Received: 8 April 2020 | Accepted: 15 August 2021

DOI: 10.1111/rssb.12471

ROYAL STATISTICAL SOCIETY
Series B Statistical Methodology

ORIGINAL ARTICLE

Functional structural equation model

Kuang-Yao Lee[1] | **Lexin Li**[2]

[1]Temple University, Philadelphia, Pennsylvania, USA

[2]University of California, Berkeley, California, USA

Correspondence
Lexin Li, Department of Biostatistics and Epidemiology, University of California, Berkeley, 2121 Berkeley Way, Berkeley, California 94720, USA.
Email: lexinli@berkeley.edu

Funding information
NSF, Grant/Award Number: CIF-2102243 and CIF-2102227; NIH, Grant/Award Number: R01AG061303, R01AG062542 and R01AG034570

Abstract

In this article, we introduce a functional structural equation model for estimating directional relations from multivariate functional data. We decouple the estimation into two major steps: directional order determination and selection through sparse functional regression. We first propose a score function at the linear operator level, and show that its minimization can recover the true directional order when the relation between each function and its parental functions is nonlinear. We then develop a sparse functional additive regression, where both the response and the multivariate predictors are functions and the regression relation is additive and nonlinear. We also propose strategies to speed up the computation and scale up our method. In theory, we establish the consistencies of order determination, sparse functional additive regression, and directed acyclic graph estimation, while allowing both the dimension of the Karhunen–Loéve expansion coefficients and the number of random functions to diverge with the sample size. We illustrate the efficacy of our method through simulations, and an application to brain effective connectivity analysis.

KEYWORDS

brain connectivity analysis, directed acyclic graph, function-on-function regression, linear operator, reproducing kernel Hilbert space, structural equation model

1 | INTRODUCTION

Multivariate functional data, where continuous observations are sampled from a vector of stochastic processes, are rapidly emerging in a broad range of applications, such as genomics (Wei & Li, 2008), finance (Tsay & Pourahmadi, 2017) and neuroimaging (Luo et al., 2019). A problem of central interest in multivariate functional data analysis is to investigate the directional relationships among the random functions. Our motivation is brain effective connectivity, which refers explicitly to the directional influence that one neural system exerts over another (Friston, 2011). Effective connectivity analysis estimates such directional influences among different brain regions through electrocorticographic imaging (ECoG) or task-based functional magnetic resonance imaging (fMRI). The imaging data are usually summarized in the form of a region by time matrix for each individual, with the rows corresponding to a set of brain regions or locations, and the columns the time points. Given the continuous nature and the short time interval between the adjacent sampling points, it is natural to treat the data at each location as a function. A central goal is to model such multivariate functions and to estimate brain effective connectivity among different regions.

In this article, we tackle the problem of learning directional relations from multivariate functional data. We develop a new functional structural equation model, which extends the classical structural equation model (SEM) from the setting of random variables to that of random functions. Motivated by Bühlmann et al. (2014) and Loh and Bühlmann (2014) for SEM of random variables, we decouple the estimation of functional SEM into two major steps: directional order determination and selection through sparse functional regression. Specifically, to determine the directional order, we introduce a new score function, which is in spirit similar to the residual sum of squares from regression, but its deployment is entirely at the operator level. We then show that, when each random function is nonlinearly associated with its parents, the true order can be identified by minimizing the proposed score function. This result echoes and extends that of Hoyer et al. (2009) in the sense that, the deviation from the linear structure between random functions can help determine the directional order. Once the directional order is found, the estimation of SEM for random variables is equivalent to variable selection that can be carried out by sparse regression (see, e.g. Shojaie & Michailidis, 2010). Analogously, we extend the sparse additive model for random variables (Meier et al., 2009) to the setting where both the response and the multivariate predictors are random functions. We introduce a least-squares type objective function at the operator level, and show that the true directional relationships in our functional SEM can be recovered by minimizing this objective function plus some sparsity regularization. Moreover, both our order determination and sparse functional regression hinge on a novel statistical quantity we propose, the functional regression operator (FRO). Finally, to speed up the computation and scale up our method to the setting of a large number of functions, we employ two additional strategies. The first is to use only the leading eigenfunctions to estimate relevant linear operators. The same strategy has been commonly used in functional data analysis (see, e.g. Yao et al., 2005). The second is to implement a pre-neighbourhood selection step, again through sparse functional regression, to reduce the number of candidate directional orders. This avoids evaluating the score function for all possible $p!$ combinations, whose computation can be huge when the total number of functions p is large.

Our proposal is built upon, but also substantially extends several lines of relevant research, including the classical SEM for random variables, directed acyclic graph (DAG) estimation, function-on-function regression and linear operator-based methods.

Structural equation modelling originated from genetic path modelling (Wright, 1923), while Pearl (2009) further developed the causal interpretation of the equation. There is a large amount of work studying SEM for random variables. The most classic is the Gaussian SEM, where the variables follow a joint Gaussian distribution (Westland, 2015). Extensions include the relaxation of the Gaussian assumption, nonlinear regression of the random variables and high-dimensional estimation through penalization (e.g. Bühlmann et al., 2014; Fu & Zhou, 2013; Li et al., 2020; Loh & Bühlmann, 2014; Peters et al., 2014; Yuan et al., 2018, among others). By contrast, we aim at SEM for random functions, and such an extension is far from incremental. Unlike the SEM setting for random variables where the dimension of each coordinate is fixed, we allow each node to be a Hilbert space-valued random element. We approximate each random function by its leading Karhunen–Loéve (KL) expansion coefficients, and allow both the dimension of the KL coefficients and the number of random functions to grow with the sample size. Luo et al. (2019) recently proposed functional SEM for twin functional data. Although our method bears the same name as that of Luo et al. (2019), we target a different problem. Luo et al. (2019) aimed to dissect functional genetic and environmental effects on twin functional data, while characterizing the varying association between functional data and covariates of interest through some varying coefficient model. We aim to recover directional influences among different functions. Consequently, our method and the associated theory are all different.

Since SEM is closely related to DAG, there are also a class of DAG estimation methods that are built on a computationally efficient PC algorithm (Kalisch & Bühlmann, 2007). These methods require a faithfulness assumption, and only estimate an equivalence class of DAG. The method we develop, on the other hand, does not require the faithfulness condition, and is fully identifiable. More recently, Qiao et al. (2019) and Li and Solea (2018) introduced graphical models for multivariate functions. However, they targeted undirected graphs, whereas we aim at directed graphs, which are two utterly different problems.

A major ingredient of our method is the function-on-function regression. Existing works mostly focus on linear functional regression, or additive functional regression with only a single predictor function (Fan et al., 2015; Luo & Qi, 2016; Müller & Yao, 2008; Reimherr et al., 2018). By contrast, we consider a model with high-dimensional functional predictors, and the relation between each function and its parent nodes is nonlinear, a property our method relies on to determine the directional order. We capture such nonlinear associations via the functional regression operator, which can be understood as the generalization of the regression coefficients to functional, additive and high-dimensional settings.

Linear operator-based estimation methods are gaining increasing attention in recent years. Lee et al. (2016b) proposed a regression operator for nonlinear variable selection, whereas Li and Song (2017) studied nonlinear function-on-function relations using the regression operator for sufficient dimension reduction. See Li (2018) for a survey of recent development of linear operator-based methods. Nevertheless, compared to those methods, our proposal aims at a completely different problem of inferring directional relationships among random functions. Moreover, while Lee et al. (2016b) and Li and Song (2017) studied the asymptotics with a fixed number of variables or functions, we derive the asymptotics under a diverging dimensionality. We have also established the error bound of our functional regression operator, which was not available in Lee et al. (2016b) and Li and Song (2017).

In summary, our proposed functional SEM expands the scope of the existing SEM to the setting of nonlinear, high-dimensional and functional data. The problem of directional learning for functional data remains largely untapped, is challenging, and requires a new set of modelling

and theoretical techniques. Methodologically, our problem requires to evaluate every regression between each functional node and its parent nodes, and for each fixed order from a large number of permutations. Theoretically, we need to take into account the error for the KL approximation of functions, as well as to derive the concentration inequalities for a number of linear operators. We establish the uniform consistency of both directional order determination and sparse functional additive regression, as well as the uniform graph consistency of DAG estimation, by allowing both the dimension of the KL coefficients and the number of functions to diverge with the sample size. To our knowledge, there has been little work studying function-on-function dependency and SEM consistency at such a level of complexity, and we consider our work to be the first involving nonlinear, high-dimensional and functional settings altogether. Moreover, the proposed model is not only designed for brain effective connectivity analysis, but is equally applicable to a variety of multivariate functional data applications. The theoretical tools we derive are also sufficiently general, and are useful for broader settings of high-dimensional, linear operator-based inference.

The rest of the article is organized as follows. We introduce the functional structural equation model in Section 2, and the functional regression operator, a new score function and a sparse functional additive regression in Section 3. We develop the estimation procedure at the operator level and under a given coordinate system in Section 4. We derive the asymptotic properties in Section 5, and study the numerical performance through simulations and a brain effective connectivity example in Section 6. We conclude with a discussion in Section 7. We relegate all technical proofs and some additional results to the Supplementary Appendix.

2 | MODEL

In this section, we first provide a simple yet concrete example to motivate our functional structural equation model. We then construct a two-layer reproducing kernel Hilbert space (RKHS), which serves as the foundation to model nonlinear relations among multivariate random functions. Lastly, we formally define the functional structural equation model.

Example 1 Suppose $\mathcal{H}$ is a generic Hilbert space spanned by an orthonormal basis $\{f^m\}_{m\in\mathbb{N}}$, and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is its inner product. Suppose $X_1 = \varepsilon_1$, and X_2, and X_3 satisfy that,

$$\begin{aligned}\langle X_2, f^m\rangle_{\mathcal{H}} &= \phi_1^m(X_1) + \langle \varepsilon_2, f^m\rangle_{\mathcal{H}},\\ \langle X_3, f^m\rangle_{\mathcal{H}} &= \phi_2^m(X_1) + \phi_3^m(X_2) + \langle \varepsilon_2, f^m\rangle_{\mathcal{H}}, \quad \text{for } m\in\mathbb{N},\end{aligned} \tag{1}$$

where $\{\phi_1^m\}_{m\in\mathbb{N}}$, $\{\phi_2^m\}_{m\in\mathbb{N}}$ and $\{\phi_3^m\}_{m\in\mathbb{N}}$ are families of functionals on $\mathcal{H}$, and ε_1, ε_2, ε_3 are i.i.d., zero-mean, $\mathcal{H}$-valued Gaussian random elements.

To gain some insight from Equation (1), note that for $i = 2, 3$, X_i has a one-to-one correspondence with $\{\langle X_i, f^m\rangle_{\mathcal{H}}\}_{m\in\mathbb{N}}$, the collection of the inner products between X_i and f^m. This means that the relation between X_1 and X_2 can be fully characterized by that between X_1 and $\{\langle X_2, f^m\rangle_{\mathcal{H}}\}_{m\in\mathbb{N}}$. Moreover, since $\langle \varepsilon_2, f^m\rangle_{\mathcal{H}}$ is a Gaussian random variable, Equation (1) implies that, for each $m \in \mathbb{N}$, the relation between X_1 and $\langle X_2, f^m\rangle_{\mathcal{H}}$ follows a standard additive regression (Buja et al., 1989), where the response $\langle X_2, f^m\rangle_{\mathcal{H}}$ is equal to an arbitrary function of the predictor X_1 plus a Gaussian error. Similarly, one can interpret the relation between X_3 and $(X_1, X_2)^\mathsf{T}$. Figure 1 visualizes the directional relations among X_1, X_2, and X_3.

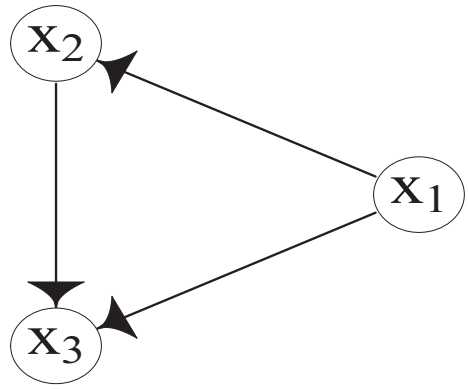

FIGURE 1 The directional relations in Example 1

The key to characterize the nonlinear relations among multivariate random functions is the two-layer RKHS we construct as follows. We begin with some notations. Given two Hilbert spaces $\mathcal{H}$ and $\mathcal{K}$, let $\mathscr{B}(\mathcal{H}, \mathcal{K})$ be the collection of all bounded linear operators from $\mathcal{H}$ to $\mathcal{K}$, $\mathscr{B}_1(\mathcal{H}, \mathcal{K})$ the collection of all trace-class operators from $\mathcal{H}$ to $\mathcal{K}$, and $\mathscr{B}_2(\mathcal{H}, \mathcal{K})$ the collection of all Hilbert–Schmidt operators from $\mathcal{H}$ to $\mathcal{K}$. Clearly, we have $\mathscr{B}_1(\mathcal{H}, \mathcal{K}) \subseteq \mathscr{B}_2(\mathcal{H}, \mathcal{K}) \subseteq \mathscr{B}(\mathcal{H}, \mathcal{K})$. Note that $\mathscr{B}_2(\mathcal{H}, \mathcal{K})$ is a Hilbert space with $\langle \cdot, \cdot \rangle_{\mathrm{HS}}$ and $\|\cdot\|_{\mathrm{HS}}$ denoting its inner product and norm, and $\mathscr{B}(\mathcal{H}, \mathcal{K})$ is a Banach space with $\|\cdot\|$ denoting its operator norm. When $\mathcal{H} = \mathcal{K}$, we simply write $\mathscr{B}_1(\mathcal{H})$, $\mathscr{B}_2(\mathcal{H})$, and $\mathscr{B}(\mathcal{H})$. For a linear operator Γ, let $\ker(\Gamma)$ denote the null space of Γ, that is, $\ker(\Gamma) = \{f : \Gamma f = 0\}$, and $\mathrm{range}(\Gamma)$ denote the range of Γ. Furthermore, let $\mathtt{G} = (\mathtt{V}, \mathtt{E})$ denote a DAG, where $\mathtt{V} = \{1, \ldots, p\}$ indexes the nodes, and $\mathtt{E} \subseteq \{(i, j) \in \mathtt{V} \times \mathtt{V} : i \neq j\}$ collects the set of directed edges with no directed cycle, with (i, j) denoting $j \to i$ and that j is a parent of i, and i is a child of j. If there is a directed path $j \to \cdots \to i$, then j is an ancestor of i, and i is a descendant of j.

First, we construct the first-layer functional space. Let $X(t) = \{X_1(t), \ldots, X_p(t)\}^{\mathsf{T}}$, $t \in T$, denote a p-dimensional $\mathcal{H}_X$-valued random function, where $\mathcal{H}_X = \mathcal{H}_{X_1} \times \cdots \times \mathcal{H}_{X_p}$ is the Cartesian product, and $\mathcal{H}_{X_i}$ is an RKHS of the $\mathbb{R}$-valued functions on an interval $T \subset \mathbb{R}$, $i \in \mathtt{V}$. For simplicity, we let $\mathcal{H}_{X_1} = \cdots = \mathcal{H}_{X_p}$. That is, we take $\mathcal{H}_{X_i}$ to be the closure of the pre-Hilbert space $\left\{ \sum_{j=1}^{n} \kappa_T(\cdot, t_j)\omega_j : \omega_1, \ldots, \omega_n \in \mathbb{R},\ t_1, \ldots, t_n \in T \right\}$, with the inner product determined by a positive kernel $\kappa_T : T \times T \to \mathbb{R}$, $i \in \mathtt{V}$. Let $\langle \cdot, \cdot \rangle_{\mathcal{H}_{X_i}}$ denote the inner product in $\mathcal{H}_{X_i}$, $i \in \mathtt{V}$, which implies the product space $\mathcal{H}_X$ is also an RKHS with the inner product $\langle f, g \rangle_{\mathcal{H}_X} = \langle f_1, g_1 \rangle_{\mathcal{H}_{X_1}} + \cdots + \langle f_p, g_p \rangle_{\mathcal{H}_{X_p}}$, for any $f = (f_1, \ldots, f_p)^{\mathsf{T}}$ and $g = (g_1, \ldots, g_p)^{\mathsf{T}}$ in $\mathcal{H}_X$. Let $X_{\mathtt{A}}(t) = \{X_i(t) : i \in \mathtt{A}\}$ for any subset $\mathtt{A} \subseteq \mathtt{V}$, and define $\mathcal{H}_{X_{\mathtt{A}}}$ similarly. Also, we write the functions $X(t)$ and $X_i(t)$ as X and X_i when there is no confusion.

Next, we introduce a second functional space on top of the first one, that is, the domain space $\mathcal{H}_X$. To formalize the idea, we first define the nested kernel and the associated RKHS. A positive definite kernel $\kappa_i : \mathcal{H}_{X_i} \times \mathcal{H}_{X_i} \to \mathbb{R}$ can be induced by $\langle \cdot, \cdot \rangle_{\mathcal{H}_{X_i}}$ if, for any $f, g \in \mathcal{H}_{X_i}$, there exists a mapping $\rho : \mathbb{R}^3 \to \mathbb{R}$ satisfying

$$\kappa_i(f, g) = \rho\left(\langle f, f \rangle_{\mathcal{H}_{X_i}}, \langle f, g \rangle_{\mathcal{H}_{X_i}}, \langle g, g \rangle_{\mathcal{H}_{X_i}} \right). \tag{2}$$

The construction of κ_i is originated from the kernel function for the classical multivariate random variable setting, whose input space is then switched from the Euclidean space to an RKHS. Since the

function κ_i is uniquely determined by the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{X_i}}$ of the nested space $\mathcal{H}_{X_i}$ and the choice of ρ, we call κ_i a nested kernel function. Examples of κ_i include the radial basis kernel $\kappa_i(f, g) = \exp(-\gamma \|f - g\|^2_{\mathcal{H}_{X_i}})$, and the polynomial kernel $\kappa_i(f, g) = (1 + \langle f, g \rangle_{\mathcal{H}_{X_i}})^\gamma$. See also Li and Song (2017).

