mathsf{S} X_\mathsf{S}} - \Sigma^d_{X_\mathsf{S} X_\mathsf{S}}\|_{\text{HS}} = \max_{(i,j,\mathsf{S}) \in \mathsf{H}(m)} \sqrt{\textstyle\sum_{i,j \in \mathsf{S}} \|\hat{\Sigma}^d_{X_i X_j} - \Sigma^d_{X_i X_j}\|^2_{\text{HS}}},$$

which is no greater than $m \times (\max_{i,j \in \mathsf{V}} \|\hat{\Sigma}^d_{X_i X_j} - \Sigma^d_{X_i X_j}\|_{\text{HS}})$. By the same derivation, we have $\max_{(i,i,\mathsf{S}) \in \mathsf{H}_1(m)} \|\hat{\Sigma}^d_{X_i X_\mathsf{S}} - \Sigma^d_{X_i X_\mathsf{S}}\|_{\text{HS}}$ is bounded by $m^{1/2} \times (\max_{i,j \in \mathsf{V}} \|\hat{\Sigma}^d_{X_i X_j} - \Sigma^d_{X_i X_j}\|_{\text{HS}})$.

Combining the bounds for $M_6(n)$, $M_7(n)$, and $M_8(n)$, and applying Theorem 7 as well as the condition that $md^{3+\gamma} (\log p)^{1/2} / (n^{1/2} \epsilon) \preceq 1$, we have

$$\begin{aligned} M_6(n) + &M_7(n) + M_8(n) \\ &\preceq \epsilon^{-1} m (\max_{i,j \in \mathsf{V}} \|\hat{\Sigma}^d_{X_i X_j} - \Sigma^d_{X_i X_j}\|_{\text{HS}}) + m (\max_{i,j \in \mathsf{V}} \|\hat{\Sigma}^d_{X_i X_j} - \Sigma^d_{X_i X_j}\|_{\text{HS}})^2 \\ &\quad + \epsilon^{-1/2} m^{1/2} (\max_{i,j \in \mathsf{V}} \|\hat{\Sigma}^d_{X_i X_j} - \Sigma^d_{X_i X_j}\|_{\text{HS}}) \\ &= O_P[md^{3+\gamma} (\log p)^{1/2} / (n^{1/2} \epsilon)] + O_p[md^{6+2\gamma} (\log p)/n] \\ &\quad + O_P[m^{1/2} d^{3+\gamma} (\log p)^{1/2} / (n^{1/2} \epsilon^{1/2})]. \end{aligned}$$

Since $d^{3+\gamma} (\log p)^{1/2} / n^{1/2} \prec 1$, the first term on the right is the dominating term. This completes the proof of Lemma 6. □

Lemma 7 *Suppose Assumptions 1, 3, and 4 hold. Then, for any* $(i,j) \in \mathsf{V} \times \mathsf{V}$ *and* $\mathsf{S} \in \mathsf{V} \backslash (i,j)$, *we have,*

$$\max_{(i,j,\mathsf{S}) \in \mathsf{H}(m)} \|\Sigma^{d,\epsilon}_{X_i X_j | X_\mathsf{S}} - \Sigma_{X_i X_j | X_\mathsf{S}}\|_{\text{HS}} = O\{m \epsilon^{-1} d^{-\gamma} + \epsilon^{1/2} s(m)\}.$$

Proof of Lemma 7: Let $\Sigma^{\epsilon}_{X_iX_j|X_\mathsf{S}} = \Sigma_{X_iX_j} - \Sigma_{X_iX_\mathsf{S}}[\Sigma_{X_\mathsf{S}X_\mathsf{S}}(\epsilon)]^{\ddagger}\Sigma_{X_\mathsf{S}X_j}$ be the intermediate operator between $\Sigma^{d,\epsilon}_{X_iX_j|X_\mathsf{S}}$ and $\Sigma_{X_iX_j|X_\mathsf{S}}$. By the triangular inequality,

$$\begin{aligned}\max_{(i,j,\mathsf{S})\in\mathsf{H}(m)}&\|\Sigma^{d,\epsilon}_{X_iX_j|X_\mathsf{S}} - \Sigma_{X_iX_j|X_\mathsf{S}}\|_{\text{HS}}\\ &\le \max_{(i,j,\mathsf{S})\in\mathsf{H}(m)}\|\Sigma^{d,\epsilon}_{X_iX_j|X_\mathsf{S}} - \Sigma^{\epsilon}_{X_iX_j|X_\mathsf{S}}\|_{\text{HS}} + \max_{(i,j,\mathsf{S})\in\mathsf{H}(m)}\|\Sigma^{\epsilon}_{X_iX_j|X_\mathsf{S}} - \Sigma_{X_iX_j|X_\mathsf{S}}\|_{\text{HS}}\\ &\equiv M_9(n) + M_{10}(n).\end{aligned}$$

We next derive the orders of magnitude for $M_9(n)$ and $M_{10}(n)$.

For $M_9(n)$, we have

$$\begin{aligned}M_9(n) \le &\max_{i,j\in\mathsf{V}}\|\Sigma^{d}_{X_iX_j} - \Sigma_{X_iX_j}\|_{\text{HS}} + \max_{(i,j,\mathsf{S})\in\mathsf{H}(m)}\|(\Sigma^{d}_{X_iX_\mathsf{S}} - \Sigma_{X_iX_\mathsf{S}})[\Sigma^{d}_{X_\mathsf{S}X_\mathsf{S}}(\epsilon)]^{\ddagger}\Sigma^{d}_{X_\mathsf{S}X_j}\|_{\text{HS}}\\ &+ \max_{(i,j,\mathsf{S})\in\mathsf{H}(m)}\|\Sigma_{X_iX_\mathsf{S}}\{[\Sigma^{d}_{X_\mathsf{S}X_\mathsf{S}}(\epsilon)]^{\ddagger} - [\Sigma_{X_\mathsf{S}X_\mathsf{S}}(\epsilon)]^{\ddagger}\}\Sigma^{d}_{X_\mathsf{S}X_j}\|_{\text{HS}}\\ &+ \max_{(i,j,\mathsf{S})\in\mathsf{H}(m)}\|\Sigma_{X_iX_\mathsf{S}}[\Sigma_{X_\mathsf{S}X_\mathsf{S}}(\epsilon)]^{\ddagger}(\Sigma^{d}_{X_\mathsf{S}X_j} - \Sigma_{X_\mathsf{S}X_j})\|_{\text{HS}},\end{aligned}$$

whose order of magnitude, by a similar argument as in the proof of Lemma 6, is no greater than that of

$$\begin{aligned}\max_{i,j\in\mathsf{V}}\|\Sigma^{d}_{X_iX_j} - \Sigma_{X_iX_j}\| + \epsilon^{-1/2}\max_{(i,j,\mathsf{S})\in\mathsf{H}(m)}(\|\Sigma^{d}_{X_iX_\mathsf{S}} - \Sigma_{X_iX_\mathsf{S}}\|_{\text{HS}} + \|\Sigma^{d}_{X_\mathsf{S}X_j} - \Sigma_{X_\mathsf{S}X_j}\|_{\text{HS}})\\ +\epsilon^{-1}\max_{(i,j,\mathsf{S})\in\mathsf{H}(m)}\|\Sigma^{d}_{X_\mathsf{S}X_\mathsf{S}} - \Sigma_{X_\mathsf{S}X_\mathsf{S}}\|.\end{aligned}$$

Let $\mathbb{N}_d = \{d+1, d+2, \ldots\}$. Then the term $\max_{(i,j,\mathsf{S})\in\mathsf{H}(m)}\|\Sigma^{d}_{X_\mathsf{S}X_\mathsf{S}} - \Sigma_{X_\mathsf{S}X_\mathsf{S}}\|_{\text{HS}}$ equals

$$\max_{(i,j,\mathsf{S})\in\mathsf{H}(m)}\sqrt{\textstyle\sum_{i,j\in\mathsf{S}}\sum_{a,b}^{\mathbb{N}_d,\mathbb{N}_d}E^2(c_i^a c_j^b)} \le \max_{(i,j,\mathsf{S})\in\mathsf{H}(m)}\sqrt{\textstyle\sum_{i,j\in\mathsf{S}}\sum_{a,b}^{\mathbb{N}_d,\mathbb{N}_d}\lambda_i^a\lambda_j^b},$$

whose order is $O(md^{-\gamma})$ by Assumption 4. Similarly, we can show that

$$\max_{(i,i,\mathsf{S})\in\mathsf{H}_1(m)}\|\Sigma^{d}_{X_iX_\mathsf{S}} - \Sigma_{X_iX_\mathsf{S}}\|_{\text{HS}} = O(m^{1/2}d^{-\gamma}).$$

Therefore, $M_9(n) \preceq m\epsilon^{-1}d^{-\gamma}$.

For $M_{10}(n)$, we note that, by the definitions of $\Sigma^{\epsilon}_{X_iX_j|X_\mathsf{S}}$, $\Sigma_{X_iX_j|X_\mathsf{S}}$, and $M_{X_\mathsf{S}X_j}$,

$$\Sigma^{\epsilon}_{X_iX_j|X_\mathsf{S}} - \Sigma_{X_iX_j|X_\mathsf{S}} = \Sigma_{X_iX_\mathsf{S}}\{[\Sigma_{X_\mathsf{S}X_\mathsf{S}}(\epsilon)]^{\ddagger}\Sigma_{X_\mathsf{S}X_\mathsf{S}} - I\}M_{X_\mathsf{S}X_j}.$$

By the definition of $[A(\epsilon)]^{\ddagger}$, we can rewrite the right-hand-side of the above equation as $\Sigma_{X_iX_\mathsf{S}}\left(\epsilon^2\{[\Sigma_{X_\mathsf{S}X_\mathsf{S}}(\epsilon)]^{\dagger}\}^2 - 2\epsilon[\Sigma_{X_\mathsf{S}X_\mathsf{S}}(\epsilon)]^{\dagger}\right)M_{X_\mathsf{S}X_j}$. Therefore,

$$M_{10}(n) \le s(m)\max_{(i,i,\mathsf{S})\in\mathsf{H}_1(m)}\left[\|\Sigma_{X_iX_\mathsf{S}}\left(\epsilon^2\{[\Sigma_{X_\mathsf{S}X_\mathsf{S}}(\epsilon)]^{\dagger}\}^2 - 2\epsilon[\Sigma_{X_\mathsf{S}X_\mathsf{S}}(\epsilon)]^{\dagger}\right)\|\right],$$

whose order, by Lemma 4, is no greater than $\epsilon^{1/2}s(m)$. Combining the orders of $M_9(n)$ and $M_{10}(n)$ completes the proof of Lemma 7. □

Proof of Theorem 3: Combining Lemmas 6 and 7, we immediately obtain the uniform convergence rate of $\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_\mathsf{S}}$.

