Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

A Nonparametric Graphical Model for Functional Data With Application to Brain Networks Based on fMRI

Bing Li & Eftychia Solea

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 524, 1637–1655: Theory and Methods
https://doi.org/10.1080/01621459.2017.1356726

A Nonparametric Graphical Model for Functional Data With Application to Brain Networks Based on fMRI

Bing Li and Eftychia Solea

Department of Statistics, Pennsylvania State University, State College, PA

ABSTRACT

We introduce a nonparametric graphical model whose observations on vertices are functions. Many modern applications, such as electroencephalogram and functional magnetic resonance imaging (fMRI), produce data are of this type. The model is based on additive conditional independence (ACI), a statistical relation that captures the spirit of conditional independence without resorting to multi-dimensional kernels. The random functions are assumed to reside in a Hilbert space. No distributional assumption is imposed on the random functions: instead, their statistical relations are characterized nonparametrically by a second Hilbert space, which is a reproducing kernel Hilbert space whose kernel is determined by the inner product of the first Hilbert space. A precision operator is then constructed based on the second space, which characterizes ACI, and hence also the graph. The resulting estimator is relatively easy to compute, requiring no iterative optimization or inversion of large matrices. We establish the consistency and the convergence rate of the estimator. Through simulation studies we demonstrate that the estimator performs better than the functional Gaussian graphical model when the relations among vertices are nonlinear or heteroscedastic. The method is applied to an fMRI dataset to construct brain networks for patients with attention-deficit/hyperactivity disorder. Supplementary materials for this article are available online

ARTICLE HISTORY
Received March 2016
Revised May 2017

KEYWORDS
Additive conditional independence; Additive correlation operator; Additive precision operator; EEG; fMRI; Gaussian graphical model; Reproducing kernel Hilbert space

1. Introduction

Functional data emerge in almost every branch of contemporary science and business, such as meteorology, medical research, longitudinal data analysis, and machine learning, where the sampling units are functions rather than numbers or vectors. Responding to these new demands, estimation and inference methods for functional data have been vigorously developed over the past decade or so. See, for example, Ramsay and Silverman (2002), Silverman and Ramsay (2005), Yao, Müller, and Wang (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012), and Hsing and Eubank (2015). The particular type of functional data concerned in this article are undirected graphs where the observations on vertices are functions, giving rise to the *functional graphical model* (Qiao, James, and Lv 2014).

These type of data are common in application, particularly in modern medical applications such as electroencephalogram (EEG) and functional magnetic resonance imaging (fMRI) data. See, for example, Lazar et al. (2002), Cheng and Herskovits (2007), and Li, Kim, and Altman (2010). Our motivating example is the fMRI data, where each vertex corresponds to a subregion of a brain, as represented by a collection of voxels. A form of brain activities called brain oxygen level-dependent, or BOLD, is recorded over a period of time at each voxel, which are then aggregated over voxels in each subregion, resulting in a vector of interdependent random functions. One of the interests in fMRI data analysis is to represent the interdependence by a network based on these random functions.

Figure 1 shows a portion of the fMRI dataset analyzed in Section 8, which consists of a sample of BOLD records in two subregions of the brains from 30 adolescents with attention-deficit/hyperactivity disorder (ADHD) patients (upper panels) and 42 adolescents who do not have ADHD. The left and right panels corresponding, respectively, to the left and right superior frontal gyrus, whose locations are indicated by the brain diagrams on top of the two columns. Colors are used to distinguish among different curves, which are based on the raw, unsmoothed data. The BOLD signals are observed at 2.5 sec intervals, which results in 74 observations over the entire time period. The goal of the analysis is to understand how these brain subregions are interconnected in ADHD patients and in healthy subjects.

Functional graphical models have been proposed and studied in several recent papers. Qiao, James, and Lv (2014) proposed a functional Gaussian graphical model (FGGM) where the random functions observed at the vertices are assumed to be Hilbert-space-valued Gaussian random elements, and developed a group-lasso estimation procedure. This model is built upon the classical Gaussian graphical model where the random elements at the vertices are Gaussian random variables (see, e.g., Meinshausen and Bühlmann 2006; Yuan and Lin 2007; Bickel and Levia 2008; Friedman, Hastie, and Tibshirani 2008; Peng et al. 2009). Zhu, Strawn, and Dunson (2016) developed a Bayesian approach to functional graphical models under the Gaussian assumption. In this article, we propose an additive

CONTACT Bing Li bing@stat.psu.edu Pennsylvania State University, State College, PA 16801, USA.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

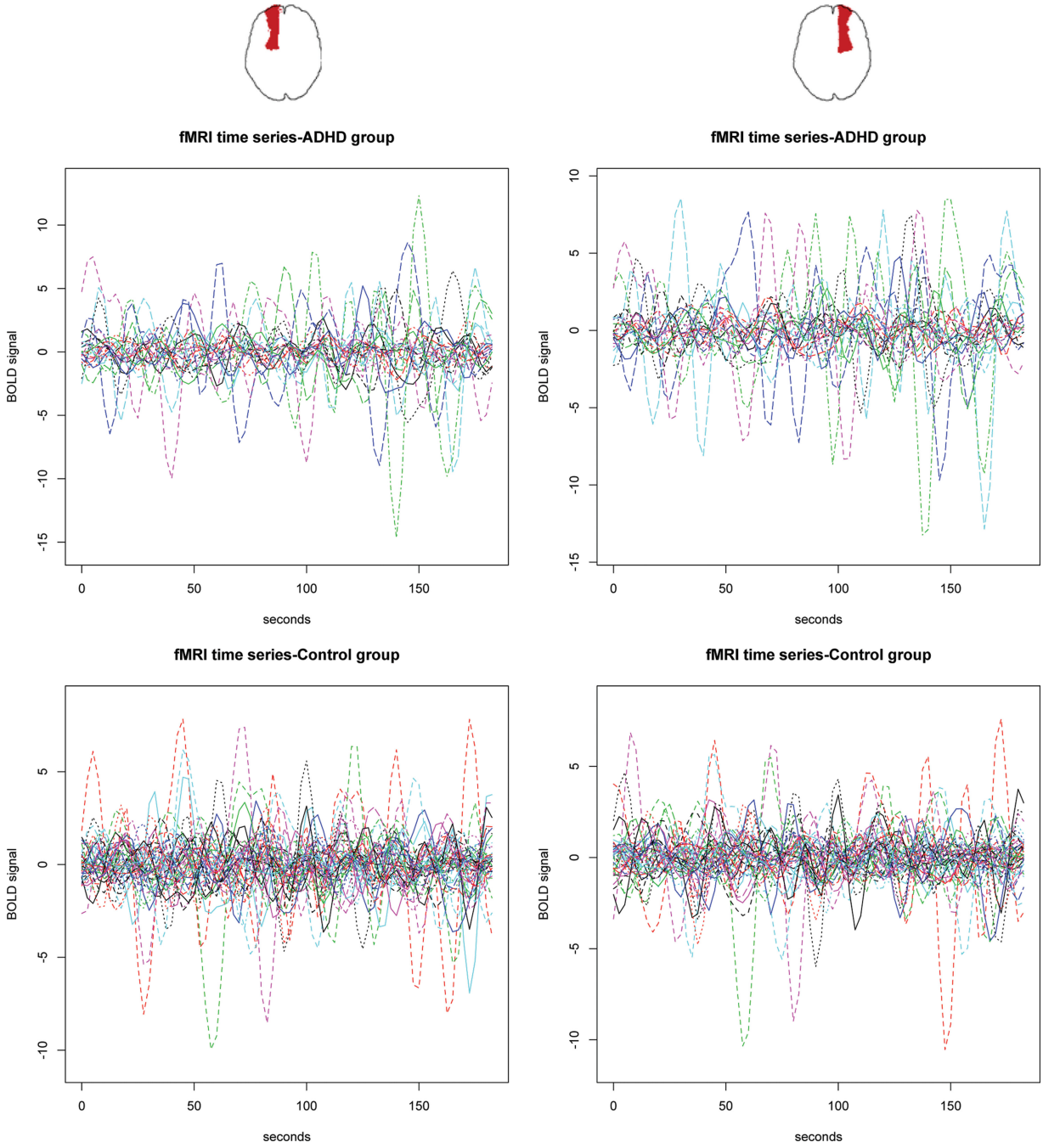

Figure 1. fMRI functional data for two brain regions. Left panels: left superior frontal gyrus; right panels: right superior frontal gyrus; upper panels: ADHD group; lower panels: control group.

nonparametric approach, which does not rely on any distribution assumption, but in some sense inherits the fundamental simplicity of a Gaussian model when characterizing the interdependence among random elements.

