Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Nonparametric Functional Graphical Modeling Through Functional Additive Regression Operator

Kuang-Yao Lee, Lexin Li, Bing Li & Hongyu Zhao

To cite this article: Kuang-Yao Lee, Lexin Li, Bing Li & Hongyu Zhao (2023) Nonparametric Functional Graphical Modeling Through Functional Additive Regression Operator, Journal of the American Statistical Association, 118:543, 1718-1732, DOI: 10.1080/01621459.2021.2006667

To link to this article: https://doi.org/10.1080/01621459.2021.2006667

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 118, NO. 543, 1718–1732: Theory and Methods
https://doi.org/10.1080/01621459.2021.2006667

Nonparametric Functional Graphical Modeling Through Functional Additive Regression Operator

Kuang-Yao Lee[a], Lexin Li[b], Bing Li[c], and Hongyu Zhao[d]

[a]Temple University, Philadelphia, PA; [b]University of California at Berkeley, Berkeley, CA; [c]Pennsylvania State University, University Park, PA; [d]Yale University, New Haven, CT

ABSTRACT

In this article, we develop a nonparametric graphical model for multivariate random functions. Most existing graphical models are restricted by the assumptions of multivariate Gaussian or copula Gaussian distributions, which also imply linear relations among the random variables or functions on different nodes. We relax those assumptions by building our graphical model based on a new statistical object—the functional additive regression operator. By carrying out regression and neighborhood selection at the operator level, our method can capture nonlinear relations without requiring any distributional assumptions. Moreover, the method is built up using only one-dimensional kernel, thus, avoids the curse of dimensionality from which a fully nonparametric approach often suffers, and enables us to work with large-scale networks. We derive error bounds for the estimated regression operator and establish graph estimation consistency, while allowing the number of functions to diverge at the exponential rate of the sample size. We demonstrate the efficacy of our method by both simulations and analysis of an electroencephalography dataset. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2019
Accepted November 2021

KEYWORDS
Additive conditional independence; Brain connectivity analysis; Graphical model; Neuroimaging analysis; Regression operator; Sparsistency.

1. Introduction

1.1. Motivation and Proposal

Multivariate functional data, where continuous observations are sampled from a vector of stochastic processes, are emerging in a wide range of scientific applications. Examples include time-course gene expression data in genomics studies (Wei and Li 2008), multivariate time series data in finance (Tsay and Pourahmadi 2017), electrocorticography and functional magnetic resonance data in neuroimaging (Zhang et al. 2015), among many others. Functional data analysis has received enormous interest recently; see, for instance, Ramsay and Silverman (2005) and Hsing and Eubank (2015), for reviews of contemporary developments. A central problem in multivariate functional data analysis is to investigate the interdependence among the multivariate random functions. This can be formulated as graphical modeling of multivariate functional data, and is the problem to be tackled in this article.

Our motivation is brain functional connectivity analysis, which is currently at the forefront of neuroscience research. Brain functional connectivity reveals intrinsic functional architecture of the brain (Varoquaux and Craddock 2013). Accumulated evidences have shown that brain connectivity network alters under different pathological conditions. Such alterations are associated with cognitive and behavioral functions, and hold crucial insights of pathologies of neurological disorders (Fox and Greicius 2010). One of the common imaging tools to study functional connectivity is electroencephalography (EEG), which measures brain activities through voltage values of electrodes placed at various scalp locations over a period of recording times. It results in multivariate functional data that take the form of a location by time data matrix for each individual subject. Based on these functional data, an undirected graph is constructed to characterize brain connectivity, where nodes represent neurological elements such as different locations and regions of the brain, and links represent interactions and dependencies among the nodes (Fornito, Zalesky, and Breakspear 2013).

In this article, we propose a new nonparametric functional graphical modeling approach. Built on a recently proposed notion of *functional additive conditional independence* (Li and Solea 2018), we formulate functional graphical estimation as a neighborhood selection problem in a nonlinear regression framework. To do so, we first introduce a new statistical object called *functional additive regression operator* (FARO), which is capable of capturing nonlinear relations without requiring distributional or linear structural assumptions, nor any conditional mean model. We next introduce an objective function that follows the spirit of least-squares but whose deployment is at the Hilbert-Schmidt operator level. This is set forth from a broad synthesis of regression with linear operators as parameters. We show that FARO is the minimizer of the proposed objective function, and thus, offers a versatile nonlinear measure of the function-to-function dependency.

CONTACT Kuang-Yao Lee *kuang-yao.lee@temple.edu* Statistical Science, Temple University, Philadelphia, PA 19122-6008.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

We estimate FARO by minimizing the objective function subject to a mix of L_1 and L_2 type penalties. The estimation algorithm is implemented efficiently via convex solvers. The functional graphical model is then recovered according to the zero operators in the coefficients of the neighborhood selection.

1.2. Related Work

Our proposal is related to but also clearly distinct from several existing lines of work, including graphical modeling for random variables, functional graphical modeling, and linear operator-based methods. Next we discuss the connections and differences in detail.

There have been a large number of proposals for graphical modeling of random variables, most notably, sparse graph estimation under an L_1 penalty or its variants (Yuan and Lin 2007; Friedman, Hastie, and Tibshirani 2008; Ravikumar et al. 2011; Cai, Liu, and Luo 2011; Fan and Lv 2016). It is also noteworthy that some solutions reformulated the problem as neighborhood selection (Meinshausen and Bühlmann 2006; Peng et al. 2009). However, most of those methods imposed a Gaussian structure, whereas some later proposals relaxed the Gaussian assumption (Liu, Lafferty, and Wasserman 2009; Liu et al. 2012; Xue and Zou 2012; Voorman, Shojaie, and Witten 2014). Besides, they all have focused on graphical modeling of random variables. When moving from random variables to random functions, it involves a different set of techniques for both the method and the theory. Moreover, when the functions are only partially observed, as we discuss in Section 4.4, we need to estimate the random functions and account for the estimation error, which in turn would lead to a slower rate of convergence as shown in Theorem 7.

Qiao, Guo, and James (2019) recently proposed a graphical model for functional data, assuming that the random functions follow a multivariate Gaussian distribution. Specifically, let $X = (X_1, \ldots, X_p)^\mathsf{T}$ be a p-dimensional random function on an interval of time T in $\mathbb{R}$. Let $\mathsf{V} = \{1, \ldots, p\}$ and $\mathsf{E} = \{(i,j) \in \mathsf{V} \times \mathsf{V}, i > j\}$ denote the sets of nodes and edges, and $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ the corresponding undirected graph. A natural way to describe the interrelations among X is via *conditional independence* (CI). That is, nodes i and j are not connected in G if and only if X_i and X_j are independent conditioning on the rest of random functions:

$$(i,j) \notin \mathsf{E} \quad \Leftrightarrow \quad X_i \perp\!\!\!\perp X_j \mid X_{-(i,j)}, \tag{1}$$

where $X_{-(i,j)}$ denotes the set of random functions $\{X_k : k \in \mathsf{V} \backslash \{i,j\}\}$. The relation embedded in (1) represents a functional graphical model. As shown in Pearl, Geiger, and Verma (1989), the three-way statistical relation of conditional independence satisfies the semi-graphoid axioms, which are the key properties of the notion of separation that underpins any graph structure. For this reason, CI is commonly used as a criterion to define a graph, and built on this notion, Qiao, Guo, and James (2019) developed a functional graphical model under the Gaussian assumption.

Although an intuitive and appealing idea, using CI to characterize separation among the nodes requires the Gaussian assumption, which is parametric and can be unrealistically strong in numerous applications. Alternatively, one can resort to a fully fledged nonparametric approach. However, it inevitably involves high-dimensional kernels, and thus, often suffers from the curse of dimensionality, a problem that is more severe for large networks. To strike a balance, Li, Chun, and Zhao (2014) proposed a new three-way statistical relation to characterize node separation: *additive conditional independence* (ACI). ACI parallels many canonical principles of CI. Like CI, ACI also satisfies the semi-graphoid axioms. But unlike CI, the estimation of ACI requires neither parametric assumption nor high-dimensional kernels. It thus, avoids the curse of dimensionality, and is able to scale to large networks.

Li and Solea (2018) extended the notion of ACI to *functional additive conditional independence* (FACI) to construct a nonparametric graphical model for multivariate random functions. Our proposal is similar to Li and Solea (2018) in that our method is built upon FACI as well. However, our proposal is also considerably different from theirs in both *methodology* and *theory*. Methodologically, Li and Solea (2018) used the pairwise Markov property induced by FACI as the criterion to construct the graph, whereas we use the local Markov property induced by FACI as such a criterion. More importantly, they used hard thresholding to achieve sparsity, whereas we use penalized minimization to achieve this purpose, which can be carried out via a range of penalty functions. Such a difference is analogous to sparsifying a partial correlation matrix for random variables by hard thresholding versus by penalized regularization (Zhu, Shen, and Pan 2014). Theoretically, Li and Solea (2018) only considered the scenario when the number of graph nodes is *fixed*, and did not derive any concentration bounds. By contrast, we allow the graph size to *diverge* at an exponential rate, establish graph estimation consistency, and derive a set of concentration inequalities and error bounds. The new asymptotic development involves considerable challenges, as little theoretical work has been done previously to investigate function-on-function dependency involving both high dimensionality and nonlinearity. In fact, even under the setting for random variables, the concentration inequalities and error bounds obtained here appear to be the first of their kinds. Furthermore, the tools and techniques we develop are sufficiently general to be applied to broader settings involving sparse estimation at the operator level.

More recently, Solea and Li (2020) proposed a copula method, and Solea and Dette (2020) proposed a nonparametric method, both for functional graphical modeling. Although we target similar problems, our solutions are completely different. Solea and Li (2020) extended the copula Gaussian idea of Liu et al. (2012) to the functional setting. Our model is more general, and can capture structures beyond the copula Gaussian distribution when using nonlinear kernels such as the radial basis function kernel. Actually, if we choose the second-layer kernel to be the inner product of the copula transformation functions, our model includes that of Solea and Li (2020) as a special case. Solea and Dette (2020) extended the joint additive model of Voorman, Shojaie, and Witten (2014). Specifically, let $\{\alpha_i^k\}_{k=1}^\infty$ be the collection of all functional principal components (fPC) from the random function X_i. The model of Solea and Dette (2020) is $E(\alpha_i^k \mid X_{-i}) = \sum_{j \neq i} \sum_{k=1}^\infty f_j^k(\alpha_j^k)$, where $f_j^k : \mathbb{R} \to \mathbb{R}$ is an arbitrary function. That is, the conditional mean of every fPCs of X_i is an additive function of all the fPCs of the rest of the random functions, for each $i \in \mathsf{V}$. Our model, on the

other hand, is of the form, $E(f(X_i) \mid \{X_j : j \neq i\}) = \sum_{j \neq i} f_j(X_j)$, for all $f \in \mathcal{H}_{X_i}$ and f_j is an arbitrary function on X_j. Our model is clearly more general, and it reduces to the model of Solea and Dette (2020) when we choose the kernel of X_i to be the linear kernel and the kernel of each component in X_{-i} to be the additive kernel on the fPCs. Another crucial difference is that, both Solea and Li (2020) and Solea and Dette (2020) were built on the classical conditional independence, whereas our method is built on the functional additive conditional independence. These differences lead to a substantially different solution.

There have also been developments of linear operator-based methods; see Li (2018) for a survey. In particular, Lee, Li, and Zhao (2016a) introduced an additive partial correlation operator to characterize ACI, which extends the partial correlation to the nonlinear setting. Our method differs from Lee, Li, and Zhao (2016a) in that we replace the hard thresholding with the penalized, operator-level neighborhood selection, allow the dimension to go to infinity, and replace random variables with random functions. Lee, Li, and Zhao (2016b) applied ACI to perform variable selection in a classical regression setting, which contains a similar idea as this article, that is, regression with operators as coefficients. On the other hand, we aim at a completely different problem: the graphical modeling of multivariate random functions. At the individual regression level, Lee, Li, and Zhao (2016b) only considered the setting when p is fixed, whereas our theory allows p to diverge at an exponential rate with the sample size. At the graph estimation level, we need to consider p regressions simultaneously with a diverging p. In addition, we establish the concentration bound of our functional additive regression operator, and this type of result is not available in Lee, Li, and Zhao (2016b). There has been some work studying the concentration bounds of the empirical covariance operator, and some on the concentration bounds in a classical regression setting where the response space is the Euclidean space; see Bosq (2000); Yao, Rosasco, and Caponnetto (2007). Our regression setting is more general than those existing results in that our response space is a reproducing kernel Hilbert space.

1.3. Organization

The rest of the article is organized as follows. We set up the nonparametric functional graphical model based on ACI and neighborhood selection in Section 2. We develop an estimation procedure at the operator-level in Section 3, and derive the asymptotic theory in Section 4. We implement the estimation procedure and present an algorithm using a coordinate representation in Section 5. We conduct simulations, and illustrate our method with an EEG data analysis in Section 6. We conclude the article in Section 7, and relegate all proofs and additional results to the supplementary material.

2. Model

In this section, we first introduce the functional additive regression operator (FARO), then develop the notions of functional additive conditional independence and functional neighborhood selection, from which we define our version of functional graphical model. Finally, we connect neighborhood selection with FARO, which lays the foundation for the subsequent development of estimating functional graphical model through sparse estimation of FARO.

2.1. Functional Additive Regression Operator

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. For each $i = 1, \ldots, p$, let Ω_{X_i} denote a Hilbert space of $\mathbb{R}$-valued functions defined on an interval $T \subseteq \mathbb{R}$, and Ω_X the Cartesian product $\Omega_{X_1} \times \cdots \times \Omega_{X_p}$. Let $X = (X_1, \ldots, X_p)^\mathsf{T}$ be a p-dimensional random function defined on Ω and taking value in Ω_X.

Definition 1. Let $\langle \cdot, \cdot \rangle_{\Omega_{X_i}}$ denote the inner product in Ω_{X_i}. Then, for any $i = 1, \ldots, p$, a positive definite kernel $\kappa_i : \Omega_{X_i} \times \Omega_{X_i} \to \mathbb{R}$ is said to be induced by $\langle \cdot, \cdot \rangle_{\Omega_{X_i}}$ if, for any $f, g \in \Omega_{X_i}$, there exists a function $\rho : \mathbb{R}^3 \to \mathbb{R}$ such that $\kappa_i(f, g) = \rho(\langle f, f \rangle_{\Omega_{X_i}}, \langle f, g \rangle_{\Omega_{X_i}}, \langle g, g \rangle_{\Omega_{X_i}})$.