For the second assertion, we first note that,

$$\begin{aligned} & P\big(\{\hat{\mathsf{E}}_{\text{CPDAG-fCCO}} \neq \mathsf{E}^0_{\text{CPDAG}}\} \cup \{\hat{\ell}_{\text{fCCO}} \neq \ell^0\}\big) \\ & \leq P[\|\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_{\mathsf{S}}}\|_{\text{HS}} > \rho_{\text{fCCO}},\ \Sigma_{X_iX_j|X_{\mathsf{S}}} = 0,\ \text{for some } (i,j,\mathsf{S}) \in \mathsf{H}_0(m)] \\ & \quad + P[\|\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_{\mathsf{S}}}\|_{\text{HS}} \leq \rho_{\text{fCCO}},\ \Sigma_{X_iX_j|X_{\mathsf{S}}} \neq 0,\ \text{for some } (i,j,\mathsf{S}) \in \mathsf{H}_0(m)]. \end{aligned} \tag{25}$$

The first term in (25) is further bounded by

$$P[\max\{\|\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_{\mathsf{S}}} - \Sigma_{X_iX_j|X_{\mathsf{S}}}\|_{\text{HS}} : (i,j,\mathsf{S}) \in \mathsf{H}_0(m)\} \geq t(m)/2] \equiv p^*(m).$$

Moreover, by the definition of $t(m)$, the second term in (25) is no greater than

$$P[\|\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_{\mathsf{S}}} - \Sigma_{X_iX_j|X_{\mathsf{S}}}\|_{\text{HS}} \geq t(m)/2,\ \Sigma_{X_iX_j|X_{\mathsf{S}}} \neq 0,\ \text{for some } (i,j,\mathsf{S}) \in \mathsf{H}_0(m)].$$

It is further bounded by $p^*(m)$, which tends to 0 by the first assertion and Assumption 5. We thus obtain the second assertion. This completes the proof of Theorem 3. □

A.3 Proof of Theorem 4

Similar as the proof of Theorem 3, to prove this theorem, we first introduce another intermediate operator $R^{\delta}_{X_iX_j|X_{\mathsf{S}}}$ between $\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_{\mathsf{S}}}$ and $R_{X_iX_j|X_{\mathsf{S}}}$,

$$R^{\delta}_{X_iX_j|X_{\mathsf{S}}} = (\Sigma_{X_iX_i|X_{\mathsf{S}}} + \delta I)^{-1/2}\Sigma_{X_iX_j|X_{\mathsf{S}}}(\Sigma_{X_iX_i|X_{\mathsf{S}}} + \delta I)^{-1/2}.$$

Then by the triangular inequality,

$$\|\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_{\mathsf{S}}} - R_{X_iX_j|X_{\mathsf{S}}}\|_{\text{HS}} \leq \|\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_{\mathsf{S}}} - R^{\delta}_{X_iX_j|X_{\mathsf{S}}}\|_{\text{HS}} + \|R^{\delta}_{X_iX_j|X_{\mathsf{S}}} - R_{X_iX_j|X_{\mathsf{S}}}\|_{\text{HS}}.$$

We next derive the order of magnitude for the differences between $\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_{\mathsf{S}}}$ and $R^{\delta}_{X_iX_j|X_{\mathsf{S}}}$ in Lemma 8, and between $R^{\delta}_{X_iX_j|X_{\mathsf{S}}}$ and $R_{X_iX_j|X_{\mathsf{S}}}$ in Lemma 9, respectively.

Lemma 8 *Suppose the conditions in Theorem 3(i) hold, $\zeta(m,d,p,\epsilon,n) \preceq 1$, and $\delta \prec 1$. Then,*

$$\max_{(i,j,\mathsf{S})\in\mathsf{H}_0(m)} \|\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_{\mathsf{S}}} - R^{\delta}_{X_iX_j|X_{\mathsf{S}}}\|_{\text{HS}} = O_P[\delta^{-3/2}\zeta(m,d,p,\epsilon,n)].$$

Moreover, if $\delta^{-3/2}\zeta(m,d,p,\epsilon,n) \prec 1$, then $\max_{(i,j,\mathsf{S})\in\mathsf{H}_0(m)} \|\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_{\mathsf{S}}} - R^{\delta}_{X_iX_j|X_{\mathsf{S}}}\|_{\text{HS}} \xrightarrow{P} 0$.

Proof of Lemma 8: Let $\Gamma_4(n) = \hat{\Sigma}^{d,\epsilon}_{X_iX_i|X_{\mathsf{S}}}$, $\Gamma_5(n) = \Sigma_{X_iX_i|X_{\mathsf{S}}}$, $\Gamma_6(n) = \hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_{\mathsf{S}}}$, $\Gamma_7(n) = \Sigma_{X_iX_j|X_{\mathsf{S}}}$, $\Gamma_8(n) = \hat{\Sigma}^{d,\epsilon}_{X_jX_j|X_{\mathsf{S}}}$, and $\Gamma_9(n) = \Sigma_{X_jX_j|X_{\mathsf{S}}}$. By the triangular inequality,

$$\begin{aligned} & \|\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_{\mathsf{S}}} - R^{d,\epsilon,\delta}_{X_iX_j|X_{\mathsf{S}}}\|_{\text{HS}} \\ & \qquad \leq \|\{\Gamma_4(n) + \delta I]^{-1/2} - [\Gamma_5(n) + \delta I]^{-1/2}\}\Gamma_6(n)[\Gamma_8(n) + \delta I]^{-1/2}\|_{\text{HS}} \\ & \qquad\quad + \|[\Gamma_5(n) + \delta I]^{-1/2}[\Gamma_6(n) - \Gamma_7(n)][\Gamma_8(n) + \delta I]^{-1/2}\|_{\text{HS}} \\ & \qquad\quad + \|[\Gamma_5(n) + \delta I]^{-1/2}\Gamma_7(n)\{[\Gamma_8(n) + \delta I]^{-1/2} - [\Gamma_9(n) + \delta I]^{-1/2}\}\|_{\text{HS}} \\ & \qquad \equiv M^{i,j,\mathsf{S}}_{11}(n) + M^{i,j,\mathsf{S}}_{12}(n) + M^{i,j,\mathsf{S}}_{13}(n). \end{aligned}$$

We next bound $M_{11}^{i,j,\mathsf{S}}(n)$, $M_{12}^{i,j,\mathsf{S}}(n)$, and $M_{13}^{i,j,\mathsf{S}}(n)$, respectively.

For $M_{11}^{i,j,\mathsf{S}}(n)$, using the identity,

$$A^{1/2} - B^{1/2} = A^{1/2}(B^{-3/2} - A^{-3/2})B^{3/2} + (A^{-1} - B^{-1})B^{3/2}$$

(Fukumizu et al., 2007), we have $M_{11}^{i,j,\mathsf{S}}(n) \leq [M_{11,1}^{i,j,\mathsf{S}}(n) + M_{11,2}^{i,j,\mathsf{S}}(n)]M_{11,3}^{i,j,\mathsf{S}}(n)$, where

$$\begin{aligned}
&M_{11,1}^{i,j,\mathsf{S}}(n) = \|[\Gamma_5(n) + \delta I]^{-1/2}\{[\Gamma_5(n) + \delta I]^{3/2} - [\Gamma_4(n) + \delta I]^{3/2}\}[\Gamma_4(n) + \delta I]^{-1}\|_{\mathrm{HS}},\\
&M_{11,2}^{i,j,\mathsf{S}}(n) = \|[\Gamma_4(n) - \Gamma_5][\Gamma_4(n) + \delta I]^{-1}\|_{\mathrm{HS}},\\
&M_{11,3}^{i,j,\mathsf{S}}(n) = \|[\Gamma_4(n) + \delta I]^{-1/2}\Gamma_6(n)[\Gamma_8(n) + \delta I]^{-1/2}\|.
\end{aligned}$$

By the inequality $\max(a^{1/2}, b^{1/2}) \leq (a+b)^{1/2}$, Lemma 4, and Fukumizu et al. (2008, Lemma 7), we have

$$\begin{aligned}
M_{11,1}^{i,j,\mathsf{S}}(n) &\leq 3\delta^{-3/2} \max(\|\Gamma_4(n) + \delta I\|^{1/2}, \|\Gamma_5(n) + \delta I\|^{1/2})\ \|\Gamma_4(n) - \Gamma_5(n)\|_{\mathrm{HS}}\\
&\leq 3\delta^{-3/2}[\|\Gamma_4(n) - \Gamma_5(n)\| + \|\Gamma_5(n) + \delta I\|]^{1/2}\ \|\Gamma_4(n) - \Gamma_5(n)\|_{\mathrm{HS}}.
\end{aligned}$$

Because $\Sigma_{X_iX_i|X_{\mathsf{S}}} \leq \Sigma_{X_iX_i}$, we have $\|\Gamma_5(n) + \delta I\| = \|\Sigma_{X_iX_i|X_{\mathsf{S}}} + \delta I\| \leq \|\Sigma_{X_iX_i} + \delta I\| \leq M_0 + \delta$. Therefore,

$$M_{11,1}^{i,j,\mathsf{S}}(n) \leq 3\delta^{-3/2}(\|\Gamma_4(n) - \Gamma_5(n)\| + M_0 + \delta)^{1/2}\ \|\Gamma_4(n) - \Gamma_5(n)\|_{\mathrm{HS}}.$$

For $M_{11,2}^{i,j,\mathsf{S}}(n)$, we have,

$$M_{11,2}^{i,j,\mathsf{S}}(n) \leq \|[\Gamma_4(n) + \delta I]^{-1}\|\ \|\Gamma_4(n) - \Gamma_5(n)\|_{\mathrm{HS}} \leq \delta^{-1}\|\Gamma_4(n) - \Gamma_5(n)\|_{\mathrm{HS}}.$$

For $M_{11,3}^{i,j,\mathsf{S}}(n)$, by Proposition 5, we have $M_{11,3}^{i,j,\mathsf{S}}(n) \leq 1$.