More specifically, let T be an interval in $\mathbb{R}$, representing time, and $X = (X^1, \ldots, X^p)^{\mathsf{T}}$ be a vector of random functions on T. Let $\mathrm{V} = \{1, \ldots, p\}$ denote the set of vertices, and $\mathrm{E} \subseteq \{(i, j) : i, j \in \mathrm{V},\ i \neq j\}$ denote the set of edges, of an undirected graph $\mathrm{G} = (\mathrm{V}, \mathrm{E})$. The FGGM is defined by

$$(i, j) \notin \mathrm{E} \quad \Leftrightarrow \quad \mathrm{cov}[X^i(s), X^j(t)|X^{-(i,j)}] = 0 \quad \forall\, s, \quad t \in T,$$

where $X^{-(i,j)} = \{X^k : k \in \mathrm{V} \setminus \{i, j\}\}$, and X is assumed to be a Gaussian random element in a Hilbert space. The basic idea of Qiao, James, and Lv (2014) is to find the first m functional principal components for each X^i, forming an m-dimensional random vector for that vertex, say $A^i(m)$. Since X is a Gaussian random element, the pm-dimensional random vector $(A^1(m)^{\mathsf{T}}, \ldots, A^p(m)^{\mathsf{T}})^{\mathsf{T}}$ itself has a multivariate Gaussian distribution. The problem then boils down to identifying the approximated graph $\mathrm{G}(m) = (\mathrm{V}, \mathrm{E}(m))$ by

$$(i, j) \notin \mathrm{E}(m) \quad \Leftrightarrow \quad \mathrm{cov}[A^i(m), A^j(m)|A^{-(i,j)}(m)] = 0. \tag{1}$$

Intuitively, $\mathrm{E} \approx \mathrm{E}(m)$ when m is sufficiently large, which is the theoretical basis of the method. Here, the fundamental simplicity of the Gaussian assumption manifests itself through the following relation. Let $\Sigma(m)$ be the $p \times p$ blocks whose

(i, j)th entry is the $m \times m$ matrix $\mathrm{cov}(A_i(m), A_j(m))$, and let $\Theta(m) = \Sigma(m)^{-1}$. Then

$$A^i(m) \perp\!\!\!\perp A^j(m) | A^{-(i,j)}(m) \quad \Leftrightarrow \quad \Theta_{ij}(m) = 0, \quad i, j \in \mathrm{V}, \tag{2}$$

where $\Theta_{ij}(m)$ is the (i, j)th block of the precision matrix $\Theta(m)$. Thus, estimating the edge set $E(m)$ reduces to sparse estimation of $\Theta(m)$, where sparse means encouraging blocks (rather than individual entries) of $\Theta(m)$ to vanish. Based on this idea Qiao, James, and Lv (2014) developed a group-lasso algorithm to estimate E, which they call functional glasso, or simply fglasso.

The Gaussian assumption in FGGM is restrictive in a number of ways. The first and the most obvious restriction is the Gaussian distribution itself. The second restriction is that the Gaussian assumption implies that the relations between vertices must be linear: relations such as $X^j = (X^k)^2 + \epsilon$ are prohibited. The third and the most subtle restriction is that the Gaussian assumption precludes any dependence that is not in the mean function: for example, relations such as $X^j = X^k \epsilon$ where ϵ is an independent random element, are prohibited. Simply put, any relations among the vertices that are nonlinear or heteroscedastic are precluded by the Gaussian assumption. Nevertheless, the Gaussian model offers an important advantage: the equivalence (2) that encodes conditional independence by the precision matrix.

We seek to remove the Gaussian assumption along with the restrictions it entails, but at the same time retain the fundamental simplicity of the Gaussian dependence structure manifested in (2). This is achieved by replacing conditional independence by additive conditional independence (ACI), which is a new statistical relation between three sets of random variables introduced by Li, Chun, and Zhao (2014). The essence of a graph is the notion of separation, that is, whether two sets of vertices can be separated by a third set of vertices. This notion turns out to have much in common with conditional independence: they satisfy the same set of axioms (Pearl, Geiger, and Verma 1989). It is through this similarity that we link a graph to conditional independence, and ultimately to the data. Li, Chun, and Zhao (2014) showed that ACI also satisfies these axioms, and proposed to use it as an alternative criterion to construct a graph. Lee, Li, and Zhao (2016b) further developed ACI for variable selection in a nonparametric regression setting.

As in Qiao, James, and Lv (2014), we assume the random functions at the vertices belong to a Hilbert space. To characterize ACI nonparametrically, we build up a second Hilbert space of functions that are defined on the first space. In the second Hilbert space, we introduce covariance operators between vertices, and then use them to define a functional additive precision operator (FAPO). This is a matrix of linear operators that inherits the relation (2) at the operator level: the (i, j)th entry of FAPO is the 0 operator if and only if X^i and X^j are additively conditionally independent given $X^{-(i,j)}$. The positions of the zero entries of FAPO then determine the edges of the graph. Our estimate is based on thresholding the small entries the estimated FAPO.

The rest of the article is organized as follows. In Section 2, we construct the additively nested Hilbert spaces, define ACI in the functional setting, and propose the nonparametric functional graphical model. In Section 3, we construct the functional additive precision operator and establish its relation with ACI. In Sections 4 and 5.1, we establish the consistency and convergence rate of the proposed estimator. In Section 6, we develop an algorithm for estimation and tuning. In Section 7, we conduct simulation studies to evaluate our estimator and compare it with FGGM. In Section 8, we apply the new estimator to an fMRI dataset. We conclude with some discussions in Section 9.

Due to the space limit we have put all the proofs in an online supplementary. While some of the lemmas are used only for the proofs in the supplementary, we retained them in the main manuscript so as to keep a complete picture of the theory.