Supplying a nested kernel κ_i, we next build a second-layer functional space $\mathscr{H}_{X_i}$ as the RKHS induced by κ_i. For a member f in $\mathcal{H}_{X_i}$, let $\kappa_i(\cdot, f) \colon g \mapsto \kappa_i(g, f)$ define a functional on $\mathscr{H}_{X_i}$. We then take $\mathscr{H}_{X_i}$ to be the closure of the space spanned by all such functionals $\kappa_i(\cdot, f)$, or equivalently, the closure of the pre-Hilbert space $\{ \sum_{j=1}^n \kappa_i(\cdot, f_j)\omega_j \colon \omega_1, \ldots, \omega_n \in \mathbb{R}, f_1, \ldots, f_n \in \mathcal{H}_{X_i}\}$ with the inner product determined by κ_i. For example, when κ_i is chosen to be the radial basis function, $\mathscr{H}_{X_i} = \overline{\text{span}}\{\exp(-\gamma \|f - \cdot\|^2_{\mathcal{H}_{X_i}}) \colon f \in \mathcal{H}_{X_i}\}$, where $\overline{\text{span}}(\cdot)$ denotes the closure of span$(\cdot)$. Lemma S1 in Section S1 of the Appendix establishes the reproducing property of $\mathscr{H}_{X_i}$. Define the direct sum $\mathscr{H}_X = \left\{ \phi_1 + \cdots + \phi_p \colon \phi_1 \in \mathscr{H}_{X_1}, \ldots, \phi_p \in \mathscr{H}_{X_p} \right\}$, whose inner product is $\langle \phi_1 + \cdots + \phi_p, \psi_1 + \cdots + \psi_p \rangle = \sum_{i=1}^p \langle \phi_i, \psi_i \rangle_{\mathscr{H}_{X_i}}$. For any subset $\mathtt{A}$ of $\mathtt{V}$, define $\mathscr{H}_{X_\mathtt{A}}$ similarly. Next, we formally define the functional structural equation model.

Definition 1 We say $X(t) = \{X_1(t), \ldots, X_p(t)\}^\mathsf{T}$ follows a functional structural equation model with respect to a DAG $\mathtt{G}^0$, if there exist $B^0_{i,j} \in \mathscr{B}_2(\mathcal{H}_{X_i}, \mathcal{H}_{X_j})$ and $\Gamma^0_{\varepsilon_i} \in \mathscr{B}_1(\mathcal{H}_{X_i})$, $i = 1, ..., p$, $j \in \mathtt{pa}(i; \mathtt{G}^0)$, such that, for all $f \in \mathcal{H}_{X_i}$,

$$\langle X_i, f \rangle_{\mathcal{H}_{X_i}} = \sum_{j \in \mathtt{pa}(i;\mathtt{G}^0)} (B^0_{i,j} f)(X_j) + \langle \varepsilon_i, f \rangle_{\mathcal{H}_{X_i}},$$

where $\mathtt{pa}(i; \mathtt{G}^0) = \{j \colon (i, j) \in \mathtt{E}^0\}$ is the set of parents of node i, ε_i is an $\mathcal{H}_{X_i}$-valued Gaussian random element with mean zero and covariance $\Gamma^0_{\varepsilon_i}$, and $\varepsilon_1, \ldots, \varepsilon_p$ are independent.

Because X_i is independent with all its non-descendant nodes when conditioning on $X_{\mathtt{pa}(i;\mathtt{G})}$, the functional SEM in Definition 1 is Markovian with respect to $\mathtt{G}^0$. In addition, when $\mathcal{H}_{X_i} = \mathbb{R}$, Definition 1 reduces to the SEM for random variables in Bühlmann et al. (2014).

For the functional SEM in Definition 1, the parameters of key interest are $B^0_{i,j}$ that captures the directional relation from node j to node i, and $\Gamma^0_{\varepsilon_i}$ that reflects the variation of the error term ε_i. Analogously, $B^0_{i,j}$ and $\Gamma^0_{\varepsilon_i}$ play the roles of the regression coefficient and error variance in a classical linear regression model. Moreover, the directional relationships among all the functions induce a directed acyclic graph $\mathtt{G}^0 = (\mathtt{V}, \mathtt{E}^0)$. Next, we discuss how to estimate the proposed functional SEM.

3 | ORDER DETERMINATION AND FUNCTIONAL REGRESSION

We break the estimation of our functional SEM into two main steps: directional order determination and selection through sparse functional regression. In this section, we first introduce the functional regression operator that is useful for both steps. We then propose a new score function, and study the conditions under which the directional order can be identified by minimizing this new score function. Finally, based on a given directional order, we develop sparse functional additive regression, which captures nonlinear associations among the random functions, and in turn the underlying directional relationships.

3.1 | Functional regression operator

In this section, we first introduce a set of covariance operators, which lead to the key operator of our method, the functional regression operator (FRO). We then derive a key property of FRO. Some regularity conditions are needed to formally define those operators.

Assumption 1 Suppose $\max_{i\in \mathtt{V}} E\|X_i\|^2_{\mathcal{H}_{X_i}} \le m_0$ for some constant m_0. Moreover, for any $f, g \in \mathcal{H}_{X_i}, |\kappa_i(f, g)| \le 1$, for all $i \in \mathtt{V}$.

Assumption 1 is a mild condition. The first part ensures the uniform boundedness of $E\|X_i\|^2_{\mathcal{H}_{X_i}}$. This is equivalent to the uniform boundedness of the trace of the covariance operator of X_i. The second part requires κ_i to be bounded, which holds for many popular kernel functions, for example, the radial basis kernel and the Laplacian kernel. Under Assumption 1, we have the square integrability of both the sample path of X_i and the members in $\mathcal{H}_{X_i}$. Moreover, combined with the representation theorem, it guarantees the existence and uniqueness of the following covariance operators.

Definition 2 For any $(i, j) \in \mathtt{V} \times \mathtt{V}, \phi_i \in \mathcal{H}_{X_i}$ and $f_i \in \mathcal{H}_{X_i}$, define

$$\begin{aligned}
&\Gamma_{X_iX_j}: \mathcal{H}_{X_j} \to \mathcal{H}_{X_i}, \quad \text{such that} \quad \langle f_i, \Gamma_{X_iX_j} f_j\rangle = E\left(\langle f_i, X_i\rangle\ \langle f_j, X_j\rangle\right),\\
&\Lambda_{X_iX_j}: \mathcal{H}_{X_j} \to \mathcal{H}_{X_i}, \quad \text{such that} \quad \langle \phi_i, \Lambda_{X_iX_j} f_j\rangle = E\left\{\langle f_j, X_j\rangle\ \langle \phi_i, \kappa_i(\cdot, X_i)\rangle\right\},\\
&\Sigma_{X_iX_j}: \mathcal{H}_{X_j} \to \mathcal{H}_{X_i}, \quad \text{such that} \quad \langle \phi_i, \Sigma_{X_iX_j} \phi_j\rangle = E\left\{\langle \phi_i, \kappa_i(\cdot, X_i)\rangle\ \langle \phi_j, \kappa_j(\cdot, X_j)\rangle\right\}.
\end{aligned}$$

Despite their similar appearance, the two operators, $\Lambda_{X_iX_j}$ and $\Sigma_{X_iX_j}$, are defined on different domains: $\Lambda_{X_iX_j}$ is defined on $\mathcal{H}_{X_j}$, whereas $\Sigma_{X_iX_j}$ on $\mathcal{H}_{X_j}$. Moreover, given any subvectors $\mathtt{A}$, $\mathtt{B}$ of $\mathtt{V}$, we define the vector of operators, $\Lambda_{X_\mathtt{A}X_j} : \mathcal{H}_{X_j} \to \mathcal{H}_{X_\mathtt{A}}$, whose ith element is $\Lambda_{X_iX_j}$, and the matrix of operators, $\Sigma_{X_\mathtt{A}X_\mathtt{B}} : \mathcal{H}_{X_\mathtt{B}} \to \mathcal{H}_{X_\mathtt{A}}$, whose (i, j)th element is $\Sigma_{X_iX_j}$.

Next, we introduce the functional regression operator, under another condition.

Assumption 2 Suppose $\ker(\Sigma_{X_\mathtt{A}X_\mathtt{A}}) = \{0\}$, $\mathrm{range}(\Lambda_{X_\mathtt{A}X_i}) \subseteq \mathrm{range}(\Sigma_{X_\mathtt{A}X_\mathtt{A}})$, for all $i \in \mathtt{V}$ and $\mathtt{A} \subseteq \mathtt{V}\setminus\{i\}$, and the operator $R_{X_i|X_\mathtt{A}} \equiv \Sigma^{-1}_{X_\mathtt{A}X_\mathtt{A}}\Lambda_{X_\mathtt{A}X_i} \in \mathcal{B}_2(\mathcal{H}_{X_i}, \mathcal{H}_{X_\mathtt{A}})$ is Hilbert–Schmidt.

Again Assumption 2 is mild. The first part is easily satisfied, because $\ker(\Sigma_{X_\mathtt{A}X_\mathtt{A}}) = \{0\}$ if and only if $\{\phi \in \mathcal{H}_{X_\mathtt{A}} : \mathrm{var}\{\phi(X_\mathtt{A})\} = 0\} = 0$. This means that we can reset $\mathcal{H}_{X_\mathtt{A}}$ to the closure of $\Sigma_{X_\mathtt{A}X_\mathtt{A}}$, which can be achieved by excluding all the constant functions in $\mathcal{H}_{X_\mathtt{A}}$. The second and third parts are mild too, and can be seen as a smoothness condition on the relation between X_i and $X_\mathtt{A}$. Note that the invertibility of $\Sigma_{X_\mathtt{A}X_\mathtt{A}}$ in $R_{X_i|X_\mathtt{A}}$ is ensured by Assumption 2, and its inverse is defined as the Moore–Penrose inverse (Li, 2018).

Definition 3 For each $i \in \mathtt{V}$ and any subvector $\mathtt{A}$ of $\mathtt{V}\setminus\{i\}$, we call $R_{X_i|X_\mathtt{A}}$ the functional regression operator from X_i to $X_\mathtt{A}$.

The form of $R_{X_i|X_\mathtt{A}}$ resembles that of the regression coefficient in the classical linear regression, and hence is called a functional regression operator. Lee et al. (2016b) proposed a similar notion of regression operator. However, it was based on additive conditional independence, a new and different

three-way statistical relation to characterize separations among the nodes in a graph (Li et al., 2014). By contrast, our functional SEM with the functional regression operator satisfies the global Markov condition. It is still built on conditional independence, which is the more classical relation for node separations. We choose to work with conditional independence here, as it is more directly associated with the conditional distribution. That being said, we speculate it is possible to develop a version of functional SEM based on additive conditional independence. Nevertheless, it would require a different set of modelling and estimation techniques, and we leave it as future research.

Next, we derive an important result that shows the equivalence between the linear operator $B^0_{i,j}$ in Definition 1, and the regression operator $R_{X_i|X_{\mathtt{A}}}$.

Proposition 1 *Suppose Assumptions 1 and 2 hold. Then* $R_{X_i|X_{\{\mathtt{pa}(i;\mathtt{G}),\mathtt{A}\}}} = (B^0_i, 0)$, *for any* $\mathtt{A} \subseteq \mathtt{nd}_i$, *where* $\mathtt{nd}_i$ *denotes the collection of non-descendant nodes of node* i, *and* $B^0_i = \{B^0_{i,j} : j \in \mathtt{pa}(i; \mathtt{G}^0)\}$, *for all* $i \in \mathtt{V}$.

To gain some further insight of this proposition, we see that $R_{X_i|X_{\{\mathtt{pa}(i;\mathtt{G}^0),\mathtt{A}\}}} = \{R_{X_i|X_{\mathtt{pa}(i;\mathtt{G}^0)}}, 0\}$, for all $\mathtt{A} \subseteq \mathtt{nd}_i$. That is, when regressing the ith function on its parents and non-descendant functions, the regression operator only depends on its parent functions, and the ith function is independent with all the non-descendant functions given its parent functions.

3.2 | Score function for order determination

To determine the directional order, we next introduce a score function that minimizes the log-determinant of the error covariance operator. Let $\sigma = (\sigma_1, \ldots, \sigma_p)$ denote a permutation of $(1, \ldots, p)$ which indicates the directional order $\sigma_1 \to \sigma_2 \to \cdots \to \sigma_p$. For any σ, we let $X_\sigma(t) = \{X_{\sigma_1}(t), \ldots, X_{\sigma_p}(t)\}^\mathsf{T}$, and $B_{\sigma_i\sigma_j}$, $\Gamma_{\varepsilon_{\sigma_i}}$, and $\mathtt{G}_\sigma = (\mathtt{V}, \mathtt{E}_\sigma)$ denote the corresponding regression operator, error covariance operator, and the associated DAG, where $\mathtt{E}_\sigma = \{(\sigma_i, \sigma_j) : i = j+1, \ldots, p,\ j = 1, \ldots, p-1\}$. We call $\mathtt{G}_\sigma$ a super-DAG of the true DAG $\mathtt{G}^0$, if the true edge set $\mathtt{E}^0$ of $\mathtt{G}^0$ is a subset of its edge set $\mathtt{E}_\sigma$. This means, $\mathtt{pa}(i, \mathtt{G}^0) \subseteq \sigma_{-i}$ and $\sigma_{-i} \setminus \mathtt{pa}(i, \mathtt{G}^0) \subseteq \mathtt{nd}(i, \mathtt{G}^0)$, where $\sigma_{-i} = (\sigma_1, \ldots, \sigma_{i-1})$, $i = 2, \ldots, p$. Let $\mathfrak{S}^0$ denote the collection of all σ such that $\mathtt{E}^0 \subseteq \mathtt{E}_\sigma$. Then any member in $\mathfrak{S}^0$ is considered a true directional order, as it correctly reflects the directional relationships among $X_1, \ldots, X_p$. In the order determination step, our goal is to identify a member σ in $\mathfrak{S}^0$.

Define the conditional covariance operator $\Gamma_{X_iX_i|X_{\mathtt{S}}} : \mathcal{H}_{X_i} \to \mathcal{H}_{X_i}$ as,

$$\Gamma_{X_iX_i|X_{\mathtt{A}}} = E\left\{X_i - \sum_{j\in\mathtt{A}} (R_{X_i|X_{\mathtt{A}}})^*_j \kappa(\cdot, X_j)\right\} \otimes \left\{X_i - \sum_{j\in\mathtt{A}} (R_{X_i|X_{\mathtt{A}}})^*_j \kappa(\cdot, X_j)\right\},$$

for $i \in \mathtt{V}$ and a subset $\mathtt{A}$ of $\mathtt{V} \setminus \{i\}$, where $(\cdot)^*$ denotes the adjoint of the designated operator. Intuitively, $\Gamma_{X_iX_i|X_{\mathtt{A}}}$ can be seen as the variation that cannot be explained by the regression relationship between a functional node X_i and a set of functional nodes $X_{\mathtt{A}}$. It resembles the classical conditional covariance, in the sense that when $\mathcal{H}_{X_i} = \mathbb{R}$, $\Gamma_{X_iX_i|X_{\mathtt{A}}}$ computes the quantity $\mathrm{cov}(X_i|X_{\mathtt{A}})$. For any permutation σ, we can similarly define $\Gamma_{X_{\sigma_i}X_{\sigma_i}|X_{\sigma_{-i}}}$ as a special case of $\Gamma_{X_iX_i|X_{\mathtt{A}}}$. The next proposition first derives a form of $\Gamma_{X_iX_i|X_{\mathtt{A}}}$ as a function of covariance operators, which is useful for deriving the score function. Moreover, it shows that $\Gamma_{X_{\sigma_i}X_{\sigma_i}|X_{\sigma_{-i}}}$ is identical, and equals $\Gamma^0_{\varepsilon_{\sigma_i}}$, the true error covariance operator, for all $\sigma \in \mathfrak{S}^0$.

Proposition 2 *Suppose Assumptions 1 and 2 hold. Then,*

$$\Gamma_{X_i X_i | X_{\mathtt{A}}} = \Gamma_{X_i X_i} - \Lambda^*_{X_{\mathtt{A}} X_i} \Sigma^{-1}_{X_{\mathtt{A}} X_{\mathtt{A}}} \Lambda_{X_{\mathtt{A}} X_i},$$

for $i \in \mathtt{V}$ and $\mathtt{A} \subseteq \mathtt{V} \setminus \{i\}$. Moreover, for any $\sigma, \sigma' \in \mathfrak{S}^0$, $\Gamma_{X_{\sigma_i} X_{\sigma_i} | X_{\sigma_{-i}}} = \Gamma_{X_{\sigma'_j} X_{\sigma'_j} | X_{\sigma'_{-j}}} = \Gamma^0_{\varepsilon_{\sigma_i}} = \Gamma^0_{\varepsilon_{\sigma'_j}} = \Gamma_{X_\ell X_\ell} - \Lambda^*_{X_{\mathrm{pa}(\ell;\mathbb{G}^0)} X_\ell} \sum^{-1}_{X_{\mathrm{pa}(\ell;\mathbb{G}^0)} X_{\mathrm{pa}(\ell;\mathbb{G}^0)}} \Lambda_{X_{\mathrm{pa}(\ell;\mathbb{G}^0)} X_\ell}$, if $\sigma_i = \sigma'_j = \ell$.

The interpretation of $\Gamma_{X_{\sigma_i} X_{\sigma_i} | X_{\sigma_{-i}}}$ suggests a natural way to determine the directional order, that is, by minimizing the magnitude of $\Gamma_{X_{\sigma_i} X_{\sigma_i} | X_{\sigma_{-i}}}$ across different orders of the nodes. Furthermore, Proposition 2 justifies why $\Gamma_{X_{\sigma_i} X_{\sigma_j} | X_{\sigma_{-i}}}$ can be used to build a score function for order determination. Accordingly, we define the score function of a permutation σ as

$$S(\sigma; q) = \sum_{i=1}^{p} \sum_{m=1}^{q} \log \left\{ \lambda_m \left(\Gamma_{X_{\sigma_i} X_{\sigma_i} | X_{\sigma_{-i}}} \right) \right\}, \quad \text{for each } q \in \mathbb{N}, \tag{3}$$

where $\lambda_m(\cdot)$ is the mth largest eigenvalue of the designated operator, for $m = 1, ..., q$. Note that in Equation (1), we only take the sum of the leading q eigenvalues of $\Gamma_{X_{\sigma_i} X_{\sigma_i} | X_{\sigma_{-i}}}$ because, by Assumption 1, $\Gamma_{X_{\sigma_i} X_{\sigma_i} | X_{\sigma_{-i}}}$ is trace class, which implies $\lambda_m(\Gamma_{X_{\sigma_i} X_{\sigma_i} | X_{\sigma_{-i}}}) \to 0$ and its logarithm tends to $-\infty$ as $m \to \infty$. In Section 4, we discuss how to specify q.