Combining the upper bounds for $M_{11,1}^{i,j,\mathsf{S}}(n)$, $M_{11,2}^{i,j,\mathsf{S}}(n)$, and $M_{11,3}^{i,j,\mathsf{S}}(n)$, and taking maximum over $\mathsf{H}_0(m)$, we obtain that,

$$\begin{aligned}
\max_{(i,j,\mathsf{S})\in\mathsf{H}_0(m)} M_{11}^{i,j,\mathsf{S}}(n) \leq &(\max_{(i,i,\mathsf{S})\in\mathsf{H}_1(m)} \|\hat{\Sigma}^{d,\epsilon}_{X_iX_i|X_{\mathsf{S}}} - \Sigma_{X_iX_i|X_{\mathsf{S}}}\|_{\mathrm{HS}})\\
&\times [3\delta^{-3/2}(\max_{(i,i,\mathsf{S})\in\mathsf{H}_1(m)} \|\hat{\Sigma}^{d,\epsilon}_{X_iX_i|X_{\mathsf{S}}} - \Sigma_{X_iX_i|X_{\mathsf{S}}}\|_{\mathrm{HS}} + M_0 + \delta)^{1/2} + \delta^{-1}].
\end{aligned}$$

By Theorem 3, the right-hand-side is dominated by $\zeta(m, d, p, \epsilon, n)\{3\delta^{-3/2}[\zeta(m, d, p, \epsilon, n) + M_0 + \delta]^{1/2} + \delta^{-1}\}$. Therefore, if $\zeta(m, d, p, \epsilon, n) \preceq 1$, then

$$\max_{(i,j,\mathsf{S})\in\mathsf{H}_0(m)} M_{11}^{i,j,\mathsf{S}}(n) = O_P[\delta^{-3/2}\zeta(m, d, p, \epsilon, n)]. \tag{26}$$

It remains to show that the orders of magnitude of the terms $\max_{(i,j,\mathsf{S})\in\mathsf{H}_0(m)} M_{12}^{i,j,\mathsf{S}}(n)$ and $\max_{(i,j,\mathsf{S})\in\mathsf{H}_0(m)} M_{13}^{i,j,\mathsf{S}}(n)$ are dominated by (26). Using similar derivations for (26),

$$\begin{aligned}
&\max_{(i,j,\mathsf{S})\in\mathsf{H}_0(m)} M_{12}^{i,j,\mathsf{S}}(n) \leq \delta^{-1}(\max_{(i,j,\mathsf{S})\in\mathsf{H}_0(m)} \|\hat{\Sigma}^{d,\epsilon}_{X_iX_j|X_{\mathsf{S}}} - \Sigma_{X_iX_j|X_{\mathsf{S}}}\|_{\mathrm{HS}}) = O_P[\delta^{-1}\zeta(m, d, p, \epsilon, n)],\\
&\max_{(i,j,\mathsf{S})\in\mathsf{H}_0(m)} M_{13}^{i,j,\mathsf{S}}(n) = O_P[\delta^{-3/2}\zeta(m, d, p, \epsilon, n)],
\end{aligned}$$

both of which are dominated by (26). This completes the proof of Lemma 8. □

Lemma 9 *Suppose Assumptions 1, 3 and 6 hold, and $\delta \prec 1$. Then,*

$$\max_{(i,j,\mathsf{S})\in\mathsf{H}_0(m)} \|R^{\delta}_{X_iX_j|X_\mathsf{S}} - R_{X_iX_j|X_\mathsf{S}}\|_{\mathrm{HS}} = O_p(\delta^{1/2}).$$

Proof of Lemma 9: By definition, $\|R^{\delta}_{X_iX_j|X_\mathsf{S}} - R_{X_iX_j|X_\mathsf{S}}\|^2_{\mathrm{HS}}$ is equal to

$$\begin{aligned}&\textstyle\sum_{a,b}^{\mathbb{M}_{i,\mathsf{S}},\mathbb{M}_{j,\mathsf{S}}} \langle \nu^a_{i,\mathsf{S}}, [(\Sigma_{X_iX_i|X_\mathsf{S}}+\delta I)^{-1/2}\Sigma_{X_iX_j|X_\mathsf{S}}(\Sigma_{X_iX_i|X_\mathsf{S}}+\delta I)^{-1/2} - R_{X_iX_j|X_\mathsf{S}}]\nu^b_{j,\mathsf{S}}\rangle^2_{\Omega_{X_i}} \\ &= \textstyle\sum_{a,b}^{\mathbb{M}_{i,\mathsf{S}},\mathbb{M}_{j,\mathsf{S}}} [(\mu^a_{i,\mathsf{S}}+\delta)^{1/2}(\mu^b_{j,\mathsf{S}}+\delta)^{1/2} - (\mu^a_{i,\mathsf{S}})^{1/2}(\mu^b_{j,\mathsf{S}})^{1/2}]^2 \dfrac{\langle \nu^a_{i,\mathsf{S}}, R_{X_iX_j|X_\mathsf{S}}\nu^b_{j,\mathsf{S}}\rangle^2_{\Omega_{X_i}}}{(\mu^a_{i,\mathsf{S}}+\delta)(\mu^b_{j,\mathsf{S}}+\delta)},\end{aligned}$$

where $\mathbb{M}_{i,\mathsf{S}}$, $\mathbb{M}_{j,\mathsf{S}}$, $\mu^a_{i,\mathsf{S}}$, $\mu^b_{j,\mathsf{S}}$, $\nu^a_{i,\mathsf{S}}$, $\nu^b_{j,\mathsf{S}}$ are as defined right before Assumption 6. Moreover,

$$\langle \nu^a_{i,\mathsf{S}}, R_{X_iX_j|X_\mathsf{S}}\nu^b_{j,\mathsf{S}}\rangle^2_{\Omega_{X_i}} = \left\langle \frac{\nu^a_{i,\mathsf{S}}}{(\mu^a_{i,\mathsf{S}})^{1/2}}, \Sigma_{X_iX_j|X_\mathsf{S}} \frac{\nu^b_{j,\mathsf{S}}}{(\mu^b_{j,\mathsf{S}})^{1/2}} \right\rangle^2_{\Omega_{X_i}} = (\rho^{a,b}_{i,j,\mathsf{S}})^2,$$

which implies that $\|R^{\delta}_{X_iX_j|X_\mathsf{S}} - R_{X_iX_j|X_\mathsf{S}}\|^2_{\mathrm{HS}}$ is upper-bounded by

$$\textstyle\sum_{a,b}^{\mathbb{M}_{i,\mathsf{S}},\mathbb{M}_{j,\mathsf{S}}} [(\mu^a_{i,\mathsf{S}}+\delta)^{1/2}(\mu^b_{j,\mathsf{S}}+\delta)^{1/2} - (\mu^a_{i,\mathsf{S}})^{1/2}(\mu^b_{j,\mathsf{S}})^{1/2}]^2(\rho^{a,b}_{i,j,\mathsf{S}})^2/(\mu^a_{i,\mathsf{S}}\mu^b_{j,\mathsf{S}}).$$

By direct calculation, for any $a \in \mathbb{M}_{i,\mathsf{S}}, b \in \mathbb{M}_{j,\mathsf{S}}$,

$$[(\mu^a_{i,\mathsf{S}}+\delta)^{1/2}(\mu^b_{j,\mathsf{S}}+\delta)^{1/2} - (\mu^a_{i,\mathsf{S}})^{1/2}(\mu^b_{j,\mathsf{S}})^{1/2}]^2 \le \delta(\mu^a_{i,\mathsf{S}} + \mu^b_{j,\mathsf{S}}) + \delta^2,$$

which is no greater than $2\delta M_0 + \delta^2$ because $\mu^a_{i,\mathsf{S}} \le \|\Sigma_{X_iX_i|X_\mathsf{S}}\| \le \|\Sigma_{X_iX_i}\| \le M_0$. Therefore, by Assumption 6,

$$\max_{(i,j,\mathsf{S})\in\mathsf{H}_0(m)} \|R^{\delta}_{X_iX_j|X_\mathsf{S}} - R_{X_iX_j|X_\mathsf{S}}\|_{\mathrm{HS}} \le c_0^{1/2}(2\delta M_0 + \delta^2)^{1/2} = O_p(\delta^{1/2}).$$

This completes the proof of Lemma 9. □

Proof of Theorem 4: Combining Lemmas 8 and 9, we immediately obtain the uniform convergence rate of $\|\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_\mathsf{S}} - R_{X_iX_j|X_\mathsf{S}}\|_{\mathrm{HS}}$. The second assertion can be obtained following a similar proof of Theorem 3. This completes the proof of Theorem 4. □