2. Additive Conditional Independence and Functional Graphical Models

2.1. Additively Nested Hilbert Spaces

Let $T \in \mathbb{R}$ be an interval. For each $i \in \mathrm{V}$, let $\mathcal{H}_i$ be a separable Hilbert space of functions on T with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_i}$. Let κ_i be a positive definite mapping from $\mathcal{H}_i \times \mathcal{H}_i \to \mathbb{R}$ that is determined by the inner product. For example, for any $f, g \in \mathcal{H}_i$,

$$\kappa_i(f, g) = \exp\left(-\gamma \|f - g\|^2_{\mathcal{H}_i}\right), \quad \kappa_i(f, g) = (c + \langle f, g \rangle_{\mathcal{H}_i})^k,$$

are the Gaussian radial basis function kernel and polynomial kernel derived from the inner product of $\mathcal{H}_i$. More generally, $\kappa_i(f, g)$ depends on (f, g) only through the inner products $\langle f, f \rangle_{\mathcal{H}_i}$, $\langle f, g \rangle_{\mathcal{H}_i}$, and $\langle g, g \rangle_{\mathcal{H}_i}$. Let $\mathfrak{M}_i$ be the reproducing kernel Hilbert space (RKHS) generated by the kernel κ_i, that is,

$$\mathfrak{M}_i = \overline{\mathrm{span}}\{\kappa_i(\cdot, f) : \ f \in \mathcal{H}_i\},$$
$$\langle \kappa_i(\cdot, f), \kappa_i(\cdot, g) \rangle_{\mathfrak{M}_i} = \kappa_i(f, g),$$

where $\overline{\mathrm{span}}$ {a set of functions} means the closure of the subspace spanned by the set of functions. We say that $(\mathcal{H}_i, \mathfrak{M}_i)$ are *nested Hilbert spaces*, because the inner product of the former determines the kernel of the latter. Note that the former is assumed to be any separable Hilbert space, but the latter is assumed to be an RKHS. We denote functions in the first-level Hilbert spaces $\mathcal{H}_i$ by f, g, and so on, and functions in the second-level Hilbert spaces $\mathfrak{M}_i$ by ϕ, ψ, and so on.

Let $\mathcal{H} = \oplus_{i=1}^p \mathcal{H}_i$ be the direct sum of $\mathcal{H}_1, \ldots, \mathcal{H}_p$. That is, $\mathcal{H}$ is the Cartesian product $\mathcal{H}_1 \times \cdots \times \mathcal{H}_p$ and its inner product is defined by

$$\langle f, g \rangle_{\mathcal{H}} = \langle f_1, g_1 \rangle_{\mathcal{H}_1} + \cdots + \langle f_p, g_p \rangle_{\mathcal{H}_p},$$

for any $f = (f_1, \ldots, f_p)^{\mathsf{T}} \in \mathcal{H}$ and $g = (g_1, \ldots, g_p)^{\mathsf{T}} \in \mathcal{H}$. Let $\mathfrak{M} = \{\phi_1 + \cdots + \phi_p : \phi_p \in \mathfrak{M}_1, \ldots, \phi_1 \in \mathfrak{M}_p\}$ with inner product defined by

$$\langle \phi, \psi \rangle_{\mathfrak{M}} = \langle \phi_1, \psi_1 \rangle_{\mathfrak{M}_1} + \cdots + \langle \phi_p, \psi_p \rangle_{\mathfrak{M}_p}$$

for any $\phi = \phi_1 + \cdots + \phi_p \in \mathfrak{M}$ and $\psi = \psi_1 + \cdots + \psi_p \in \mathfrak{M}$. We note the subtle difference between the constructions of $\mathcal{H}$ and $\mathfrak{M}$: the former puts $\mathcal{H}_1, \ldots, \mathcal{H}_p$ together by Cartesian product; the latter puts $\mathfrak{M}_1, \ldots, \mathfrak{M}_p$ together by adding. We say that $(\mathcal{H}, \mathfrak{M})$ are *additively nested Hilbert spaces*. Li and Song (2017) employed a similar structure of nested Hilbert spaces for nonlinear sufficient dimension reduction for functional data in a nonadditive framework.

Having laid out the geometric groundwork we are ready to introduce random functions. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\mathcal{B}$ be the Borel σ-field generated by the open sets in $\mathcal{H}$. Let $X : \Omega \to \mathcal{H}$ be a random element in $\mathcal{H}$ that is measurable with respect to $\mathcal{F}/\mathcal{B}$. Let $P_X = P \circ X^{-1}$ be the measure on $(\mathcal{H}, \mathcal{B})$ induced by X, that is, P_X is the distribution of X. Hence $X = (X^1, \ldots, X^p)^\intercal$ where X^i is a random element in $\mathcal{H}_i$. Let $P_{X^i} = P \circ (X^i)^{-1}$ be the distribution of X^i.

2.2. Functional Additive Conditional Independence

The most direct way to state additive conditional independence is in terms of $L_2(P)$-geometry, which is not the same as the RKHS-geometry natural to $\mathfrak{M}$. However, the asymptotic results are more easily derived under the RKHS geometry. For this reason, we employ both geometries in this article. To avoid ambiguity, we indicate the concepts such as orthogonality and direct difference in the $L_2(P)$-geometry by dotted notations such as $\dot{\perp}$ and $\dot{\ominus}$. Specifically, if $\mathfrak{N}_1$ and $\mathfrak{N}_2$ are two subspaces of $\mathfrak{M}$, we write $\mathfrak{N}_1 \dot{\perp} \mathfrak{N}_2$ if $\operatorname{cov}(\phi(X), \psi(X)) = 0$ for all $\phi \in \mathfrak{N}_1$ and $\psi \in \mathfrak{N}_2$. If $\mathfrak{N}_1 \subseteq \mathfrak{N}_2$, then $\mathfrak{N}_2 \dot{\ominus} \mathfrak{N}_1$ denotes the set $\{\phi \in \mathfrak{N}_2 : \phi \dot{\perp} \mathfrak{N}_1\}$. For a subvector U of X, let $\operatorname{supp}(U)$ denote index set of U, and let $\mathfrak{M}_{(U)}$ denote the subspace $\mathfrak{M}_{i_1} + \cdots + \mathfrak{M}_{i_s}$, where $(i_1, \ldots, i_s) \in \operatorname{supp}(U)$.

Definition 1. Let U, V, and W be subvectors of X. We say that random functions U and V are additively conditionally independent given random function W iff

$$\mathfrak{M}_{(U \cup W)} \dot{\ominus} \mathfrak{M}_{(W)} \dot{\perp} \mathfrak{M}_{(V \cup W)} \dot{\ominus} \mathfrak{M}_{(W)}, \tag{3}$$

we denote this relation by $U \perp\!\!\!\perp_A V | W$.

This is an extension of the definition of ACI in Li, Chun, and Zhao (2014) to random functions. Li, Chun, and Zhao (2014) showed that ACI, like conditional independence, satisfies the same set of axioms shared by conditional independence and the notion of separation, as developed in Pearl and Verma (1987) and Pearl, Geiger, and Verma (1989). See also Dawid (1979). The same conclusion applies to the functional setting; the proof is essentially the same and is omitted.

Replacing conditional independence by ACI, we now define a new functional graphical model which we call the Functional Additive Semigraphoid Model.

Definition 2. We say that X follows a functional additive semigraphoid model *(FASG)* with respect to an undirected graph $\mathrm{G} = (\mathrm{V}, \mathrm{E})$ iff

$$X^i \perp\!\!\!\perp_A X^j | X^{-(i,j)}, \quad \forall\, (i, j) \notin \mathrm{E}.$$

If this relation holds then we write $X \sim \mathrm{FASG}(\mathrm{G})$.

2.3. Additive Conditional Independence and Conditional Independence

In this subsection, we explore the relations between ACI and conditional independence (CI). In the case where $X^1, \ldots, X^p$ are random variables, Li, Chun, and Zhao (2014) demonstrated some relations between ACI and CI under the copula Gaussian model assumption. We now show that the similar relations hold in our context under a *functional copula Gaussian model* recently introduced by Solea and Li (2016). To do so, we first outline the definition of the functional copula Gaussian model.