The same construction was used in Li and Solea (2018). This definition is inspired by kernel mapping in the multivariate random variable setting, except that the domain of the kernel has been changed from a Euclidean space to an infinite-dimensional functional space. There are many types of kernels, for instance, the radial basis kernel $\kappa_i(f, g) = \exp(-\gamma \|f - g\|^2_{\Omega_{X_i}})$, and the polynomial kernel $\kappa_i(f, g) = (1 + \langle f, g \rangle_{\Omega_{X_i}})^\gamma$.

Given the kernel κ_i, we build up a second-level Hilbert space $\mathcal{H}_{X_i}$, which is the reproducing kernel Hilbert space generated by κ_i. Let $\mathcal{H}_X$ be the direct sum $\oplus_{i=1}^p \mathcal{H}_{X_i}$, which is the linear space $\mathcal{H}_X = \left\{ f_1 + \cdots + f_p : f_1 \in \mathcal{H}_{X_1}, \ldots, f_p \in \mathcal{H}_{X_p} \right\}$ with the inner product $\langle f_1 + \cdots + f_p, g_1 + \cdots + g_p \rangle = \sum_{i=1}^p \langle f_i, g_i \rangle_{\mathcal{H}_{X_i}}$. For any subset A of $\{1, \ldots, p\}$, we define $\mathcal{H}_{X_A}$ to be the direct sum $\oplus_{i \in A} \mathcal{H}_{X_i}$.

Suppose, for each $i = 1, \ldots, p$, every member of $\mathcal{H}_{X_i}$ is square-integrable. Then by Theorem 2.2 of Conway (1990), for each pair (i, j), there exists a unique linear operator $\Sigma_{X_i X_j} : \mathcal{H}_{X_j} \to \mathcal{H}_{X_i}$ such that $\langle f, \Sigma_{X_i X_j} g \rangle_{\mathcal{H}_{X_i}} = \operatorname{cov}[f(X_i), g(X_j)]$, for all $f \in \mathcal{H}_{X_i}$ and $g \in \mathcal{H}_{X_j}$. We then define an operator $\Sigma_{XX} : \mathcal{H}_X \to \mathcal{H}_X$ via

$$\Sigma_{XX} f = \sum_{i=1}^p \sum_{j=1}^p \Sigma_{X_i X_j} f_j,$$

where $f = f_1 + \cdots + f_p \in \mathcal{H}_\mathcal{X}$. In other words, Σ_{XX} is a matrix of operators, whose (i, j)th entry is $\Sigma_{X_i X_j}$. This operator is called the functional additive variance operator in Li and Solea (2018). Similarly, for any subvectors U, V of X, we define the functional additive covariance operator Σ_{UV} as the matrix of operators whose entries are $\Sigma_{U_i V_j}$.

We next define the Moore-Penrose inverse of an operator. For a linear operator T, let (T) denote the kernel space (or null space) of T; that is, $\ker(T) = T^{-1}(\{0\}) = \{f : Tf = 0\}$. Let $\operatorname{ran}(T)$ denote the range of T, and $\overline{\operatorname{ran}}(T)$ denote the closure of $\operatorname{ran}(T)$. Note that Σ_{XX} is not invertible if $\ker(\Sigma_{XX}) \neq \{0\}$. However, if we let $\{\Sigma_{XX} | \ker(\Sigma_{XX})^\perp\}$ to be Σ_{XX} restricted on $\ker(\Sigma_{XX})^\perp$, then the restricted operator is invertible. We call the inverse of this restricted operator the Moore-Penrose inverse of Σ_{XX}, and denote it by $\Sigma^\dagger_{XX}$. Since Σ_{XX} is a self-adjoint operator, we have $\ker(\Sigma_{XX})^\perp = \overline{\operatorname{ran}}(\Sigma_{XX})$. Thus, $\Sigma^\dagger_{XX}$ is a mapping from

$\text{ran}(\Sigma_{XX})$ to $\overline{\text{ran}}(\Sigma_{XX})$ that maps the member y of $\text{ran}(\Sigma_{XX})$ to the unique member $x \in \overline{\text{ran}}(\Sigma_{XX})$ that satisfies $\Sigma_{XX}x = y$.

Assumption 1. Suppose $\text{ran}(\Sigma_{UV}) \subseteq \text{ran}(\Sigma_{UU})$ for all disjoint subvectors U, V of X. Moreover, the linear operator $B^0_{X_{-i}X_i} \equiv \Sigma^{\dagger}_{X_{-i}X_{-i}}\Sigma_{X_{-i}X_i}$ is Hilbert-Schmidt, for all $i \in \mathsf{V}$.

We extend the definition of $B^0_{X_{-i}X_i}$ to $B^0_{UV} \equiv \Sigma^{\dagger}_{UU}\Sigma_{UV}$, for any disjoint subvectors U, V of X. Note that B^0_{UV} is well-defined when $\text{ran}(\Sigma_{UV}) \subseteq \text{ran}(\Sigma_{UU})$, which is ensured by Assumption 1. Moreover, the condition $\text{ran}(\Sigma_{UV}) \subseteq \text{ran}(\Sigma_{UU})$ means the operator Σ_{UV} sends the function in $\mathcal{H}_V$ to the low-frequency part of Σ_{UU}, which is a type of collective smoothness in the relation between U and V. This condition is not directly related to the dimension of V, as $\mathcal{H}_V$ consists of scalar-valued functions of V. Nevertheless, when the dimension of V increases, we expect the class of $\mathcal{H}_V$ to be more complex, and the condition $\text{ran}(\Sigma_{UV}) \subseteq \text{ran}(\Sigma_{UU})$ would impose a stronger smoothness on the relation between U and V. This agrees with the spirit of typical nonparametric estimations: the more complex the function is, the stronger the penalty should be. Later in Section 2.2, we discuss a concrete setting under which Assumption 1 is satisfied.

Definition 2. Let U, V be any subvectors of X, and suppose Assumption 1 holds. Then we call the operator B^0_{UV} the functional additive regression operator (FARO) from $\mathcal{H}_V$ to $\mathcal{H}_U$.

This concept of FARO plays a central role in our proposal. It can be viewed as the functional extension of the additive regression operator, which was developed in Lee, Li, and Zhao (2016b) for nonparametric variable selection. The term "regression operator" was motivated by the similarity of $\Sigma^{\dagger}_{UU}\Sigma_{UV}$ to the regression coefficient matrix in multivariate linear regression.

2.2. Functional Additive Conditional Independence

We first give an alternative but equivalent definition of FACI to that in Li and Solea (2018).

Definition 3. Let U, V, W be subvectors of X, and suppose Assumption 1 holds. We say that the random elements U and V are additively conditionally independent given the random element W, denoted by $U \perp\!\!\!\perp_A V \mid W$, if and only if, for any $f \in \mathcal{H}_U$, $g \in \mathcal{H}_V$,

$$\text{cov}[f(U) - (B^0_{WU}f)(W), g(V) - (B^0_{WV}g)(W)] = 0.$$

Li and Solea (2018) defined FACI using orthogonality between subspaces. The above alternative definition directly involves FARO, which helps greatly to simplify the subsequent theoretical and computational developments.

Next, we discuss how different choices of kernels can lead to different FACI relations. We first show that a stronger FACI is implied by larger RKHSs.

Proposition 1. Suppose U, V are subvectors of X with the corresponding RKHSs $\mathcal{H}^{(1)}_U, \mathcal{H}^{(2)}_U$ and $\mathcal{H}^{(1)}_V, \mathcal{H}^{(2)}_V$, and W is another subvector of X with the corresponding RKHS $\mathcal{H}_W$. If $\mathcal{H}^{(1)}_U \subseteq \mathcal{H}^{(2)}_U$, $\mathcal{H}^{(1)}_V \subseteq \mathcal{H}^{(2)}_V$, $\perp\!\!\!\perp_{A_1}$ is defined via $(\mathcal{H}^{(1)}_U, \mathcal{H}^{(1)}_V, \mathcal{H}_W)$, and $\perp\!\!\!\perp_{A_2}$ is defined via $(\mathcal{H}^{(2)}_U, \mathcal{H}^{(2)}_V, \mathcal{H}_W)$, then $U \perp\!\!\!\perp_{A_2} V \mid W \Rightarrow U \perp\!\!\!\perp_{A_1} V \mid W$.

The next result shows different relations between FACI and CI under the functional copula Gaussian model. Specifically, suppose $X_i = E(X_i) + \sum_{k=1}^{\infty} \alpha_{ik}\phi_{ik}$ is the Karhunen-Loéve expansion of X_i, for each $i \in \mathsf{V}$. We say X_i follows a copula Gaussian distribution if there exists a collection of monotone functions $\{f_{ik} : k = 1, \ldots\}$, such that $F_i(X_i) \equiv \sum_{k=1}^{\infty} f_{ik}(\alpha_{ik})\phi_{ik}$ is an Ω_{X_i}-valued Gaussian element. Furthermore, $X = (X_i, \ldots, X_p)^{\mathsf{T}}$ follows a joint copula Gaussian distribution if $F(X) \equiv [F_1(X_1), \ldots, F_p(X_p)]^{\mathsf{T}}$ is an $\Omega_{X_1} \times \cdots \times \Omega_{X_p}$-valued Gaussian element; see also Li and Solea (2018) for such definitions.

Proposition 2. Suppose X follows a copula Gaussian distribution with the sequences of copula transformation functions $F(\cdot) = [F_1(\cdot), \ldots, F_p(\cdot)]^{\mathsf{T}}$.

(a) If there exists $M > 0$ such that $E\|F_i(X_i)\|^2_{\Omega_{X_i}} \le M$ for all $i \in \mathsf{V}$, and that $\mathcal{H}_{X_i}$ is dense in $L_2(P_{X_i})$, then for any subvectors U, V, W of X, we have $U \perp\!\!\!\perp_A V \mid W \Rightarrow U \perp\!\!\!\perp V \mid W$.
(b) If $\mathcal{H}_{X_i} = \overline{\text{span}}\{\kappa_i(\cdot, x) : x \in \Omega_{X_i}\}$ with $\kappa_i(x', x) = \langle F_i(x'), F_i(x)\rangle_{\Omega_{X_i}}$ for any $x, x' \in \Omega_{X_i}$, then for any subvectors U, V, W of X, we have $U \perp\!\!\!\perp_A V \mid W \Leftrightarrow U \perp\!\!\!\perp V \mid W$.

The proof of this proposition follows Li and Solea (2018, Theorem 1) and is omitted. Proposition 2(a) shows that FACI implies CI, but not vice versa, when the kernel is characteristic, for example, the radial basis kernel, while Proposition 2(b) shows that FACI and CI are equivalent when the kernel induces the same space as spanned by the copula transformation functions. Both Propositions 1 and 2 suggest that it is beneficial to choose a kernel that is dense in the ambient L_2-space. In practice, we suggest choosing a kernel function that satisfies this condition, such as the radial basis kernel function.

Next, to further understand the probabilistic mechanism underlying FACI, as well as its relation with CI, we introduce a new distribution, the Additive Gaussian Distribution, under which FACI and CI are equivalent. This distribution is much more general than the Gaussian distribution. More importantly, it provides a concrete probability model that generates the FACI relation for multivariate random functions.

If the p-variate random function $X = (X_1, \ldots, X_p)^{\mathsf{T}}$ satisfies: (a) Ω_{X_i} is a finite-dimensional Hilbert space spanned by an orthonormal basis $\{\eta_1, \ldots, \eta_m\}$ with inner product $\langle\cdot,\cdot\rangle_{\Omega_{X_i}}$, and (b) $X_i = \sum_{k=1}^{m} U_{ik}\eta_k$, where $U_{ik} = \langle X_i, \eta_k\rangle_{\Omega_{X_i}}$, then the Ω_{X_i}-valued random function X_i has a one-to-one correspondence with the random vector $U_i \equiv (U_{i1}, \ldots, U_{im})^{\mathsf{T}}$. Let $\mathbb{N} = \{1, 2, \ldots, \}$ be the collection of natural numbers. Given $\alpha \in \mathbb{N}$, let Q_α be the set of all monomials in $U_{i1}, \ldots, U_{im}$ of orders between 1 and α, that is, $Q_\alpha(U_i) = \{U_{i1}^{\alpha_1} \cdots U_{im}^{\alpha_m} : \alpha_1, \ldots, \alpha_m \in \{0\} \cup \mathbb{N}, 1 \le \sum_{j=1}^{m} \alpha_j \le \alpha\}$. Let $U = (U_1^{\mathsf{T}}, \ldots, U_p^{\mathsf{T}})^{\mathsf{T}}$, $R_\alpha(U) = (Q_\alpha^{\mathsf{T}}(U_1), \ldots, Q_\alpha^{\mathsf{T}}(U_p))^{\mathsf{T}}$, $\mu_\alpha = E[R_\alpha(U)]$, and $\Sigma_\alpha = \text{var}[R_\alpha(U)]$. By simple combinatorics, it can be shown that the

dimension of $Q_\alpha(U_i)$ is $c(m,\alpha) = \sum_{\beta=1}^{\alpha} \binom{m+\beta-1}{m-1}$. Henceforth, we abbreviate the elements of $Q_\alpha(u_i)$ by $q_1(u_i), \ldots, q_{c(m,\alpha)}(u_i)$.

We say that U follows an additive Gaussian distribution if there exists $c > 0$ such that the density of U satisfies

$$f(U) = f(U_1, \ldots, U_p) = c \exp\left\{-\frac{1}{2}[R_\alpha(U) - \mu_\alpha]^\mathsf{T} \Sigma_\alpha^{-1} [R_\alpha(U) - \mu_\alpha]\right\}. \quad (2)$$

We write it as $U \sim \mathrm{AN}(\mu_\alpha, \Sigma_\alpha)$. Note that $\mathrm{AN}(\mu_1, \Sigma_1)$ is the p-variate Gaussian distribution; that is, the Gaussian distribution is the first-order additive Gaussian distribution. If $U \sim \mathrm{AN}(\mu_\alpha, \Sigma_\alpha)$, then we call X an additive Gaussian random function in $\Omega_{X_1} \times \cdots \times \Omega_{X_p}$. Since there is a one-to-one correspondence between X and U, we also write $X \sim \mathrm{AN}(\mu_\alpha, \Sigma_\alpha)$.