A.4 Proofs of other theoretical results

Proof of Proposition 1: By Assumption 2, for $f \in \Omega_{X_\mathsf{B}}$, $g \in \Omega_{X_\mathsf{A}}$, and $(t_1, t_2) \in \mathbb{R}^2$,

$$\begin{aligned}&E\exp\left\{\iota\left[t_1(\langle f, X_\mathsf{B}\rangle_{\Omega_{X_\mathsf{B}}} - \langle M_{X_\mathsf{A}X_\mathsf{B}}f, X_\mathsf{A}\rangle_{\Omega_{X_\mathsf{A}}}) + t_2\langle g, X_\mathsf{A}\rangle_{\Omega_{X_\mathsf{A}}}\right]\right\} \\ &= \exp\left(-\frac{1}{2}\left\langle \begin{pmatrix} t_1 f \\ t_2 g - t_1 M_{X_\mathsf{A}X_\mathsf{B}}f\end{pmatrix}, \begin{pmatrix}\Sigma_{UU} & \Sigma_{UV} \\ \Sigma_{VU} & \Sigma_{VV}\end{pmatrix}\begin{pmatrix} t_1 f \\ t_2 g - t_1 M_{X_\mathsf{A}X_\mathsf{B}}f\end{pmatrix}\right\rangle_{\Omega_{X_\mathsf{B}}\oplus\Omega_{X_\mathsf{A}}}\right),\end{aligned}$$

where $\iota = \sqrt{-1}$. By direct calculation, the inner product in the above expression equals

$$\begin{aligned}&t_1^2\langle f, \Sigma_{UU}f\rangle_{\Omega_{X_\mathsf{B}}} - 2t_1^2\langle f, \Sigma_{UV}M_{X_\mathsf{A}X_\mathsf{B}}f\rangle_{\Omega_{X_\mathsf{B}}} + 2t_1t_2\langle f, \Sigma_{UV}g\rangle_{\Omega_{X_\mathsf{B}}} \\ &- 2t_1t_2\langle g, \Sigma_{VV}M_{X_\mathsf{A}X_\mathsf{B}}f\rangle_{\Omega_{X_\mathsf{A}}} + t_1^2\langle M_{X_\mathsf{A}X_\mathsf{B}}f, \Sigma_{VV}M_{X_\mathsf{A}X_\mathsf{B}}f\rangle_{\Omega_{X_\mathsf{A}}} + t_2^2\langle g, \Sigma_{VV}g\rangle_{\Omega_{X_\mathsf{A}}} \\ &= t_1^2\langle f, (\Sigma_{UU} - \Sigma_{UV}\Sigma^{\dagger}_{VV}\Sigma_{VU})f\rangle_{\Omega_{X_\mathsf{B}}} + t_2^2\langle g, \Sigma_{VV}g\rangle_{\Omega_{X_\mathsf{A}}} \equiv t_1^2\sigma_f^2 + t_2^2\sigma_g^2,\end{aligned}$$

where we have used the relations $\langle M_{X_\mathsf{A}X_\mathsf{B}} f, \Sigma_{VV} M_{X_\mathsf{A}X_\mathsf{B}} f\rangle_{\Omega_{X_\mathsf{A}}} = \langle f, \Sigma_{UV} M_{X_\mathsf{A}X_\mathsf{B}} f\rangle_{\Omega_{X_\mathsf{B}}} = \langle f, \Sigma_{UV}\Sigma_{VV}^{\dagger}\Sigma_{VU} f\rangle_{\Omega_{X_\mathsf{B}}}$, and $\langle f, \Sigma_{UV} M_{X_\mathsf{A}X_\mathsf{B}} f\rangle_{\Omega_{X_\mathsf{B}}} = \langle f, \Sigma_{UV} g\rangle_{\Omega_{X_\mathsf{B}}}$. This implies, $\langle f, X_\mathsf{B}\rangle_{\Omega_{X_\mathsf{B}}} - \langle M_{X_\mathsf{A}X_\mathsf{B}} f, X_\mathsf{A}\rangle_{\Omega_{X_\mathsf{A}}}$ and $\langle g, X_\mathsf{A}\rangle_{\Omega_{X_\mathsf{A}}}$ are independent Gaussian variables with variances σ_f^2 and σ_g^2, respectively. Since this holds for all $f \in \Omega_{X_\mathsf{B}}$ and $g \in \Omega_{X_\mathsf{A}}$, we have $\langle f, X_\mathsf{B}\rangle_{\Omega_{X_\mathsf{B}}} - \langle M_{X_\mathsf{A}X_\mathsf{B}} f, X_\mathsf{A}\rangle_{\Omega_{X_\mathsf{A}}} \perp\!\!\!\perp X_\mathsf{A}$, which completes the proof of Proposition 1. □

Proof of Theorem 1: Following the proof of Proposition 1, we can show that, for any $(f, g) \in \Omega_{X_i} \times \Omega_{X_j}$, the conditional distribution of $(\langle f, X_i\rangle, \langle g, X_j\rangle)$ given X_S is

$$N\left(\begin{pmatrix}\langle M_{X_\mathsf{S}X_i} f, X_\mathsf{S}\rangle \\ \langle M_{X_\mathsf{S}X_j} g, X_\mathsf{S}\rangle\end{pmatrix}, \begin{pmatrix}\langle f, \Sigma_{X_iX_i|X_\mathsf{S}} f\rangle & \langle f, \Sigma_{X_iX_j|X_\mathsf{S}} g\rangle \\ \langle g, \Sigma_{X_jX_i|X_\mathsf{S}} f\rangle & \langle g, \Sigma_{X_jX_j|X_\mathsf{S}} g\rangle\end{pmatrix}\right),$$

which implies (i).

Because, by Assumption 3, $\mathrm{ran}(\Sigma_{X_iX_j|X_\mathsf{S}}) \subseteq \mathrm{ran}(\Sigma_{X_iX_i})$, and $\mathrm{ran}(\Sigma_{X_jX_i|X_\mathsf{S}}) \subseteq \mathrm{ran}(\Sigma_{X_jX_j})$, it suffices to show that, for any $f \in \mathrm{ran}(\Sigma_{X_iX_i})$ and $g \in \mathrm{ran}(\Sigma_{X_jX_j})$,

$$\langle f, \Sigma_{X_iX_j|X_\mathsf{S}} g\rangle_{\Omega_{X_i}} = \langle f, \left[\textstyle\sum_{a,b\in\mathbb{N}} \mathrm{cov}(c_i^a, c_j^b \mid X_\mathsf{S})(\eta_i^a \otimes \eta_j^b)\right] g\rangle_{\Omega_{X_i}}. \tag{27}$$

Since $f \in \mathrm{ran}(\Sigma_{X_iX_i})$ and $g \in \mathrm{ran}(\Sigma_{X_jX_j})$, we have $f = \sum_{a\in\mathbb{N}} \langle f, \eta_i^a\rangle_{\Omega_{X_i}} \eta_i^a$, and $g = \sum_{b\in\mathbb{N}} \langle g, \eta_j^b\rangle_{\Omega_{X_j}} \eta_j^b$. Substituting these into the left-hand-side of (27), we can obtain the right-hand-side of (27). Thus, (ii) holds. This completes the proof of Theorem 1. □

Proof of Proposition 2: Note that, for any $a, b \in \mathrm{span}(\mathcal{B}_r)$,

$$[a \otimes b] = ([(a \otimes b)b_1], \ldots, [(a \otimes b)b_r]) = ([a]\,\langle b, b_1\rangle, \ldots, [a]\,\langle b, b_r\rangle),$$

which equals $\big([a]\;[b]^\mathsf{T} e_1, \ldots, [a]\;[b]^\mathsf{T} e_r\big)$, where e_i is the r-dimensional vector with its ith element equal to 1 and other elements equal to 0. Noting that $\hat{\Sigma}_{X_iX_j} = E_n[(X_i - E_nX_i) \otimes (X_j - E_nX_j)]$ completes the proof of Proposition 2. □

Proof of Proposition 4: We have, by the faithfulness, $X_i \perp\!\!\!\perp X_j \mid X_\mathsf{S}$ if and only if i and j are d-separated by S. Following the proof in Kalisch and Bühlmann (2007, Proposition 1), we can show that the above equivalence implies that the output of Step 1 of the functional-PC^0 is the true skeleton E_{SKE}, which further implies $\ell^0 \le m$ by the definition of m. Therefore (ii) holds. Moreover, by Meek (1995), the output from Step 2 of functional-PC^0 is the CPDAG of G, which implies (i). This completes the proof of Proposition 4. □

Proof of Proposition 5: By the definition of $\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_\mathsf{S}}$,

$$\begin{aligned}
\mathrm{ran}(\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_\mathsf{S}}) &\subseteq \mathrm{ran}(\hat{\Sigma}^{d,\epsilon}_{X_iX_i|X_\mathsf{S}}) = \mathrm{span}\{\hat{\eta}_i^a : a = 1, \ldots, d\} = \mathrm{span}\{\mathcal{B}_i^*\}, \\
\ker(\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_\mathsf{S}}) &\supseteq \ker(\hat{\Sigma}^{d,\epsilon}_{X_jX_j|X_\mathsf{S}}) = \mathrm{span}\{\hat{\eta}_j^a : a = 1, \ldots, d\}^\perp = \mathrm{span}\{\mathcal{B}_j^*\}^\perp.
\end{aligned}$$

This implies that the operator norm of $\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_\mathsf{S}}$ is the same as the largest singular value of the coordinate representation of $\hat{R}^{d,\epsilon,\delta}_{X_iX_j|X_\mathsf{S}}$ with respect to $\mathcal{B}_j^*$ and $\mathcal{B}_i^*$. Therefore, by