For each $i = 1, \ldots, p$, let $X^i = EX^i + \sum_{r=1}^{\infty} \lambda_{ir} \xi_{ir} \phi_{ir}$ be the Karhunen–Loeve expansion of X^i (Bosq 2000, p. 25). That is, $\{(\lambda_{ir}, \phi_{ir}) : r = 1, 2, \ldots\}$ are the eigenvalues and eigenfunctions of the linear operator $E[(X^i - EX^i) \otimes (X^i - EX^i)]$, and $\{\xi_{ir} : r = 1, 2, \ldots\}$ are random variables with $E(\xi_{ir}) = 0$, $\operatorname{var}(\xi_{ir}) = 1$, and $\operatorname{cov}(\xi_{ir}, \xi_{is}) = 0$ for $r \neq s$. We say that X^i follows a functional copula Gaussian model if there exists a sequence of injective transformations from $\mathbb{R}$ to $\mathbb{R}$, $\{c_{ir} : r = 1, 2, \ldots\}$, such that $c_{ir}(\xi_{ir})$ is distributed as $N(0, 1)$. This implies that $C_i(X^i) = \sum_{r=1}^{\infty} \lambda_{ir} c_{ir}(\xi_{ir}) \phi_{ir}$ is a Gaussian random element in $\mathcal{H}_i$. Furthermore, we say that $(X^1, \ldots, X^p)$ is a copula Gaussian random element in $\mathcal{H}$ if $(C_1(X^1), \ldots, C_p(X^p))$ is a Gaussian random element in $\mathcal{H}$.

In the following, for each $i = 1, \ldots, p$, let $\mathfrak{F}_i = \operatorname{span}\{\langle b, C_i(X^i) \rangle_{\mathcal{H}_i} : b \in \mathcal{H}_i\}$. For a subset $\mathfrak{A}$ of $L_2(P)$, let $\overline{\mathfrak{A}}$ denote the $L_2(P)$-closure of $\mathfrak{A}$. The next theorem is a generalization of Theorems 3 and 4 in Li, Chun, and Zhao (2014).

Theorem 1. Suppose $\mathfrak{F}_i \subseteq L_2(P)$ and $\mathfrak{M}_i \subseteq L_2(P)$.

1. If $\overline{\mathfrak{F}}_\ell \subseteq \overline{\mathfrak{M}}_\ell$ for each $\ell \in \mathrm{V}$, then $X^i \perp\!\!\!\perp_A X^j | X^{-(i,j)} \Rightarrow X^i \perp\!\!\!\perp X^j | X^{-(i,j)}$ for each $(i, j) \in \mathrm{E}$.
2. If $\overline{\mathfrak{F}}_\ell = \overline{\mathfrak{M}}_\ell$ for each $\ell \in \mathrm{V}$, then $X^i \perp\!\!\!\perp_A X^j | X^{-(i,j)} \Leftrightarrow X^i \perp\!\!\!\perp X^j | X^{-(i,j)}$ for each $(i, j) \in \mathrm{E}$.

We would also like to point out that ACI is justified as a criterion for constructing graphical models regardless of its relation with CI, because ACI satisfies the four axioms of a semigraphoid that characterize any relation of separation, which was indeed one of the main reasons conditional independence was used as a criterion to construct graphical models in the first place (Dawid 1979; Pearl and Verma 1987).

3. Functional Additive Precision Operator and Its Estimation

In this section, we introduce the FAPO, which is the extension of the additive precision operator (APO; Li, Chun, and Zhao 2014) to the functional case. We also introduce its sample estimator in the operator form. Henceforth the norm of a linear operator is always taken to be the operator norm, and is denoted by $\|\cdot\|$. The norm of a function is denoted by $\|\cdot\|_{\mathcal{H}}$, where $\mathcal{H}$ is the Hilbert space where the function resides. Since we will frequently make references to relevant classical results where X is a random vector rather than a vector of random functions, we will refer to the former setting as the *multivariate setting*, and the letter setting as the *functional setting*.

3.1. Population-Level Definition

For two Hilbert spaces $\mathfrak{N}_1, \mathfrak{N}_2$, let $\mathcal{B}(\mathfrak{N}_1, \mathfrak{N}_2)$ denote the class of all bounded linear operators from $\mathfrak{N}_1$ to $\mathfrak{N}_2$. The class $\mathcal{B}(\mathfrak{N}_1, \mathfrak{N}_1)$ is abbreviated by $\mathcal{B}(\mathfrak{N}_1)$. For $A \in \mathcal{B}(\mathfrak{N}_1, \mathfrak{N}_2)$, let $\ker(A)$ denote the kernel of A, that is, $\{\phi \in \mathfrak{N}_1 : A(\phi) = 0\}$, let $\operatorname{ran}(A)$ denote the range of A, and let $\overline{\operatorname{ran}}(A)$ denote closure of $\operatorname{ran}(A)$.

Assumption 1. $E[\kappa_i(X^i, X^i)] < \infty, \quad i = 1, \ldots, p.$

This is a mild assumption satisfied by most kernels. Under this assumption, the bilinear form $\mathfrak{M}_i \times \mathfrak{M}_j \mapsto \mathbb{R}$, $(\phi, \psi) \mapsto \mathrm{cov}[\phi(X^i), \psi(X^j)]$ is bounded. By Riesz's representation theorem, there is $\Sigma_{X^iX^j} \in \mathcal{B}(\mathfrak{M}_j, \mathfrak{M}_i)$ such that

$$\langle \phi, \Sigma_{X^iX^j}\psi\rangle_{\mathfrak{M}_i} = \mathrm{cov}[\phi(X^i), \psi(X^i)]. \tag{4}$$

This is called the covariance operator between X^i and X^j (see, e.g., Baker 1973; Fukumizu, Bach, and Jordan 2009). Note that $f \in \ker(\Sigma_{X^iX^i})$ if and only if $\mathrm{var}[f(X^i)] = 0$. Thus, $\ker(\Sigma_{X^iX^i})$ contains only constant functions. Since constants are irrelevant to ACI, we exclude them from $\mathfrak{M}_i$, which amounts to resetting $\mathfrak{M}_i$ to $\overline{\mathrm{ran}}(\Sigma_{X^iX^i})$ or, equivalently, making the following assumption.

Assumption 2. $\ker(\Sigma_{X^iX^i}) = \{0\}$ for all $i \in \mathrm{V}$.

We now introduce the notion of a matrix of operators, which underpins our construction.

Definition 3. An operator $A \in \mathcal{B}(\mathfrak{M})$ is called a matrix of operators with respect to $\{\mathfrak{M}_1, \ldots, \mathfrak{M}_p\}$ if there exist $A_{ij} \in \mathcal{B}(\mathfrak{M}_j, \mathfrak{M}_i)$ such that, for any $\phi = \phi_1 + \cdots + \phi_p \in \mathfrak{M}$, $A\phi = \sum_{i,j=1}^p A_{ij}\phi_j$. The collection of all such operators is written as $\times_{i,j=1}^p \mathcal{B}(\mathfrak{M}_i, \mathfrak{M}_j)$.

Note that $\times_{i,j=1}^p \mathcal{B}(\mathfrak{M}_i, \mathfrak{M}_j)$ is a subset of $\mathcal{B}(\mathfrak{M})$. We denote the matrix of operators by $A = \{A_{ij}\}_{i,j=1}^p$. We now introduce three key operators in our theory, all of which are members of $\times_{i,j=1}^p \mathcal{B}(\mathfrak{M}_i, \mathfrak{M}_j)$.

Definition 4. The operator $\{\Sigma_{X^iX^j}\}_{i,j=1}^p$ is called the functional additive variance operator (FAVO), and is written as Σ_{XX}.