Next, we give a concrete example to justify the existence of the density in (2).

Example 1. Define $u(x) = (x, x^2)^\mathsf{T}$, and $\mu = (\mu_1, \mu_2)^\mathsf{T} \in \mathbb{R}^2$. Consider the function,

$$f(x) = \exp\left(-\frac{1}{2}(u(x) - \mu)^\mathsf{T}(u(x) - \mu)\right).$$

In Section S4 of the supplementary material, we show the existence of the density f in Example 1. We can also extend the same construction to the multivariate case of $X = (X_1, \ldots, X_p)^\mathsf{T}$.

The construction of X is in line with the original motivation of ACI in Li, Chun, and Zhao (2014), and reveals the probabilistic mechanism that generalizes partial correlation and Gaussian-like behavior to nonlinear features. When $X \sim \mathrm{AN}(\mu_\alpha, \Sigma_\alpha)$, FACI and CI are equivalent, as shown in the next theorem.

Theorem 1. If $X \sim \mathrm{AN}(\mu_\alpha, \Sigma_\alpha)$, and $\kappa_i(x, x') = (1 + \langle x, x' \rangle_{\Omega_{X_i}})^\alpha$, for all $i \in \mathsf{V}$, then the three statements, (a) $(\Sigma_\alpha^{-1})_{i,j} = 0$, (b) $X_i \perp\!\!\!\perp_A X_j \mid X_{-(i,j)}$, (c) $X_i \perp\!\!\!\perp X_j \mid X_{-(i,j)}$, are all equivalent, where $(\Sigma_\alpha^{-1})_{i,j}$ is the (i,j)th block of Σ_α^{-1}.

We note that, when X is an additive Gaussian random function, Assumption 1 holds, because all operators involved are finite-rank operators.

2.3. Functional Neighborhood Selection

Based on the definition of FACI, we now define the neighborhood N_i of node $i = 1, \ldots, p$.

Definition 4. Suppose Assumption 1 holds. A subset N_i of V is called the neighborhood of node i with respect to FACI, if (a) $\mathsf{N}_i \subseteq \mathsf{V} \setminus \{i\}$ and $X_i \perp\!\!\!\perp_A X_{-i} \mid X_{\mathsf{N}_i}$, and (b) for any A that satisfies (i), $\mathsf{N}_i \subseteq \mathsf{A}$. Moreover, we say $X = (X_1, \ldots, X_p)^\mathsf{T}$ is associated with $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ with respect to FACI if $\mathsf{N}_i = \{j : (i,j) \in \mathsf{E}\}$.

By definition, N_i is the smallest subset of $\mathsf{V}\backslash\{i\}$ such that, conditioning on X_{N_i}, X_i is additively independent of the rest of random functions. Note that an equivalent way of saying $X_i \perp\!\!\!\perp_A X_{-i} \mid X_{\mathsf{N}_i}$ is $X_i \perp\!\!\!\perp_A X_{\mathsf{V}\setminus(\mathsf{N}_i \cup \{i\})} \mid X_{\mathsf{N}_i}$. Definition 4 essentially gives a formal definition of our version of the functional graphical model. That is, we aim to find the graph G, a functional additive semi-graphoid, such that X is associated with G with respect to FACI.

Li and Solea (2018) introduced their version of functional additive semi-graphoid model based on the following equivalence

$$(i,j) \notin \mathsf{E} \quad \Leftrightarrow \quad X_i \perp\!\!\!\perp_A X_j \mid X_{-(i,j)}. \quad (3)$$

It is interesting to note that, the relations characterized in Definition 4 and Equation (3) are closely related but *not* identical. Specifically, they are related to different Markov properties for undirected graphs. Equation (3) requires the *pairwise* additive Markov property to hold, meaning that if two nodes are not connected, they are additively conditionally independent given the rest of nodes. Definition 4 relies on the *local* additive Markov property, which indicates that, given its neighbors, a node is additively conditionally independent with every nonadjacent node. Lauritzen (1996, chap. 3) has shown that the local Markov condition is generally stronger than the pairwise Markov condition. The next proposition provides a parallel relation between the local and pairwise additive Markov conditions.

Proposition 3. If X is associated with $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ with respect to FACI, then we have $X_i \perp\!\!\!\perp_A X_j \mid X_{-(i,j)}$ for any $(i,j) \notin \mathsf{E}$.

We next give a counterexample to show that the local additive Markov property is *not* implied by the pairwise additive Markov property.

Example 2. Let $X = (X_1, X_2, X_3) \in \Omega^3$ be a 3-variate random function, and satisfy that $P(X_1 = X_2 = X_3) = 1$. Let $\mathcal{H}_X = \oplus_{i=1}^3 \overline{\mathrm{span}}\{\kappa_i(\cdot, x_i) : x_i \in \Omega_{X_i}\}$, with κ_i being a positive kernel function. Let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$, where $\mathsf{V} = \{1, 2, 3\}$ and E is an empty set. Then X satisfies Equation (3) with respect to G: for example, $X_1 \perp\!\!\!\perp X_2 \mid X_3$ because X_1 and X_2 are both fixed given X_3. However, X does not satisfy Definition 4 with respect to G: for example, $\mathsf{N}_1 = \varnothing$ and $X_1 \perp\!\!\!\perp_A (X_2, X_3) \mid \varnothing$ does not hold.

Although the pairwise additive Markov property does not always imply the local additive Markov property, we next show that this relation can still hold under some conditions.

Proposition 4. Suppose $\ker(\Sigma_{XX}) = \{0\}$, and the correlation operator $C_{X_i X_j}$ is compact for any $(i,j) \in \mathsf{V} \times \mathsf{V}$ with $i \neq j$. Then statement (a) implies statement (b), where

(a) $(i,j) \notin \mathsf{E} \Rightarrow X_i \perp\!\!\!\perp_A X_j \mid X_{-(i,j)}$ (pairwise additive Markov condition);
(b) $X_i \perp\!\!\!\perp_A X_{\mathsf{V}\setminus(\mathsf{N}_i \cup \{i\})} \mid X_{\mathsf{N}_i}$ (local additive Markov condition).

Next we show that, the interdependency defined by functional neighborhood selection in Definition 4 can be fully captured by FARO. The next theorem involves the regression operator $B^0_{X_{-i}X_i}$. Since this is an operator from $\mathcal{H}_{X_i}$ to $\mathcal{H}_{X_{-i}}$, it can be written as a vector of operators $(B_1, \ldots, B_{i-1}, B_{i+1}, \ldots, B_p)$, where each B_k is a mapping from $\mathcal{H}_{X_i}$ to $\mathcal{H}_{X_k}$. For a subset $\mathsf{A} \subseteq \mathsf{V} \setminus \{i\}$, we use $(B^0_{X_{-1}X_i})_\mathsf{A}$ to denote the vector of operators $\{B_k : k \in \mathsf{A}\}$.

Theorem 2. Suppose Assumption 1 holds. Then, for any $\mathsf{N}_i \subseteq \mathsf{V}\backslash\{i\}$, $i = 1, \ldots, p$, we have $(B^0_{X_{-i}X_i})_{\mathsf{V}\setminus(\mathsf{N}_i \cup \{i\})} = 0$ if and only if $X_i \perp\!\!\!\perp_A X_{\mathsf{V}\setminus(\mathsf{N}_i \cup \{i\})} \mid X_{\mathsf{N}_i}$.

Theorem 2 implies that one can recover the graphical model through functional neighborhood selection, by estimating the set of active predictors in regressions with X_i as the response and $X_{\mathsf{V}\setminus\{i\}}$ as the predictors. Also, by the rule of operator inversion, we have $[(B_{X_{-i}X_i})_j(\Sigma_{X_iX_i} - \Sigma_{X_iX_{-i}}\Sigma^{\dagger}_{X_{-i}X_{-i}}\Sigma_{X_{-i}X_i})^{-1}]^* = (B_{X_{-j}X_j})_i(\Sigma_{X_jX_j} - \Sigma_{X_jX_{-j}}\Sigma^{\dagger}_{X_{-j}X_{-j}}\Sigma_{X_{-j}X_j})^{-1}$, which implies that $(B_{X_{-i}X_i})_j = 0$ if and only if $(B_{X_{-j}X_j})_i = 0$. Therefore, by Theorem 2, $j \in \mathsf{N}_i \Leftrightarrow i \in \mathsf{N}_j$.

3. Estimation

In this section, we first develop a population-level objective function whose minimizer is FARO. We then add a mixture of the L_1 and L_2 penalties to the sample-level objective function to obtain a sparse estimate of the FARO, then the functional graphical model.

3.1. The Objective Function

We first introduce some notation. Given two Hilbert spaces $\mathcal{H}$ and $\mathcal{K}$, let $\mathscr{B}(\mathcal{H},\mathcal{K})$ denote 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 $\mathcal{K}$. It can be shown that $\mathscr{B}_1(\mathcal{H},\mathcal{K}) \subseteq \mathscr{B}_2(\mathcal{H},\mathcal{K}) \subseteq \mathscr{B}(\mathcal{H},\mathcal{K})$. Note that $\mathscr{B}_2(\mathcal{H},\mathcal{K})$ is a Hilbert space, and we use $\langle\cdot,\cdot\rangle_{\mathrm{HS}}$ and $\|\cdot\|_{\mathrm{HS}}$ to denote its inner product and norm; $\mathscr{B}(\mathscr{H},\mathscr{K})$ is a Banach space with respect to the operator norm denoted by $\|\cdot\|$. For convenience, $\mathscr{B}_1(\mathcal{H})$, $\mathscr{B}_2(\mathcal{H})$, and $\mathscr{B}(\mathcal{H})$ are used whenever $\mathcal{H} = \mathcal{K}$.

In the classical setting, the neighborhood of each node can be determined by the nonzero linear regression coefficients. To capture the nonlinear relations, we extend linear regression to regression at the operator level. Toward that end, we define the following objective function, $L_i : \mathscr{B}_2(\mathcal{H}_{X_i},\mathcal{H}_{X_{-i}}) \to \mathbb{R}$, for $i = 1,\ldots,p$,

$$L_i(B) = -2\langle\Sigma_{X_{-i}X_i}, B\rangle_{\mathrm{HS}} + \langle B, \Sigma_{X_{-i}X_{-i}}B\rangle_{\mathrm{HS}}. \tag{4}$$

This objective function is motivated by the least squares idea which minimizes $\mathrm{var}(Y - \beta^{\mathsf{T}}X)$ over β. As we show in Theorem 3 below, minimizing (4) is equivalent to minimizing $\sum_{a=1}^{\infty}\mathrm{var}[f_a(X_i) - (Bf_a)(X_{-i})]$, where $\{f_a\}_{a=1}^{\infty}$ is an orthonormal basis in $\mathcal{H}_{X_i}$. As such, (4) is a generalization of least squares by regressing a class of functions of X_i on a class of functions of X_{-i}. We next show that $B_{X_{-i}X_i}$ is the minimizer of $L_i(B)$. We need an additional condition.

Assumption 2. There exists a constant $m > 0$, such that $|\kappa_i(f,g)| \le m$, for any $f,g \in \Omega_{X_i}, i = 1,\ldots,p$.

This condition requires the kernel κ_i to be uniformly bounded, which is satisfied by many commonly used kernels such as the radial basis kernel and the Laplacian kernel. Without loss of generality, we take $m = 1$ in Assumption 2 for all subsequent analyses.

Theorem 3. Suppose Assumptions 1 and 2 hold. Then, for $i = 1,\ldots,p$,

$$\begin{aligned} B^0_{X_{-i}X_i} &= \arg\min\{L_i(B) : B \in \mathscr{B}_2(\mathcal{H}_{X_i},\mathcal{H}_{X_{-i}})\} \\ &= \arg\min\{\sum_{a=1}^{\infty}\mathrm{var}[f_a(X_i) - (Bf_a)(X_{-i})] : B \in \mathscr{B}_2(\mathcal{H}_{X_i},\mathcal{H}_{X_{-i}}), \\ &\qquad \{f_a\}_{a=1}^{\infty} \text{ is an orthonormal basis in } \mathcal{H}_{X_i}\}. \end{aligned}$$

Lee, Li, and Zhao (2016b, Theorem 2) showed a similar result. However, we only require bounded kernels, while Lee, Li, and Zhao (2016b) required the covariance operator to be trace-class.

3.2. Sample-Level Regularized Estimation

We next derive the sample version of $L_i(B)$, and further introduce a mix of L_1 and L_2 penalties. Let $\{(X_1^k,\ldots,X_p^k)^{\mathsf{T}} : k = 1,\ldots,n\}$ be independent copies of $(X_1,\ldots,X_p)^{\mathsf{T}}$. We estimate the mean element μ_{X_i} of $\mathcal{H}_{X_i}$ via $\hat{\mu}_{X_i} = n^{-1}\sum_{k=1}^{n}\kappa_i(\cdot,X_i^k)$, and estimate the covariance operator $\Sigma_{X_iX_j}$ by $\hat{\Sigma}_{X_iX_j} = n^{-1}\sum_{k=1}^{n}[\kappa_i(\cdot,X_i^k) - \hat{\mu}_{X_i}] \otimes [\kappa_j(\cdot,X_j^k) - \hat{\mu}_{X_j}]$, where $\otimes$ denotes the tensor product. We then use $\hat{\Sigma}_{X_iX_j}$ to build up $\hat{\Sigma}_{UV}$ for any subvectors U, V of X. With the covariance operators in (4) substituted by their empirical counterparts, the sample version of $L_i(B)$ becomes

$$\hat{L}_i(B) = -2\langle\hat{\Sigma}_{X_{-i}X_i}, B\rangle_{\mathrm{HS}} + \langle B, \hat{\Sigma}_{X_{-i}X_{-i}}B\rangle_{\mathrm{HS}}. \tag{5}$$

To encourage sparsity and enhance prediction, we introduce two penalty terms on $\hat{L}_i(B)$:

$$\widehat{PL}_i(B) = \hat{L}_i(B) + \lambda_n\|B\|^2_{\mathrm{HS}} + \lambda_n(\sum_{j\in\mathsf{V}\setminus\{i\}}\|(B)_j\|_{\mathrm{HS}})^2. \tag{6}$$

The first penalty term $\|B\|^2_{\mathrm{HS}}$ is the Hilbert-Schmidt norm and is similar to an L_2 penalty, while the second is equivalent to $(\sum_{j\in\mathsf{V}\setminus\{i\}}\|(B)_j\|_{\mathrm{HS}})$ that resembles an L_1 penalty. This type of mixture of L_1 and L_2 penalties are often employed in high-dimensional regressions (Zou and Hastie 2005; Lee, Li, and Zhao 2016b). We use it here to achieve desired asymptotic properties. Also, to simplify tuning and theoretical development, we impose the same parameter λ_n for both penalty terms, but in principle they can be different.