To estimate the directional ordering, we seek the order σ that minimizes the score function $S(\sigma; q)$. Next, we study the conditions under which the true order minimizes $S(\sigma; q)$. We first define a quantity that measures the level of the separation between the model with the true directional order and the one with a wrong order. For $q \in \mathbb{N}$, define

$$\theta_q = p^{-1} \min_{\sigma \notin \mathfrak{S}^0} \left\{ S(\sigma; q) - S(\sigma^0; q) \right\}, \tag{4}$$

where $S(\sigma^0; q)$ is the score function evaluated under the true order $\sigma^0 \in \mathfrak{S}^0$. If $\theta_q > 0$, the score based on the wrong order is strictly greater than that based on the true order, which in turn implies one can estimate the true order by minimizing the score $S(\sigma; q)$. The next theorem provides the sufficient conditions under which θ_q tends to a strictly positive number as $q \to \infty$. For generic Hilbert spaces $\mathcal{H}_1, \ldots, \mathcal{H}_J$, let $\bigoplus_{j=1}^J \mathcal{H}_j$ denote their direct sum.

Theorem 1 *Suppose $X(t) = \{X_1(t), \ldots, X_p(t)\}^\mathsf{T}$ follows a functional structural equation model with respect to $\mathbb{G}^0$, the function ρ in Equation (2) is three times differentiable, and $\mathcal{H}_{X_i}$ is spanned by an orthonormal basis $\{f_i^m\}_{m \in \mathbb{N}}$ for all $i \in \mathtt{V}$. Moreover, suppose there exist a subset $\mathbb{N}_i = \{m_{i,1}, \ldots, m_{i,q_i}\} \subset \mathbb{N}$, and a subspace $\Omega_{X_i} = \mathrm{span}\{f_i^m : m \in \mathbb{N}_i\} \subseteq \mathcal{H}_{X_i}$ such that, for each permutation σ,*

1. $R_{X_{\sigma_i} | X_{\sigma_{-i}}} f^m_{\sigma i}(\cdot) = \bigoplus_{j=1}^{i-1} [b^m_{\sigma_i, \sigma_j}(P_{\Omega_{X_{\sigma_j}}}(\cdot))]$, *for $m \in \mathbb{N}_i$ and $j = 1, ..., i-1$, where $b^m_{\sigma_i, \sigma_j}$ is a nonlinear function in $\mathscr{H}_{U_{\sigma_j}}$, $\mathscr{H}_{U_i}$ is the RKHS induced by the positive kernel $\kappa_{U_i}(\cdot, \cdot) = \kappa_i(\cdot, \cdot) : \Omega_{X_i} \times \Omega_{X_i} \to \mathbb{R}$, and $P_{\Omega_{X_i}}$ is the projection onto Ω_{X_i};*
2. $R_{X_{\sigma_i} | X_{\sigma_{-i}}} f = 0$, *for any $f \in \Omega^\perp_{X_{\sigma_i}}$, where $\Omega^\perp_{X_i}$ is the orthogonal complement of Ω_{X_i};*
3. $\mathrm{cov}(r^m_{\sigma_i | \sigma_{-i}}, r^\ell_{\sigma_i | \sigma_{-i}}) = 0$, *for any $(m, \ell) \in \mathbb{N}_i \times \mathbb{N}_i$ with $m \neq \ell$, where $r^m_{\sigma_i | \sigma_{-i}} = \langle f^m_{\sigma i}, X_{\sigma_i} \rangle_{\mathcal{H}_{X_{\sigma_i}}} - R_{X_{\sigma_i} | X_{\sigma_{-i}}} f^m_{\sigma i}(X_{\sigma_i}), m \in \mathbb{N}_i$.*

Then, there exists $\theta^* > 0$, such that $\theta_q \to \theta^*$, as $\mathrm{q} \to \infty$.

Suppose $U_i = P_{\Omega_{X_i}} X_i$ is the projection of X_i onto Ω_{X_i}. Conditions (a) and (b) imply that, the mapping of the regression operator $R_{X_{\sigma_i}|X_{\sigma_{-i}}} f$ is a function in $\oplus_{j=1}^{i-1} \mathscr{H}_{U_{\sigma_j}}$, if f is in $\Omega_{X_{\sigma_i}}$, and 0, if f is in $\Omega^{\perp}_{X_{\sigma i}}$. This means, the relation between the node X_{σ_i} and its parent nodes $(X_{\sigma_1}, \ldots, X_{\sigma_{i-1}})^{\mathsf{T}}$, depends solely on the relation between their projections U_{σ_i} and $(U_{\sigma_1}, \ldots, U_{\sigma_{i-1}})^{\mathsf{T}}$. This is the key to establish the identifiability in the infinite-dimensional space: we allow the dimension of $\mathcal{H}_{X_i}$ to be infinite, but the dimension of 'effective space', that is, Ω_{X_i}, is finite. Moreover, condition (c) ensures there is a connection between the KL divergence and the score function. These three conditions are satisfied, for example, if there exists $b^m_{i,j} \in \mathscr{H}_{U_j}$, such that

$$\langle X_i, f_i^m \rangle_{\mathcal{H}_{X_i}} = \begin{cases} \sum_{j \in \mathrm{pa}(i;\mathbb{G}^0)} b^{0,m}_{i,j}(U_j) + \gamma_i^m \varepsilon_i^m, \text{for } m \in \mathbb{N}_i, \\ \gamma_i^m \varepsilon_i^m, \qquad\qquad \text{for } m \notin \mathbb{N}_i, \end{cases}$$

where γ_i^m's are positive constants, and ε_i^m's are i.i.d. standard normal variables.

Theorem 1 also provides a justification to the estimating procedure we develop in Section 4. Suppose $\{g_i^m\}_{m=1}^d$ are the eigenfunctions associated with the leading d eigenvalues of $\Gamma_{X_i X_i}$, which we will use later to estimate the truncated covariance operators. By Theorem 1, there would be no loss of information as long as the effective space Ω_{X_i} is contained in the space spanned by $\{g_i^m\}_{m=1}^d$. Empirically, our simulations suggest that we usually only need the top $d = 2$ or 3 eigenfunctions to obtain satisfactory estimation results.

We also note that, Bühlmann et al. (2014, Lemma 3) obtained a similar result as Theorem 1 under the setting of random variables, that is, when $\mathcal{H}_{X_i} = \mathbb{R}$ or $q = 1$. In contrast, Theorem 1 extends their result to the setting of random functions, where we consider a more complex setting, and also allow q to diverge. Note that, for nested kernels, κ_i is three times Fréchet differentiable when the ρ function is properly selected in Equation (2), for example, the radial basis function. Correspondingly, $R_{X_{\sigma_i}|X_{\sigma_{-i}}} f$ is also three times differentiable. When this condition is not satisfied, or when $R_{X_{\sigma_i}|X_{\sigma_{-i}}} f$ is actually linear, we expect the separation constant θ_q to become zero, indicating that the identifiability may no longer hold. Nevertheless, we can still use the score function $S(\sigma; q)$ to estimate the equivalence class, that is, the collection of all $(B^0_{i,j}, \Gamma^0_{\varepsilon_i})$ that lead to the same distribution under Definition 1 (see, e.g. Peters et al., 2014). To avoid digression, we follow a similar strategy as Bühlmann et al. (2014), Yuan et al. (2018) and focus on the identifiable model in this paper.

3.3 | Functional additive regression

For the classical SEM for random variables, once the directional order is determined, the estimation of SEM can be carried out through variable selection in sparse regression (Shojaie & Michailidis, 2010). The same holds true for our function-based SEM. Next, we introduce the functional additive regression model where both the response and the predictor are functions. We first introduce a least-squares type objective function at the operator level. We then show that the minimizer of this objective function is actually the regression operator in Definition 3, and hence can be used to estimate the true regression operator $B^0_{i,j}$ in our functional SEM model.

For all $i \in \mathtt{V}$ and $\mathtt{A} \subseteq \mathtt{V} \setminus \{i\}$, we consider the objective function of $B_i : \mathcal{H}_{X_i} \to \mathscr{H}_{X_{\mathtt{A}}}$:

$$L_i(B_i) = E\left(\operatorname{tr}\left[\left\{X_i - \sum_{j\in\mathtt{A}} B_{i,j}^{*}\kappa_j(\cdot, X_j)\right] \otimes \left\{X_i - \sum_{j\in\mathtt{A}} B_{i,j}^{*}\kappa_j(\cdot, X_j)\right\}\right]\right),$$

where tr(·) denotes the trace. The next proposition derives an explicit form of $L_i(B_i)$ as a function of B_i and the covariance operators $\sum_{X_iX_j}$ and $\Lambda_{X_iX_j}$.

Proposition 3 *For any $i \in \mathtt{V}$ and $\mathtt{A} \subseteq \mathtt{V}\setminus\{i\}$, there exists a constant $c > 0$ independent of B_i, such that*

$$L_i(B_i) = \left\langle \Sigma_{X_{\mathtt{A}}X_{\mathtt{A}}} B_i, B_i \right\rangle_{\mathrm{HS}} - 2\left\langle \Lambda_{X_{\mathtt{A}}X_i}, B_i \right\rangle_{\mathrm{HS}} + c. \tag{5}$$

The next proposition shows the minimizer of $L_i(B_i)$ is the functional regression operator.

Proposition 4 *If Assumptions 1 and 2 hold, then we have* $R_{X_i|X_{\mathtt{A}}} = \arg\min\{L_i(B_i): B_i \in \mathscr{B}_2(\mathcal{H}_{X_i}, \mathscr{H}_{X_{\mathtt{A}}})\}$.

When $\mathcal{H}_{X_i} = \mathbb{R}^q$ and $\mathscr{H}_{X_{\mathtt{A}}} = \mathbb{R}^p$, Proposition 4 reduces to the classical multivariate regression. Therefore, it can be viewed as a generalization of the linear multivariate regression to both functional and nonlinear settings. Moreover, combined with Proposition 1, Proposition 4 implies that, minimizing the objective function under the proper directional order, can recover the underlying directional relations of our functional SEM. This opens the door to a two-step penalized estimation procedure, which we discuss next.

4 | ESTIMATION

In this section, we first present the estimation at the operator level, followed by two strategies to accelerate the computation and help scale up our method to large graphs. We then develop the estimation under a coordinate system for discretely observed functional data. Finally, we provide a step-by-step algorithm to summarize the estimation procedure.

4.1 | Estimation at the operator level

We begin with the estimation at the operator level. Let $\{X_1^m(t), \ldots, X_p^m(t)\}^{\mathsf{T}}$, $m = 1, \ldots, n$, be i.i.d. samples of the multivariate function $X(t) = \{X_1(t), \ldots, X_p(t)\}^{\mathsf{T}}$. For a sample $\{\omega^m\}_{m=1}^n$ from $\omega \in \mathcal{H}$, let $E_n(\omega) = n^{-1}\sum_{m=1}^n \omega^m$ be the mean operator. We first estimate the operators in Definition 2 by

$$\widehat{\Gamma}_{X_iX_i} = E_n X_i \otimes X_i, \quad \widehat{\Lambda}_{X_iX_j} = E_n \kappa_i(\cdot, X_i) \otimes X_j, \quad \widehat{\Sigma}_{X_iX_j} = E_n \kappa_i(\cdot, X_i) \otimes \kappa_j(\cdot, X_j), \tag{6}$$

for $(i, j) \in \mathtt{V} \times \mathtt{V}$. for $(i, j) \in \mathtt{V} \times \mathtt{V}$. For any $\mathtt{A}, \mathtt{B} \subseteq \mathtt{V}$, we then stack the sample operators and form the matrices of operators $\widehat{\Lambda}_{X_{\mathtt{A}}X_j}$ and $\widehat{\Sigma}_{X_{\mathtt{A}}X_{\mathtt{B}}}$, such that $[\widehat{\Lambda}_{X_{\mathtt{A}}X_j}]_i = \widehat{\Lambda}_{X_iX_j}$ and $[\widehat{\Sigma}_{X_{\mathtt{A}}X_{\mathtt{B}}}]_{ij} = \widehat{\Sigma}_{X_iX_j}$.

For a permutation σ, using Proposition 2 we can estimate the conditional covariance operator $\Gamma_{X_{\sigma_i}X_{\sigma_i}|X_{\sigma_{-i}}}$ by $\widehat{\Gamma}_{X_{\sigma_i}X_{\sigma_i}|X_{\sigma_{-i}}} = \widehat{\Gamma}_{X_{\sigma_i}X_{\sigma_i}} - \widehat{\Lambda}^{*}_{X_{\sigma_{-i}}X_{\sigma_i}}(\widehat{\Sigma}_{X_{\sigma_{-i}}X_{\sigma_{-i}}} + \epsilon I)^{-1}\widehat{\Lambda}_{X_{\sigma_{-i}}X_{\sigma_i}}$, where ε is a prespecified ridge parameter to ensure the invertibility of $\widehat{\Sigma}_{X_{\sigma_{-i}}X_{\sigma_{-i}}}$, and I is the identify mapping. We then estimate the directional order by,

$$\widehat{\sigma} = \arg\min_{\sigma} \widehat{S}(\sigma; q), \quad \text{where } \widehat{S}(\sigma; q) = \sum_{i=1}^{p} \sum_{m=1}^{q} \log\left\{ \lambda_m \left(\widehat{\Gamma}_{X_{\sigma_i} X_{\sigma_i} | X_{\sigma_{-i}}} \right) \right\}.$$

Next, letting $i = \sigma_i$ and $\mathtt{A} = \sigma_{-i}$ in Equation (5), and plugging the estimated operators and the directional order, we obtain the sample version of the objective function, for $i \in \mathtt{V}$,

$$\tilde{L}_i(B_{\widehat{\sigma}_i}) = \left\langle \widehat{\Sigma}_{X_{\widehat{\sigma}_{-i}} X_{\widehat{\sigma}_{-i}}} B_{\widehat{\sigma}_i}, B_{\widehat{\sigma}_i} \right\rangle_{\mathrm{HS}} - 2 \left\langle \widehat{\Lambda}_{X_{\widehat{\sigma}_{-i}} X_{\widehat{\sigma}_i}}, B_{\widehat{\sigma}_i} \right\rangle_{\mathrm{HS}}.$$

We further introduce an L_1-type penalty to encourage the sparsity of $B_{\widehat{\sigma}_i}$,

$$\widehat{L}_i(B_{\widehat{\sigma}_i}) = \tilde{L}_i(B_{\widehat{\sigma}_i}) + \lambda \left\{ \sum_{j=1}^{i-1} \|B_{\widehat{\sigma}_i, \widehat{\sigma}_j}\|_{\mathrm{HS}} \right\}^2,$$

where λ is a sparsity tuning parameter. The above problem can be solved by some standard convex optimization, for example, the iterative shrinkage thresholding algorithm (Beck & Teboulle, 2009). Denote $\widehat{B}_{\widehat{\sigma}_i} = \arg\min \{\widehat{L}_i(B_{\widehat{\sigma}_i}) : B_{\widehat{\sigma}_i} \in \mathscr{B}_2(\mathcal{H}_{X_{\widehat{\sigma}_i}}, \mathscr{H}_{X_{\widehat{\sigma}_{-i}}})\}$, for $i \in \mathtt{V}$.

Finally, let $\widehat{B}_{\widehat{\sigma}_i, \widehat{\sigma}_j} = (\widehat{B}_{\widehat{\sigma}_i})_{\widehat{\sigma}_j}$ and we can estimate the edges of the corresponding DAG of our functional SEM by

$$\widehat{\mathrm{E}}_{\widehat{\sigma}} = \left\{ (\widehat{\sigma}_i, \widehat{\sigma}_j) : \|\widehat{B}_{\widehat{\sigma}_i, \widehat{\sigma}_j}\|_{\mathrm{HS}} \neq 0, \ \ i = j+1, \ldots, p, j = 1, \ldots, p-1 \right\}.$$

4.2 | Computational acceleration

Still at the operator level, we discuss two strategies to accelerate the computation of our estimation procedure, which is particularly useful when the number of functions p is large.

The first strategy is to consider a reduced RKHS spanned by only the leading eigenfunctions, a technique commonly used in functional data analysis, for example, the functional principal component analysis (fPCA; Yao et al., 2005). We recognize that the sample covariance operators $\widehat{\Gamma}_{X_i X_i}$, $\widehat{\Lambda}_{X_i X_j}$ and $\widehat{\Sigma}_{X_i X_j}$ in Equation (6) are the key building blocks of our estimation, and thus we seek to replace them by their truncated counterparts under the reduced RKHS. Specifically, let $\{(a_i^m, f_i^m)\}_{m \in \mathbb{N}}$ and $\{(\mu_i^m, \phi_i^m)\}_{m \in \mathbb{N}}$ denote the eigenvalues and eigenfunctions of $\Gamma_{X_i X_i}$ and $\Sigma_{X_i X_i}$ respectively. By the KL expansion (Bosq, 2000), the random elements $X_i \in \mathcal{H}_{X_i}$ and $\kappa_i(\cdot, X_i) \in \mathscr{H}_{X_i}$ can be almost surly represented as $X_i = \sum_{m \in \mathbb{N}} x_i^m f_i^m$ and $\kappa_i(\cdot, X_i) = \sum_{m \in \mathbb{N}} \alpha_i^m \phi_i^m$, where $x_i^m = \langle X_i, f_i^m \rangle_{\mathcal{H}_{X_i}}$, and $\alpha_i^m = \langle \kappa_i(\cdot, X_i), \phi_i^m \rangle_{\mathscr{H}_{X_i}}$ are the KL coefficients. Then, the un-truncated covariance operators, for $(i, j) \in \mathtt{V} \times \mathtt{V}$, are,

$$\Gamma_{X_i X_i} = \sum_{m \in \mathbb{N}} a_i^m \left(f_i^m \otimes f_i^m \right), \quad \Lambda_{X_i X_j} = \sum_{m \in \mathbb{N}} \sum_{\ell \in \mathbb{N}} E\left(\alpha_i^m x_j^\ell \right) \left(\phi_i^m \otimes f_j^\ell \right),$$
$$\Sigma_{X_i X_j} = \sum_{m \in \mathbb{N}} \sum_{\ell \in \mathbb{N}} E\left(\alpha_i^m \alpha_j^\ell \right) \left(\phi_i^m \otimes \phi_j^\ell \right).$$

Given the data, we first obtain $\widehat{\Gamma}_{X_i X_i}$, $\widehat{\Lambda}_{X_i X_j}$ and $\widehat{\Sigma}_{X_i X_j}$ using Equation (6). We then obtain the eigenvalue–eigenfunction pairs $\{(\widehat{a}_i^m, \widehat{f}_i^m)\}_{m=1}^n$ and $\{(\widehat{\mu}_i^m, \widehat{\phi}_i^m)\}_{m=1}^n$ from the decomposition of $\widehat{\Gamma}_{X_i X_i}$ and $\widehat{\Sigma}_{X_i X_i}$ respectively. We also compute the sample KL coefficients $\widehat{x}_i^m = \langle X_i, \widehat{f}_i^m \rangle_{\mathcal{H}_{X_i}}$, and

$\widehat{\alpha}_j^{\ell} = \langle \kappa_j(\cdot, X_j), \widehat{\phi}_j^{\ell} \rangle_{\mathscr{H}_{X_i}}$. Then, for a pre-specified truncation number d, the truncated sample covariance operators $\widehat{\Gamma}_{X_iX_i}$, $\widehat{\Lambda}_{X_iX_j}$ and $\widehat{\Sigma}_{X_iX_j}$ in Equation Equation (6) are given by,

$$\begin{aligned} \widehat{\Gamma}^d_{X_iX_i} &= \sum_{m=1}^{d} \widehat{a}_i^m \left(\widehat{f}_i^m \otimes \widehat{f}_i^m \right), \quad \widehat{\Lambda}^d_{X_iX_j} = \sum_{m,\ell=1}^{d} E_n \left(\widehat{\alpha}_i^m \widehat{x}_j^{\ell} \right) \left(\widehat{\phi}_i^m \otimes \widehat{f}_j^{\ell} \right), \\ \widehat{\Sigma}^d_{X_iX_j} &= \sum_{m,\ell=1}^{d} E_n \left(\widehat{\alpha}_i^m \widehat{\alpha}_j^{\ell} \right) \left(\widehat{\phi}_i^m \otimes \widehat{\phi}_j^{\ell} \right). \end{aligned} \tag{7}$$

The second strategy is to employ a pre-neighbourhood selection step to narrow down the neighbourhood of each node. This avoids evaluations over all possible permutations, whose computation can be enormous when p is large. A similar technique was also used in Bühlmann et al. (2014) and Loh and Bühlmann (2014). Furthermore, we combine this strategy with the first strategy of using a reduced RKHS.