Proposition 3, $\|\hat{R}^{d,\epsilon,\delta}_{X_i X_j | X_{\mathsf{S}}}\|$ can be computed via the optimization:

$$\begin{aligned} \text{maximize}: \quad & [f_i]^\intercal A^\intercal_{i,\mathsf{S}} A_{j,\mathsf{S}} [f_j], \\ \text{subject to}: \quad & [f_i]^\intercal (A^\intercal_{i,\mathsf{S}} A_{i,\mathsf{S}} + \delta I_n)[f_i] = [f_j]^\intercal (A^\intercal_{j,\mathsf{S}} A_{j,\mathsf{S}} + \delta I_n)[f_j] = 1, \\ & [f_i] \in \mathbb{R}^d, [f_j] \in \mathbb{R}^d, \end{aligned}$$

where $A_{i,\mathsf{S}} = [I_n - D(\mathsf{S})]^{1/2} C_i^{1:d}$. By the Cauchy-Schwarz inequality, we have

$$[f_i]^\intercal A^\intercal_{i,\mathsf{S}} A_{j,\mathsf{S}} [f_j] \le ([f_i]^\intercal A^\intercal_{i,\mathsf{S}} A_{i,\mathsf{S}} [f_i])^{1/2} ([f_j]^\intercal A^\intercal_{j,\mathsf{S}} A_{j,\mathsf{S}} [f_j])^{1/2},$$

which is no greater than $\left([f_i]^\intercal (A^\intercal_{i,\mathsf{S}} A_{i,\mathsf{S}} + \delta I_n)[f_i]\right)^{1/2} \left([f_j]^\intercal (A^\intercal_{j,\mathsf{S}} A_{j,\mathsf{S}} + \delta I_n)[f_j]\right)^{1/2} = 1$. This completes the proof of Proposition 5. □

Proof of Theorem 5: First, we show (a) ⇒ (b). We pick a permutation π, such that, for any $i = 2, \ldots, p$, $\mathrm{pa}(i) \subseteq \pi([i-1])$. For convenience, we reset $\pi(1), \ldots, \pi(p)$ to $1, \ldots, p$. By Lauritzen et al. (1990, Corollary 2), the global Markov property is equivalent to

$$X_i \perp\!\!\!\perp X_{[i-1]} \mid X_{\mathrm{pa}(i)},$$

which means that the conditional distributions of $X_i \mid X_{[i-1]}$ and $X_i \mid X_{\mathrm{pa}(i)}$ are identical. Moreover, following a similar proof as that of Proposition 1, we can show that,

$$\begin{aligned} X_i \mid X_{[i-1]} &\sim N(M^*_{X_{[i-1]} X_i} X_{[i-1]}, \Sigma_{X_i X_i | X_{[i-1]}}), \\ X_i \mid X_{\mathrm{pa}(i)} &\sim N(M^*_{X_{\mathrm{pa}(i)} X_i} X_{\mathrm{pa}(i)}, \Sigma_{X_i X_i | X_{\mathrm{pa}(i)}}). \end{aligned}$$

Since the above two Gaussian distributions are the same, we have $(M_{X_{[i-1]} X_i})_j = 0$, for $j \notin \mathrm{pa}(i)$. Therefore, by Lemma 1, X satisfies the functional linear structural equation model with respect to G.

Next, we show (b) ⇒ (a). This holds because the global Markov property is implied by the local Markov property, which, under the Gaussian distribution, is implied by (b).

This completes the proof of Theorem 5. □

A.5 Additional numerical results

Additional combinations of (p, q, n): We carry out an additional simulation study for Model I with more combinations of (p, q, n). In our simulation, the true DAG is generated by a random graph, whose level of sparsity is controlled by the expected neighborhoods size q. Meanwhile, as q increases, we expect the maximal degree m to increase as well. Table 2 reports the average values of m for different combinations of (p, q, n). This new study thus allows us to further examine the empirical performance of the proposed PCO method with denser graphs. Moreover, it offers new insight into the condition that Theorem 3 requires about m that,

$$m d^{3+\gamma} (\log p)^{1/2} / (n^{1/2} \epsilon) \preceq 1. \tag{28}$$

This condition implies that m can grow at most at a polynomial rate, and thus in turn imposes a level of sparsity on the graph. We generate (p, q, n) following the relation,

$$q(n) = c_1 + c_2 \times n, \quad \log_{c_3}[p(n)] = c_4 \times n^{c_5},$$

n	p	q	ave. m	TDR	TPR	FM
20	4	1.18	2.150	0.956	0.139	0.195
40	8	1.26	3.237	0.742	0.399	0.505
60	17	1.34	4.400	0.730	0.498	0.586
80	32	1.42	5.150	0.687	0.530	0.594
100	57	1.50	5.937	0.689	0.568	0.619
120	100	1.58	6.600	0.684	0.574	0.622

Table 2: The true discovery rate (TDR), true positive rate (TPR) and F-measure (FM) between the estimated and true CPDAG, as p grows in an exponential order of n, and q grows in a polynomial order of n.

with $c_1 = 0.77, c_2 = 0.003$, $c_3 = 6$, $c_4 = 0.09$, and $c_5 = 0.7$. As such, p grows in an exponential order of n, and q grows in a polynomial order of n. The resulting combinations are $n = \{20, 40, 60, 80, 100, 120\}, p = \{4, 8, 17, 32, 57, 100\}$, and $q = \{0.82, 0.88, 0.93, 0.99, 1.05, 1.10\}$.

For the evaluation criteria, in addition to the structure Hamming distance (SHD) and the true discovery rate (TDR), we employ two additional criteria, the true positive rate (TPR) and the F-measure (FM), which are defined as,

$$\begin{aligned}\text{TPR}(\hat{\mathsf{E}}_{\text{CPDAG}}, \mathsf{E}_{\text{CPDAG}}) &= |\{(i,j) \in \hat{\mathsf{E}}_{\text{CPDAG}} : (i,j) \in \mathsf{E}_{\text{CPDAG}}\}|/|\mathsf{E}_{\text{CPDAG}}|,\\ \text{FM}(\hat{\mathsf{E}}_{\text{CPDAG}}, \mathsf{E}_{\text{CPDAG}}) &= 2 \times \frac{\text{TDR}(\hat{\mathsf{E}}_{\text{CPDAG}}, \mathsf{E}_{\text{CPDAG}}) \times \text{TPR}(\hat{\mathsf{E}}_{\text{CPDAG}}, \mathsf{E}_{\text{CPDAG}})}{\text{TDR}(\hat{\mathsf{E}}_{\text{CPDAG}}, \mathsf{E}_{\text{CPDAG}}) + \text{TPR}(\hat{\mathsf{E}}_{\text{CPDAG}}, \mathsf{E}_{\text{CPDAG}})},\end{aligned}$$

where $\text{FM}(\hat{\mathsf{E}}_{\text{CPDAG}}, \mathsf{E}_{\text{CPDAG}})$ is computed as the harmonic mean of $\text{TDR}(\hat{\mathsf{E}}_{\text{CPDAG}}, \mathsf{E}_{\text{CPDAG}})$ and $\text{TPR}(\hat{\mathsf{E}}_{\text{CPDAG}}, \mathsf{E}_{\text{CPDAG}})$. A motivation of adding the new criteria is that an estimator may sometimes have a high true discovery rate (TDR) but a low TPR (Rijsbergen, 1979). A higher TPR and a higher FM indicate a more accurate estimator.

Figures 5 to 8 report the box plots of the SHD, TDR, TPR and FM, respectively, between the true CPDAG and the PCO estimate across 80 data replications. We see that, as sample size increase, the SHD decreases, and both TPR and FM increase with a stabilizing TDR. This shows that our method performs well for these new combinations, and, in particular, for the larger m's, as long as the sample size is reasonably large.

Table 2 reports the average TDR, TPR, and FM for six combinations of (p, q, n), along with the average m. We see that, with the increasing sample size, while TDR slightly deceases, TPR and FM both increase substantially. This results indicates that the overall performance of our PCO method improves as n increases. It also provides additional support for our theoretical consistency and why we need condition (28).

Undirected screening: We also investigate the performance of our method when coupled with an undirected graph estimation as a starting point. In the classical setting of random variables, the undirected graph is a supergraph of the skeleton E_{SKE} of the DAG G, and it is easy to show this statement continues to hold in the setting of random functions.

Specifically, we employ Qiao et al. (2019) to estimate an undirected graph, then feed it as an initial graph into our proposed PCO-based PC-algorithm. We have chosen Qiao et al. (2019) over Li and Solea (2018), since both our method and Qiao et al. (2019) consider the