Let $\hat{B}_i = \arg\min\{\widehat{PL}_i(B) : B \in \mathscr{B}_2(\mathcal{H}_{X_i},\mathcal{H}_{X_{-i}})\}$. We then use $\hat{\mathsf{N}}_i = \{\|(\hat{B}_i)_j\|_{\mathrm{HS}} \neq 0 : j \in \mathsf{V}\setminus\{i\}\}$ to estimate the neighborhood of i. Subsequently, we use $\hat{\mathsf{N}}_i$ to estimate E via

$$\begin{aligned} \hat{\mathsf{E}}_{\mathrm{OR}} &= \{(i,j) : i \in \hat{\mathsf{N}}_j \text{ or } j \in \hat{\mathsf{N}}_i\} \quad \text{or} \\ \hat{\mathsf{E}}_{\mathrm{AND}} &= \{(i,j) : i \in \hat{\mathsf{N}}_j \text{ and } j \in \hat{\mathsf{N}}_i\}. \end{aligned} \tag{7}$$

These are two slightly different ways to construct an estimate of E in (7), because it may happen that $j \in \hat{\mathsf{N}}_i$ but $i \notin \hat{\mathsf{N}}_j$, and vise versa. However, as we show later in Section 4.3, this type of discrepancy, and thus, the difference between $\hat{\mathsf{E}}_{\mathrm{OR}}$ and $\hat{\mathsf{E}}_{\mathrm{AND}}$, is asymptotically negligible. In our implementation in Section 5.2, we choose $\hat{\mathsf{E}}_{\mathrm{OR}}$ as our final estimate of the functional graphical model.

4. Theory

In this section, we derive the asymptotic property of the sparse FARO estimator $\hat{B}_i$. We then derive the consistency of the subsequent neighborhood estimator $\hat{\mathsf{N}}_i$ and the graph estimators $\hat{\mathsf{E}}_{\mathrm{OR}}, \hat{\mathsf{E}}_{\mathrm{AND}}$ based on $\hat{B}_i$. We achieve our goal in four steps. First, we derive a concentration bound on the sample covariance operator. Second, we introduce an intermediate estimator, $\hat{B}_i^0$, and derive the concentration bound for the distance between the intermediate estimator and the true FARO. Third, we show that we can construct a minimizer of the original objective function (6) based on $\hat{B}_i^0$ with a high probability. Combined with the second step, this in effect establishes the consistency and convergence rate of the minimizer $\hat{B}_i$ of (6). Finally, we establish the desired neighborhood selection and the graph estimation consistency built on $\hat{B}_i$. We mostly assume that the trajectory of X_i is fully observed on $t \in T$, for $i = 1, \ldots, p$. We briefly discuss the scenario when X_i is partially observed in Section 4.4. In the interest of space, we relegate some technical results to the supplementary material.

4.1. A Key Lemma and an Intermediate Estimator

We begin with a lemma that establishes a key concentration inequality for the norm $\|\hat{\Sigma}_{X_{\mathsf{A}}X_{\mathsf{B}}} - \Sigma_{X_{\mathsf{A}}X_{\mathsf{B}}}\|_{\mathrm{HS}}$. Let $|\mathsf{A}|$ denote the cardinality of a set A.

Lemma 1. Suppose Assumption 2 holds. Then, for any $\delta > 0$,

$$P\left(\|\hat{\Sigma}_{X_{\mathsf{A}}X_{\mathsf{B}}} - \Sigma_{X_{\mathsf{A}}X_{\mathsf{B}}}\|_{\mathrm{HS}} \geq \delta\right) = O\left\{\exp\left(-\frac{n\delta^2}{|\mathsf{A}|\,|\mathsf{B}|}\right)\right\}.$$

Since both $|\mathsf{A}|$ and $|\mathsf{B}|$ can be arbitrary, this lemma extends the bound for a high-dimensional covariance matrix to the operator level. Since the estimation of covariance plays a crucial role in many topics in high-dimensional statistics, we expect this result to be useful in other contexts involving high-dimensional matrices of linear operators. We note that Bosq (2000) has studied the concentration bound of the empirical covariance operator for functional data too. However, our result extends his to both nonlinear and high-dimensional settings.

We next derive the concentration bound for the FARO estimator $\hat{B}_i$. Toward that end, we introduce an intermediate estimator. Consider the following objective function,

$$\begin{aligned}\widehat{PL}_i^{\,0}(B) = &-2\langle\hat{\Sigma}_{X_{\mathsf{N}_i}X_i}, B\rangle_{\mathrm{HS}} + \langle B, \hat{\Sigma}_{X_{\mathsf{N}_i}X_{\mathsf{N}_i}}B\rangle_{\mathrm{HS}} \\ &+ \lambda_n\|B\|_{\mathrm{HS}}^2 + \lambda_n\|B\|_{\mathrm{hb}}^2, \end{aligned} \tag{8}$$

where $\|B\|_{\mathrm{HB}} = \sum_{j\in\mathsf{N}_i}\|(B)_j\|_{\mathrm{HS}}$. Let $\hat{B}_i^0 = \arg\min\{\widehat{PL}_i^{\,0}(B) : B \in \mathscr{B}_2(\mathcal{H}_{X_i}, \mathcal{H}_{X_{\mathsf{N}_i}})\}$. Note that the objective function (8) differs from (6) in that (8) treats $\mathsf{N}_1, \ldots, \mathsf{N}_p$ as known. Besides, the dimension of the minimizer $\hat{B}_i^0$ of (8) is $|\mathsf{N}_i|$, whereas the dimension of the minimizer $\hat{B}_i$ of (6) is $(p-1)$. We first establish the concentration bound for $\hat{B}_i^0$. Let $C_{\mathsf{N}_i\mathsf{N}_i}$ be the correlation operator from $\mathcal{H}_{X_{\mathsf{N}_i}}$ to $\mathcal{H}_{X_{\mathsf{N}_i}}$ that satisfies $\Sigma_{X_{\mathsf{N}_i}X_{\mathsf{N}_i}} = D_{\mathsf{N}_i}C_{\mathsf{N}_i\mathsf{N}_i}D_{\mathsf{N}_i}$, where D_{N_i} is a diagonal matrix of the operators with diagonal entries $(D_{\mathsf{N}_i})_{kk} = \Sigma_{X_kX_k}^{1/2}$, for $k \in \mathsf{N}_i$. The existence and uniqueness of the correlation operator $C_{\mathsf{N}_i\mathsf{N}_i}$ was established by Baker (1973). We require another two assumptions.

Assumption 3. There exists a constant $c > 0$ such that, for all $i = 1, \ldots, p$, $cI_{\mathsf{N}_i} \leq C_{\mathsf{N}_i\mathsf{N}_i}$, where $I_{\mathsf{N}_i} : \mathcal{H}_{X_{\mathsf{N}_i}} \to \mathcal{H}_{X_{\mathsf{N}_i}}$ is the identity mapping.

Assumption 3 holds for all $i \in \mathsf{V}$, if there exists $c > 0$ such that the joint correlation operator $C_{\mathsf{V}\mathsf{V}}$ is bounded below by cI. This is a fairly general condition, and it holds when $C_{\mathsf{V}\mathsf{V}}$ is invertible and all its off-diagonal elements $C_{i,j}$, $i, j \in \mathsf{V}\times\mathsf{V}$ with $i \neq j$, are compact; see also Solea and Li (2020, Proposition 2). Fukumizu, Bach, and Gretton (2007) has studied the condition for the compact operators, and showed that the correlation operator is compact when the mean square contingency of the associated random elements is finite, which in general requires that there cannot be too strong dependency between the random elements. We also note that Zhao and Yu (2006); Wainwright (2009); Ravikumar et al. (2009) all imposed a similar condition in the linear model or the generalized additive model settings to derive the consistency of LASSO. Moreover, under the Additive Gaussian distribution setting as discussed in Section 2.2, Assumption 3 holds, because all the pairwise correlation operators have finite ranks.

Suppose H_i^0 is an $|\mathsf{N}_i| \times |\mathsf{N}_i|$ diagonal matrix of operators with the diagonal entries $(H_i^0)_{jj} = (1 + \|B_i^0\|_{\mathrm{HB}}/\|(B_i^0)_j\|_{\mathrm{HS}})\mathbf{1}_j$, and $\mathbf{1}_j : \mathcal{H}_{X_j} \to \mathcal{H}_{X_j}$ is the identity mapping, $j \in \mathsf{N}_i$.

Assumption 4. For $i = 1, \ldots, p$, there exists an operator $\mathsf{C}_i^0 \in \mathscr{B}_2(\mathcal{H}_{X_i}, \mathcal{H}_{X_{\mathsf{N}_i}})$, such that $B_i^0 = (H_i^0)^{-1}\Sigma_{X_{\mathsf{N}_i}X_{\mathsf{N}_i}}\mathsf{C}_i^0$. Moreover, $\|(\mathsf{C}_i^0)_j\|_{\mathrm{HS}} \leq c_1$ for all $j \in \mathsf{N}_i$ and some constant $c_1 > 0$.

Note that Assumption 4 is satisfied, if $\|[\Sigma_{X_{\mathsf{N}_i}X_{\mathsf{N}_i}}(H_i^0)^{-1}\Sigma_{X_{\mathsf{N}_i}X_{\mathsf{N}_i}}]^{\dagger}\Sigma_{X_{\mathsf{N}_i}X_i}\|_{\mathrm{HS}} \leq c_1$, for all $i \in \mathsf{V}$. Under the additive Gaussian distribution setting, this is equivalent to

$$\begin{aligned}&\|\mathrm{diag}\{L^{1/2}, \ldots, L^{1/2}\}\left[\{(\Sigma_\alpha)_{j,k} : j,k \in \mathsf{N}_i\}(H_i^0)^{-1}\{(\Sigma_\alpha)_{j,k} : j,k \in \mathsf{N}_i\}\right]^{-1} \\ &\qquad \times \{(\Sigma_\alpha)_{j,i} : j \in \mathsf{N}_i\}L^{1/2}\|_{\mathrm{F}} \leq c_1, \end{aligned} \tag{9}$$

where $(\Sigma_\alpha)_{i,j}$ is the (i,j)th block of Σ_α, $[H_i^0]$ is a $\{c(m,\alpha)|\mathsf{N}_i|\} \times \{c(m,\alpha)|\mathsf{N}_i|\}$ diagonal block matrix with $[H_i^0]_{jj} = (1 + \|B_i^0\|_{\mathrm{HB}}/\|(B_i^0)_j\|_{\mathrm{HS}})I_{c(m,\alpha)}$, for $j \in \mathsf{N}_i$, L is the Gram kernel matrix $\{\langle q_s, q_t\rangle_{\mathcal{H}_{X_i}}\}_{s,t=1}^{c(m,\alpha)}$, and $\|\cdot\|_F$ is the Frobenius norm. Condition (9) is essentially a form of smoothness in the relation between X_i and its neighborhood X_{N_i}. To see this, note that (9) implies that $\|\{(\Sigma_\alpha)_{j,k} : j,k \in \mathsf{N}_i\}^{-1}\{(\Sigma_\alpha)_{j,i} : j \in \mathsf{N}_i\}\|_{\mathrm{F}}$ is uniformly bounded, which means the Frobenius norm of the regression coefficient for regressing $Q_\alpha(U_i)$ on $\{Q_\alpha(U_j) : j \in \mathsf{N}_i\}$ is uniformly bounded. Li and Song (2017) used a similar condition and referred it as "collective smoothness" in the context of nonlinear dimension reduction. The next proposition shows that, if X follows an Additive Gaussian distribution with some additional conditions, then Assumption 4 is satisfied for each $i \in \mathsf{V}$.

Proposition 5. Suppose the random function $X = (X_1, \ldots, X_p)^{\mathsf{T}} \sim \mathrm{AN}(\mu_\alpha, \Sigma_\alpha)$, with (a) $\sigma_{\min}(\Sigma_\alpha) \geq c_1$, (b) $\max\{|\mathsf{N}_i| : i \in \mathsf{V}\} \leq c_2$, and (c) $\min\left\{\|(\Sigma_\alpha^{-1})_{i,j}\|_{\mathrm{F}} : (\Sigma_\alpha^{-1})_{i,j} \neq 0, (i,j) \in \mathsf{V}\times\mathsf{V}, i \neq j\right\} \geq c_3$, where c_1, c_2, c_3 are positive constants, and $\sigma_{\min}(\cdot)$ is the minimum eigenvalue of the designated matrix. Then, for each $i \in \mathsf{V}$, there exists an operator $\mathsf{C}_i^0 \in \mathscr{B}_2(\mathcal{H}_{X_i}, \mathcal{H}_{X_{\mathsf{N}_i}})$ and

$c_0 > 0$, such that $B_i^0 = (H_i^0)^{-1}\Sigma_{X_{\mathsf{N}_i}X_{\mathsf{N}_i}}\mathsf{C}_i^0$ with $\|(\mathsf{C}_i^0)_j\|_{\mathrm{HS}} \le c_0$ for all $j \in \mathsf{N}_i$.

Now we are ready to derive the concentration bound and the convergence rate of the intermediate estimator $\hat{B}_i^0$. Hereafter, for two positive sequences $\{a_n\}$ and $\{b_n\}$, let $a_n \preceq b_n$ represent $a_n = O(b_n)$; let $a_n \prec b_n$ represent $a_n = o(b_n)$; $a_n \wedge b_n = a_n$ and $a_n \vee b_n = b_n$ if $a_n \preceq b_n$. Similarly, if c_n is a third sequence and $\preceq$ orders $\{a_n\}$, $\{b_n\}$ and $\{c_n\}$, then we use the notations $a_n \wedge b_n \wedge c_n = (a_n \wedge b_n) \wedge c_n$, and $a_n \vee b_n \vee c_n = (a_n \vee b_n) \vee c_n$. Let $\mathsf{b}_i^0 \equiv \min\{\|(B_i^0)_j\|_{\mathrm{HS}} : j \in \mathsf{N}_i\}$.