Specifically, for $i \in \mathtt{V}$, we define the neighbourhood $\mathtt{N}_i$ of node i as the smallest subset in $\mathtt{V} \setminus \{i\}$ satisfying $R_{X_i|X_{-i}} = (R_{X_i|X_{\mathtt{N}_i}}, 0)$; that is, for any subset $\mathtt{N}_i' \subseteq \mathtt{V} \setminus \{i\}$ satisfying $R_{X_i|X_{-i}} = (R_{X_i|X_{\mathtt{N}_i'}}, 0)$, we have $\mathtt{N}_i \subseteq \mathtt{N}_i'$. With a slight abuse of notation, we let $(R_{X_i|X_{\mathtt{N}_i}}, 0)$ denote the operator in $\mathscr{B}_2(\mathcal{H}_{X_i}, \mathscr{H}_{X_{-i}})$ such that $\{(R_{X_i|X_{\mathtt{N}_i}}, 0)_j : j \in \mathtt{N}_i\} = R_{X_i|X_{\mathtt{N}_i}}$, and $(R_{X_i|X_{\mathtt{N}_i}}, 0)_j = 0$ for all $j \in \mathtt{V} \setminus \{\mathtt{N}_i, i\}$. We then estimate $\mathtt{N}_i$ by first carrying out the minimization,

$$\check{L}_i(A_i) = \left\langle \widehat{\Sigma}^d_{X_{-i}X_{-i}} A_i, A_i \right\rangle_{\mathrm{HS}} - 2 \left\langle \widehat{\Lambda}^d_{X_{-i}X_i}, A_i \right\rangle_{\mathrm{HS}} + \check{\lambda} \left\{ \sum_{j \neq i} \left\| (A_i)_j \right\|_{\mathrm{HS}} \right\}^2, \tag{8}$$

for $A_i \in \mathscr{B}_2(\mathcal{H}_{X_i}, \mathscr{H}_{X_{-i}})$. Letting $\widehat{A}_i = \arg\min\{\check{L}_i(A_i) : A_i \in \mathscr{B}_2(\mathcal{H}_{X_i}, \mathscr{H}_{X_{-i}})\}$, we then use $\widehat{\mathtt{N}}_i = \{j \in \mathtt{V} \setminus \{i\} : \|(\widehat{A}_i)_j\|_{\mathrm{HS}} \neq 0\}$ to estimate the neighbourhood $\mathtt{N}_i$.

Letting $\widehat{\mathbb{N}} = \{\widehat{\mathtt{N}}_i : i \in \mathtt{V}\}$, we next define a refined collection of permutations $\mathfrak{S}_{\widehat{\mathbb{N}}}$, which satisfies that, if $\sigma, \sigma' \in \mathfrak{S}_{\widehat{\mathbb{N}}}$, and $\sigma \neq \sigma'$, then $(\sigma_{-i} \cap \widehat{\mathtt{N}}_{\sigma_i}) \neq (\sigma'_{-i} \cap \widehat{\mathtt{N}}_{\sigma'_i})$ for all $i \in \mathtt{V}$. We then carry out a refined search of the directional order as,

$$\widehat{\sigma}(\widehat{\mathbb{N}}) = \arg\min\{\widehat{S}^d(\sigma; q) : \sigma \in \mathfrak{S}_{\widehat{\mathbb{N}}}\}, \quad \text{where } \widehat{S}^d(\sigma; q) = \sum_{i=1}^{p} \sum_{m=1}^{q} \log \left\{ \lambda_m \left(\widehat{\Gamma}_{X_{\sigma_i}X_{\sigma_i}|X_{\mathtt{R}_{\sigma,i}}} \right) \right\}, \tag{9}$$

and $\widehat{\Gamma}_{X_{\sigma_i}X_{\sigma_i}|X_{\mathtt{R}_{\sigma,i}}} = \widehat{\Gamma}_{X_{\sigma_i}X_{\sigma_i}} - \widehat{\Lambda}^*_{X_{\mathtt{R}_{\sigma,i}}X_{\sigma_i}} (\widehat{\Sigma}_{X_{\mathtt{R}_{\sigma,i}}X_{\mathtt{R}_{\sigma,i}}} + \epsilon I)^{-1} \widehat{\Lambda}_{X_{\mathtt{R}_{\sigma,i}}X_{\sigma_i}}$ with $\mathtt{R}_{\sigma,i} = \sigma_{-i} \cap \widehat{\mathtt{N}}_{\sigma_i}$ being the refined neighbourhood.

Note that $\mathfrak{S}_{\widehat{\mathbb{N}}}$ may not be unique; this means, there exists $\mathfrak{S}'_{\widehat{\mathbb{N}}} \neq \mathfrak{S}_{\widehat{\mathbb{N}}}$, such that $\mathfrak{S}'_{\widehat{\mathbb{N}}}$ satisfies the same condition as $\mathfrak{S}_{\widehat{\mathbb{N}}}$. Nevertheless, $\mathfrak{S}_{\widehat{\mathbb{N}}}$ and $\mathfrak{S}'_{\widehat{\mathbb{N}}}$ are equivalent in the sense that, for every $\sigma \in \mathfrak{S}_{\widehat{\mathbb{N}}}$, there exists $\sigma' \in \mathfrak{S}'_{\widehat{\mathbb{N}}}$ such that $\widehat{S}^d(\sigma; q) = \widehat{S}^d(\sigma'; q)$ for every $q \in \mathbb{N}$; see also Bühlmann et al. (2014). Consequently, this does not affect our search for the directional order. Moreover, the cardinality of $\mathfrak{S}_{\widehat{\mathbb{N}}}$ can be substantially smaller than that of all possible permutations, which leads to a considerably faster computation.

We continue with the refined functional additive regression, using the pre-neighbourhood selection $\widehat{\mathbb{N}}$ from Equation (8), and the estimated order $\widehat{\sigma}(\widehat{\mathbb{N}})$ from Equation (9). Define $\mathtt{r}_i = \{\widehat{\sigma}(\widehat{\mathbb{N}})\}_{-j} \cap \widehat{\mathtt{N}}_i$, with $\{\widehat{\sigma}(\widehat{\mathbb{N}})\}_j = i$, as the refined parents of the ith node. We solve,

$$\widehat{L}_i(B_i) = \left\langle \widehat{\Sigma}^d_{X_{\mathtt{r}_i} X_{\mathtt{r}_i}} B_i, B_i \right\rangle_{\mathrm{HS}} - 2\left\langle \widehat{\Lambda}^d_{X_{\mathtt{r}_i} X_i}, B_i \right\rangle_{\mathrm{HS}} + \lambda \left\{ \sum_{j=1}^{i-1} \left\| (B_i)_j \right\|_{\mathrm{HS}} \right\}^2. \tag{10}$$

Denote $\widehat{B}_i = \arg\min_{B_i} \widehat{L}_i(B_i)$, $i \in \mathtt{V}$, and $\widehat{B}_{i,j} = (\widehat{B}_i)_j$. We estimate the edges of the true DAG $\mathtt{E}^0$ by $\widehat{\mathtt{E}} = \{(i, j) \in \mathtt{V} \times \mathtt{V} : \|\widehat{B}_{i,j}\|_{\mathrm{HS}} \neq 0,\ j \in \mathtt{r}_i,\ i \in \mathtt{V}\}$.

4.3 | Estimation under a coordinate system

Next, we discuss the estimation under a chosen coordinate system. This is because the random functions $X^m(t)$ can only be observed at finite set of time points, $t_{m1}, \ldots, t_{mn_m}$, for $m = 1, ..., n$. To enable the computation, we need to approximate the random functions using the partially observed data $\{X^m(t_{m\ell})\}_{\ell=1}^{n_m}$. For simplicity, we develop our estimation procedure under the balanced design; that is, the set $\{t_{m1}, \ldots, t_{mn_m}\}$ is the same for all $m = 1, ..., n$, and is denoted as $\tau = \{t_1, \ldots, t_r\}$, where r is the number of distinct time points. Nevertheless our method readily applicable to the unbalanced design too. Moreover, we assume the sample path of $X_i^m(t)$, for all $i \in \mathtt{V}$, lies on a reproducing kernel Hilbert space spanned by a finite number of some functional bases. Such a technique is common in functional data analysis; see, for example, Li and Song (2017).

The computation of the coordinates consists of four steps. We first develop the coordinate system for the first-layer RKHS $\mathcal{H}_{X_i}$, then obtain the coordinate representation of X_i^m. We next develop the coordinate system for the second-layer nested RKHS $\mathscr{H}_{X_i}$, then derive the coordinates of the truncated operators $\widehat{\Gamma}^d_{X_i X_i}$, $\widehat{\Lambda}^d_{X_i X_j}$, and $\widehat{\Sigma}^d_{X_i X_j}$. After these four steps, we plug the coordinates into Equation (8) to (10) to estimate the directional relations.

First, we develop the coordinate system for the first-layer RKHS $\mathcal{H}_{X_i}$. For the first-layer kernel κ_T, let $(K_T)^r_{s,t=1} = \kappa_T(t_s, t_t)$ denote the corresponding $r \times r$ Gram kernel matrix, and $\mathcal{H}^r$ denote the RKHS spanned by $\mathcal{B}_T(\cdot) = \{\kappa_T(\cdot, t_1), \ldots, \kappa_T(\cdot, t_r)\}^\mathsf{T}$. Let $\lfloor \cdot \rfloor_{\mathcal{B}_T}$ represent the coordinates with respect to $\mathcal{B}_T$; that is, for each $f \in \mathcal{H}^r$, f can be expressed as $f = \lfloor f \rfloor_{\mathcal{B}_T}^\mathsf{T} \mathcal{B}_T(\cdot)$. Let $X_i^m(\tau) = \{X_i^m(t_1), \ldots, X_i^m(t_r)\}^\mathsf{T}$. We have,

$$X_i^m(\tau) = \left\{ \lfloor X_i^m \rfloor_{\mathcal{B}_T}^\mathsf{T} \mathcal{B}_T(t_1), \ldots, \lfloor X_i^m \rfloor_{\mathcal{B}_T}^\mathsf{T} \mathcal{B}_T(t_r) \right\}^\mathsf{T} = K_T \lfloor X_i^m \rfloor_{\mathcal{B}_T}.$$

Consider the eigen-decomposition $K_T = U_T D_T U_T^\mathsf{T}$, where D_T is a diagonal matrix, and U_T collects all the eigenvectors. If the Gram matrix K_T is of a full rank, we build an orthonormal basis $\mathcal{B}_T^* = D_T^{-\frac{1}{2}} U_T^\mathsf{T} \mathcal{B}(\cdot)$. If K_T is not of a full rank, we simply use the leading non-zero eigenvalues and their corresponding eigenvectors to form the orthonormal basis.

Next, we compute the coordinate of X_i^m with respect to $\mathcal{B}_T^*$ as,

$$\lfloor X_i^m \rfloor_{\mathcal{B}_T^*} = D_T^{-\frac{1}{2}} U_T^\mathsf{T} X_i^m(\tau), \tag{11}$$

from which we compute the inner product $\langle X_i^m, X_i^\ell \rangle_{\mathcal{H}^r} = \lfloor X_i^m \rfloor_{\mathcal{B}_T^*}^\mathsf{T} \lfloor X_i^\ell \rfloor_{\mathcal{B}_T^*}$, for $X_i^m, X_i^\ell \in \mathcal{H}_{X_i}$.

Next, we develop the coordinate system for the second-layer RKHS $\mathscr{H}_{X_i}$. Let $\mathscr{H}^n_{X_i}$ be the linear span of the sets of functions $\{\kappa_i(\cdot, X_i^m) - E_n \kappa_i(\cdot, X_i) : m = 1, \ldots, n\}$. Because $\mathrm{range}(\widehat{\Sigma}_{X_i X_i}) = \ker(\widehat{\Sigma}_{X_i X_i})^\perp = \mathscr{H}^n_{X_i}$, for all $i \in \mathtt{V}$, we can restrict our development to the finite-dimensional space $\mathscr{H}^n_{X_i}$. Let $G_{X_i} = n^{-1} Q_n K_i Q_n$, where $Q_n = (\mathbf{I}_n - n^{-1} \mathbf{1}_n^\mathsf{T} \mathbf{1}_n)$, $\mathbf{I}_n$ is the $n \times n$ iden-

tity matrix, $\mathbf{1}_n$ is the $n \times 1$ vector of ones, and $(K_i)_{s,t=1}^n = \kappa_i(X_i^s, X_i^t)$ is the Gram kernel matrix of X_i. Compute the eigen decomposition, $G_{X_i} = U_{X_i} D_{X_i} U_{X_i}^\top + U_{X_i}^0 D_{X_i}^0 (U_{X_i}^0)^\top$, where $U_{X_i} D_{X_i} U_{X_i}^\top$ and $U_{X_i}^0 D_{X_i}^0 (U_{X_i}^0)^\top$ correspond to the top d and tail $n-d$ eigenvalues of G_{X_i}. Therefore, for each $i \in \mathtt{V}$, we construct an orthonormal basis for $\mathscr{H}_{X_i}^n$ as,

$$C_i^*(\cdot) = D_{X_i}^{-1/2} U_{X_i}^\top \left[\kappa_i(\cdot, X_i^1) - E_n \kappa_i(\cdot, X_i), \ldots, \kappa_i(\cdot, X_i^n) - E_n \kappa_i(\cdot, X_i)\right]^\top. \tag{12}$$

Finally, we derive the coordinates of the truncated operators $\widehat{\Gamma}_{X_i X_i}^d$, $\widehat{\Lambda}_{X_i X_j}^d$, and $\widehat{\Sigma}_{X_i X_j}^d$ in Equation (7). Towards that end, we first derive the coordinates of the un-truncated operators $\widehat{\Gamma}_{X_i X_i}$ and $\widehat{\Sigma}_{X_i X_i}$. Let $\mathcal{H}$ and $\mathcal{H}'$ denote two generic Hilbert spaces spanned by $\mathcal{B}$ and $\mathcal{B}'$, $A: \mathcal{H} \to \mathcal{H}'$ a bounded linear operator, and ${}_{\mathcal{B}}\lfloor A \rfloor_{\mathcal{B}'}$ the coordinate of A with respect to $\mathcal{B}$ and $\mathcal{B}'$.

Lemma 1 *For each $i \in \mathtt{V}$, ${}_{\mathcal{B}_T^*}\lfloor \widehat{\Gamma}_{X_i X_i} \rfloor_{\mathcal{B}_T^*} = E_n\left(\lfloor X_i \rfloor \lfloor X_i \rfloor^\top\right)$ and $_{C_i^*}\lfloor \widehat{\Sigma}_{X_i X_i} \rfloor_{C_i^*} = n^{-1} D_{X_i}$.*

Letting $\{(\widehat{\mu}_i^m, \lfloor \widehat{\phi}_i^m \rfloor_{C_i^*})\}_{m=1}^n$ be the pairs of eigenvalues and eigenfunctions of $_{C_i^*}\lfloor \widehat{\Sigma}_{X_i X_i} \rfloor_{C_i^*}$, Lemma 1 implies $\lfloor \widehat{\phi}_i^m \rfloor_{C_i^*} = e_m$, where e_m is the unit vector whose mth position is one and the rest zero. Moreover, by the reproducing property of κ_i in Lemma S1,

$$\widehat{\phi}_i^m(X_i^\ell) = \langle \kappa_i(\cdot, X_i^\ell), \widehat{\phi}_i^m \rangle_{\mathcal{H}_{X_i}} = e_m^\top [C_i^*(X_i^\ell)], \tag{13}$$

for $\ell = 1, \ldots, n$, and $m = 1, \ldots, d$, which implies $\lfloor \kappa_i(\cdot, X_i^\ell) - E_n \kappa_i(\cdot, X_i) \rfloor_{C_i^*}$, the coordinate of $\kappa_i(\cdot, X_i^\ell) - E_n \kappa_i(\cdot, X_i)$ with respect to C_i^*, is equal to $C_i^*(X_i^\ell) - E_n[C_i^*(X_i)]$. Therefore, we can compute the mth sample KL coefficient of $\kappa_i(\cdot, X_i^\ell) - E_n \kappa_i(\cdot, X_i)$ via

$$\widehat{\alpha}_i^{m,\ell} = \lfloor \kappa_i(\cdot, X_i^\ell) - E_n \kappa_i(\cdot, X_i) \rfloor_{C_i^*}^\top \lfloor \widehat{\phi}_i^m \rfloor_{C_i^*} = e_m^\top \{C_i^*(X_i^\ell) - E_n[C_i^*(X_i)]\}. \tag{14}$$

Let $\{(\widehat{a}_i^m, \lfloor \widehat{f}_i^m \rfloor)\}_{m=1}^r$ denote the pairs of eigenvalues and eigenfunctions of ${}_{\mathcal{B}_T^*}\lfloor \widehat{\Gamma}_{X_i X_i} \rfloor_{\mathcal{B}_T^*}$, and $\mathcal{B}_i^* = \{\widehat{f}_i^1, \ldots, \widehat{f}_i^d\}$ be the basis consisting of the d leading eigenfunctions of $\widehat{\Gamma}_{X_i X_i}$. The next result provides the coordinates of the truncated sample covariance operators. Its proof follows immediately from Equations (7), (14), and Lemma 1, and isomitted.