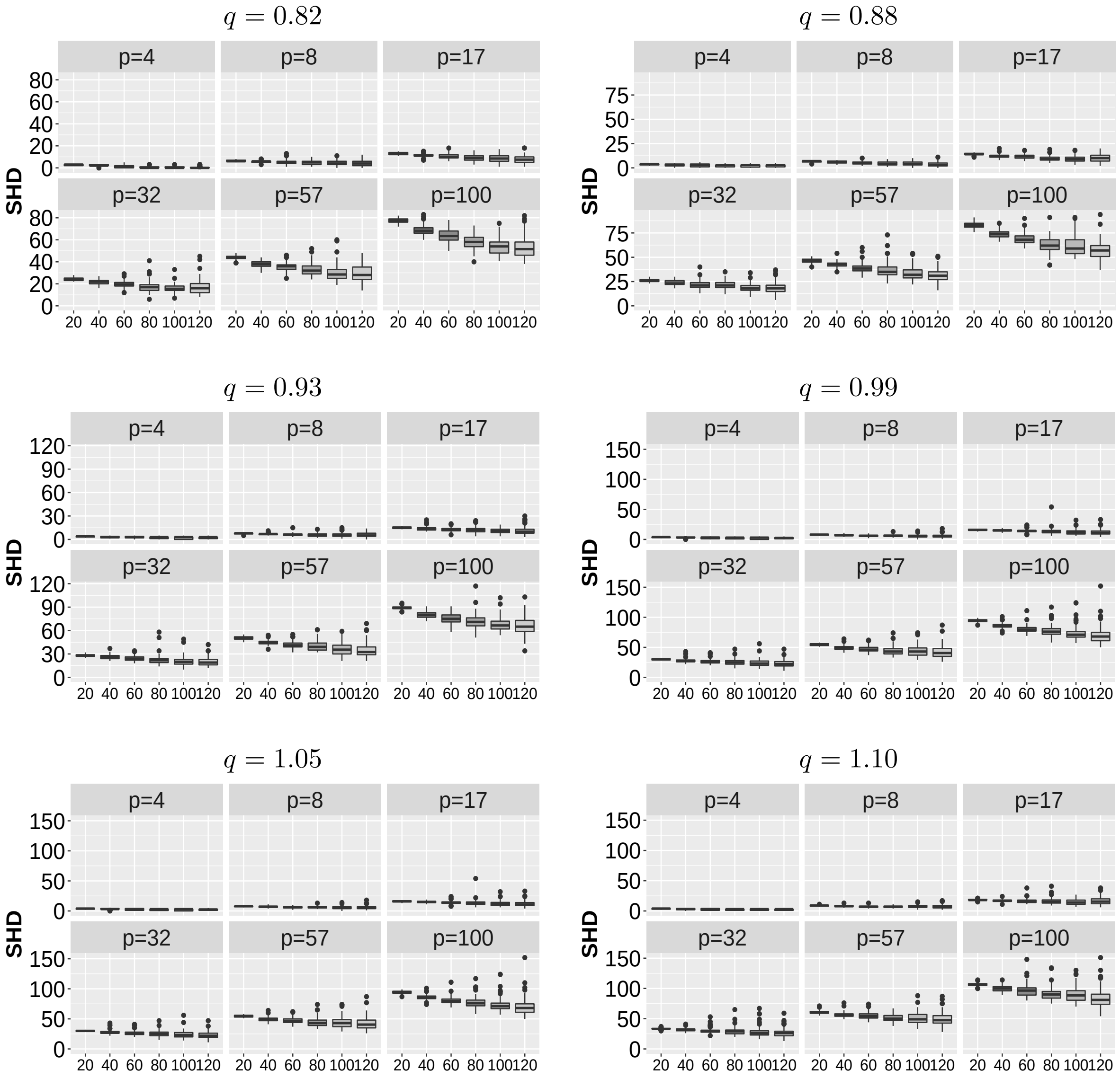


Figure 5: The structure Hamming distance (SHD) between the estimated and true CPDAG for combinations of the sample size n, graph size p, and sparsity rate q.

Gaussian distribution. There is a penalty constant in Qiao et al. (2019) that determines the level of sparsity of the estimated undirected graph. We have experimented with a range of penalty values, resulting in different percentage values of the selected edges among all edges, from 100% to 4%, for the initial estimation. When the percentage is 100%, there is no penalty in the undirected graph estimation, or effectively, no pre-screening for our method.

Table 3 reports the average SHD and TDR and the standard error (in the parenthesis) based on 80 data replications for Model I with $(n, p, q) = (50, 50, 0.7)$. We see that, as the

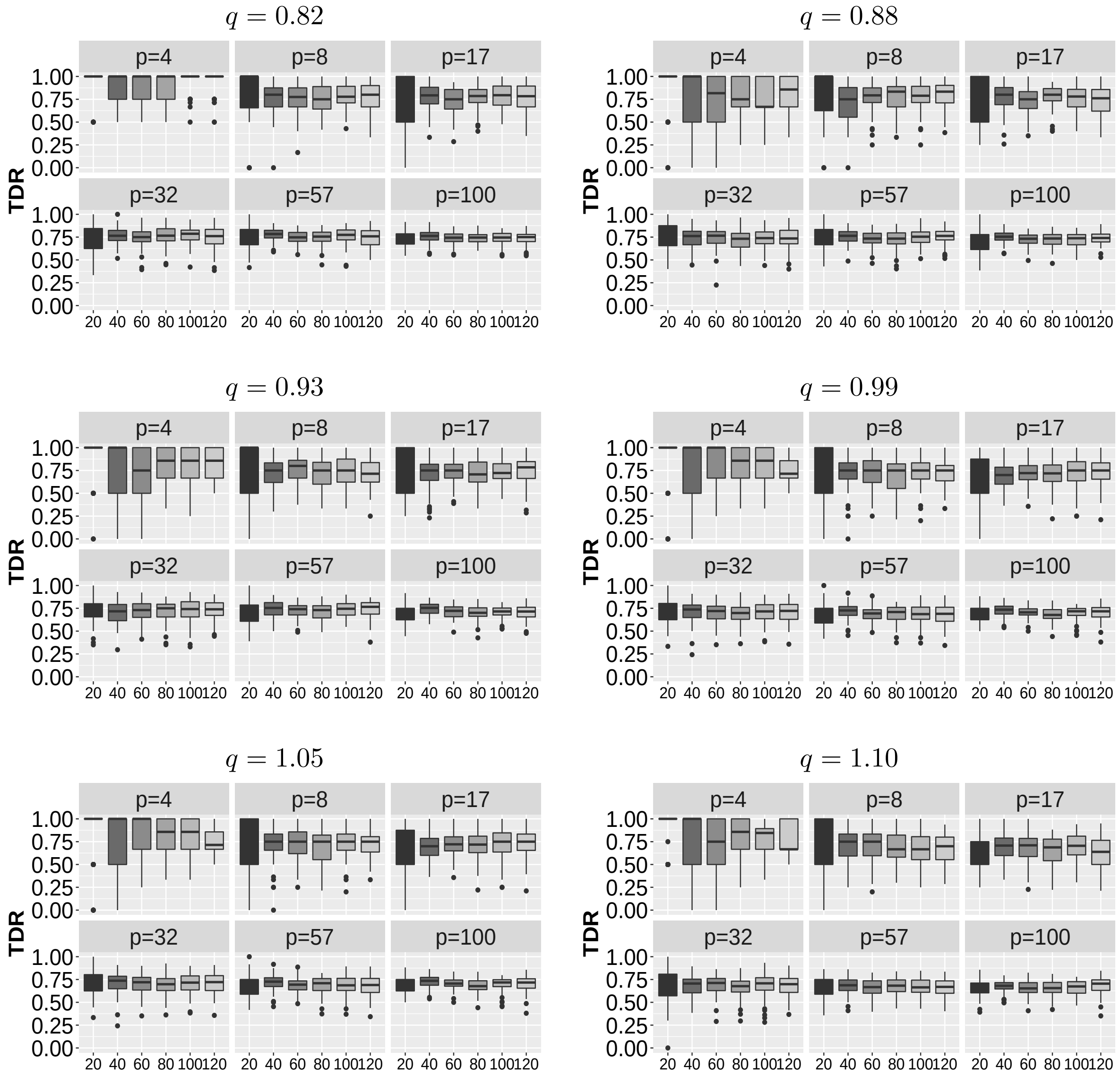


Figure 6: The true discovery rate (TDR) between the estimated and true CPDAG for combinations of the sample size n, graph size p, and sparsity rate q.

percentage of the pre-selected edges decreases to 10%, the performance of the combined algorithm improves. When this percentage drops below 10%, the performance begins to decline. This example shows the potential advantage of coupling an initial undirected graph estimation with our proposed DAG estimation method.

Effect of the kernel function: In our simulations in Section 6.1, we employ the Brownian motion covariance function (BMC) kernel, $\kappa_T = \min(s,t)$. To investigate whether our method is sensitive to the choice of kernels, here we use the radial basis function (RBF) kernel, $\kappa_T = \exp\{\gamma_G(s-t)^2\}$, where the bandwidth parameter is computed as