By this definition, for any $\phi = \phi_1 + \cdots + \phi_p$, $\psi = \psi_1 + \cdots + \psi_p \in \mathfrak{M}$,

$$\begin{aligned}\langle \phi, \Sigma_{XX}\psi\rangle_{\mathfrak{M}} &= \textstyle\sum_{i=1}^p \sum_{j=1}^p \langle \phi_i, \Sigma_{X^iX^j}\psi_j\rangle_{\mathfrak{B}_i} \\ &= \textstyle\sum_{i=1}^p \sum_{j=1}^p \mathrm{cov}[\phi_i(X^i), \psi_j(X^j)] = \mathrm{cov}[\phi(X), \psi(X)].\end{aligned}$$

By Baker (1973), for every $\Sigma_{X^iX^j} \in \mathcal{B}(\mathfrak{M}_j, \mathfrak{M}_i)$, there is a unique operator $C_{X^iX^j} \in \mathcal{B}(\mathfrak{M}_j, \mathfrak{M}_i)$ with $\|C_{X^iX^j}\| \le 1$ such that $\Sigma_{X^iX^j} = \Sigma_{X^iX^i}^{1/2} C_{X^iX^j} \Sigma_{X^jX^j}^{1/2}$. It is easy to see that $C_{X^iX^i}$ is simply the identity mapping in $\mathcal{B}(\mathfrak{M}_i)$.

Definition 5. The operator $\{C_{X^iX^j}\}_{i,j=1}^p$ is called the functional additive correlation operator (FACO), and is written as C_{XX}. When C_{XX} is invertible, the operator C_{XX}^{-1} is called the functional additive precision operator (FAPO), and is written as Θ_{XX}.

Note that the invertibility of C_{XX} is guaranteed by Assumption 2. In fact, C_{XX}^{-1} is a bounded operator under the following reasonably mild assumption.

Assumption 3. For any $i \ne j$, $C_{X^iX^j}$ is a compact operator.

This assumption ensures that C_{XX} is the sum of an identity and a compact operator. Such operators are bounded from below by cI for some $c > 0$ (Bach 2008). Hence $\Theta_{XX} \in \mathcal{B}(\mathfrak{M})$. Note that we define the precision operator as the inverse of C_{XX} rather than Σ_{XX}, and the inverse of the latter is the direct analog of the precision matrix in the classical setting. We choose to work with C_{XX}^{-1} instead of Σ_{XX}^{-1} because Σ_{XX} is a compact operator under mild conditions, and its inverse is unbounded. What is interesting is that a pattern resembling the Gaussian dependence structure (2) reemerges at the operator level without the Gaussian assumption, as shown in the next theorem.

Theorem 2. Under Assumptions 1 through 3,

$$X^i \perp\!\!\!\perp_A X^j | X^{-(i,j)} \text{ iff } \Theta_{X^iX^j} = 0. \tag{5}$$

The proof is similar to the multivariate case (Li, Chun, and Zhao 2014), and is omitted. Similar to the Gaussian relation (2), the above relation encodes additive conditional independence by FAPO, but without any restrictive distributional assumption. Moreover, by the additive nature of our setting, FAPO can be estimated by a kernel defined marginally on $\mathfrak{M}_i$ rather than jointly on $\mathfrak{M}$. For convenience, henceforth we let C denote the complete graph $\{(i, j) : i, j \in \mathrm{V},\ i \ne j\}$.

Note that Theorem 2 holds regardless of the geometry in which ACI is stated. Relation (3) can be represented in terms of the RKHS-geometry, as was done in Lee, Li, and Zhao (2016b). The alternative representation does not affect the validity of (5).

Corollary 1. Under Assumptions 1 through 3, a vector-valued random function X follows an *FASG*(G) model iff $\mathrm{E} = \{(i, j) \in \mathsf{C} : \|\Theta_{X^iX^j}\| > 0\}$.

3.2. Estimation

We now describe the estimator of Θ_{XX} in an operator form, which is not yet usable as an algorithm but can facilitate the asymptotic development. The concrete algorithm for this estimator in matrix form will be given in Section 6.

For two Hilbert spaces $\mathfrak{N}_1$ and $\mathfrak{N}_2$ and $\phi \in \mathfrak{N}_1$ and $\psi \in \mathfrak{N}_2$, the tensor product $\phi \otimes \psi$ is the operator $\phi \otimes \psi : \mathfrak{N}_1 \to \mathfrak{N}_2$, $\omega \mapsto \phi\langle \psi, \omega\rangle_{\mathfrak{N}_1}$. Using tensor product we can define $\Sigma_{X^jX^i}$ alternatively as $E\{[\kappa_j(\cdot, X^j) - \mu_{X^j}] \otimes [\kappa_i(\cdot, X^i) - \mu_{X^i}]\}$ where μ_{X^i} is the Riesz's representation of the linear map $\mathfrak{M}_i \to \mathbb{R}$, $\phi \mapsto E\phi(X^i)$. Thus, for any $\phi \in \mathfrak{M}_i$,

$$\begin{aligned}&E\{[\kappa_j(\cdot, X^j) - \mu_{X^j}] \otimes [\kappa_i(\cdot, X^i) - \mu_{X^i}]\}\phi \\ &\quad = E\{[\kappa_j(\cdot, X^j) - \mu_X^j][\phi(X^i) - E\phi(X^i)]\}.\end{aligned}$$

This form is special to the RKHS setting and can be mimicked at the sample level. In comparison, construction (4) applies to general Hilbert spaces.

At sample level, we replace the true distribution by the empirical distribution, yielding

$$\hat{\Sigma}_{X^jX^i} = E_n\{[\kappa_j(\cdot, X^j) - E_n\kappa_j(\cdot, X^j)] \otimes [\kappa_i(\cdot, X^i) - E_n\kappa_i(\cdot, X^i)]\}, \tag{6}$$

where $E_n\kappa_i(\cdot, X^i) = n^{-1}\sum_{a=1}^n \kappa_i(\cdot, X_a^i)$. We then define the estimator of $C_{X^jX^i}$ as

$$\hat{C}_{X^jX^i} = (\hat{\Sigma}_{X^jX^j} + \epsilon_n I)^{-1/2} \hat{\Sigma}_{X^jX^i} (\hat{\Sigma}_{X^iX^i} + \epsilon_n I)^{-1/2},\ i \ne j; \quad \hat{C}_{X^iX^i} = I,$$

where $\epsilon_n > 0$ is a regularization constant, whose convergence rate will be discussed in Section 5.1. The estimator $\hat{C}_{XX}$ of C_{XX} is defined as $\{\hat{C}_{X^iX^j}\}_{i,j=1}^p$. Finally, the estimator of the precision operator Θ_{XX} is defined as $\hat{\Theta}_{XX} = \hat{C}_{XX}^{-1}$.

In several aspects, the construction here differs from the direct analog of the estimator of the additive precision operator proposed by Li, Chun, and Zhao (2014) in the multivariate

setting, which is

$$\tilde{\Theta}_{XX} = \mathrm{diag}\left(\hat{\Sigma}_{X^iX^i}^{1/2}\right)(\hat{\Sigma}_{XX} + \epsilon_n I)^{-1}\mathrm{diag}\left(\hat{\Sigma}_{X^iX^i}^{1/2}\right),$$

where $\mathrm{diag}(\hat{\Sigma}_{X^iX^i}^{1/2})$ is the diagonal operator in $\times_{i,j=1}^{p}\mathcal{B}(\mathfrak{M}_i, \mathfrak{M}_j)$ whose diagonal entries are $\{\hat{\Sigma}_{X^iX^i}^{1/2} : i = 1, \ldots, p\}$. First, the regularization constants appear in different places in $\hat{\Theta}_{XX}$ and $\tilde{\Theta}_{XX}$. The new arrangement is guided by asymptotic analysis in Sections 4 and 5.1. The second difference is due to the change of geometry: the coordinates of relevant operators are different in the $L_2(P)$ and RKHS settings. Finally, as can be seen in Section 6, $\hat{\Theta}_{XX}$ is easier to implement than $\tilde{\Theta}_{XX}$ because the latter involves an additional regularization constant. Our experiences indicate that the numerical behavior of the two versions is similar.