Lemma 2 *For each $(i, j) \in \mathtt{V} \times \mathtt{V}$,*

$$\begin{aligned} &{}_{\mathcal{B}_i^*}\lfloor \widehat{\Gamma}_{X_i X_i}^d \rfloor_{\mathcal{B}_i^*} = n^{-1}[(\mathcal{X}_i^m)^\top \mathcal{X}_i^\ell]_{m,\ell=1}^d, \; {}_{C_i^*}\lfloor \widehat{\Lambda}_{X_i X_j}^d \rfloor_{\mathcal{B}_j^*} = n^{-1}[(\mathfrak{X}_i^m)^\top \mathcal{X}_j^\ell]_{m,\ell=1}^d, \\ &{}_{C_i^*}\lfloor \widehat{\Sigma}_{X_i X_j}^d \rfloor_{C_j^*} = n^{-1}[(\mathfrak{X}_i^m)^\top \mathfrak{X}_j^\ell]_{m,\ell=1}^d, \end{aligned} \tag{15}$$

where $\mathcal{X}_i^m = (\widehat{x}_i^{m,1}, \ldots, \widehat{x}_i^{m,n})^\top$, $\mathfrak{X}_i^m = (\widehat{\alpha}_i^{m,1}, \ldots, \widehat{\alpha}_i^{m,n})^\top$, $\widehat{x}_i^{m,\ell} = \langle X_i^\ell, \widehat{f}_i^m \rangle_{\mathcal{H}^m} = \lfloor X_i^\ell \rfloor^\top \lfloor \widehat{f}_i^m \rfloor$, and $\widehat{\alpha}_i^{m,\ell} = e_m^\top \{C_i^*(X_i^\ell) - E_n[C_i^*(X_i)]\}$.

Using Lemma 2, we obtain the coordinates of $\widehat{\Lambda}_{X_{\mathtt{A}} X_{\mathtt{B}}}^d$ and $\widehat{\Sigma}_{X_{\mathtt{A}} X_{\mathtt{B}}}^d$, for any $\mathtt{A}, \mathtt{B} \subseteq \mathtt{V}$. Let $\mathcal{B}_{\mathtt{A}}^* = \{\mathcal{B}_i^* : i \in \mathtt{A}\}$ and $C_{\mathtt{A}}^* = \{C_i^* : i \in \mathtt{A}\}$. Then we have $(_{C_{\mathtt{A}}^*}\lfloor \widehat{\Lambda}_{X_{\mathtt{A}} X_j}^d \rfloor_{\mathcal{B}_j^*})_i = {}_{C_i^*}\lfloor \widehat{\Lambda}_{X_i X_j}^d \rfloor_{\mathcal{B}_j^*}$ and $[{}_{C_{\mathtt{A}}^*}\lfloor \widehat{\Sigma}_{X_{\mathtt{A}} X_{\mathtt{B}}}^d \rfloor_{C_{\mathtt{B}}^*}]_{i,j} = {}_{C_i^*}\lfloor \widehat{\Sigma}_{X_i X_j}^d \rfloor_{C_j^*}$. Hereafter, we abbreviate ${}_{\mathcal{B}}\lfloor A \rfloor_{\mathcal{B}'}$ as $\lfloor A \rfloor$ when it is appropriate.

4.4 | Algorithm

Below we provide a step-by-step algorithm to summarize the entire estimation procedure.

Step 1. Choose the first-layer kernel κ_T; for example, the Brownian motion function $\kappa_T(s,t) = \min(s,t)$, or the radial basis function $\kappa_T(s,t) = \exp\{-\gamma(s-t)^2\}$, for $(s,t) \in \mathbb{R}^2$; the parameter γ is determined via $1/\sqrt{\gamma} = \sum_{1\le i<j\le r} |t_i - t_j|/\binom{r}{2}$. Compute the Gram matrix K_T, and its eigen decomposition to get U_T and D_T.

Step 2. Compute the coordinate of X_i^m using Equation (11) for each $i \in \mathbb{V}$ and $m = 1, \ldots, n$.

Step 3. Choose the second-layer kernel κ_i. Compute the Gram matrix K_i based on the estimated coordinates of X_i^m, and its eigen-decomposition. We then construct the orthonormal basis using Equation (12).

Step 4. Compute the coordinates of the truncated operators $\widehat{\Gamma}^d_{X_iX_i}$, $\widehat{\Lambda}^d_{X_iX_j}$, and $\widehat{\Sigma}^d_{X_iX_j}$, for all $(i,j) \in \mathbb{V} \times \mathbb{V}$, using Equation (15). For the truncation number d, we set $d = O(n^{1/5})$, following the rule in Ravikumar et al. (2009).

Step 5. Perform the pre-neighbourhood selection as in Equation (8), by minimizing the objective function at the coordinate level, for all $i \in \mathbb{V}$,

$$\check{L}_i(\lfloor A_i \rfloor) = \mathrm{tr}\left(\lfloor A_i \rfloor^\top \lfloor \widehat{\Sigma}^d_{X_{-i}X_{-i}} \rfloor \lfloor A_i \rfloor - 2\lfloor \widehat{\Lambda}^d_{X_{-i}X_i} \rfloor^\top \lfloor A_i \rfloor\right) + \check{\lambda}\left\{\sum_{j\neq i} \|(\lfloor A_i \rfloor)_j\|_{\mathrm{F}}\right\}^2,$$

where $\lfloor A_i \rfloor = (\lfloor (A_i)_1 \rfloor^\top, \ldots, \lfloor (A_i)_{i-1} \rfloor^\top, \lfloor (A_i)_{i+1} \rfloor^\top, \ldots, \lfloor (A_i)_p \rfloor^\top)^\top$, $\lfloor (A_i)_j \rfloor \in \mathbb{R}^{d\times d}$ is the coordinate of $(A_i)_j$, and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. For the sparsity parameter $\check{\lambda}$, we set $\check{\lambda} = O(\sqrt{\log p/n})$, following the choice as in Rothman et al. (2008). Denote the solution as $\lfloor \widehat{A}_i \rfloor$.

Step 6. Estimate the neighbourhood of the ith node by $\widehat{\mathbb{N}}_i = \{j \in \mathbb{V}\setminus\{i\} : \|\lfloor (\widehat{A}_i)_j \rfloor\|_{\mathrm{F}} \neq 0\}$, and obtain the refined collection of permutations $\mathfrak{S}_{\widehat{\mathbb{N}}}$ from the estimated neighbourhoods. Then estimate the directional order as in Equation (9), where the score function at the coordinate level with the truncated operators, $\widehat{S}^d(\sigma; q)$, is

$$\sum_{i=1}^{p}\sum_{l=1}^{d} \log\left[\lambda_l\left\{\lfloor \widehat{\Gamma}^d_{X_{\sigma_i}X_{\sigma_i}} \rfloor - \lfloor \widehat{\Lambda}^d_{X_{\mathrm{R}_{\sigma,i}}X_{\sigma_i}} \rfloor^\top \left(\lfloor \widehat{\Sigma}^d_{X_{\mathrm{R}_{\sigma,i}}X_{\mathrm{R}_{\sigma,i}}} \rfloor + \epsilon I\right)^{-1} \lfloor \widehat{\Lambda}^d_{X_{\mathrm{R}_{\sigma,i}}X_{\sigma_i}} \rfloor\right\}\right].$$

For the ridge parameter ε, we set $\epsilon = \epsilon_0 \times \lambda_1(\lfloor \widehat{\Sigma}^d_{X_{\mathrm{R}_{\sigma,i}}X_{\mathrm{R}_{\sigma,i}}} \rfloor)$ (Lee et al., 2016a). Moreover, with the adoption of a reduced basis, q is no larger than d, and we often set $q = d$. Denote $\widehat{\sigma}(\widehat{\mathbb{N}}) = \arg\min\{\widehat{S}^d(\sigma; q) : \sigma \in \mathfrak{S}_{\widehat{\mathbb{N}}}\}$.

Step 7. Given the estimated order $\widehat{\sigma}(\widehat{\mathbb{N}})$, perform a sparse functional additive regression as in Equation (10), by minimizing the objective function at the coordinate level, for all $i \in \mathbb{V}$,

$$\widehat{L}_i(\lfloor B_i \rfloor) = \mathrm{tr}\left(\lfloor B_i \rfloor^\top \lfloor \widehat{\Sigma}^d_{X_{\mathrm{r}_i}X_{\mathrm{r}_i}} \rfloor \lfloor B_i \rfloor - 2\lfloor \widehat{\Lambda}^d_{X_{\mathrm{r}_i}X_i} \rfloor^\top \lfloor B_i \rfloor\right) + \lambda\left\{\sum_{j\neq i} \|\lfloor (B_i)_j \rfloor\|_{\mathrm{F}}\right\}^2,$$

where $\lfloor B_i \rfloor$ is the coordinate of B_i. For the sparsity parameter λ, we again set $\lambda = O(\sqrt{\log p/n})$ as in Step 5. Denote the solution as $\lfloor \widehat{B}_i \rfloor, i \in \mathbb{V}$.

Step 8. Estimate the DAG of the functional SEM by $\widehat{\mathtt{E}} = \{(i,j) \in \mathtt{V} \times \mathtt{V} : \| \lfloor \widehat{B}_{i,j} \rfloor \|_{\mathrm{F}} \neq 0,\ j \in \mathtt{r}_i,\ i \in \mathtt{V}\}$, with $\lfloor \widehat{B}_{i,j} \rfloor = (\lfloor \widehat{B}_i \rfloor)_j$.

5 | THEORY

In this section, we first derive some useful concentration bounds for the key operators. We then establish the uniform consistency of the order determination. Finally, we establish the uniform consistency of the functional regression operator and the graph consistency of the DAG estimation. All results are derived at the operator level, and allow p and d to diverge with n. The proofs and many supporting lemmas are relegated to the Appendix.

5.1 | Concentration inequalities

We first derive the concentration bounds for the sample truncated covariance operators, $\widehat{\Gamma}^d_{X_iX_i}$, $\widehat{\Lambda}^d_{X_iX_j}$, and $\widehat{\Sigma}^d_{X_iX_j}$, which are the estimates of their population counterparts:

$$\Gamma^d_{X_iX_i} = \sum_{m=1}^{d} a_i^m f_i^m \otimes f_i^m, \quad \Lambda^d_{X_iX_j} = \sum_{m,\ell=1}^{d} E(\alpha_i^m x_j^\ell)\phi_i^m \otimes f_j^\ell, \quad \Sigma^d_{X_iX_j} = \sum_{m,\ell=1}^{d} E(\alpha_i^m \alpha_j^\ell)\phi_i^m \otimes \phi_j^\ell.$$

For positive sequences $\{a_n\}$, $\{b_n\}$ and $\{c_n\}$, denote $a_n \prec b_n$ if $a_n = o(b_n)$, and $a_n \preceq b_n$ if $a_n = O(b_n)$. Denote $a_n \wedge b_n = b_n$, or $a_n \vee b_n = a_n$, if $b_n \preceq a_n$. Similarly, denote $a_n \wedge b_n \wedge c_n = (a_n \wedge b_n) \wedge c_n$, or $a_n \vee b_n \vee c_n = (a_n \vee b_n) \vee c_n$. Define $\nu_d = \min\{\min(|a_i^m - a_i^\ell|, |\mu_i^m - \mu_i^\ell|) : 1 \leq m < \ell \leq d+1, i \in \mathtt{V}\}$, where $|a_i^m - a_i^\ell|$ and $|\mu_i^m - \mu_i^\ell|$ are the distances among all $d+1$ leading eigenvalues of $\Gamma_{X_iX_i}$ and $\Sigma_{X_iX_i}$, for all $i \in \mathtt{V}$. Let $\mathrm{card}(\cdot)$ denote the cardinality.

We require two additional assumptions, one about the degree of the DAG, and the other about B_i^0, the true regression operator. Let $\|B_i^0\|_{\mathrm{HB}} = \sum_{j \in \mathtt{pa}_i} \|(B_i^0)_j\|_{\mathrm{HS}}$ be the hybrid norm of the L_1-norm and Hilbert–Schmidt norm, and F_i^0 be a diagonal matrix of operators with $(F_i^0)_{j,j} = (\|B_i^0\|_{\mathrm{HB}} / \|(B_i^0)_j\|_{\mathrm{HS}})I$, for $j \in \mathtt{pa}_i$.

Assumption 3 There exists $m < 0$ such that $\max_{i \in \mathtt{V}} \mathrm{card}\{\mathtt{pa}(i; \mathtt{G}^0)\} \leq m$.

Assumption 4 There exist $H_i^0 \in \mathscr{B}(\mathscr{H}_{X_i}, \mathscr{H}_{X_{\mathtt{pa}_i}})$ and a constant $c_H > 0$, such that $B_i^0 = (F_i^0)^{-1}\Sigma_{X_{\mathtt{pa}_i}X_{\mathtt{pa}_i}} H_i^0$ with $\max_{i \in \mathtt{V}, j \in \mathtt{pa}(i;\mathtt{G}^0)} \|(H_i^0)_j\|_{\mathrm{HS}} \leq c_H$.

Assumption 3 holds when the maximum degree of the true DAG is bounded. This condition is satisfied when the number of neighbourhoods $\mathtt{N}_i$ is finite. Bühlmann et al. (2014) imposed a similar condition on the neighbourhoods under the scalar-based SEM. Meanwhile, it is possible to relax this condition by allowing $\mathrm{card}\left\{\mathtt{pa}(i; \mathtt{G}^0)\right\}$ to diverge with the same size, which we leave for future research. Assumption 4 imposes a level of smoothness on the true regression operator B_i^0. It extends the range condition in Bach (2008) and Lee et al. (2016b) to the functional setting. We require a stronger condition here to bound the Hilbert–Schmidt norms of all the elements of

H_i^0, which is to ensure the desired convergence rate. Li and Song (2017) also imposed a similar assumption.

Theorem 2 *Suppose Assumptions 1, 3 and 4 hold, and* $\log p/n \prec 1$. *Then there exist positive constants* C *and* C' *such that, for any* $t > 0$, $P(\max_{i\in\mathbb{V}} \|\widehat{\Gamma}^d_{X_iX_i} - \Gamma^d_{X_iX_i}\|_{\mathrm{HS}} > t)$, $P(\max_{i,j\in\mathbb{V}} \|\widehat{\Lambda}^d_{X_iX_j} - \Lambda^d_{X_iX_j}\|_{\mathrm{HS}} > t)$, *and* $P(\max_{i,j\in\mathbb{V}} \|\widehat{\Sigma}^d_{X_iX_j} - \Sigma^d_{X_iX_j}\|_{\mathrm{HS}} > t)$ *are bounded by*

$$C\exp\left\{-C'\frac{n}{\log p}\left(1\wedge\frac{\sqrt{t}\nu_d}{d}\wedge\frac{t\nu_d^2}{d^2}\wedge\frac{t^2\nu_d^2}{d^4}\right)\right\}.$$

Theorem 2 plays a critical role in developing the consistency for both order estimation and functional sparse regression. It extends the concentration bound of high-dimensional covariance matrix (Bickel & Levina, 2008), to that of high-dimensional matrix of covariance operators. Bach (2009) studied the concentration bound of the covariance operator in RKHS, but with a fixed dimension. However, even with p fixed, our result cannot be directly implied by Bach (2009). This is because Bach (2009) only considered the un-truncated covariance operator, whereas Theorem 2 considers the truncated covariance operator, which is more complicated. We need to deal with a diverging number of KL expansions of random functions, and need to develop additional tools such as the perturbation theory of the covariance operator in Lemma S3 in Section S1 of the Appendix.

5.2 | Order consistency

We next establish the uniform consistency of the estimated directional order obtained by minimizing the score function (9). We consider both cases where we do or do not carry out preneighbourhood selection. We begin with two additional assumptions.

Assumption 5 There exists $\beta > 0$, such that $\max_{i\in\mathbb{V}}\left(\sum_{m=d+1}^{\infty} a_i^m\right) \preceq d^{-\beta}$, and $\max_{i\in\mathbb{V}}\left(\sum_{m=d+1}^{\infty} \mu_i^m\right) \preceq d^{-\beta}$.

Assumption 6 For any $\ell \in \mathrm{pa}(i, \mathbb{G}^0)$, if $X_i' = X_i - \sum_{j\in\mathbb{N}_i\setminus\{\ell\}} (R^*_{X_i|X_{\mathbb{N}_i\setminus\{\ell\}}})_j \kappa_j(\cdot, X_j)$, then $R_{X_i'|X_\ell} \neq 0$.