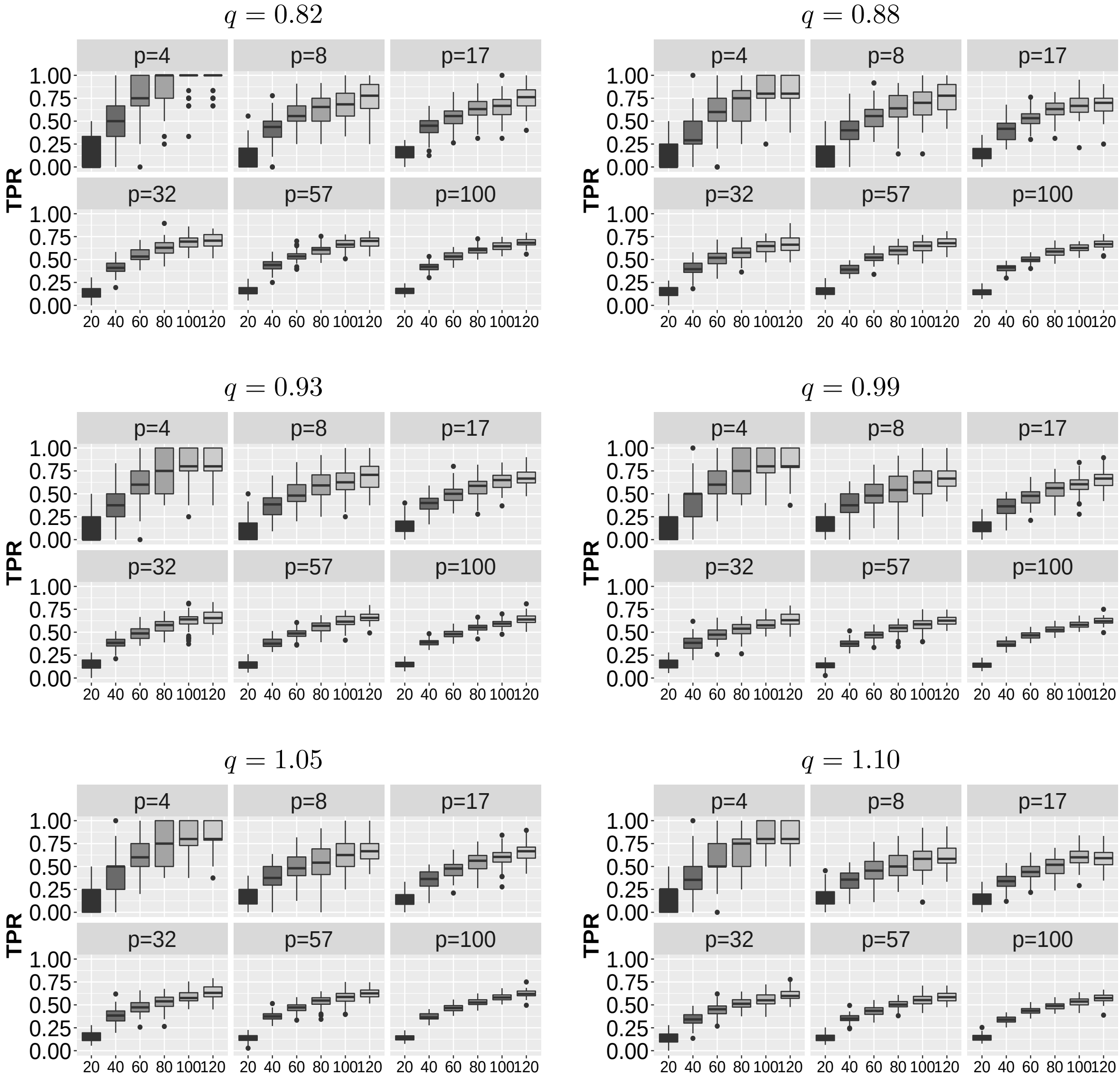


Figure 7: The true positive rate (TPR) between the estimated and true CPDAG for combinations of the sample size n, graph size p, and sparsity rate q.

$\gamma_T = \left\{\sum_{s<t}|\tau_s - \tau_t|/\binom{\ell}{2}\right\}^{-2}$ following Li and Song (2017). We compare the estimation results using these two different kernels. We still use the simulation Model I in Section 6.1, with $(n, p, q) = (50, 50, 1.05)$. Figure 9 shows the box plots of structure Hamming distance (SHD) and true positive rate (TPR) based on 80 data replications. We see that our method displays a relatively stable performance under the change of kernels.

Comparison with Lee and Li (2022) We analytically compare our proposed method with the SEM method of Lee and Li (2022) in Section 1; see the last paragraph of page 3. Here we numerically compare the two methods. We adopt the simulation Model I

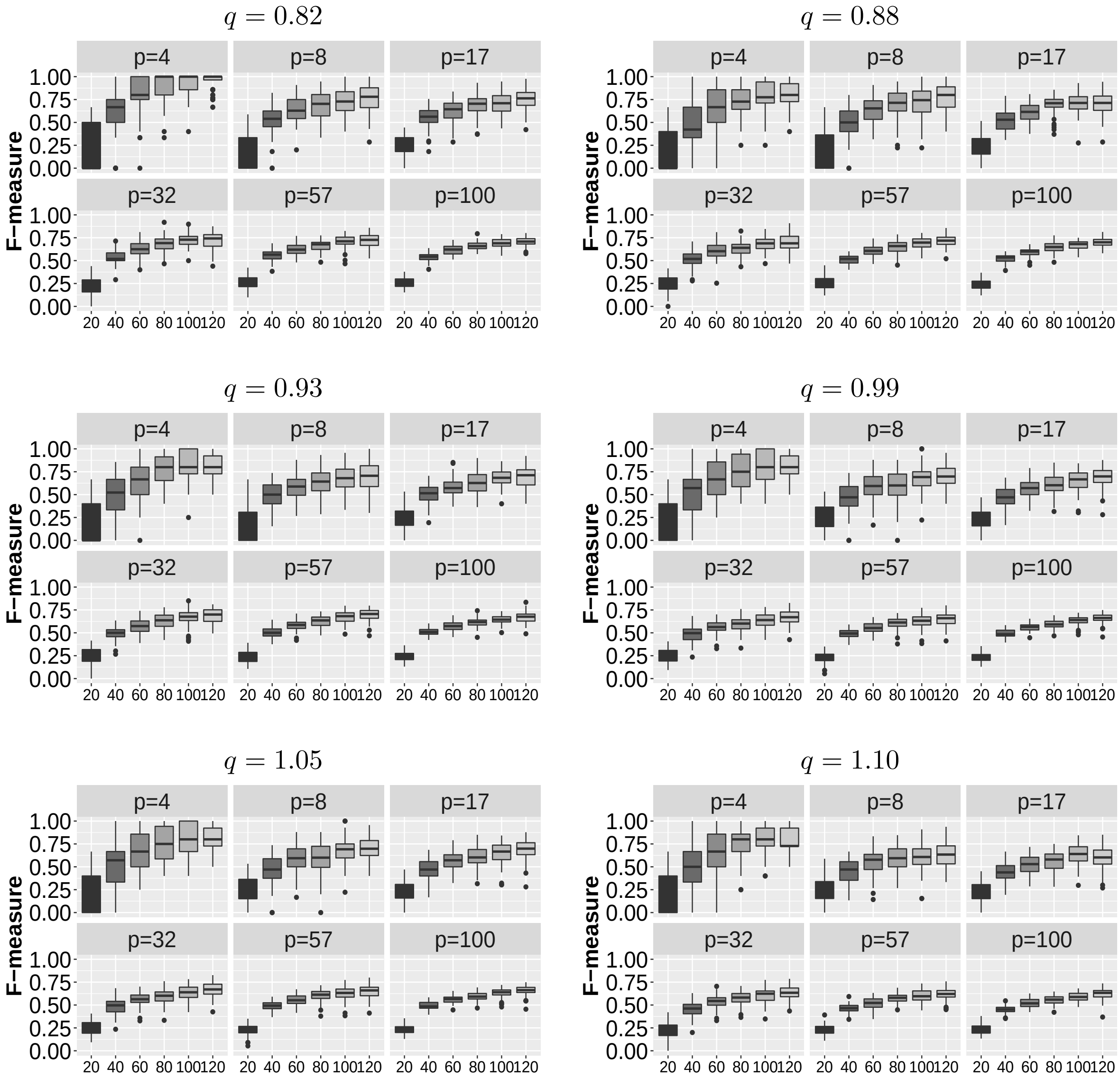


Figure 8: The F-measure (FM) between the estimated and true CPDAG for combinations of the sample size n, graph size p, and sparsity rate q.

in Section 6.1, with $(n, p, q) = (50, 50, 1.05)$. Figure 10 shows the box plots of structure Hamming distance (SHD) and true positive rate (TPR) based on 80 data replications. We see that our method performs better in this example, partly because our method is built upon the Gaussian assumption, which is satisfied in this simulated model. By comparison, the SEM method of Lee and Li (2022) does not require the Gaussian assumption. We also point out that, in this example, the computation of our method is much faster than SEM. On a 2 x E5-2630 v4 workstation, the average running time of our method is 6.31 seconds, and that of SEM is 18.64 seconds.

ave. sparsity	100%	17%	15%	12%	**10%**	8%	5 %	6%	4%
SHD	23.73	22.34	20.95	19.71	**18.51**	18.69	19.96	22.56	31.75
(s.e.)	6.60	6.29	6.11	6.14	6.19	5.98	5.99	5.82	3.98
TDR	0.71	0.75	0.77	0.79	**0.81**	0.81	0.80	0.77	0.62
(s.e.)	0.08	0.08	0.08	0.08	0.08	0.08	0.09	0.10	0.09

Table 3: The average and standard error (in the parenthesis) of the structure Hamming distance (SHD) and the true discovery rate (TDR), for the undirected graph pre-screening as the initialization + the proposed DAG estimation.

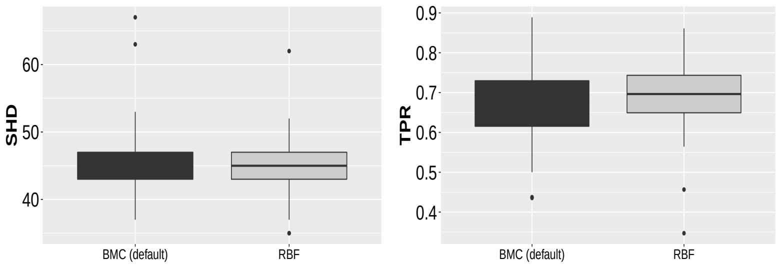


Figure 9: Empirical performance, in terms of structure Hamming distance (SHD) and true positive rate (TPR), under Brownian motion covariance function (BMC) kernel and radial basis function (RBF) kernel.

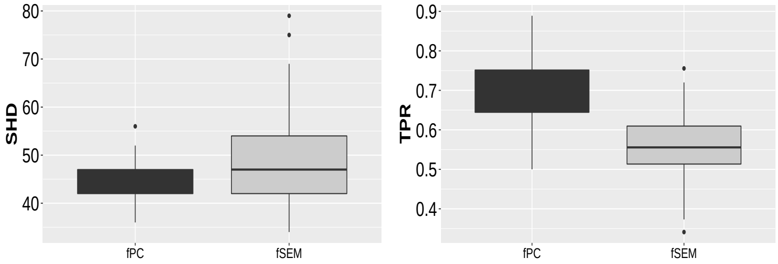


Figure 10: Empirical performance, in terms of structure Hamming distance (SHD) and true positive rate (TPR), between the proposed PCO method and the SEM method of Lee and Li (2022).

Plot of the proteomic data: Figure 11 plots the time-course measurements for all 20 protein levels in the DREAM breast cancer proteomic dataset.

References

Rita Giuliano Antonini. Subgaussian random variables in hilbert spaces. *Rendiconti del Seminario Matematico della Università di Padova*, 98:89–99, 1997.

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

P. J. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

D Bosq. *Linear Processes in Function Spaces.* Springer, 2000. ISBN 0387950524.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, 2013.