Theorem 4. Suppose Assumptions 1–4 hold. Then,

(a) The concentration bound: $P(\|\hat{B}_i^0 - B_i^0\|_{\mathrm{HS}} \ge \delta) = O\{\exp\left(-n|\mathsf{N}_i|^{-7}(\lambda_n \mathsf{b}_i^0 \delta \|B_i^0\|_{\mathrm{HB}}^{-1})^2\right)\}$, for $\delta > 0$ and λ_n with $\lambda_n \preceq \mathsf{b}_i^0 \delta\, |\mathsf{N}_i|^{-3/2}\|B_i^0\|_{\mathrm{HB}}^{-1}$.

(b) The convergence rate: $\|\hat{B}_i^0 - B_i^0\|_{\mathrm{HS}} = O_p\{(\mathsf{b}_i^0)^{-1}|\mathsf{N}_i|^{3/2}\|B_i^0\|_{\mathrm{HB}}(\lambda_n \vee n^{-1/2}\lambda_n^{-1}\,|\mathsf{N}_i|^2)\}$.

4.2. Consistency and Convergence Rate of the FARO Estimator

We next construct an estimator based on the intermediate estimator $\hat{B}_i^0$, and show that, with a high probability, it minimizes the objective function $\widehat{PL}_i(B)$ in (6). Coupled with Theorem 4, this in effect establishes the concentration bound and convergence rate of the FARO estimator $\hat{B}_i$ of (6). Specifically, we construct an operator in $\mathscr{B}_2(\mathcal{H}_{X_i}, \mathcal{H}_{X_{-i}})$ whose corresponding N_i-subvector is equal to $\hat{B}_i^0 \in \mathscr{B}(\mathcal{H}_{X_i}, \mathcal{H}_{X_{\mathsf{N}_i}})$, and the rest is a $(p-1-|\mathsf{N}_i|)$-dimensional zero-operator $\mathbf{0}$. To avoid overly complicated notation, we denote this operator by $(\hat{B}_i^0, \mathbf{0})$, keeping in mind that $\hat{B}_i^0$ doesn't have to occupy the first $|\mathsf{N}_i|$ positions. We derive a series of inequalities under which $(\hat{B}_i^0, \mathbf{0})$ satisfies the Karush-Kuhn-Tucker (KKT) conditions with a high probability. This is equivalent to saying that $(\hat{B}_i^0, \mathbf{0})$ and the FARO estimator $\hat{B}_i$ are asymptotically equivalent. In the interest of space, we give the full details in Section S3 in the supplementary material. We also introduce another assumption,

Assumption 5. There exists $0 < \eta \le 1$ such that, for any $j \in \mathsf{V}\backslash(\mathsf{N}_i \cup \{i\})$, and $i = 1, \ldots, p$,

$$\|D_j C_{j\mathsf{N}_i} C_{\mathsf{N}_i\mathsf{N}_i}^{-1}\|_{\mathrm{HS}} \le \frac{\|B_i^0\|_{\mathrm{HB}}}{\|C_{\mathsf{N}_i\mathsf{N}_i} D_{\mathsf{N}_i} \mathsf{C}_i^0\|_{\mathrm{HS}}}(1-\eta).$$

Assumption 5 can be viewed as a generalized version of the irrepresentable condition usually imposed in the classical regression setting to establish the consistency of LASSO (Zhao and Yu 2006; Wainwright 2009). Under the Additive Gaussian distribution, we have $\|\left[\{(\Sigma_\alpha)_{j,k} : j,k \in \mathsf{N}_i\}\right]^{-1}\{(\Sigma_\alpha)_{j,j'} : j \in \mathsf{N}_i\}\|_{\mathrm{F}} \le 1-\eta$, for all $j' \in \mathsf{V}\backslash(\mathsf{N}_i \cup \{i\})$. This simply means that the Frobenius norm of the regression coefficient for regressing $Q_\alpha(X_{j'})$, which are the random elements at the nonneighboring nodes, on $\{Q_\alpha(X_j) : j \in \mathsf{N}_i\}$, which are the random elements at neighboring nodes, is uniformly bounded by $1-\eta$. Since $Q_\alpha(X_j)$ are random vectors, this condition is similar in spirit to its counterpart in the random variable case; that is, it avoids strong dependency between the nonneighboring and neighboring nodes. Besides, Qiao, Guo, and James (2019, Condition 5) used a similar condition for the Gaussian functional graphical model. Our Assumption 5 can be seen as a nonlinear extension of their condition. We also give an example under which Assumption 5 is satisfied for each $i \in \mathsf{V}$.

Example 3. Suppose $X = (X_1, X_2, X_3)^{\mathsf{T}}$ follows an Additive Gaussian distribution, such that, for each $i = 1, 2, 3$, (a) $X_i = U_{i1}\eta_1 + U_{i2}\eta_2 \in \Omega_{X_i}$, where Ω_{X_i} is the linear span of the orthonormal basis $\{\eta_1, \eta_2\}$; (b) $\kappa_i(x, x') = 1 + \langle x, x'\rangle_{\Omega_{X_i}}$; and (c) the covariance matrix of $U = (U_{11}, U_{12}, U_{21}, U_{22}, U_{31}, U_{32})^{\mathsf{T}}$ is, with $-1 < \beta_1,\ \beta_2 < 1$ being some constants,

$$\Sigma_U = \begin{pmatrix} 1 & 0 & \beta_1 & 0 & \beta_1 & 0 \\ 0 & 1 & 0 & \beta_2 & 0 & \beta_2 \\ \beta_1 & 0 & 1 & 0 & \beta_1^2 & 0 \\ 0 & \beta_2 & 0 & 1 & 0 & \beta_2^2 \\ \beta_1 & 0 & \beta_1^2 & 0 & 1 & 0 \\ 0 & \beta_2 & 0 & \beta_2^2 & 0 & 1 \end{pmatrix}.$$

The proof for this example is given in the supplementary material. Similar to the classical neighborhood selection, when Assumption 5 does not hold, we still expect the support of our functional neighborhood selection to recover the true graph to a certain extent, in the sense that the probability of erroneous selection converges to a small positive constant instead of zero. Moreover, in the usual regression setting, alternative regularization methods, such as adaptive LASSO and SCAD, may be employed to relax the irrepresentable condition. We expect that similar modifications can be made to FARO, so that Assumption 5 can be removed.

The next corollary provides the convergence rate of $\hat{B}_i$, and the connection between the convergence rate and the λ_n and p.

Corollary 1. Suppose Assumptions 1–5 hold, $|\mathsf{N}_i|$ does not depend on n, and $\{(\log p)/n\}^{1/3} \prec \lambda_n \prec 1$. Then, $\|\hat{B}_i - (B_i, \mathbf{0})\|_{\mathrm{HS}} = O_p(\lambda_n \wedge n^{-1/2}\lambda_n^{-1})$.

We remark that the convergence rate of FARO in terms of the Hilbert-Schmidt norm depends on p through λ_n, whose order of magnitude can be arbitrarily close to $(\log p/n)^{1/3}$. In the classical linear regression, the convergence rate of the estimated regression coefficient from LASSO and Dantzig estimators in terms of the L_2 norm depends on $(\log p/n)^{1/2}$ (Bickel, Ritov, and Tsybakov 2009; Fan and Lv 2010). This discrepancy is somehow expected though, as our setting is more general and involves both nonlinearity and high-dimensional functional data.

4.3. Neighborhood Selection and Graph Estimation Consistency

We now establish the asymptotics of neighborhood selection. The next theorem provides an upper bound for the probability of incorrectly selecting the neighbors.

Theorem 5. Suppose Assumptions 1–5 hold, $\lambda_n \prec 1$, $\lambda_n^{1/2} \preceq (\mathsf{b}_i^0)^3\,|\mathsf{N}_i|^{-11/4}\,\|B_i^0\|_{\mathrm{HB}}^{-2}$. Then,

$$P(\hat{\mathsf{N}}_i \neq \mathsf{N}_i) = (p - |\mathsf{N}_i|) \left(O\left\{ \exp\left(-\frac{n\Delta_1^2}{|\mathsf{N}_i|^2} \right) \right\} + O\left\{ \exp\left(-\frac{n(\lambda_n \mathsf{b}_i^0 \Delta_2)^2}{|\mathsf{N}_i|^7 \, \|B_i^0\|_{\mathrm{HB}}^2} \right) \right\} \right),$$

where $\Delta_1 = \lambda_n^{3/2} |\mathsf{N}_i|^{-3/4}$, and $\Delta_2 = \lambda_n^{1/2} \, |\mathsf{N}_i|^{-5/4} \, (\mathsf{b}_i^0)^2 \, \|B_i^0\|_{\mathrm{HB}}^{-1}$.

In the classical regression setting, the sparsistency of LASSO has been studied in Zhao and Yu (2006); Wainwright (2009). Theorem 5 establishes the sparsistency in a much more general setting where both response and predictors are random functions, and no structural assumptions such as linearity are imposed on their relationships.

The next theorem shows that, using (7), E can be correctly recovered with probability tending to one. Moreover, the difference between $\hat{\mathsf{E}}_{\mathrm{OR}}$ and $\hat{\mathsf{E}}_{\mathrm{AND}}$ is asymptotically negligible. Its proof follows immediately from Theorem 5, and is omitted.

Theorem 6. Suppose the same conditions of Theorem 5 hold. Moreover, suppose $|\mathsf{N}_1|, \ldots, |\mathsf{N}_p|$ do not depend on n, and $\hat{\mathsf{E}}$ can be either $\hat{\mathsf{E}}_{\mathrm{OR}}$ or $\hat{\mathsf{E}}_{\mathrm{AND}}$ defined in (7). Then, there exists $c_2 > 0$ such that $P(\hat{\mathsf{E}} \neq \mathsf{E}) = O\left\{ p^2 \, \exp\left(-c_2 n \lambda_n^3 \right) \right\}$.

We make some remarks regarding Corollary 1 and Theorem 6. First, by Corollary 1, the convergence rate of our FARO estimator depends on $\lambda_n + n^{-1/2}\lambda_n^{-1}$, with the parameter λ_n satisfying that $(\log p/n)^{1/3} \prec \lambda_n \prec 1$. As a comparison, Li and Solea (2018, Theorem 4) showed that the rate of convergence for their FACI estimator was $n^{-1/6}$. However, they treated the number of functions p as fixed, while we allow p to grow with n in an exponential order. If we also treat p as fixed, then our rate can be made arbitrarily close to $n^{-1/6}$. Second, in the same vein, our consistency of graph estimation in Theorem 6 holds while allowing the graph size to diverge at an exponential order, whereas Li and Solea (2018) treated p as fixed. In fact, their estimator did not take advantage of the sparsity of the graph, and needed to estimate all the off-diagonal elements on their precision operator. Since the cardinality of all the off-diagonal elements grows in the order of p^2, this means p can grow only in a polynomial rate of n, but not in an exponential rate as in our result. Finally, we note that, in both Corollary 1 and Theorem 6, we assume the number of neighborhoods for each node fixed. This condition can be relaxed by carefully choosing the rates of p, $|\mathsf{N}_i|$, $\|B_i^0\|_{\mathrm{HS}}$, and b_i^0, which we leave as potential future research.

4.4. Consistency for Partially Observed Random Functions

We have so far assumed that X_i is fully observed. Next, we briefly study the scenario when the function is only partially observed. See Wang, Chiou, and Muller (2016) for discussions on different schedules on which functional data are collected. Note that Theorems 5 and 6 only rely on the concentration bound of the sample covariance operator in Lemma 1. In the following, we allow the convergence rate under a partially observed schedule to be slower than the one under a fully observed schedule. However, we do not pursue any specific measurement schedule or smoothing setting to avoid digressing too much from the main theme.

Suppose $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)^{\mathsf{T}}$ is an estimator of X. We then estimate the sample covariance operator by

$$\tilde{\Sigma}_{X_j X_i} = E_n[\kappa_i(\cdot, \tilde{X}_j) - E_n \kappa_i(\cdot, \tilde{X}_j)] \otimes [\kappa_i(\cdot, \tilde{X}_i) - E_n \kappa_i(\cdot, \tilde{X}_i)].$$

for $i, j = 1, \ldots, p$. For any subvectors U, V of X, let $\tilde{\Sigma}_{UV}$ be the matrix of operators whose elements are composed of $\tilde{\Sigma}_{X_j X_i}$ with $X_j \in U$ and $X_i \in V$. Then we compute the new penalized estimator of B by $\tilde{B}_i = \arg\min\{\widetilde{PL}_i(B) : B \in \mathscr{B}_2(\mathcal{H}_{X_i}, \mathcal{H}_{X_{-i}})\}$, where $\widetilde{PL}_i(B)$ is obtained by substituting $\hat{\Sigma}_{X_{-i}X_i}$ and $\hat{\Sigma}_{X_{-i}X_{-i}}$ with $\tilde{\Sigma}_{X_{-i}X_i}$ and $\tilde{\Sigma}_{X_{-i}X_{-i}}$ in (5) and (6) accordingly. Finally, we estimate the neighborhood and the graph by

$$\tilde{\mathsf{N}}_i = \{\|(\tilde{B}_i)_j\|_{\mathrm{HS}} > 0 : j \in \mathsf{V} \backslash \{i\}\},$$
$$\tilde{\mathsf{E}}_{\mathrm{OR}} = \{(i,j) : \, i \in \tilde{\mathsf{N}}_j \text{ or } j \in \tilde{\mathsf{N}}_i\} \quad \text{or}$$
$$\tilde{\mathsf{E}}_{\mathrm{AND}} = \{(i,j) : \, i \in \tilde{\mathsf{N}}_j \text{ and } j \in \tilde{\mathsf{N}}_i\}.$$

The next theorem shows the consistency of our method under a partially observed schedule, which is a generalization of Theorems 5 and 6. Its proof follows immediately from that of Theorem 5 and is thus, omitted.