4. Consistency

In this and the next sections, we develop the asymptotic theory of our method, which includes consistency, convergence rate, and optimal regularization. This theory is based on the assumption that the random functions X^i are fully observed on $t \in T$. A more careful analysis would take into account that X^i is observed on a finite set of time points and preliminary smoothing is often performed to approximate X^i, which may affect the final asymptotic results. However, this is beyond the scope of the current article and will be left to future research.

The following lemma can be proved similarly to the parallel result in the multivariate setting in Fukumizu, Bach, and Gretton (2007, Lemma 5).

Lemma 1. If Assumption 1 holds, then, for any $i, j \in \mathrm{V}$, $\|\hat{\Sigma}_{X^jX^i} - \Sigma_{X^jX^i}\| = O_P(n^{-1/2})$.

The next three lemmas are needed for proving Theorem 3. Lemma 3 is easily verified by computation and its proof is omitted. Lemmas 2 and 4 are proved in the online supplementary.

Lemma 2. If $A \in \times_{i,j=1}^{p}\mathcal{B}(\mathfrak{M}_i, \mathfrak{M}_j)$, then $\|A\| \le \sum_{i,j=1}^{p}\|A_{ij}\|$.

Lemma 3. Suppose $A, B \in \mathcal{B}(\mathfrak{M})$ are self-adjoint and invertible operators. Then

$$\begin{aligned} A^{-1} - B^{-1} = {} & (A-B)A^{-2} + B^{-2}(A-B) \\ & - B^{-2}A^2(A-B)A^{-2} - B^{-2}(A-B)B^2A^{-2} \\ & - B^{-2}A(A-B)BA^{-2}. \end{aligned}$$

As we mentioned before, under Assumption 3, $C_{X^iX^j} \ge cI$ for some $c > 0$ for any $i \ne j$. This implies C^{-1} and its consistent estimators are bounded or bounded in probability.

Lemma 4. Suppose $A \in \mathcal{B}(\mathfrak{M})$ is a self-adjoint operator such that $A \ge cI$ for some $c > 0$, and $\{\hat{A}^{(n)}\} \subseteq \mathcal{B}(\mathfrak{M})$ is a sequence of self-adjoint random operators such that $\hat{A}^{(n)} \xrightarrow{P} A$. Then

(1) $\|A^{-1}\| < \infty$;
(2) $P(\|(\hat{A}^{(n)})^{-1}\| \le (c-\epsilon)^{-1}) \to 1$ for any $c > \epsilon > 0$.

The next theorem establishes the consistency of $\hat{\Theta}_{XX}$. To prove the consistency of $\hat{\Theta}_{XX}$, we first need the consistency of $\hat{C}_{X^jX^i}$ for each $i, j \in \mathrm{V}$, whose proof is similar to the proof of consistency of the correlation operator in Fukumizu, Bach, and Gretton (2007). So our proof in the online supplementary focuses on the passage from the consistency of $\hat{C}_{X^iX^j}$ to the consistency of $\hat{\Theta}_{XX}$.

Theorem 3. Under Assumptions 1 through 3, $\|\hat{\Theta}_{XX} - \Theta_{XX}\| \xrightarrow{P} 0$.

The above consistency implies the following consistency of the estimated edge set. For a $\rho > 0$, let $\hat{\mathrm{E}}(\rho, \epsilon_n) = \{(i, j) \in \mathsf{C} : \|\hat{\Theta}_{X^iX^j}(\epsilon_n)\| > \rho\}$.

Corollary 2. If Assumptions 1 through 3 hold, then, for all sufficiently small $\rho > 0$,

$$P(\hat{\mathrm{E}}(\rho, \epsilon_n) = \mathrm{E}) \to 1.$$

5. Convergence Rate and Optimal Regularization

In this section, we derive the convergence rate of $\hat{\Theta}_{XX}$. In the process we also derive the convergence rate of $\hat{C}_{X^iX^j}$. Because the convergence rate for the correlation operator was previously unknown even in the multivariate setting, the results here also refine the asymptotic theory of Fukumizu, Bach, and Gretton (2007).

5.1. Convergence Rate

For simplicity, let $X^i = U$ and $X^j = V$. It is easy to see that the following results also apply to the cases where U and V are random vectors rather than random functions. Let

$$\begin{aligned} \hat{C}_{UV} &= (\hat{\Sigma}_{UU} + \epsilon_n I)^{-1/2}\hat{\Sigma}_{UV}(\hat{\Sigma}_{VV} + \epsilon_n I)^{-1/2}, \\ \tilde{C}_{UV} &= (\Sigma_{UU} + \epsilon_n I)^{-1/2}\Sigma_{UV}(\Sigma_{VV} + \epsilon_n I)^{-1/2}. \end{aligned}$$

The next lemma can be verified by simple computation (one of them was given in Fukumizu, Bach, and Gretton 2007). The proof is omitted.

Lemma 5. If A and B are self-adjoint and invertible linear operators, then

$$\begin{aligned} A^{-1/2} - B^{-1/2} &= A^{-3/2}(B^{3/2} - A^{3/2})B^{-1/2} + A^{-3/2}(A-B) \\ &= A^{-1/2}(B^{3/2} - A^{3/2})B^{-3/2} + (A-B)B^{-3/2}. \end{aligned}$$

If ϵ_n and δ_n are two sequences of positive numbers such that $\epsilon_n/\delta_n \to 0$, then we write $\epsilon_n \prec \delta_n$ or $\delta_n \succ \epsilon_n$. If the sequence ϵ_n/δ_n either goes to 0 or is bounded, then we write $\epsilon_n \preceq \delta_n$ or $\delta_n \succeq \epsilon_n$. The next lemma reveals the role played by Tychonoff regularization in the asymptotic order of magnitude.

Lemma 6. For any self-adjoint operator A, $\epsilon_n \prec 1$, and $a > 0$, $b > 0$, we have

$$\|(A + \epsilon_n I)^{-b}A^a\| = O(\epsilon_n^{\min\{0, a-b\}}).$$

If $\hat{A}_n$ is a sequence of self-adjoint random operator with $\|\hat{A}_n\| = O_P(1)$, then

$$\|(\hat{A}_n + \epsilon_n I)^{-b}\hat{A}_n^a\| = O_P(\epsilon_n^{\min\{0, a-b\}}).$$

The next assumption has to do with the smoothness in the relation between U and V. Since we are, in effect, regressing a space of functions on another space of functions, smoothness

in our context is less intuitive than, say, in the nonparametric regression setting.