Assumption 5 regulates the tail behaviours of both X_i and $\kappa_i(\cdot, X_i)$, and is essentially a smoothness condition on the marginal distribution of X_i. Recall that, in our estimation, we truncate the covariance operator and keep only its d leading eigenfunctions. At the population level, this truncation loses no information, as long as the covariance operator is of a finite rank, for example, $\mathrm{rank}(\mathcal{H}_{X_i}) = d$, or $\Gamma_{X_iX_i}$ has the decomposition $\Gamma_{X_iX_i} = \sum_{m=1}^{d} a_i^m (f_i^m \otimes f_i^m)$. However, we feel the finite-rank assumption would be overly strong. Instead, we let the tail eigenvalues of the covariance operator to decay in a polynomial rate of d as $d^{-\beta}$ in Assumption 5. In other words, it requires the eigenfunctions from the leading eigenvalues explain a sufficiently large portion of the variation. The parameter β imposes a level of complexity on the marginal distribution of the random function: a larger β indicates a faster decaying rate. Moreover, the effects of d and β have nice interpretations in our asymptotic theory. Basically, a smaller d and a larger β represent a simpler structure which in turn leads to a faster convergence. This can be seen by examining the uniform convergence rate of the conditional covariance operator in Lemma S7 in Section S4 of the Appendix, where we show the

convergence rate grows in the order of d^2 and $d^{-\beta}$. Assumption 6 ensures that, if we do pre-neighbourhood selection, the neighbourhood $\mathtt{N}_i$ contains the parent nodes of the ith node. The quantity X_i' in Assumption 6 can be viewed as the 'residual' after regressing X_i on $X_{\mathtt{N}_i \setminus \{\ell\}}$, where $X_{\mathtt{N}_i \setminus \{\ell\}}$ is the neighbourhood of the ith node excluding the ℓth node. Thus, Assumption 6 requires the regression operator of X_i' on X_ℓ to be non-zero, which indicates that $\mathtt{pa}(i, \mathtt{G}^0) \subseteq \mathtt{N}_i$. A similar condition is also imposed for classical SEM (Bühlmann et al., 2014, Condition B1) (Loh & Bühlmann, 2014, Assumption 1).

Let $\mathtt{M}_i$ be a subset of $\mathtt{V} \setminus \{i\}$ such that $\mathtt{N}_i \subseteq \mathtt{M}_i, i \in \mathtt{V}$. Note that there are two apparent choices of $\mathtt{M}_i$: if we carry out pre-neighbourhood selection, then $\mathtt{M}_i = \mathtt{N}_i$; if we do not, then $\mathtt{M}_i = \mathtt{V} \setminus \{i\}$. Let $\mathtt{M} = \{\mathtt{M}_i : i \in \mathtt{V}\}$ be the collection of $\mathtt{M}_i$. Then we define the refined collection of permutations $\mathfrak{S}_\mathtt{M}$, and similar to Equation (9), use it to solve $\arg\min\{\widehat{S}^d(\sigma; q) : \sigma \in \mathfrak{S}_\mathtt{M}\}$, from which we obtain the estimated order $\widehat{\sigma}(\mathtt{M})$. Moreover, for $i \in \mathtt{V}$ and $\mathtt{S} \subseteq \mathtt{V} \setminus \{i\}$, let $\lambda_{i,\mathtt{S}}^1 > \lambda_{i,\mathtt{S}}^2 \cdots > \lambda_{i,\mathtt{S}}^{q+1} > 0$ denote the $q+1$ leading non-zero eigenvalues of the conditional covariance operator $\Gamma_{X_i X_i | X_\mathtt{S}}$. For any $\mathtt{M}$, let $M^0 = \max_{i \in \mathtt{V}} \mathrm{card}(\mathtt{M}_i)$, $\lambda_q(\mathtt{M}) = \min\{\lambda_{i,\mathtt{S}}^m : m = 1, \ldots, q, \mathtt{S} \subseteq \mathtt{M}_i, i \in \mathtt{V}\}$, and $\nu_q(\mathtt{M}) = \min[\min\{\lambda_{i,\mathtt{S}}^m - \lambda_{i,\mathtt{S}}^{m+1} : m = 1, \ldots, q\} : \mathtt{S} \subseteq \mathtt{M}_i, i \in \mathtt{V}]$. Let $R_0 = \max\{\|R_{X_i|X_\mathtt{S}}\| : i \in \mathtt{V}, \mathtt{S} \subseteq \mathtt{V} \setminus \{i\}\}$. The next theorem establishes the uniform consistency of the estimated order $\widehat{\sigma}(\mathtt{M})$.

Theorem 3 *Suppose Assumptions 1 to 5 hold, and the separation quantity θ_q in Equation (4) and $\mathtt{M}_i$ satisfy $\inf_q \theta_q \geq 0$, $\mathtt{N}_i \subseteq \mathtt{M}_i$ for $i \in \mathtt{V}$, and $\eta_1(\mathtt{M}) \prec [\{\theta_q q^{-1} \lambda_q(\mathtt{M})\} \wedge \nu_q(\mathtt{M})]$, with $\eta_1(\mathtt{M}) = (\log p/n)^{1/2} \nu_d^{-1} \epsilon^{-1} M^0 d^2 + \epsilon^{-1} M^0 d^{-\beta} + \epsilon^{1/2} R_0$. Then, $P\{\widehat{\sigma}(\mathtt{M}) \in \mathfrak{S}^0\} \to 1$, as $\mathrm{n} \to \infty$.*

Bühlmann et al. (2014) established the consistency of the order estimation when the nodes are random variables and linked by additive regression models. Theorem 3 extends their result to a broader setting with nodes being a vector of random functions. Nevertheless, this generalization is far beyond routine. Unlike the random variable setting where the score function is based on the scalar conditional covariance, our score function is computed via the log-determinant of the conditional covariance operator $\Gamma_{X_{\sigma_i} X_{\sigma_i} | X_{\sigma_{-i}}}$. We need to derive a number of inequalities, as given in Lemmas S5 and S6 in Section S4 of the Appendix, and to associate the conditional covariance operator with the covariance operators. To show Theorem 3, we also require the concentration bounds in Theorem 2 to derive the uniform convergence of $\Gamma_{X_{\sigma_i} X_{\sigma_i} | X_{\sigma_{-i}}}$, as summarized in Lemma S7 of the Appendix.

5.3 | Functional additive regression consistency

We next establish the consistency of our functional regression estimation. Our main target is to derive the error bound between the true regression operator B_i^0 and its estimate $\widehat{B}_i$, obtained by solving the objective function $\widehat{L}_i(B_i)$ in Equation (10) with the estimated order $\widehat{\sigma}(\widehat{\mathtt{N}})$ and neighbourhood $\widehat{\mathtt{N}}$. Towards that goal, we first introduce an intermediate objective function $\widehat{L}_i'(B_i)$ and its minimizer $\widehat{B}_i'$, by treating the true order $\sigma^0 \in \mathfrak{S}^0$ and neighbourhood $\mathtt{N} = \{\mathtt{N}_i : i \in \mathtt{V}\}$ as known. We then establish the error bound of this intermediate estimator $\widehat{B}_i'$ in Theorem 4. Next, we establish the pre-neighbourhood selection consistency in Theorem 5. Combining Theorems 4 and 5 with the order consistency established in Theorem 3, we obtain the parameter consistency of the regression operator $\widehat{B}_i$ and the graph consistency of the estimated DAG $\widehat{\mathtt{E}}$ in Theorem 6.

We begin with introduction of the intermediate loss function, $\widehat{L}_i'(B_i)$, for $i \in \mathtt{V}$,

$$\widehat{L}_i'(B_i) = \left\langle \widehat{\Sigma}^d_{X_{\mathtt{r}_{i,0}} X_{\mathtt{r}_{i,0}}} B_i, B_i \right\rangle_{\mathrm{HS}} - 2\left\langle \widehat{\Lambda}^d_{X_{\mathtt{r}_{i,0}} X_i}, B_i \right\rangle_{\mathrm{HS}} + \lambda \left\{ \sum_{j \in \mathtt{r}_{i,0}} \left\| (B_i)_j \right\|_{\mathrm{HS}} \right\}^2, \tag{16}$$

where $\mathtt{r}_{i,0} = \sigma^0_{-j} \cap \mathtt{N}_i$ with $i = \sigma^0_j$ is the refined neighbourhood of X_i based on the true order σ and true neighbourhood $\mathtt{N}_i$. Comparing the loss function $\widehat{L}_i'(B_i)$ in Equation (16) to $\widehat{L}_i(B_i)$ in Equation (10), the main difference is that we regress X_i on $X_{\mathtt{r}_i}$ in Equation (10), whereas we regress X_i on the $X_{\mathtt{r}_{i,0}}$ in Equation (16). In other words, we treat the true order σ^0 and neighbourhood $\mathtt{N}$ as if they were known in Equation (16). Let $\widehat{B}_i'$ denote the minimizer of $\widehat{L}_i'(B_i)$, and $\widehat{\mathtt{pa}}'(i; \mathtt{G}^0) = \{j \in \mathtt{r}_{i,0} : \|\widehat{B}_{ij}'\|_{\mathrm{HS}} \neq 0\}$ denote the estimate of the true parent nodes $\mathtt{pa}(i; \mathtt{G}^0)$.

For any subvectors $\mathtt{A}, \mathtt{B}$ of $\mathtt{V}$, let $C_{\mathtt{AB}} : \mathscr{H}_{X_{\mathtt{B}}} \to \mathscr{H}_{X_{\mathtt{A}}}$ be the correlation operator such that $D_{\mathtt{A}} C_{\mathtt{AB}} D_{\mathtt{B}} = \Sigma_{X_{\mathtt{A}} X_{\mathtt{B}}}$, where $D_{\mathtt{A}}$ is the diagonal matrix of the operators with $[D_{\mathtt{A}}]_{i,i} = \Sigma^{1/2}_{X_i X_i}$ for $i \in \mathtt{A}$, see Baker (1973). We impose two conditions on the correlation operator.

Assumption 7 There exist constants $c_{\min}, c_{\max} > 0$, such that $c_{\min} I \preceq C_{\mathtt{VV}} \preceq c_{\max} I$.

Assumption 8 There exists $\xi \in (0, 1]$ such that, $\|D_j C_{j\mathtt{pa}_i} C^{-1}_{\mathtt{pa}_i \mathtt{pa}_i}\|_{\mathrm{HS}} \leq (1 - \xi) \|B^0_i\|_{\mathrm{HB}} \|C_{\mathtt{pa}_i \mathtt{pa}_i} D_{\mathtt{pa}_i} H^0_i\|^{-1}_{\mathrm{HS}}$, for all $i \in \mathtt{V}$ and all $j \in \mathtt{r}_{i,0} \setminus \mathtt{pa}_i$.

Assumption 7 regulates the dependency among the multivariate functions, by placing bounds on the correlation operator $C_{\mathtt{VV}}$. In the random variable setting, it is often required to restrict the correlation matrix to ensure the convergence of the L_1 penalized estimator Ravikumar et al. (2011). Assumption 7 is an extension to the random function setting. Similarly, Assumption 8 generalizes the random variable-based irrepresentable condition (Zhao & Yu, 2006) to the functional setting, and is required to guarantee the selection consistency.

We next establish the consistency of $\widehat{B}_i'$ and the sparsistency of $\widehat{\mathtt{pa}}'(i; \mathtt{G}^0)$. Let $a^0 = \min_{i \in \mathtt{V}, j \in \mathtt{pa}_i} \|(B^0_i)_j\|_{\mathrm{HS}}$, $\eta_2 = (a^0 \lambda^0_d) \wedge 1$, and $\eta_3 = \nu^0_d \wedge \lambda^0_d$, where $\nu^0_d = \min\{\lambda_m(\sum_{X_{\mathtt{pa}_i} X_{\mathtt{pa}_i}}) - \lambda_{m+1}(\sum_{X_{\mathtt{pa}_i} X_{\mathtt{pa}_i}}) : m = 1, \ldots, d,\ i \in \mathtt{V}\}$, and $\lambda^0_d = \min_{i \in \mathtt{V}} \lambda_d(\sum_{X_{\mathtt{pa}_i} X_{\mathtt{pa}_i}})$.

Theorem 4 *Suppose Assumptions 1 to 8 hold, and* λ *satisfies that* $d^{-\beta} \preceq \lambda^{3/2}$, $\lambda^{1/2} \preceq a^0$, $(\lambda^{-1} d^{-\beta} \vee \lambda) \preceq \eta_2 \lambda^{1/2} (a^0)^2$, *and* $(\log p/n)^{1/2} \prec d^{-2} \nu_d [\eta_3 \wedge \eta_2 \lambda^{3/2} (a^0)^2]$. *Then, as* $\mathrm{n} \to \infty$, *(a)* $\max_{i \in \mathtt{V}} \|\widehat{B}_i' - (B^0_i, 0)\|_{\mathrm{HS}} \to 0$ *in probability; (b)* $P\{\widehat{\mathtt{pa}}'(i; \mathtt{G}^0) \neq \mathtt{pa}(i; \mathtt{G}^0),$ *for some* $i \in \mathtt{V}\} \to 0$.

Lee et al. (2016b) obtained the consistency of a sparse nonparametric regression using linear operators. Although we both employ the idea of regression with linear operators as the parameters, our theoretical development is considerably different. Lee et al. (2016b) only considered a single regression and the point-wise convergence with a fix dimension. In comparison, we show the uniform convergence in Theorem 4 for the DAG estimation, for which we need to consider p regressions simultaneously, and we also allow p to diverge. To address these issues, we introduce a set of intermediate operators, and derive a sequence of concentration bounds to associate the functional regression operator with the covariance operator; see Lemmas S8, S9, S10 and S11 in Section 5 of the Appendix. None of these results are available in Lee et al. (2016b). We also need the concentration bound in Theorem 2 again to bound the probability of erroneous selections in Theorem 4.

Next, we establish the consistency of the pre-neighbourhood selection described in Section 4.2. Recall that $\widehat{A}_i$ is the minimizer of the objective function in Equation (8), and $\widehat{\mathtt{N}}_i$ is the estimate of the true neighbourhood $\mathtt{N}_i$, for all $i \in \mathtt{V}$. Let $\check{a}^0 = \min_{i\in\mathtt{V}, j\in\mathtt{N}_i} \|(R_{X_i|X_{\mathtt{N}_i}})_j\|_{\mathrm{HS}}$, $\check{\eta}_2 = (\check{a}^0 \check{\lambda}_d^0) \wedge 1$, $\check{\eta}_3 = \check{\nu}_d^0 \wedge \check{\lambda}_d^0$, with $\check{\nu}_d^0 = \min\{\lambda_m(\Sigma_{X_{\mathtt{N}_i}X_{\mathtt{N}_i}}) - \lambda_{m+1}(\Sigma_{X_{\mathtt{N}_i}X_{\mathtt{N}_i}}) : m = 1, \ldots, d,\ i \in \mathtt{V}\}$, and $\check{\lambda}_d^0 = \min_{i\in\mathtt{V}} \lambda_d(\Sigma_{X_{\mathtt{N}_i}X_{\mathtt{N}_i}})$. The next theorem establishes the consistency of both $\widehat{A}_i$ and $\widehat{\mathtt{N}}_i$. Its proof is similar to that of Theorem 4, and is thus omitted. We need one more condition.

Assumption 9 There exist constants $\check{m}, \check{c}_H > 0$, $\check{\xi} \in (0, 1]$, and $\check{H}_i^0 \in \mathscr{B}(\mathcal{H}_{X_i}, \mathscr{H}_{X_{\mathtt{N}_i}})$, such that: (a) $\max_{i\in\mathtt{V}} \{\mathrm{card}(\mathtt{N}_i)\} \preceq \check{m}$; (b) $R_{X_i|X_{\mathtt{N}_i}} = (\check{F}_i^0)^{-1}\Sigma_{X_{\mathtt{N}_i}X_{\mathtt{N}_i}}\check{H}_i^0$ with $\max_{i\in\mathtt{V}, j\in\mathtt{N}_i} \|(\check{H}_i^0)_j\|_{\mathrm{HS}} \preceq \check{c}_H$; and (c) $\|D_j C_{j\mathtt{N}_i} C_{\mathtt{N}_i\mathtt{N}_i}^{-1}\|_{\mathrm{HS}} \le (1 - \check{\xi})\|R_{X_i|X_{\mathtt{N}_i}}\|_{\mathrm{u}} \|C_{\mathtt{N}_i\mathtt{N}_i} D_{\mathtt{N}_i} H_i^0\|_{\mathrm{HS}}^{-1}$, for $j \in \mathtt{V}\backslash(\mathtt{N}_i \cup \{i\})$, where $\check{F}_i^0$ is the diagonal matrix of operators, $(\check{F}_i^0)_{jj} = (\|R_{X_i|X_{\mathtt{N}_i}}\|_{\mathrm{HB}} / \|(R_{X_i|X_{\mathtt{N}_i}})_j\|_{\mathrm{HS}})I$, $j \in \mathtt{N}_i,\ i \in \mathtt{V}$.

Assumptions 9(a), 9(b) and 9(c) are the extensions of Assumptions 3, 4, and 8, respectively, and have similar interpretations. That is, Assumption 9(a) assumes the number of neighbourhood is finite, whereas Assumption 9(b) and 9(c) are the smoothness and irrepresentable conditions for $R_{X_i|X_{\mathtt{N}_i}}$.

Theorem 5 *Suppose Assumptions 1, 2, 5, 6, 7, and 9 hold, and $\check{\lambda}$ satisfies that $d^{-\beta} \preceq \check{\lambda}^{3/2}$, $\lambda^{1/2} \preceq \check{a}^0$, $(\check{\lambda}^{-1} d^{-\beta} \vee \check{\lambda}) \preceq \check{\eta}_2 \lambda^{1/2}(\check{a}^0)^2$, and $(\log p/n)^{1/2} \prec d^{-2}\nu_d[\check{\eta}_3 \wedge \check{\eta}_2 \check{\lambda}^{3/2}(\check{a}^0)^2]$. Then, as $n \to \infty$, (a) $\max_{i\in\mathtt{V}} \|\widehat{A}_i - (R_{X_i|X_{\mathtt{N}_i}}, 0)\|_{\mathrm{HS}} \to 0$ in probability; (b) $P\{\widehat{\mathtt{N}}_i \ne \mathtt{N}_i,\ \text{for } i \in \mathtt{V}\} \to 0$.*

Finally, we obtain the consistency of the regression operator $\widehat{B}_i$ and the estimated DAG $\widehat{\mathtt{E}}$, by combining Theorems 3, 4 and 5. The proof is straightforward and is omitted.

Theorem 6 *Suppose Assumptions 1–8 hold, and $\inf_q \theta_q \ge 0$, $d^{-\beta} \preceq \lambda^{3/2}$, $\lambda^{1/2} \preceq a^0$, $(\lambda^{-1} d^{-\beta} \vee \lambda) \preceq \eta_2 \lambda^{1/2}(a^0)^2$, $(\log p/n)^{1/2} \prec d^{-2}\nu_d[\eta_3 \wedge \eta_2 \lambda^{3/2}(a^0)^2]$, and $\eta_1(\mathtt{N}) \prec [\{\theta_q q^{-1} \lambda_q(\mathtt{N})\} \wedge \nu_q(\mathtt{N})]$ hold for θ_q, λ, η_2 and η_3. Suppose the conditions in Theorem 5 also hold. Then we have, as $n \to \infty$, (a) $\max_{i\in\mathtt{V}} \|\widehat{B}_i - (B_i^0, 0)\|_{\mathrm{HS}} \to 0$ in probability; and (b) $P\left(\widehat{\mathtt{E}} \ne \mathtt{E}^0\right) \to 0$.*

6 | NUMERICAL STUDIES

In this section, we first examine the empirical performance of the proposed method through simulations, where we consider different graph structures, association models, graph sizes and sample sizes. We also compare with some potential alternative methods. We then apply our method to a real-world application of a brain effective connectivity analysis. All the relevant computer code and the dataset can be downloaded from https://sites.google.com/site/kuangyaolee1217/software, and more information is given in Section S10 of the Appendix.