Theorem 7. Suppose Assumptions 1–5 hold, $\lambda_n \prec 1$, and $\lambda_n^{1/2} \preceq (\mathsf{b}_i^0)^3 \, |\mathsf{N}_i|^{-11/4} \, \|B_i^0\|_{\mathrm{HB}}^{-2}$. Moreover, suppose there exists $0 < \alpha \leq 1$ such that, for any $\delta > 0$,

$$P\left(\|\tilde{\Sigma}_{X_{\mathsf{A}} X_{\mathsf{B}}} - \Sigma_{X_{\mathsf{A}} X_{\mathsf{B}}}\|_{\mathrm{HS}} \geq \delta \right) = O\left\{ \exp\left(-\frac{n^\alpha \delta^2}{|\mathsf{A}| \, |\mathsf{B}|} \right) \right\}.$$

Then we have,

$$P(\hat{\mathsf{N}}_i \neq \mathsf{N}_i) = (p - |\mathsf{N}_i|) \left(O\left\{ \exp\left(-\frac{n^\alpha \Delta_1^2}{|\mathsf{N}_i|^2} \right) \right\} + O\left\{ \exp\left(-\frac{n^\alpha (\lambda_n \mathsf{b}_i^0 \Delta_2)^2}{|\mathsf{N}_i|^7 \, \|B_i^0\|_{\mathrm{HB}}^2} \right) \right\} \right),$$

where $\Delta_1 = \lambda_n^{3/2} |\mathsf{N}_i|^{-3/4}$, and $\Delta_2 = \lambda_n^{1/2} \, |\mathsf{N}_i|^{-5/4} \, (\mathsf{b}_i^0)^2 \, \|B_i^0\|_{\mathrm{HB}}^{-1}$. Moreover, if $|\mathsf{N}_1|, \ldots, |\mathsf{N}_p|$ do not depend on n, and $\tilde{\mathsf{E}}$ can be either $\tilde{\mathsf{E}}_{\mathrm{OR}}$ or $\tilde{\mathsf{E}}_{\mathrm{AND}}$, then, there exists $c_3 > 0$ such that

$$P(\tilde{\mathsf{E}} \neq \mathsf{E}) = O\left\{ p^2 \, \exp\left(-c_3 n^\alpha \lambda_n^3 \right) \right\}.$$

5. Implementation

In this section, we introduce a coordinate system to implement the estimator developed at the operator-level in Section 3.

5.1. Coordinate Representation

We first develop the coordinate system for $\Omega_{X_i}, \Omega_X, \mathcal{H}_{X_i}$, and $\mathcal{H}_X$. For a generic finite-dimensional Hilbert space Ω spanned by $\mathcal{B} = \{b_1, \ldots, b_m\}$, any $x \in \Omega$ can be written as $\sum_{u=1}^m \alpha_u b_u$. We call the vector $(\alpha_1, \ldots, \alpha_m)^{\mathsf{T}}$ the coordinate of x relative to the spanning system $\mathcal{B}$, and write it as $[x]_{\mathcal{B}} = ([x]_{\mathcal{B},1}, \ldots, [x]_{\mathcal{B},m})^{\mathsf{T}}$. For any pair $(x_1, x_2) \in \Omega$, the inner product $\langle x_1, x_2 \rangle_\Omega = [x_1]_{\mathcal{B}}^{\mathsf{T}} K_{\mathcal{B}} [x_2]_{\mathcal{B}}$, where $K_{\mathcal{B}} = [\langle b_u, b_v \rangle_\Omega]_{u,v=1}^m$ is the Gram matrix of $\mathcal{B}$.

Let $(X^1, \ldots, X^n)$ denote iid samples from X of size n and X_i^k denote the ith component of X^k, $k = 1, \ldots, n$. Suppose X_i^k is observed on a finite subset $T_k = \{t_{k1}, \ldots, t_{km_k}\}$ of T, where m_k is the number of time points observed for subject k. Let $(\tau_1, \ldots, \tau_M) = \cup_{k=1}^n T_k$ denote all the unique time points ordered from the smallest to the largest, where M is its cardinality. Let $\kappa_T : T \times T \to \mathbb{R}$ be a positive definite kernel. We consider the reproducing kernel Hilbert space $\Omega^n = \text{span}\{\kappa_T(\cdot, \tau_1), \ldots, \kappa_T(\cdot, \tau_M)\} \equiv \text{span}\{B_u^n : u = 1, \ldots, M\}$, with the inner product determined by $\langle B_u^n, B_v^n \rangle_{\Omega^n} = \kappa_T(\tau_u, \tau_v)$ for $u, v = 1, \ldots, M$. Since X_i^k is only observed at the m_k time points in T_k, we use $\kappa_T(\cdot, T_k) \equiv \{\kappa_T(\cdot, t_{k1}), \ldots, \kappa_T(\cdot, t_{km_k})\}$ to construct X_i^k. That is, $X_i^k = \sum_{u=1}^{m_k} [X_i^k]_u \kappa_T(\cdot, t_{ku})$, implying that

$$X_i^k(T_k) \equiv \left(X_i^k(t_{k1}), \ldots, X_i^k(t_{km_k})\right)^\mathsf{T} = (K_T^{k,k})[X_i^k], \tag{10}$$

where $K_T^{k,k}$ is the $m_k \times m_k$ matrix $[\kappa_T(t_{ku}, t_{kv})]_{u,v=1,\ldots,m_k}$.

From (10), we estimate the coordinate $[X_i^k]$ by $[X_i^k] = (K_T^{k,k} + \epsilon_T^k I_{m_k})^{-1} X_i^k(T_k)$, where ϵ_T^k is a ridge-regression-type tuning parameter. Note that the coordinates are not unique when the spanning system is linearly dependent. Nevertheless, both the inner product and the norm it induces are unique. This is because the inner product, like eigenvalues and eigenfunctions, is coordinate-free. We then compute the inner product between X_i^k and X_i^ℓ by,

$$\langle X_i^k, X_i^\ell \rangle_{\Omega^n} = X_i^k(T_k)^\mathsf{T} (K_T^{k,k} + \epsilon_T I_{m_k})^{-1} K_T^{k,\ell} (K_T^{\ell,\ell} + \epsilon_T I_{m_k})^{-1} X_i^\ell(T_\ell), \tag{11}$$

where $K_T^{k,\ell} = [\kappa_T(t_{ku}, t_{\ell v})]_{u=1,\ldots,m_k, v=1,\ldots,m_\ell}$, for $k, \ell \in \{1, \ldots, n\}$.

Having constructed X_i^k, we next proceed to the construction of the sample version of $\mathcal{H}_{X_i}$, which we denote by $\mathcal{H}_{X_i}^n$. Letting $\kappa_i(\cdot, \cdot)$ be the second-level kernel that can be computed via Definition 1 and Equation (11), we define $\phi_i^k = \kappa_i(\cdot, X_i^k) - n^{-1} \sum_{k=1}^n \kappa_i(\cdot, X_i^k)$, $k = 1, \ldots, n$. Let $\mathcal{H}_{X_i}^n$ be the RKHS spanned by $\{\phi_i^k : k = 1, \ldots, n\}$, and K_i be the Gram matrix $[\kappa_i(X_i^k, X_i^\ell)]_{k,\ell=1,\ldots,n}$. Let $G_i = Q_n K_i Q_n$ be the centered version of K_i, where $Q_n = I_n - n^{-1} \mathbf{1}_n^\mathsf{T} \mathbf{1}_n$ with I_n and $\mathbf{1}_n$ being the identity matrix and the n-dimensional vector $(1, \ldots, 1)^\mathsf{T}$.

It is often the case that the important features of a kernel are concentrated on leading eigenfunctions (Lee and Huang 2007; Chen et al. 2010). So, without losing much efficiency, we may use the leading eigenfunctions to construct the empirical RKHS, which can bring substantial saving of computing time. Suppose G_i has the eigen-decomposition,

$$G_i = V_i D_i V_i^\mathsf{T} + \tilde{V}_i \tilde{D}_i \tilde{V}_i^\mathsf{T}, \tag{12}$$

where $V_i D_i V_i^\mathsf{T}$ and $\tilde{V}_i \tilde{D}_i \tilde{V}_i^\mathsf{T}$ correspond to the first n_i and the last $n - n_i$ eigenvalues of G_i. Let $\psi_i^k = (D_i)_{kk}^{-1/2} V_i^\mathsf{T} (\phi_i^1, \ldots, \phi_i^n)^\mathsf{T}$. We then use $\{\psi_i^1, \ldots, \psi_i^{n_i}\}$ as a basis of the reduced space $\mathcal{H}_{X_i}^{n_i}$, by which we replace $\mathcal{H}_{X_i}^n$ to save computing time. We lose no information as long as $\text{ran}(\Sigma_{X_i X_{-i}}) \subseteq \text{span}\{\psi_i^1, \ldots, \psi_i^{n_i}\}$. In the following, we denote $(\psi_i^1, \ldots, \psi_i^{n_i})^\mathsf{T}$ by ψ_i.

Using the coordinate representations and the reduced space derived above, we next provide the numerical procedure to implement the constrained optimization problem in Section 3.2. Let B_i be an operator in $\mathscr{B}_2(\mathcal{H}_{X_i}^{n_i}, \oplus_{j \neq i} \mathcal{H}_{X_j}^{n_i})$. Since $\{\psi_i^k : k = 1, \ldots, n_i\}$ is an orthonormal basis in $\mathcal{H}_{X_i}^{n_i}$, we can rewrite the objective function $\hat{L}_i(B_i)$ in (5) as

$$\hat{L}_i(B_i) = -2 \sum_{k=1}^{n_i} \left\{ \langle \hat{\Sigma}_{X_{-i} X_i} \psi_i^k, B_i \psi_i^k \rangle_{\mathcal{H}_{X_{-i}}^{n_i}} + \langle B_i \psi_i^k, \hat{\Sigma}_{X_{-i} X_{-i}} B_i \psi_i^k \rangle_{\mathcal{H}_{X_{-i}}^{n_i}} \right\}.$$

By the definition of empirical covariance operator, the penalized objective function $\widehat{PL}_i(B)$ in (6) can be written as

$$\begin{aligned} \widehat{PL}_i(A_i) = & -2 \sum_{j \neq i} \text{tr} \left\{ \left(V_i D_i^{1/2}\right)^\mathsf{T} \left(V_j D_j^{1/2} (A_i)_j\right) \right\} \\ & + \| \sum_{j \neq i} \left(V_j D_j^{1/2} (A_i)_j\right) \|_F^2 \\ & + \lambda_n \{ (\sum_{j \neq i} \|(A_i)_j\|_F)^2 + \|A_i\|_F^2 \}. \end{aligned} \tag{13}$$

where $A_i = [(A_i)_1^\mathsf{T}, (A_i)_{i-1}^\mathsf{T}, (A_i)_{i+1}^\mathsf{T}, \ldots, (A_i)_p^\mathsf{T}]^\mathsf{T}$ with $(A_i)_j \equiv [(B_i)_j] \in \mathbb{R}^{n_j \times n_i}$ being the coordinate expression of $(B_i)_j$ with respect to ψ_i and ψ_j.

5.2. Algorithm and Tuning

We next summarize our estimation algorithm, followed by a discussion on parameter tuning and the computation complexity.

Step 1: Choose the kernel κ_T. One option is the Brownian motion function $\kappa_T(s, t) = \min(s, t)$. Another option is the radial basis function (RBF) $\kappa_T(s, t) = \exp\{-\gamma_T |s - t|^2\}$, for $s, t \in \mathbb{R}$, where γ_T is determined by $(\sum_{s<t}^M |\tau_s - \tau_t|)^2 \gamma_T = M^2(M-1)^2/4$.

Step 2: Compute the first-level Gram matrices $K_T^{k,\ell}$, for $k, \ell = 1, \ldots, n$.

Step 3: Determine the ridge parameter ϵ_T^k via $\epsilon_T^k = c_T \times \sigma_{\max}(K_T^k)$, $k = 1, \ldots, n$, where $\sigma_{\max}(\cdot)$ denotes the largest eigenvalue of the associated matrix; c_T is to control the level of smoothness, which we fix at $c_T = 0.04$. Then use (11) to calculate the inner product $\langle X_i^k, X_i^\ell \rangle_{\Omega^n}$, for $k, \ell = 1, \ldots, n$ and $i = 1, \ldots, p$.

Step 4: Select the second-level kernel function, and compute the second-level Gram matrix K_i and its centered version G_i for $i = 1, \ldots, p$. If the RBF kernel is used, compute the width parameter γ by $(\sum_{k<\ell}^n \|X_i^k - X_i^\ell\|_{\Omega^n})^2 \gamma = n^2(n-1)^2/4$, where $\|\cdot\|_{\Omega^n}$ is the norm induced by the inner product in (11).

Step 5: Conduct the eigen-decomposition on G_i in (12). For the selection of n_i, we follow the rule in Ravikumar et al. (2009) and choose it adaptively based on the sample size as $n_i = O(n^{1/5})$, $i = 1, \ldots, p$.

Step 6: For a given λ_n and each $i = 1, \ldots, p$, minimize $\widehat{PL}_i(A_i)$ in (13) over $A_i \in \mathbb{R}^{n_{-i} \times n_i}$, where $n_{-i} = \sum_{j \neq i} n_i$, using, for example, the disciplined convex programming method of Boyd and Vandenberghe (2004).

Step 7: Let $\hat{A}_i^{\lambda_n}$ denote the resulting minimizer in the previous step. Estimate the neighbors $\hat{\mathsf{N}}_i^{\lambda_n} = \{j \in \mathsf{V} \backslash \{i\} : \text{tr}[(\hat{A}_i^{\lambda_n})_j^\mathsf{T} (\hat{A}_i^{\lambda_n})_j] \neq 0\}$. Then estimate the graph $\hat{\mathsf{E}}^{\lambda_n}$ by

$\hat{\mathsf{E}}^{\lambda_n}_{\mathrm{OR}}$ or $\hat{\mathsf{E}}^{\lambda_n}_{\mathrm{AND}}$ in (7). We use $\hat{\mathsf{E}}^{\lambda_n}_{\mathrm{OR}}$ as the final estimate for all the numerical analyses.