Assumption 4. There exists a bounded operator D_{UV} and $\beta > 0$ such that

$$C_{UV} = \Sigma_{UU}^{1/2+\beta} D_{UV} \Sigma_{VV}^{1/2+\beta}.$$

This assumption requires, for example, that $\Sigma_{UU}^{-1/2-\beta} C_{UV}$ to be a bounded operator. Since Σ_{UU} is a Hilbert-Schmidt operator under mild conditions (Fukumizu, Bach, and Gretton 2007), the sequence of its eigenvalues tends to 0. Thus, in order for $\Sigma_{UU}^{-1/2-\beta} C_{UV}$ to be bounded, the range space of C_{UV} needs to be sufficiently concentrated on the eigen-spaces of Σ_{UU} with large eigenvalues, that is, the low-frequency components of Σ_{UU}. Moreover, the degree of concentration increases as β increases. For this reason, Assumption 4 can be interpreted as a type of smoothness, with β characterizing the degree of smoothness.

Lemma 7. If Assumptions 1 through 4 hold and $n^{-2/5} \prec \epsilon_n \prec 1$, then

$$\|\hat{C}_{UV} - \tilde{C}_{UV}\| = O_P(n^{-1}\epsilon_n^{-5/2} + n^{-1/2}\epsilon_n^{-1}).$$

The next lemma gives the convergence rate of $\|\tilde{C}_{UV} - C_{UV}\|$.

Lemma 8. If Assumptions 1 through 4 hold and $\epsilon_n \prec 1$, then

$$\|\tilde{C}_{UV} - C_{UV}\| = O_P\left(\epsilon_n^{\min\{1,\beta\}}\right).$$

Combining the convergence rates in Lemma 7 and Lemma 8, we can easily derive the convergence rate of $\hat{\Theta}_{XX}$, as given in the next theorem. The simple proof is omitted.

Theorem 4. If Assumptions 1 through 4 hold and $n^{-2/5} \prec \epsilon_n \prec 1$, then

$$\|\hat{\Theta}_{XX} - \Theta_{XX}\| = O_P\left(n^{-1}\epsilon_n^{-5/2} + n^{-1/2}\epsilon_n^{-1} + \epsilon_n^{\min\{1,\beta\}}\right).$$

Since increasing β above 1 would no longer improve the convergence rate, for the rest of the article we assume $0 < \beta \le 1$, in which case $\epsilon_n^{\min\{1,\beta\}} = \epsilon_n^{\beta}$.

5.2. Optimal Regularization

The convergence rate in Theorem 4 depends on both the smooth index β and the rate of convergence of ϵ_n. The next theorem gives the optimal choice of ϵ_n for a given β, assuming ϵ_n is of the form $n^{-\alpha}$ for some $\alpha > 0$, as well as the optimal rate of $\|\hat{\Theta}_{\hat{X}\hat{X}} - \Theta_{XX}\|$ under such a choice.

Theorem 5. Suppose Assumptions 1 through 4 hold with β in Assumption 4 satisfying $0 < \beta \le 1$, and $\epsilon_n = n^{-\alpha}$ for some $\alpha > 0$. Then the optimal convergence rate is

$$\|\hat{\Theta}_{\hat{X}\hat{X}} - \Theta_{XX}\| \preceq \begin{cases} n^{-\frac{\beta}{2+2\beta}} & \text{if} \quad 1/2 \le \beta \le 1 \\ n^{-\frac{2\beta}{5+2\beta}} & \text{if} \quad 0 < \beta \le 1/2 \end{cases} \tag{7}$$

and the optimal rate for ϵ_n that achieves the above optimal rate is

$$\epsilon_n \propto \begin{cases} n^{-\frac{1}{2+2\beta}} & \text{if} \quad 1/2 \le \beta \le 1 \\ n^{-\frac{2}{5+2\beta}} & \text{if} \quad 0 < \beta \le 1/2. \end{cases}$$

We should mention that the rate in (7) is best that can be achieved by our asymptotic machinery. It may be possible to improve this rate by more complex asymptotic arguments, or modified versions of the estimator.

Note that, when β takes its largest value 1, the optimal convergence rate is $n^{-1/4}$, and the optimal regularization constant is $\epsilon_n = n^{-1/4}$. Also note that the optimal rate of ϵ_n ranges from $n^{-2/5}$ to $n^{-1/4}$. The smoother the relation the larger the optimal penalty. Using Theorem 5 we can also refine Corollary 2 to allow the threshold ρ to goes to 0 with n, which would increase discriminating power of the estimate. The proof is omitted.

Corollary 3. Suppose Assumptions 1 through 4 hold with β being the smooth index.

1. If $0 < \beta \le 1/2$, then for any $\rho_n \succ n^{-\frac{2\beta}{5+2\beta}}$, $P(\hat{\mathrm{E}}(\rho_n, \epsilon_n) = \mathrm{E}) \to 1$.
2. If $1/2 \le \beta \le 1$, then for any $\rho_n \succ n^{-\frac{\beta}{2+2\beta}}$, $P(\hat{\mathrm{E}}(\rho_n, \epsilon_n) = \mathrm{E}) \to 1$.

5.3. Incompletely Observed Random Functions

The above asymptotic results are derived under the premise that each X^i is observed in its entirety, but in reality it can only be observed at a finite set of time points—or a *measurement schedule*—and must be estimated from its observed values at the sampled time points. The convergence rate of the estimator of X^i, say $\hat{X}^i$, depends on both the measurement schedule and the smoothers employed. Wang, Chiou, and Muller (2016) classified these schedules as "dense" or "sparse" according to whether $\hat{X}_i$ has an $n^{-1/2}$ convergence rate. Following this convention, we refer the measurement schedules that are sufficiently frequent so that covariance operators $\Sigma_{X^iX^j}$ can be estimated by the sample covariance operators based on $\hat{X}_1^i, \ldots, \hat{X}_n^i$ at the $n^{-1/2}$ rate as *dense schedules*. Applications involving automated measurements by instruments, such as fMRI, EEG, and smart wearable records, may be regarded as belonging to this category. The asymptotic results such as Theorem 3, Corollary 2, Lemma 7, Theorem 4, Theorem 5, and Corollary 3 still hold under dense schedules.

At the other extreme, the measurement schedules where the number of time points does not go to infinity are referred to as *sparse schedules*. For example, some longitudinal studies belong to this category. For such schedules, consistency can still be achieved if the number of pooled time points converges to infinity in some fashion. For simplicity, we refer to measurement schedules for which the operators $\Sigma_{X^iX^j}$ cannot be estimated at the $n^{-1/2}$ rate as *nondense*. Under such schedules the above asymptotic results in Sections 5.1 and 5.2 need to be modified accordingly.

To take into account the effect of measurement schedule, we allow the estimation rate of the estimator of $\Sigma_{X^iX^j}$ based on approximated sample $\hat{X}^1, \ldots, \hat{X}_n$ to be an arbitrary sequence δ_n satisfying $n^{-1/2} \preceq \delta_n \prec 1$, and rederive convergence rates of Θ_{XX}, ϵ_n, and ρ_n that reflect the rate δ_n. In this way, our asymptotic results are sufficiently flexible to adapt to specific smoothers and measurement schedules. Let

$$\begin{aligned} \hat{\Sigma}_{\hat{X}^i\hat{X}^j} &= E_n[(\hat{X}^i - E_n\hat{X}^i) \otimes (\hat{X}^j - E_n\hat{X}^j)], \\ \hat{C}_{\hat{X}^i\hat{X}^j} &= (\hat{\Sigma}_{\hat{X}^i\hat{X}^i} + \epsilon_n I)^{-1/2} \hat{\Sigma}_{\hat{X}^i\hat{X}^j} (\hat{\Sigma}_{\hat{X}^j\hat{X}^j} + \epsilon_n I)^{-1/2}, \end{aligned}$$

$$\hat{C}_{\hat{X}\hat{X}} = \{\hat{C}_{\hat{X}^i\hat{X}^j}\}_{i,j=1}^{p}, \quad \hat{\Theta}_{\hat{X}\hat{X}} = \hat{C}_{\hat{X}\hat{X}}^{-1}.$$

The next theorem extends Theorem 4. Its proof is similar that of Theorem 4, and is omitted.