6.1 | Simulations

We consider two graph sizes $p = 50, 100$, two sample sizes $n = 100, 200$, and three graph structures with a random graph, a hub graph, and a tree graph. For the random graph, we generate

the true edge set $\mathbb{E}^0$ from independent Bernoulli variables with the probability of selecting an edge equal to $3/(p-1)$; this results in the mean number of edges $E\left\{\text{card}(\mathbb{E}^0)\right\} = 3p/2$. For the hub structure, we generate $p/10$ independent hubs. For the tree structure, we expand the tree until the graph size reaches p. Given a DAG $\mathbb{G}^0$ with the true directional order $1 \to 2 \to \cdots \to p$, we sequentially generate the p-dimensional random function $X(t) = \{X_1(t), \ldots, X_p(t)\}^\top$ as,

$$X_1(t) = \sum_{m=1}^{J} x_{1,m} f_m(t),$$
$$X_i(t) = \sum_{m=1}^{J} x_{i,m} f_m(t) \equiv \sum_{m=1}^{J} \left\{ \sum_{j \in \text{pa}(i;\mathbb{G}^0)} b_{i,j}^m(x_{j,m}) + e_{i,m} \right\} f_m(t), \quad i = 2, \ldots, p,$$

where $b_{i,j}^m : \mathbb{R} \to \mathbb{R}$ is the regression function associating node i, the child node, with node j, the parent node. Let $x_{1,m}$ and the errors $e_{2,m}, \ldots, e_{p,m}$ be independent copies from a standard normal distribution. The function $f_m(t) = \sqrt{2}\sin\{(m-1/2)\pi t\}$ is the mth leading eigenfunction of the Brownian motion kernel, $m = 1, ..., J$, and J is the number of KL coefficients with $J = 20$. We consider two association models, one linear and the other nonlinear, between the child and the parent nodes for $b_{i,j}^m$, for $j \in \text{pa}(i; \mathbb{G}^0)$ and $i = 2, ..., p$,

$$\text{Linear:} \quad b_{i,j}^m(x) = a_{i,j}x, \quad \text{for } m = 1, \ldots, J;$$
$$\text{Nonlinear:} \quad b_{i,j}^m(x) = a_{i,j}x, \quad \text{if } m \text{ is even, and } a_{i,j}x^2 \text{ otherwise,}$$

where $a_{i,j}$'s are independent Bernoulli draws from $\{-1, 1\}$ with $P(a_{i,j} = 1) = 0.5$. We note that the joint distribution of $X(t)$ is non-Gaussian for the nonlinear model. We take the observed time points $t_1, \ldots, t_r$ to be equally spaced grids between $[0, 1]$, and set $r = 50$.

We apply the algorithm in Section 4.4 to the simulated data, using the Brownian motion kernel and the radial basis kernel as the first- and the second-layer kernel functions respectively. The parameters are tuned using the rules described in the algorithm. In Section S7 of the Appendix, we give more implementation details. Moreover, in Section S8, we carry out a sensitivity analysis to study the effects of choices of different kernel functions.

We also compare with some potential alternative solutions. To the best of our knowledge, our proposal is the first structural equation model to learn directional relations among multivariate random functions, and there is no direct competitor. Therefore, we consider two methods that were designed for multivariate random variables: the method of Kalisch and Bühlmann (2007) based on the PC algorithm and a partial correlation test (abbreviated as `linear-PC`), and the method of Harris and Drton (2013) based on the PC algorithm and a rank correlation test (`rank-PC`). To make these methods applicable to the functional data, we couple them with functional principal components (fPCs). Specifically, for each $m = 1, ..., 20$ and $i \in \mathbb{V}$, we compute the mth estimated fPC for the ith function, and denote it by $\hat{x}_i^m$. Let $\mathcal{X}^m = (\hat{x}_1^m, \ldots, \hat{x}_p^m)^\top$ be the p-dimensional vector of the estimated fPCs. We then apply both `linear-PC` and `rank-PC` to $\mathcal{X}^m$, for $m = 1, ..., 20$. This results in a total of 20 estimated graphs, which we merge into a single graph, by claiming an edge if it is present in at least one of the 20 graphs.

To evaluate the performance of each method, we compute the structured Hamming distance (Tsamardinos et al., 2006), which is defined as the smallest number of single operations, including deletions, insertions and re-orientations, that are required to go from the estimated

DAG to the true DAG. Moreover, we compute the true discovery rate of the estimated DAG as $\text{card}[\{(i,j) \in \widehat{\mathrm{E}} : (i,j) \in \mathrm{E}^0\}]/\text{card}[\{(i,j) \in \widehat{\mathrm{E}}\}]$. Note that, because the underlying DAG is sparse, the proportions of true positives and negatives, that is, an edge is present or not, are highly imbalanced. Therefore, we choose to report the true discovery rate in place of the true positive rate or false positive rate. The smaller the distance, and the larger the true discovery rate indicate better performance.

Figures 2–4 report the box plots of the structured Hamming distance and the true discovery rate based on 500 data replications for various combinations of graph structures, association models, graph sizes and sample sizes. We also report the corresponding averages and standard errors of the two criteria in Tables S1 to S3 in Section S9 of the Appendix. We see that our method consistently outperforms the competing methods across all settings. We also note that the improvement increases as the sample size increases. Moreover, under the linear model setting for $b_{i,j}^m$, our model can only determine the DAG up to its equivalence class; however, our method still performs the best under this setting.

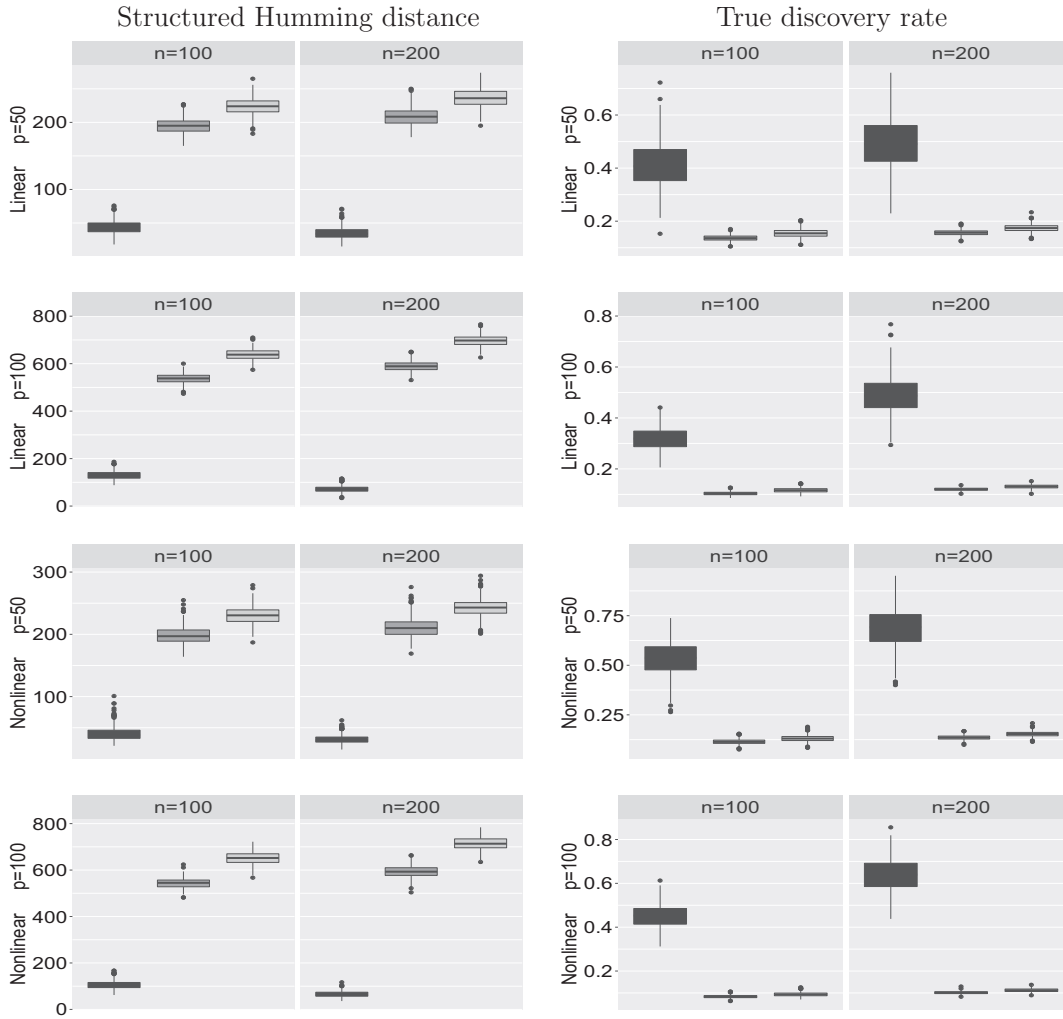

FIGURE 2 Simulations for the random graph structure. The boxes from left to right show the three methods under comparison: our method, `linear-PC`, and `rank-PC`

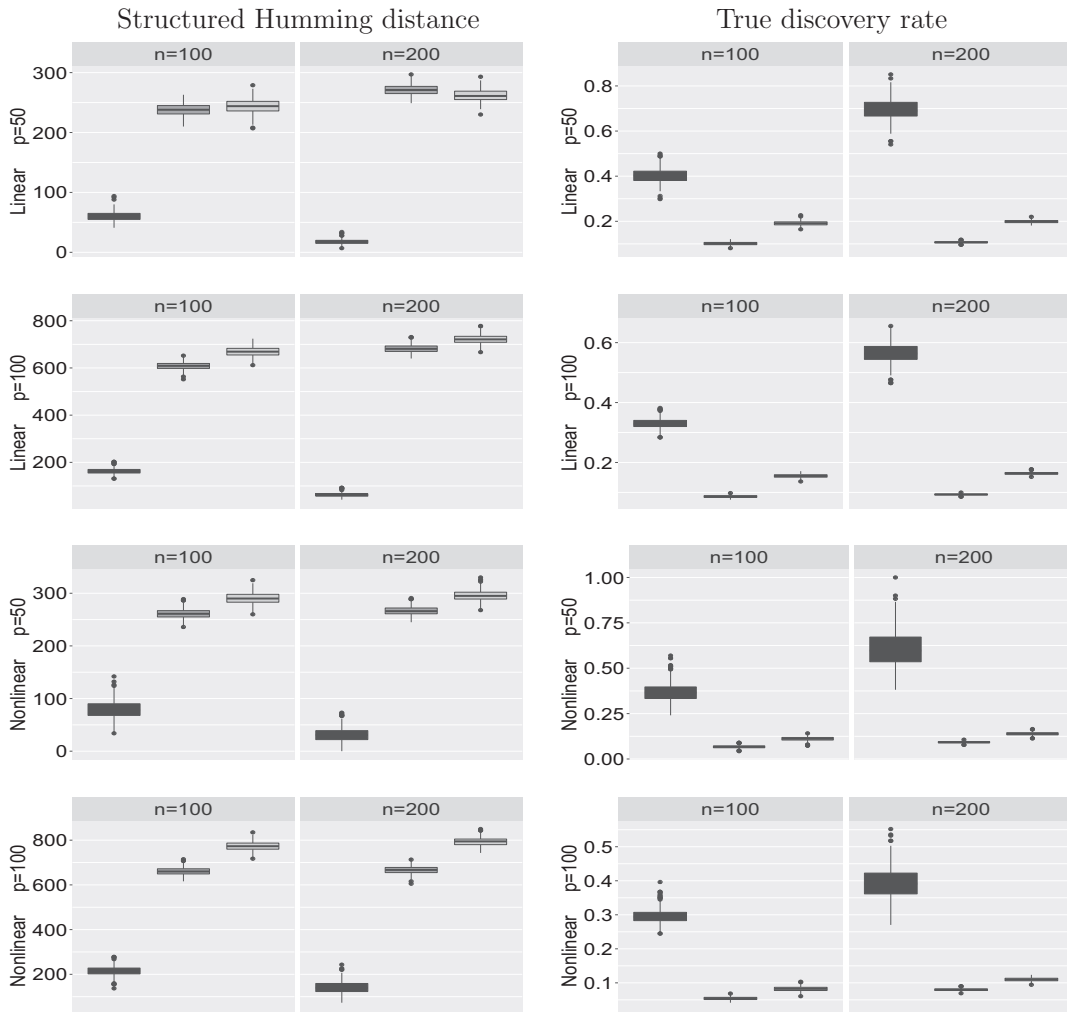

FIGURE 3 Simulations for the hub graph structure. The boxes from left to right show the three methods under comparison: our method, `linear-PC` and `rank-PC`

6.2 | Brain effective connectivity analysis

Next, we illustrate our method with an application to a brain effective connectivity analysis using ECoG. A scientific question of central interest in neuroscience is to understand the directional relations among the neural elements under different activities. The data we analyse is an ECoG study of the brain during decision making (Saez et al., 2018). It consists of the ECoG recordings of 61 electrodes placed in the orbitofrontal cortex (OFC) region of an epilepsy patient when performing gambling tasks with varying levels of winning risk. The patient performed 72 rounds of gambling games, half of which are low-risk games, and half high-risk games. The length of the ECoG signals at each electrode location is 3001. See Saez et al. (2018) for more details about the data collection and processing.

We apply the proposed method to these data, and analyse the low-risk and high-risk games separately. This leads to $p = 61$, $n = 36$, $r = 3001$. We are interested in the electrodes where the estimated connectivity patterns differ considerably between the low-risk and high-risk games.

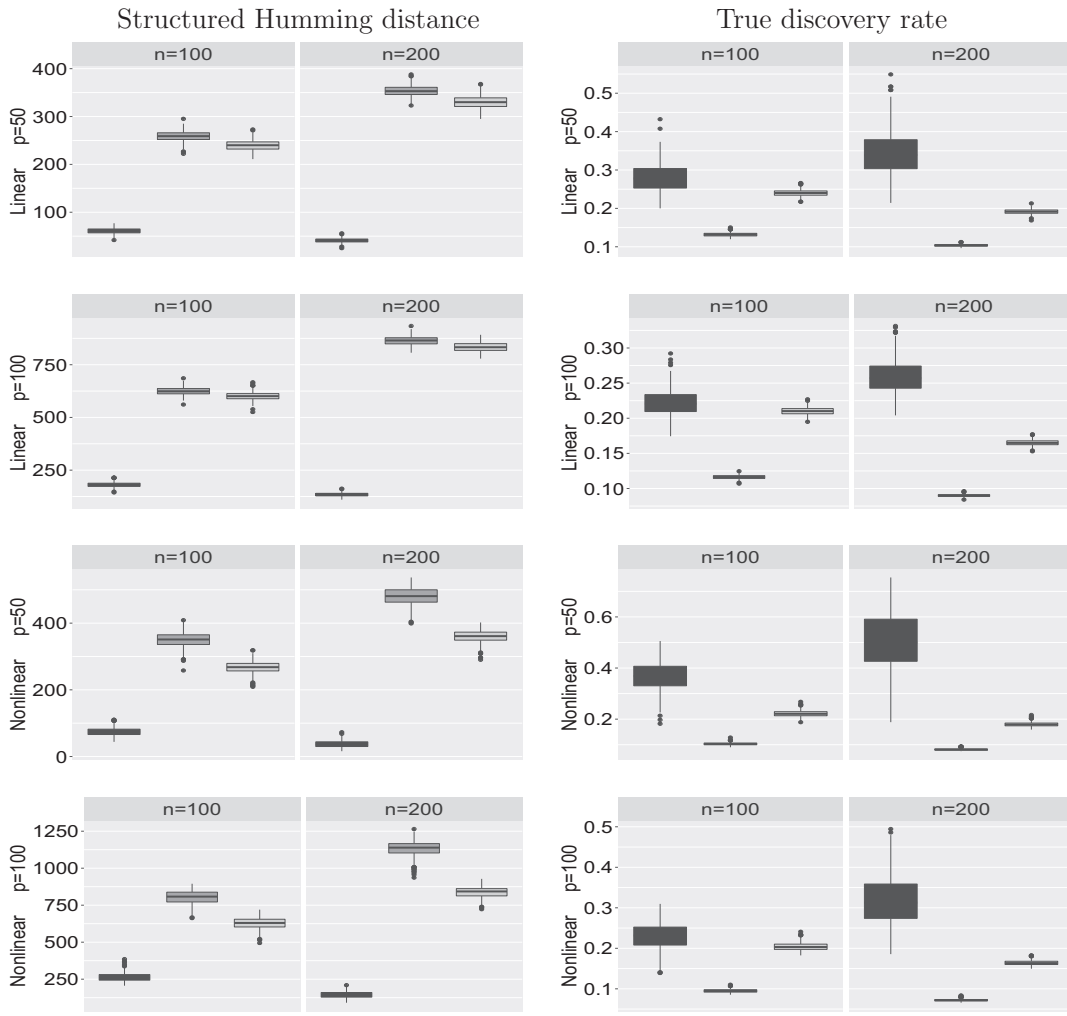

FIGURE 4 Simulations for the tree graph structure. The boxes from left to right show the three methods under comparison: our method, `linear-PC` and `rank-PC`

Towards that goal, we identify those nodes whose total number of inward and outward edges is no more than 2 in one risk type, but no fewer than 9 in the other risk type, which results in nodes 1, 4 and 54. Figure 5 plots the estimated connectivity patterns of those three nodes, first on the entire brain, then in an amplified area. We observe that, for nodes 1 and 4, there are many more outward edges during the low-risk games than the high-risk games, whereas for node 54, the pattern is reversed. These findings are particularly interesting, since according to the cytoarchitecture of OFC (Henssen et al., 2016), node 1 locates in a region called Fo2, node 4 in a region called Fo3, and node 54 in a region called Fo1. Among them, both Fo2 and Fo3 belong to posterior OFC, which is more involved in simple reward type decision making, whereas Fo1 belongs to anterior OFC, which is involved in abstract reward (Kringelbach & Rolls, 2004). Our results suggest that, during the low-risk games where the reward is relatively simple and clear, the posterior OFC is more active. Meanwhile, during the high-risk games that naturally involve more calculations and harder decisions, the anterior OFC tends to more actively influence other nodes.