We next discuss the tuning of the penalty parameter λ_n in Step 6. Specifically, we consider a BIC-type criterion,

$$\mathrm{BIC}(\lambda) = \sum_{i=1}^{p} \mathrm{BIC}_i(\lambda), \quad \text{where}$$
$$\mathrm{BIC}_i(\lambda) = n \log\{\mathrm{RSS}_i(\lambda)\} + \log(n)\, \mathrm{DF}_i(\lambda), \tag{14}$$

and $\mathrm{RSS}_i(\lambda) = \|[I_n - \mathbf{U}_i(\lambda)]\mathbf{V}_i\|_{\mathrm{F}}^2$, with $\mathbf{V}_i = (V_i D_i^{1/2})^{\mathsf{T}}$, $\mathbf{U}_i(\lambda) = \mathbf{W}_i(\lambda)\{\mathbf{W}_i^{\mathsf{T}}(\lambda)\mathbf{W}_i(\lambda) + r(\lambda) I_{R_i(\lambda)}\}^{-1}\mathbf{W}_i^{\mathsf{T}}(\lambda)$, $\mathbf{W}_i(\lambda) = (\mathbf{V}_{i_1}, \ldots, \mathbf{V}_{i_{M_i(\lambda)}})$, $r(\lambda) = n^{-1/5} \times \sigma_{\max}[\mathbf{W}_i^{\mathsf{T}}(\lambda)\mathbf{W}_i(\lambda)]$, $R_i(\lambda) = \sum_{j=1}^{M_i(\lambda)} n_{i_j}$, $\{i_1, \ldots, i_{M_i(\lambda)}\} = \hat{\mathsf{N}}_i^{\lambda}$, $M_i(\lambda) = |\hat{\mathsf{N}}_i^{\lambda}|$ and $\mathrm{DF}_i(\lambda) = \mathrm{tr}[\mathbf{U}_i(\lambda)]$. Our idea is to regress each random function on its neighboring functions, and thereby calculate the sum of squared errors $\mathrm{RSS}_i(\lambda)$. We then select λ that minimizes $\mathrm{BIC}(\lambda)$ over a grid of candidate values.

In Section S5.3 of the supplementary material, we further investigate the effect of the choice of the tuning parameters and the kernel functions. In general, we have found that our algorithm is relatively robust, as long as the choices are within reasonable ranges.

Finally, we discuss the computation complexity of our algorithm. This complexity can be divided into two parts: the pre-convex optimization part and the convex optimization part. The first part involves the construction of both first-layer and second-layer kernels, and the eigen-decomposition of the Gram matrix G_i. Its complexity grows in the order of $p[\sum_{k,\ell=1}^{n}(m_k^3 + m_k m_\ell) + n^3]$ for the unbalanced setting, and $p(n^2 m^3 + n^3)$ for the balanced setting, assuming that $m_k = m$ for all $k = 1, \ldots, n$. For the convex optimization part, for graphs of small to moderate sizes, we recommend `CVX`, a Matlab toolbox for constrained minimization, whose default solver is based on the semidefinite-quadratic-linear program (Tutuncu, Toh, and Todd 2003). For each iteration and each regression, the complexity of this part is of the order $(pn^*)^3$ (Sra, Nowozin, and Wright 2011, chap. 3), where $n^* = \max(n_1, \ldots, n_n)$. For large graphs, we can reduce the computation by keeping only the L_1 penalty in the optimization. This simplified algorithm is equivalent to the group Lasso, and can be easily implemented by the iterative shrinkage thresholding algorithm (ISTA, Beck and Teboulle 2009), whose complexity grows in the order of $p(n^*)^2(n + n^*)$. In our numerical experiments, we have found this simplification loses little estimation accuracy, but brings substantial gain in computation. Moreover, in our simulation with $p = 100$ and $n = 100$, the average number of iterations for our algorithm to converge was 33, and the average running time of a single optimization was 2.48 sec. The implementation was done on a `2 x E5-2630 v4` workstation.

6. Numerical Studies

In this section, we first investigate the finite-sample performance of our proposed method by simulations. We then illustrate our method with an EEG data analysis.

6.1. Simulations

We generate multivariate random functions using the structural equation model of Pearl (2009). We consider both the Gaussian case (Model I), and the non-Gaussian case with nonlinear relations among the nodes (Models II and III). Specifically, given a directed edge set D with the ordering $1 \to \cdots \to p$, we generate $(X_1(t), \ldots, X_p(t))^{\mathsf{T}}$ sequentially via

$$X_i(t) = f_i\big[\{X_j(t) : (j,i) \in \mathsf{D}\}, \varepsilon_i(t)\big], \quad i = 1, \ldots, p,$$

for some functions $f_1, \ldots, f_p$ we specify later. We use the Brownian motion covariance as the kernel to construct the error function $\varepsilon_i(t)$, $i = 1, \ldots, p$; that is, $\varepsilon_i(t)$ is generated by $\sum_{u=1}^{J} \xi_u \kappa_T(t, t_u)$, where t_u and ξ_u are independently generated from Uniform$(0,1)$ and Normal$(0,1)$. For the observed time points $\{t_{k1}, \ldots, t_{km_k} : k = 1, \ldots, n\}$, we consider two different scenarios: the balanced case with $m_k = 10$ equally spaced points between $[0,1]$, and the unbalanced case with $m_k = 10$ time points independently drawn from the discrete uniform distribution on $\{0.01, 0.02, \ldots, 1\}$. We consider the following choices of f_i:

$$\text{Model I}: \ f_i\big[\{X_j(t) : (j,i) \in \mathsf{D}\}, \varepsilon_i(t)\big] = \sum_{(j,i)\in\mathsf{D}} x_j + \varepsilon_i,$$
$$\text{Model II}: \ f_i\big[\{X_j(t) : (j,i) \in \mathsf{D}\}, \varepsilon_i(t)\big] = \sum_{(j,i)\in\mathsf{D}} x_j^2 + \varepsilon_i,$$
$$\text{Model III}: \ f_i\big[\{X_j(t) : (j,i) \in \mathsf{D}\}, \varepsilon_i(t)\big] = \{\sum_{(j,i)\in\mathsf{D}} x_j\}\, \varepsilon_i.$$

The edge set D is generated from a hub structure. We then use the moral graph of D as the undirected graph E (Lauritzen 1996). Given a graph of size p, ten independent hubs are generated so that the module in each hub is of degree $p/10 - 1$. We set the sample size as $n = 100$, and the number of nodes as $p = 100$. In Section S5.1 of the supplementary material, we further consider additional network structures, including the tree and the chain structures, and additional network sizes, including $p = 50, 200$.

For each model, the proposed method is applied with the second layer kernel being a Gaussian kernel (denoted as FARO), and a linear kernel (denoted as Linear). The latter shares the same spirit as the Gaussian graphical model method of Qiao, Guo, and James (2019), while we further compare with Qiao, Guo, and James (2019) in Section S5.2 of the supplementary material. We also compare with the functional additive precision operator of Li and Solea (2018) (denoted as Li and Solea). We first calculate the false positive (FP) rate and true positive (TP) rate,

$$\mathrm{TP}(\lambda) = \frac{\sum_{1\le i<j\le p} I[(i,j) \in \mathsf{E}^0,\ (i,j) \in \hat{\mathsf{E}}(\lambda)]}{\sum_{1\le i<j\le p} I\big[(i,j) \in \mathsf{E}^0\big]},$$
$$\mathrm{FP}(\lambda) = \frac{\sum_{1\le i<j\le p} I[(i,j) \notin \mathsf{E}^0,\ (i,j) \in \hat{\mathsf{E}}(\lambda)]}{\sum_{1\le i<j\le p} I\big[(i,j) \notin \mathsf{E}^0\big]},$$

for a given parameter λ, where $I(\cdot)$ denotes the indicator function, E^0 the true graph, and $\hat{\mathsf{E}}(\lambda)$ the estimated graph under λ. We then compute the receiver operating characteristic (ROC) curve for a gird of values of λ. Figure 1 shows the average ROC curves for the three methods under different models. Each

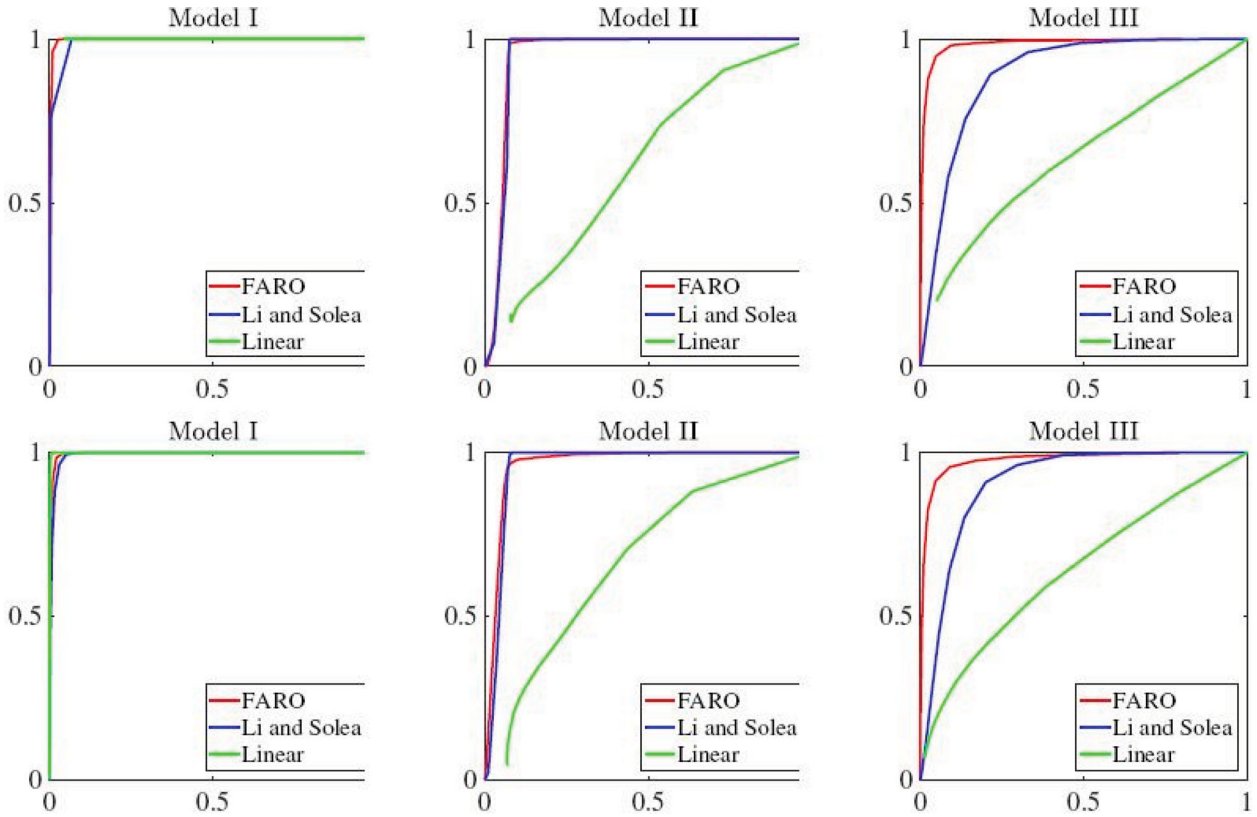

Figure 1. ROC curves for Models I to III, and for the balanced case (top panels) and the unbalanced case (bottom panels).

curve is averaged over 40 replications. It is seen that our method performs about the same as the Gaussian estimation method when the true model is indeed Gaussian (Model I), but performs much better when the underlying model is non-Gaussian (Models II and III). Moreover, FARO performs significantly better than Li and Solea (2018) in Model III for both the balanced and unbalanced settings, indicating the benefit of penalized optimization over hard thresholding.

6.2. EEG Data Analysis

We next apply our method in an EEG data analysis. The data are available at *https://kdd.ics.uci.edu/databases/eeg/eeg.data.html*. The goal of the study is to investigate the EEG relatedness of genetic predisposition with alcoholism. The data were collected from two groups of individuals: 77 subjects from the alcoholic group and 45 from the control group. Each subject was asked to wear a 64-channel electrode cap and shown one of three types of stimuli, while the voltage value from each electrode was recorded every second for a total of 256 sec. Each subject completed 120 trials from each of the three stimuli.

The same dataset has been analyzed before (Li, Kim, and Altman 2010; Xia and Li 2017; Qiao, Guo, and James 2019). Following Li, Kim, and Altman (2010), a preprocessing step was carried out by taking the average of measurements from single-stimulus trials. This results in a 64×256 data matrix for each individual. Our goal is to estimate the brain connectivity network with $p = 64$ for the alcoholic group and the control group, respectively. We employ a Gaussian kernel for this data analysis, and use the BIC criterion in (14) for parameter tuning.

Figure 2 shows the estimated graphs, $\hat{\mathrm{E}}_{\mathrm{ALC}}$ and $\hat{\mathrm{E}}_{\mathrm{CTL}}$, for the alcoholic group and the control group, as well as the difference graphs, $\hat{\mathrm{E}}_{\mathrm{ALC}} \setminus \hat{\mathrm{E}}_{\mathrm{CTL}}$ and $\hat{\mathrm{E}}_{\mathrm{CTL}} \setminus \hat{\mathrm{E}}_{\mathrm{ALC}}$. It is seen that both $\hat{\mathrm{E}}_{\mathrm{ALC}}$ and $\hat{\mathrm{E}}_{\mathrm{CTL}}$ are relatively sparse, with a 10.2% and 9.9 % sparsity rate, respectively, which also indicates a decrease of connectivity in the alcoholic group. Moreover, compared to the control individuals, the alcoholic individuals reveal asymmetric patterns, in which the left frontal, central and parietal regions have more connections than their right counterparts. We also note that the electrodes in regions other than frontal and parietal are connected only sparsely. These findings in general agree with the literature (Hayden et al. 2006).