P Bühlmann, J Peters, and J Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer Publishing Company, 1st edition, 2011. ISBN 3642201911.

T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Kehui Chen and Jing Lei. Localized functional principal component analysis. *Journal of the American Statistical Association*, 110(511):1266–1275, 2015.

Xiaohui Chen and Yun Yang. Hanson–Wright inequality in Hilbert spaces with application to K-means clustering for non-Euclidean data. *Bernoulli*, 27(1):586 – 614, 2021.

D M Chickering. Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2(3):445–498, 2002.

D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, pages 3921–3962, 2014.

Y. Fan and J. Lv. Innovated scalable efficient estimation in ultra-large Gaussian graphical models. *The Annals of Statistics*, 44(5):2098–2126, 2016.

Y. Fan, G.M. James, and P. Radchenko. Functional additive regression. *The Annals of Statistics*, 43:2296–2325, 10 2015.

Jerome H. Friedman, Trevor J. Hastie, and Robert J. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.

Karl J. Friston. Functional and effective connectivity: A review. *Brain Connectivity*, 1(1): 13–36, 2011.

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

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

Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, mar 2011.

Ana María Estrada Gómez, Kamran Paynabar, and Massimo Pacella. Functional directed graphical models and applications in root-cause analysis and diagnosis. *Journal of Quality Technology*, 53(4):421–437, 2021.

Naftali Harris and Mathias Drton. PC algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(1):3365–3383, 2013.

Alain Hauser and Peter Bühlmann. Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society, Series B.*, 77:291–318, 2015.

Hill et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature Methods*, 13(4):310–318, February 2016.

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

T. Kato. *Perturbation Theory of Linear Operators*. Springer, 1980.

J. L. Kazdan. Perturbation of complete orthonormal sets and eigen-function expansions. *Proceedings of the American Mathematical Society*, 27:506–510, 1971.

John L. Kelley. *General topology*. Springer-Verlag, 1955.

S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H. G. Leimer. Independence properties of directed markov fields. *Networks*, 20(5):491–505, 1990.

Steffen L Lauritzen. *Graphical Models*. Oxford: Oxford University Press, 1996.

Kuang-Yao Lee and Lexin Li. Functional structural equation model. *Journal of the Royal Statistical Society: Series B*, 84(2):600–629, April 2022.

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

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.

Kuang-Yao Lee, Dingjue Ji, Lexin Li, Todd Constable, and Hongyu Zhao. Conditional functional graphical models. *Journal of the American Statistical Association*, 2021.

B. Li and E. 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. Linear operator-based statistical analysis: A useful paradigm for big data. *Canadian Journal of Statistics*, 46:79–103, 2018.

Bing Li and Jun Song. Nonlinear sufficient dimension reduction for functional data. *Annals of Statistics*, 45(3):1059–1095, 2017. ISSN 00905364.

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

Chunlin Li, Xiaotong Shen, and Wei Pan. Likelihood ratio tests for a large directed acyclic graph. *Journal of the American Statistical Association*, 115(531):1304–1319, 2020.

Yehua Li and Yongtao Guan. Functional principal component analysis of spatiotemporal point processes with applications in disease surveillance. *Journal of the American Statistical Association*, 109(507):1205–1215, 2014.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

Qingyang Liu, Yuping Zhang, and Zhengqing Ouyang. Structural inference of time-varying mixed graphical models. *Stat*, 10(1):e414, 2021.

Weidong Liu. Structural similarity and difference testing on multiple sparse Gaussian graphical models. *The Annals of Statistics*, 45(6):2680–2707, December 2017.

R Luo and X Qi. Function-on-function linear regression by signal compression. *Journal of the American Statistical Association*, 112(518):690–705, 2016.

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

Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6):3133–3164, 2009.

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

Ardalan Mirshani and Matthew Reimherr. Adaptive function-on-scalar regression with a smoothing elastic net. *Journal of Multivariate Analysis*, 185:104765, 2021.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press, 2nd Edition, 2009.

J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

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

Xinghao Qiao, Shaojun Guo, and Gareth M. James. Functional graphical models. *Journal of the American Statistical Association*, 114(525):211–222, 2019.

Xinghao Qiao, Cheng Qian, Gareth M James, and Shaojun Guo. Doubly functional graphical models in high dimensions. *Biometrika*, 107(2):415–431, 02 2020. ISSN 0006-3444. doi: 10.1093/biomet/asz072.

Shaojun Guo Qin Fang and Xinghao Qiao. Adaptive functional thresholding for sparse covariance function estimation in high dimensions. *Journal of the American Statistical Association*, 0(0):1–13, 2023. doi: 10.1080/01621459.2023.2200522. URL `https://doi.org/10.1080/01621459.2023.2200522`.

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980, 2011.

Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H. Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, 2nd edition, 1979.

Eftychia Solea and Bing Li. Copula gaussian graphical models for functional data. *Journal of the American Statistical Association, To appear*, 2020.

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

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):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, April 2013.

Sara van de Geer and Peter Bühlmann. l0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.

Kartik G. Waghmare, Tomas Masak, and Victor M. Panaretos. The functional graphical lasso. *arXiv*, 2023.

F. Yao and H.-G. Müller. Functional quadratic regression. *Biometrika*, 97(1):49–64, 2010.

F. Yao, H.-G. Müller, and J.-L. Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903, 2005.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

J Zapata, S Y Oh, and A Petersen. Partial separability and functional graphical models for multivariate Gaussian processes. *Biometrika*, 109(3):665–681, 10 2021. ISSN 1464-3510. doi: 10.1093/biomet/asab046. URL `https://doi.org/10.1093/biomet/asab046`.

Boxin Zhao, Y. Samuel Wang, and Mladen Kolar. Fudge: A method to estimate a functional differential graph in a high-dimensional setting. *Journal of Machine Learning Research*, 23(82):1–82, 2022.

Hongxiao Zhu, Nate Strawn, and David B. Dunson. Bayesian graphical models for multivariate functional data. *Journal of Machine Learning Research*, 17(204):1–27, 2016. URL `http://jmlr.org/papers/v17/16-164.html`.

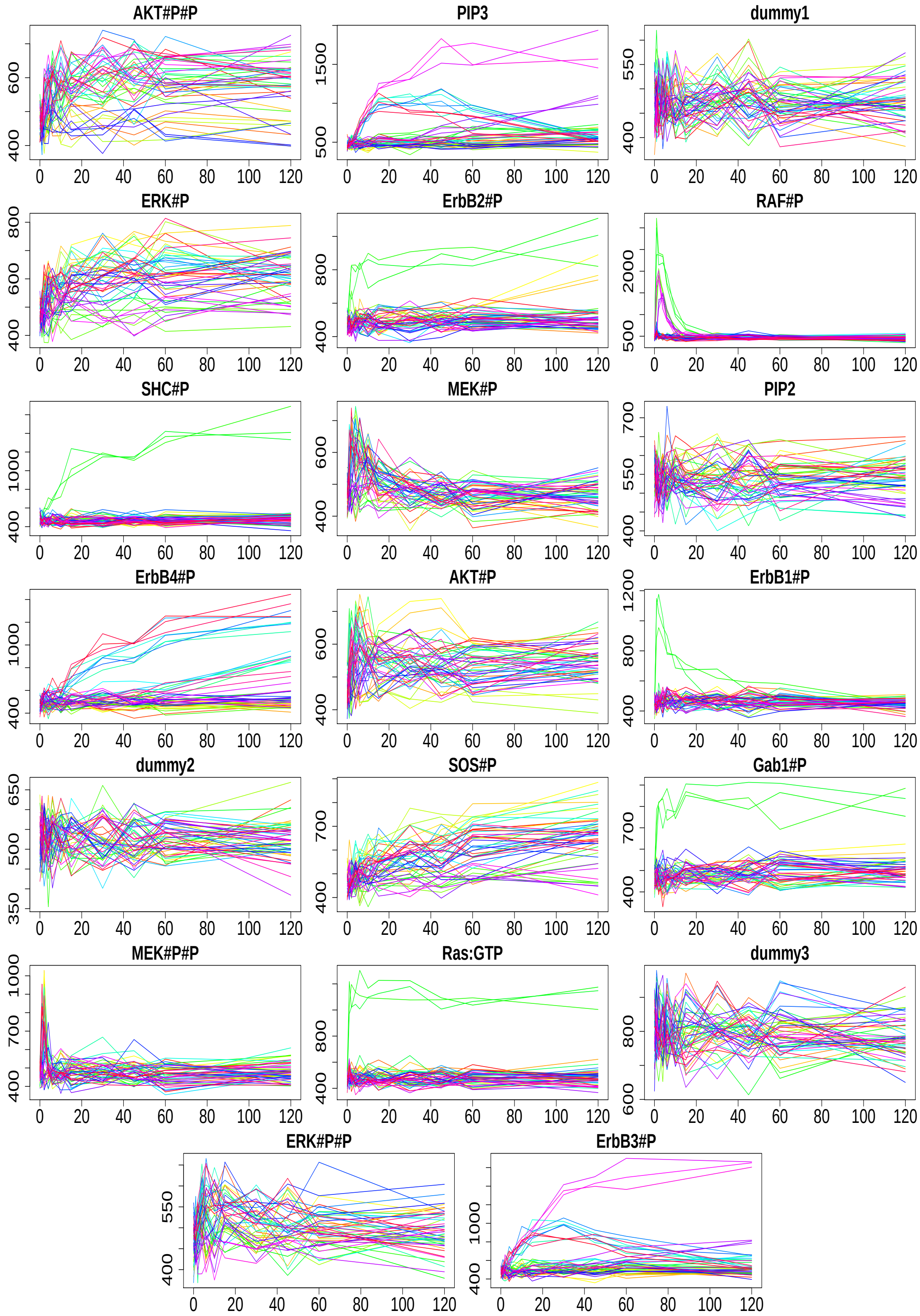


Figure 11: Plots of 20 protein levels for the proteomic data.