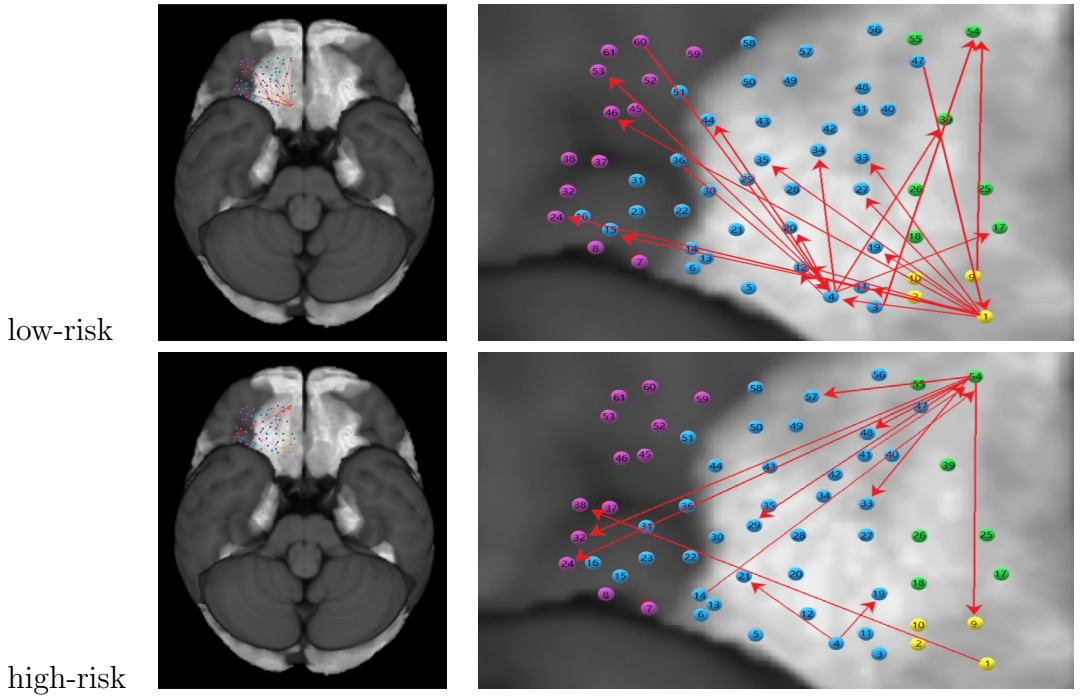

FIGURE 5 ECoG study of the orbitofrontal cortex during gambling games. Green: Fo1; yellow: Fo2; blue: Fo3; purple: other regions [Colour figure can be viewed at wileyonlinelibrary.com]

7 | DISCUSSIONS

We conclude the paper with discussions on some alternative formulations of the functional SEM problem, as well as the normality assumption of the error random function.

7.1 | Alternative model formulations

We consider two related formulations of the functional SEM.

We begin with an orthonormal basis formulation. By the reproducing property of $\mathscr{H}_{X_j}$, we have $(B^0_{i,j}f)(X_j) = \langle B^0_{i,j}f, \kappa_j(\cdot, X_j)\rangle_{\mathscr{H}_{X_j}} = \langle f, (B^0_{i,j})^*\kappa_j(\cdot, X_j)\rangle_{\mathcal{H}_{X_i}}$, for any $f \in \mathcal{H}_{X_i}$. Therefore, we can rewrite the functional structural equation model in Definition 1 as

$$X_i = \sum_{j \in \mathrm{pa}(i;\mathbb{G}^0)} (B^0_{i,j})^*\kappa_j(\cdot, X_j) + \varepsilon_i.$$

Moreover, because $(B^0_{i,j})^*\kappa_j(\cdot, X_j)$ is an element in $\mathcal{H}_{X_i}$,

$$(B^0_{i,j})^*\kappa_j(\cdot, X_j) = \sum_{m\in\mathbb{N}} \langle (B^0_{i,j})^*\kappa_j(\cdot, X_j), f^m\rangle_{\mathcal{H}_{X_i}} f^m \equiv \sum_{m\in\mathbb{N}} b^{0,m}_{i,j}(X_j) f^m,$$

where $\{f^m\}_{m\in\mathbb{N}}$ is an orthonormal basis of $\mathcal{H}_{X_i}$, and $b_{i,j}^{0,m} = B_{i,j}^0 f^m$ is a function in $\mathscr{H}_{X_j}$. This means we can further rewrite Definition 1 as

$$X_i = \sum_{m\in\mathbb{N}} \left\{ \sum_{j\in \mathrm{pa}(i;\mathcal{G}^0)} b_{i,j}^{0,m}(X_j) \right\} f^m + \varepsilon_i. \tag{17}$$

Compared to Definition 1, which is defined via the operator $B_{i,j}^0$, the formulation in Equation (17) is defined via the function $b_{i,j}^{0,m}$ in $\mathscr{H}_{X_j}$ and has a natural interpretation. It is straightforward to show that, $E(\langle X_i, f^m\rangle_{\mathcal{H}_{X_i}} | X_{\mathrm{pa}(i;\mathcal{G}^0)})$ is equal to the additive function $\sum_{j\in\mathrm{pa}(i;\mathcal{G}^0)} b_{i,j}^{0,m}(X_j)$. This means that, when we regress $\langle X_i, f^m\rangle_{\mathcal{H}_{X_i}}$ on its parent nodes $X_{\mathrm{pa}(i;\mathcal{G}^0)}$, $b_{i,j}^{0,m}(X_j)$ represents the additive regression function of X_j for each $j \in \mathrm{pa}(i;\mathcal{G}^0)$. Moreover, Equation (17) provides a concrete way to generate samples that follows our functional SEM.

Nevertheless, using Definition 1 has some advantages over using Equation (17). First, the additive function $b_{i,j}^{0,m}$ defined through Equation (17) is not unique, as it depends on the selection of the orthonormal bases $\{f^m\}_{m\in\mathbb{N}}$. By contrast, the formulation in Definition 1 has no such issue. Second, as implied by Proposition 1, the operator B_i^0 is the same as the regression operator $R_{X_i|X_{\mathrm{pa}(i;\mathcal{G}^0)}}$, which, by definition, is equal to $\Sigma^{-1}_{X_{\mathrm{pa}(i;\mathcal{G}^0)}X_{\mathrm{pa}(i;\mathcal{G}^0)}} \Lambda_{X_{\mathrm{pa}(i;\mathcal{G}^0)}X_i}$. In other words, the parameter of interest B_i^0 can be expressed as a function of the covariance operators $\Sigma_{X_{\mathrm{pa}(i;\mathcal{G}^0)}X_{\mathrm{pa}(i;\mathcal{G}^0)}}$ and $\Lambda_{X_{\mathrm{pa}(i;\mathcal{G}^0)}X_i}$. This is the key property that allows us to derive and study our estimator, at both the population level and at the sample level, using the covariance operators. Moreover, there have been tools developed recently to study the properties of the covariance operators, which simplify our theoretical development (see, e.g. Bach, 2008, 2009; Fukumizu et al., 2009; Lee et al., 2016b; Li & Song, 2017).

We next discuss another related formulation. Kadri et al. (2016) has recently introduced an operator-valued kernel, from which they constructed an RKHS to study nonlinear, function-on-function regressions. We give an example to further discuss the connection between our model and Kadri et al. (2016). We show that, when taking the reproducing kernel in Kadri et al. (2016) to be the product of an $\mathbb{R}$-valued kernel function and the identity operator, its function-valued RKHS can be constructed using our second-layer RKHS $\mathscr{H}_{X_i}$. Moreover, the functional SEM can be reformulated using an operator-valued kernel and the function-valued RKHS induced by it.

Example 2 Let $\chi_{j,i}(\cdot,\cdot): \mathcal{H}_{X_j} \times \mathcal{H}_{X_j} \to \mathcal{B}(\mathcal{H}_{X_i})$ be an operator-valued kernel function such that $\chi_{j,i}(g, g') = \kappa_j(g, g')I_{X_i}$, for all $g, g' \in \mathcal{H}_{X_j}$, where I_{X_i} is the identity mapping from $\mathcal{H}_{X_i}$ to $\mathcal{H}_{X_i}$. By Kadri et al. (2016), the RKHS induced by $\chi_{j,i}$ is

$$\mathcal{K}_{j,i} = \overline{\mathrm{span}}\left\{ \chi_{j,i}(\cdot, g)f : g \in \mathcal{H}_{X_j}, f \in \mathcal{H}_{X_i} \right\} = \overline{\mathrm{span}}\left\{ \kappa_j(\cdot, g)f : g \in \mathcal{H}_{X_j}, f \in \mathcal{H}_{X_i} \right\}.$$

Note that if $\mathcal{H}_{X_i}$ contains the constant functions, then $\mathcal{K}_{j,i} = \overline{\mathrm{span}}\{\phi(\cdot)f : \phi \in \mathscr{H}_{X_j}, f \in \mathcal{H}_{X_i}\}$. Moreover, because $b_{i,j}^{0,m} = B_{i,j}^0 f^m$ is in $\mathscr{H}_{X_j}$, we have that $\sum_{m\in\mathbb{N}} b_{i,j}^{0,m}(\cdot) f^m$ is a member of $\mathcal{K}_{j,i}$, where $b_{i,j}^{0,m}$ is from Equation (17). Therefore, Equation (17) or Definition 1 can be reformulated as,

$$X_i = \sum_{j\in\mathrm{pa}(i;\mathcal{G}^0)} h_{i,j}^0(X_j) + \varepsilon_i,$$

where $h_{i,j}^0(\cdot) = \sum_{m\in\mathbb{N}} b_{i,j}^{0,m}(\cdot) f^m \in \mathcal{K}_{j,i}$.

Motivated by the least squares, one can estimate the function $h_{i,j}^0$ by

$$\min\left\{E\|X_i - \sum_{j\in \mathtt{pa}(i;\mathbb{G}^0)} h_{i,j}(X_j)\|_{\mathcal{H}_{X_i}}^2 : h_{i,j}\in\mathcal{K}_{j,i}, j\in \mathtt{pa}(i;\mathbb{G}^0)\right\}. \tag{18}$$

Note that Kadri et al. (2016) studied a special case of Equation (18) with a nonlinear functional regression of a single response function on a single predictor function. By contrast, our model in Definition 1 is more general, as we allow multiple predictor functions. We thus extend their functional simple regression model to the functional additive regression model.

7.2 | Non-Gaussian error functions

In this article, we have assumed that the error function ϵ_i follows a Gaussian distribution. In the random variable setting, Bühlmann et al. (2014) considered an additive model with Gaussian errors to study the identifiability of the directional order. Analogous to their setting, we adopt the Gaussian assumption in our functional SEM. This allows us to establish a connection between the proposed score function and the KL divergence, a requirement we need to establish the identifiability in Theorem 1. Moreover, under the Gaussian assumption, our proposed score function is the summation of the log-determinant of the error covariance operator, which is relatively simple to compute.

On the other hand, although we require the error function ϵ_i to be Gaussian, the joint distribution $X = (X_1, \ldots, X_p)^\intercal$ induced by our functional SEM can still be non-Gaussian; see, for instance, our simulation example when $b_{i,j}^m$ is nonlinear. Moreover, most of the results in our asymptotic development can be extended to the non-Gaussian setting. Note that both the order consistency in Theorem 3 and the DAG consistency in Theorem 6 depend on a key Lemma S4 in Section S3 of the Appendix, where we derive the concentration bounds of the sample covariance operators. Although the Gaussian assumption is currently imposed in Lemma S4, it can be replaced by some variants of the moment conditions, for example, Bühlmann et al. (2014 Condition B3). Finally, we note that, in the random variable setting, there have been some studies of identifiability without requiring the Gaussian assumption (e.g. Peters et al., 2014; Pfister et al., 2018). It is possible to adopt some of those ideas and extend them to the functional setting. However, we choose to focus on the Gaussian setting in this article, as it is already a very interesting and highly nontrivial problem, and we leave the non-Gaussian setting for future research.

ACKNOWLEDGEMENTS

Lee's research was partially supported by the NSF grant CIF-2102243, and the Seed Funding grant from Fox School of Business, Temple University. Li's research was partially supported by the NSF grant CIF-2102227, and the NIH grants R01AG061303, R01AG062542, and R01AG034570. The authors thank the Editor, the Associate Editor and two referees for their constructive comments and suggestions.

ORCID

Kuang-Yao Lee http://orcid.org/0000-0002-7647-438X
Lexin Li http://orcid.org/0000-0003-2962-1989

REFERENCES

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

Bach, F. (2009) High-dimensional non-linear variable selection through hierarchical kernel learning. *HAL 00413473*.

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

Beck, A. & Teboulle, M. (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1), 183–202.

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

Bosq, D. (2000) *Linear processes in function spaces*. New York: Springer.

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

Buja, A., Hastie, T. & Tibshirani, R. (1989) Linear smoothers and additive models. *The Annals of Statistics*, 17(2), 453–555.

Fan, Y., James, G. & Radchenko, P. (2015) Functional additive regression. *The Annals of Statistics*, 43, 2296–2325.

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

Fu, F. & Zhou, Q. (2013) Learning sparse causal Gaussian networks with experimental intervention: Regularization and coordinate descent. *Journal of the American Statistical Association*, 108(501), 288–300.

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

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

Henssen, A., Zilles, K., Palomero-Gallagher, N., Schleicher, A., Mohlberg, H., Gerboga, F., et al. (2016) Cytoarchitecture and probability maps of the human medial orbitofrontal cortex. *Cortex*, 75, 87–112.

Hoyer, P.O., Janzing, D., Mooij, J.M., Peters, J. & Schölkopf, B. (2009) Nonlinear causal discovery with additive noise models. In: Koller, D., Schuurmans, D., Bengio, Y. & Bottou, L. (Eds.) *Advances in Neural Information Processing Systems 21 (NIPS)*. New York, NY: Curran Associates Inc.

Kadri, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A. & Audiffren, J. (2016) Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17(20), 1–54.

Kalisch, M. & Bühlmann, P. (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8, 613–636.

Kringelbach, M.L. & Rolls, E.T. (2004) The functional neuroanatomy of the human orbitofrontal cortex: Evidence from neuroimaging and neuropsychology. *Progress in Neurobiology*, 72(5), 341–372.

Lee, K.-Y., Li, B. & Zhao, H. (2016a) On an additive partial correlation operator and nonparametric estimation of graphical models. *Biometrika*, 103, 513–530.

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

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

Li, B. & Solea, E. (2018) 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.

Li, B. & Song, J. (2017) Nonlinear sufficient dimension reduction for functional data. *The Annals of Statistics*, 45, 1059–1095.

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

Li, C., Shen, X. & Pan, W. (2020) Likelihood ratio tests for a large directed acyclic graph. *Journal of the American Statistical Association*, 115(531), 1304–1319.

Loh, P.-L. & Bühlmann, P. (2014) High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(1), 3065–3105.
Luo, R. & Qi, X. (2016) Function-on-function linear regression by signal compression. *Journal of the American Statistical Association*, 112(518), 690–705.
Luo, S., Song, R., Styner, M., Gilmore, J. & Zhu, H. (2019) FSEM: Functional structural equation models for twin functional data. *Journal of the American Statistical Association*, 114(525), 344–357.
Meier, L., van de Geer, S. & Bühlmann, P. (2009) High-dimensional additive modeling. *The Annals of Statistics*, 37(6B), 3779–3821.
Müller, H.-G. & Yao, F. (2008) Functional additive models. *Journal of the American Statistical Association*, 103(484), 1534–1544.
Pearl, J. (2009) *Causality: Models, reasoning and inference*, 2nd edn. Cambridge: Cambridge University Press.
Peters, J., Mooij, J.M., Janzing, D. & Schölkopf, B. (2014) Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15, 2009–2053.
Pfister, N., Bühlmann, P., Schölkopf, B. & Peters, J. (2018) Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B*, 80(1), 5–31.
Qiao, X., Guo, S. & James, G.M. (2019) Functional graphical models. *Journal of the American Statistical Association*, 114(525), 211–222.
Ravikumar, P., Lafferty, J., Liu, H. & Wasserman, L. (2009) Sparse additive models. *Journal of the Royal Statistical Society: Series B*, 71, 1009–1030.
Ravikumar, P., Wainwright, M.J., Raskutti, G. & Yu, B. (2011) High-dimensional covariance estimation by minimizing 'ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5, 935–980.
Reimherr, M., Sriperumbudur, B. & Taoufik, B. (2018) Optimal prediction for additive function-on-function regression. *Electronic Journal of Statistics*, 12(2), 4571–4601.
Rothman, A.J., Bickel, P.J., Levina, E. & Zhu, J. (2008) Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2(0), 494–515.
Saez, I., Lin, J., Stolk, A., Chang, E., Parvizi, J., Schalk, G. et al. (2018) Encoding of multiple reward-related computations in transient and sustained high-frequency activity in human OFC. *Current Biology*, 28, 2889–2899.e3.
Shojaie, A. & Michailidis, G. (2010) Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3), 519–538.
Tsamardinos, I., Brown, L.E. & Aliferis, C.F. (2006) The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1), 31–78.
Tsay, R.S. & Pourahmadi, M. (2017) Modelling structured correlation matrices. *Biometrika*, 104(1), 237–242.
Wei, Z. & Li, H. (2008) A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data. *The Annals of Applied Statistics*, 2(1), 408–429.
Westland, J.C. (2015) *Structural equation models: From paths to networks*. Switzerland: Springer International Publishing.
Wright, S. (1923) The theory of path coefficients a reply to Niles's criticism. *Genetics*, 8(3), 239–255.
Yao, F., Müller, H.-G. & Wang, J.-L. (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470), 577–590.
Yuan, Y., Shen, X., Pan, W. & Wang, Z. (2018) Constrained likelihood for reconstructing a directed acyclic Gaussian graph. *Biometrika*, 106(1), 109–125.
Zhao, P. & Yu, B. (2006) On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7, 2541–2563.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Lee K-Y, Li L. Functional structural equation model. *J R Stat Soc Series B*. 2022;84:600–629. Available from: https://doi.org/10.1111/rssb.12471