7. Discussion

In this article, we have proposed a new nonparametric functional graphical model. A key and novel feature of this work is the estimation of a large number of regressions at the operator level, where both the number of predictors in each regression and the number of the regressions increase with the sample size at an exponential rate. This versatile framework yields flexible, accurate and computationally feasible estimate of the non-Gaussian and nonlinear functional graphical model with a large number of nodes. To the best of our knowledge, our consistency result is the first of its kind for neighborhood selection where both the response and the predictor are functions, the relation between them can be nonlinear, and the dimension of the functions can outgrow the sample size. Additional novel

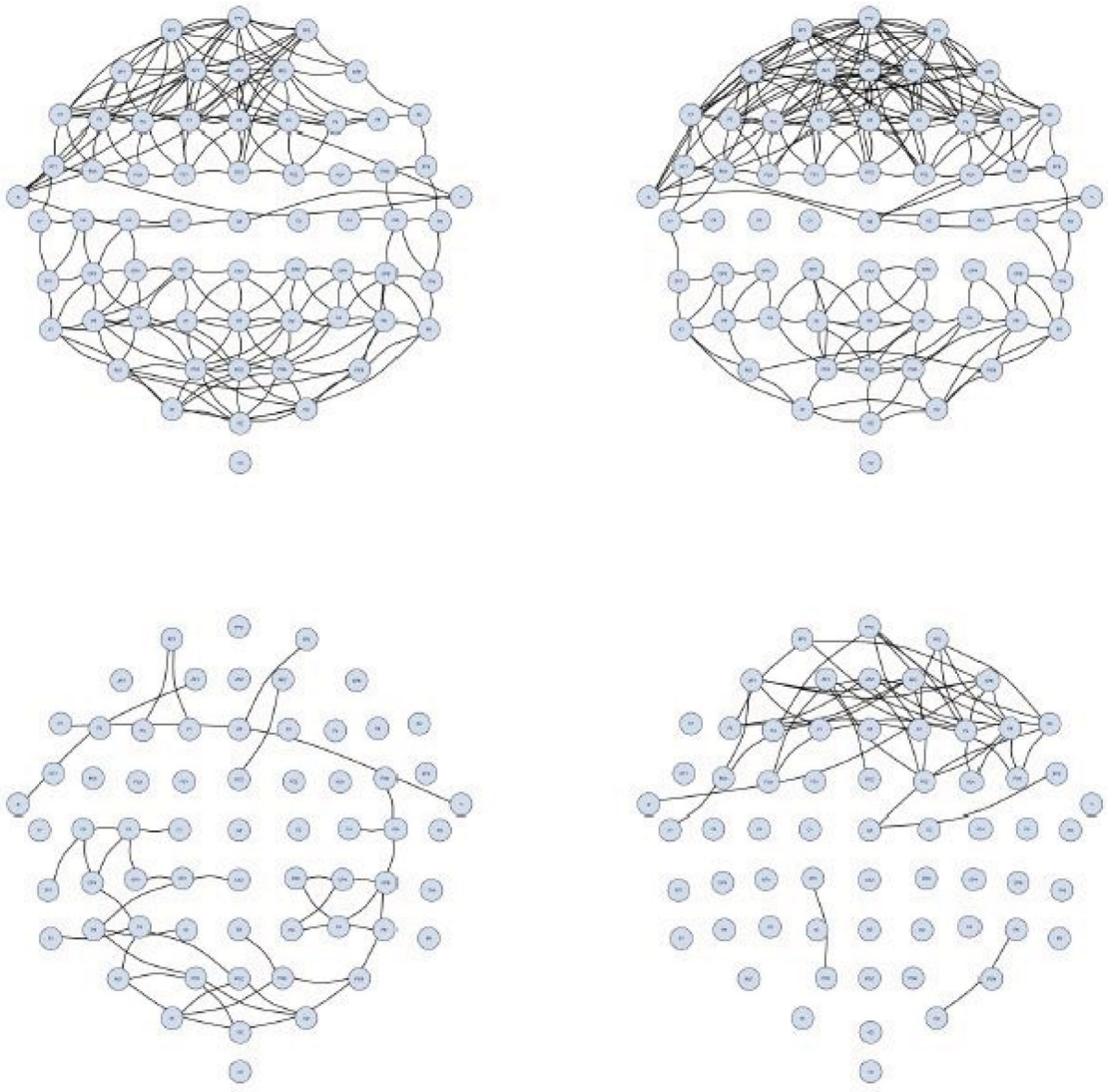

Figure 2. Estimated graphs by BIC: the alcoholic group $\hat{E}_{ALC}$ (upper left); the control group $\hat{E}_{CTL}$ (upper right); the difference $\hat{E}_{ALC} \setminus \hat{E}_{CTL}$ (lower left); and $\hat{E}_{ctl} \setminus \hat{E}_{ALC}$ (lower right).

features include the use of ACI as the selection criterion in the functional setting, the new concept of local additive Markovian property and its relation with the pairwise additive Markovian property, as well as the introduction of the Additive Gaussian distribution that puts ACI in solid footing.

This new framework involves considerable asymptotic developments. Specifically, we need to extend the individual convergence in Lee, Li, and Zhao (2016b) to the uniform convergence. Toward that end, we derive a series of concentration bounds for the sample covariance operator and the sample mean elements in RKHS, as shown in Lemma 1 and Lemma S3. From these, we show that the tail probability of the estimation error in terms of the Hilbert-Schmidt norm behaves like a sub-Gaussian variable. These concentration bounds and tail probabilities are the key elements for developing the uniform consistency in the high-dimensional setting. Li and Solea (2018) and Lee, Li, and Zhao (2016b) did not have such results.

Moreover, in order to make these concentration bounds applicable to our setting, we need to develop several relations to link the covariance operator with various key quantities in our method. For example, in Proposition S2, we derive an inequality between the estimated FARO and an intermediate operator, based on which we obtain the concentration bound of the FARO estimator. In Proposition S3, we derive a series of inequalities under which the KKT conditions are satisfied, which leads to an upper bound of the probability of erroneous neighborhood selection. These developments are far beyond routine. Besides, they are sufficiently general to be useful for other problems in functional data analysis.

Although the current form of the Additive Gaussian distribution is of a finite dimension, we expect that it can be extended to the infinite-dimensional setting. In addition, our experience suggests that, often in practice, there exist only a few dominating eigenfunctions in the empirical covariance operator. For the truly infinite-dimensional setting, we expect that some additional regularization is needed, such as the fast decaying of the tail eigen-structures. We leave such an extension to future research.

Acknowledgments

The authors are grateful to the two anonymous referees, the Associate Editor and the Editor for their constructive comments that improved the quality of this article.

Funding

Kuang-Yao Lee's research was partially supported by the NSF grant CIF-2102243, and the Seed Funding grant from Fox School of Business, Temple University. Lexin Li's research was partially supported by the NSF grant CIF-2102227, and the NIH grant R01AG061303.

Supplementary Materials

The Supplementary Appendix contains all the proofs, results of additional theoretical and numerical analysis, and a downloading site of our computer code.

References

Baker, C. R. (1973), "Joint Measures and Cross-Covariance Operators," *Transactions of the American Mathematical Society*, 186, 273–289. [1724]

Beck, A., and Teboulle, M. (2009), "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM Journal on Imaging Sciences*, 2, 183–202. [1728]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732. [1725]

Bosq, D. (2000), *Linear Processes in Function Spaces. Theory and Applications*, New York: Springer-Verlag. [1720,1724]

Boyd, S., and Vandenberghe, L. (2004), *Convex Optimization*, New York, NY: Cambridge University Press. [1727]

Cai, T., Liu, W., and Luo, X. (2011), "A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation," *Journal of the American Statistical Association*, 106, 594–607. [1719]

Chen, P.-C., Lee, K.-Y., Lee, T.-J., Lee, Y.-J., and Huang, S.-Y. (2010), "Multiclass Support Vector Classification via Coding and Regression," *Neurocomputing*, 73, 1501–1512. [1727]

Conway, J. (1990), *A Course in Functional Analysis*, (Vol. 96), New York: Springer. [1720]

Fan, J., and Lv, J. (2010), "A Selective Overview of Variable Selection in High Dimensional Feature Space," *Statistica Sinica*, 20, 101–148. [1725]

Fan, Y., and Lv, J. (2016), "Innovated Scalable Efficient Estimation in Ultra-Large Gaussian Graphical Models," *The Annals of Statistics*, 44, 2098–2126. [1719]

Fornito, A., Zalesky, A., and Breakspear, M. (2013), "Graph Analysis of the Human Connectome: Promise, Progress, and Pitfalls," *NeuroImage*, 80, 426–444. [1718]

Fox, M. D., and Greicius, M. (2010), "Clinical Applications of Resting State Functional Connectivity," *Frontiers in Systems Neuroscience*, 4, 19. [1718]

Friedman, J. H., Hastie, T. J., and Tibshirani, R. J. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [1719]

Fukumizu, K., Bach, F. R., and Gretton, A. (2007), "Statistical Consistency of Kernel Canonical Correlation Analysis," *Journal of Machine Learning Research*, 8, 361–383. [1724]

Hayden, E. P., Wiegand, R. E., Meyer, E. T., Bauer, L. O., O'Connor, S. J., Nurnberger, J. I., Chorlian, D. B., Porjesz, B., and Begleiter, H. (2006), "Patterns of Regional Brain Activity in Alcohol-Dependent Subjects," *Alcoholism: Clinical and Experimental Research*, 30, 1986–1991. [1729]

Hsing, T., and Eubank, R. (2015), *Theoretical Foundations of Functional Data Analysis, With an Introduction to Linear Operators*, Hoboken, NJ: Wiley. [1718]

Lauritzen, S. L. (1996), *Graphical Models*, Oxford: Oxford University Press. [1722,1728]

Lee, K.-Y., Li, B., and Zhao, H. (2016a), "On an Additive Partial Correlation Operator and Nonparametric Estimation of Graphical Models," *Biometrika*, 103, 513–530. [1720]

Lee, K.-Y., Li, B., and Zhao, H. (2016b), "Variable Selection via Additive Conditional Independence," *Journal of the Royal Statistical Society*, Series B, 78, 1037–1055. [1720,1721,1723,1730]

Lee, Y.-J., and Huang, S.-Y. (2007), "Reduced Support Vector Machines: A Statistical Theory," *IEEE Transactions on Neural Networks and Learning Systems*, 18, 1–13. [1727]

Li, B. (2018), "Linear Operator-Based Statistical Analysis: A Useful Paradigm for Big Data," *Canadian Journal of Statistics*, 46, 79–103. [1720]

Li, B., Chun, H., and 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. [1719,1722]

Li, B., Kim, M. K., and Altman, N. (2010), "On Dimension Folding of Matrix- or Array-Valued Statistical Objects," *Annals of Statistics*, 38, 1094–1121. [1729]

Li, B., and 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, 1637–1655. [1718,1719,1720,1721,1722,1726,1728,1729,1730]

Li, B., and Song, J. (2017), "Nonlinear Sufficient Dimension Reduction for Functional Data," *The Annals of Statistics*, 45, 1059–1095. [1724]

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012), "High-Dimensional Semiparametric Gaussian Copula Graphical Models," *The Annals of Statistics*, 40, 2293–2326. [1719]

Liu, H., Lafferty, J., and Wasserman, L. (2009), "The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs," *Journal of Machine Learning Research*, 10, 2295–2328. [1719]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462. [1719]

Pearl, J. (2009), *Causality: Models, Reasoning and Inference*, (2nd ed.), Cambridge, UK: Cambridge University Press. [1728]

Pearl, J., Geiger, D., and Verma, T. (1989), "Conditional Independence and its Representations," *Kybernetika*, 25, 33–44. [1719]

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial Correlation Estimation by Joint Sparse Regression Models," *Journal of the American Statistical Association*, 104, 735–746. [1719]

Qiao, X., Guo, S., and James, G. M. (2019), "Functional Graphical Models," *Journal of the American Statistical Association*, 114, 211–222. [1719,1725,1728,1729]

Ramsay, J., and Silverman, B. (2005), *Functional Data Analysis*, (2nd ed.), New York: Springer. [1718]

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009), "Sparse Additive Models," *Journal of the Royal Statistical Society*, Series B, 71, 1009–1030. [1724,1727]

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011), "High-Dimensional Covariance Estimation by Minimizing ℓ_1-Penalized Log-Determinant Divergence," *Electronic Journal of Statistics*, 5, 935–980. [1719]

Solea, E., and Dette, H. (2020), "Nonparametric and High-Dimensional Functional Graphical Models," *Journal of the American Statistical Association*. [1719,1720]

Solea, E., and Li, B. (2020), "Copula Gaussian Graphical Models for Functional Data," *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2020.1817750. [1719,1720,1724]

Sra, S., Nowozin, S., and Wright, S. J. (2011), *Optimization for Machine Learning*, Cambridge, MA: The MIT Press. [1728]

Tsay, R. S., and Pourahmadi, M. (2017), "Modelling Structured Correlation Matrices," *Biometrika*, 104, 237–242. [1718]

Tutuncu, R. H., Toh, K. C., and Todd, M. J. (2003), "Solving Semidefinite-Quadratic-Linear Programs Using SDPT3," *Mathematical Programing*, 95, 189–217. [1728]

Varoquaux, G., and Craddock, R. C. (2013), "Learning and Comparing Functional Connectomes Across Subjects," *NeuroImage*, 80, 405–415. [1718]

Voorman, A., Shojaie, A., and Witten, D. (2014), "Graph Estimation with Joint Additive Models," *Biometrika*, 101, 85–101. [1719]

Wainwright, M. J. (2009), "Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using-Constrained Quadratic Programming

(Lasso)," *IEEE Transactions on Information Theory*, 55, 2183–2202. [1724,1725,1726]

Wang, J.-L., Chiou, J.-M., and Muller, H.-G. (2016), "Functional Data Analysis," *Annual Review of Statistics and Its Application*, 3, 257–295. [1726]

Wei, Z., and 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, 408–429. [1718]

Xia, Y., and Li, L. (2017), "Hypothesis Testing of Matrix Graph Model with Application to Brain Connectivity Analysis," *Biometrics*, 73, 780–791. [1729]

Xue, L., and Zou, H. (2012), "Regularized Rank-Based Estimation of High-Dimensional Nonparanormal Graphical Models," *The Annals of Statistics*, 40, 2541–2571. [1719]

Yao, Y., Rosasco, L., and Caponnetto, A. (2007), "On Early Stopping in Gradient Descent Learning," *Constructive Approximation*, 26, 289–315. [1720]

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [1719]

Zhang, T., Wu, J., Li, F., Caffo, B., and Boatman-Reich, D. (2015), "A Dynamic Directional Model for Effective Brain Connectivity Using Electrocorticographic (ECoG) Time Series," *Journal of the American Statistical Association*, 110, 93–106. [1718]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *The Journal of Machine Learning Research*, 7, 2541–2563. [1724,1725,1726]

Zhu, Y., Shen, X., and Pan, W. (2014), "Structural Pursuit Over Multiple Undirected Graphs," *Journal of the American Statistical Association*, 109, 1683–1696. [1719]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [1723]