Theorem 6. Suppose

1. Assumptions 1 through 4 hold with $0 < \beta \leq 1$.
2. There is a sequence $n^{-1/2} \preceq \delta_n \prec 1$ such that, for each $(i, j) \in \mathrm{V} \times \mathrm{V}$,

$$\|\hat{\Sigma}_{\hat{X}^i\hat{X}^j} - \Sigma_{X^iX^j}\| = O_P(\delta_n).$$

Then, for any tuning parameter ϵ_n satisfying $\delta_n^{4/5} \prec \epsilon_n \prec 1$, we have

$$\|\hat{\Theta}_{\hat{X}\hat{X}} - \Theta_{XX}\| = O_P\left(\delta_n^2\epsilon_n^{-5/2} + \delta_n\epsilon_n^{-1} + \epsilon_n^{\min\{1,\beta\}}\right). \tag{8}$$

If we take $\delta_n = n^{-\alpha}$ for some $0 < \alpha \leq 1/2$, and $\epsilon_n = n^{-c}$ for some $c > 0$, then the rate ϵ_n that achieves optimal convergence rate is derived as follows.

Theorem 7. If the conditions 1 and 2 in Theorem 4 hold with δ_n and ϵ_n being defined in the last paragraph, then the optimal convergence rate is

$$\|\hat{\Theta}_{XX} - \Theta_{XX}\| \preceq \begin{cases} n^{-\frac{\alpha\beta}{1+\beta}} & \text{if} \quad 1/2 \leq \beta \leq 1 \\ n^{-\frac{4\alpha\beta}{5+2\beta}} & \text{if} \quad 0 < \beta \leq 1/2 \end{cases} \tag{9}$$

which is achieved by

$$\epsilon_n \propto \begin{cases} n^{-\frac{\alpha}{1+\beta}} & \text{if} \quad 1/2 \leq \beta \leq 1 \\ n^{-\frac{4\alpha}{5+2\beta}} & \text{if} \quad 0 < \beta \leq 1/2. \end{cases}$$

Consequently,

1. if $0 < \beta \leq 1/2$ and $\epsilon_n \propto n^{-\frac{4\alpha}{5+2\beta}}$, then for any $\rho_n \succ n^{-\frac{4\alpha\beta}{5+2\beta}}$, $P(\hat{\mathrm{E}}(\rho_n, \epsilon_n) = \mathrm{E}) \to 1$;
2. if $1/2 \leq \beta \leq 1$ and $\epsilon_n \propto n^{-\frac{\alpha}{1+\beta}}$, then for any $\rho_n \succ n^{-\frac{\alpha\beta}{1+\beta}}$, $P(\hat{\mathrm{E}}(\rho_n, \epsilon_n) = \mathrm{E}) \to 1$.

This theorem sums up how the convergence optimal rate of $\|\hat{\Theta}_{\hat{X}\hat{X}} - \Theta_{XX}\|$ and the corresponding tuning constant ϵ_n are affected by the smoothness index β, and the measurement schedule index α:

1. the optimal $\hat{\Theta}_{\hat{X}\hat{X}}$ converges to Θ_{XX} faster if the measurement schedule is denser and the relations among X^i and X^j are smoother;
2. the optimal ϵ_n converges to 0 faster if the measurement schedule is denser and slower if the relations among X^i and X^j are smoother.

Again, if we take α and β as their maximum values, then the optimal convergence rate of $\|\hat{\Theta}_{\hat{X}\hat{X}} - \Theta_{XX}\| \to 0$ is $n^{-1/4}$, which is achieved when $\epsilon_n \propto n^{-1/4}$.

6. Implementation

6.1. Coordinate Mapping

In this section, we implement the estimating procedures for FAPO and FASG, which involves representing linear operators as matrices. We carry this out using the following notational system (Horn and Johnson 1985). Let $\mathfrak{N}$ be a finite-dimensional Hilbert space spanned by $\mathfrak{B} = \{b_1, \ldots, b_m\}$, which need not be linearly independent. Any $h \in \mathfrak{N}$ can be written as a linear combination of $b_1, \ldots, b_m$. The vector of coefficients is denoted by $[h]_{\mathfrak{B}}$ and is called the coordinate of h with respect to $\mathfrak{B}$.

We define the power of a self-adjoint operator $A \in \mathcal{B}(\mathfrak{N})$ as follows. Let r be the rank of A. Let $\{(\lambda_i, f_i) : i = 1, \ldots, r\}$ be the eigenvalue-eigenvector pairs of A with nonzero eigenvalues. For any $c \in \mathbb{R}$, let $A^c = \sum_{i=1}^r \lambda_i^c (f_i \otimes f_i)$. Note that, when $c = -1$, we only take the reciprocal of nonzero eigenvalues. Hence A^{-1} has a different meaning from the inverse of A. To avoid confusion, when $c < 0$ we write T^c as $A^{\dagger|c|}$, so that we can reserve A^{-1} as notation for inverse. We write $A^{\dagger 1}$ as $A^{\dagger}$, which is simply the Moore–Penrose inverse of the operator A (Hsing and Eubank 2015, p. 158). The next theorem sums up six important properties of coordinate mapping, of which the first three are well known (see, e.g., Horn and Johnson 1985, p. 31); the last three are tailored for our use. Since these properties are very useful for constructing estimators of linear operators in a systematic fashion, we collect them below as a theorem.

Let $\mathfrak{N}_\alpha$, $\alpha = 1, 2, 3$, be finite-dimensional Hilbert spaces spanned by $\mathfrak{B}_\alpha = \{b_1^{(\alpha)}, \ldots, b_{m_\alpha}^{(\alpha)}\}$. For $A \in \mathcal{B}(\mathfrak{N}_1, \mathfrak{N}_2)$, the $m_2 \times m_1$ matrix $([Ab_1^{(1)}]_{\mathfrak{B}_2}, \ldots, [Ab_n^{(1)}]_{\mathfrak{B}_2})$ is called the coordinate of A relative to $\mathfrak{B}_1$ and $\mathfrak{B}_2$, and is denoted by ${}_{\mathfrak{B}_2}[A]_{\mathfrak{B}_1}$. We call the function ${}_{\mathfrak{B}_2}[\cdot]_{\mathfrak{B}_1} : \mathcal{B}(\mathfrak{N}_1, \mathfrak{N}_2) \to \mathbb{R}^{m_2 \times m_1}$ the *coordinate mapping*.

Theorem 8. Let $(\mathfrak{N}, \mathfrak{B})$ and $\{(\mathfrak{N}_\alpha, \mathfrak{B}_\alpha) : \alpha = 1, 2, 3\}$ be as defined in the last paragraph. The coordinate mapping has the following properties.

1. (evaluation) For any $h \in \mathfrak{N}_1$, $[Ah]_{\mathfrak{B}_2} = ({}_{\mathfrak{B}_2}[A]_{\mathfrak{B}_1})[h]_{\mathfrak{B}_1}$.
2. (linearity) If $A_1, A_2 \in \mathcal{B}(\mathfrak{N}_1, \mathfrak{N}_2)$ and $c_1, c_2 \in \mathbb{R}$, then
$${}_{\mathfrak{B}_2}[c_1A_1 + c_2A_2]_{\mathfrak{B}_1} = c_1\,({}_{\mathfrak{B}_2}[A_1]_{\mathfrak{B}_1}) + c_2\,({}_{\mathfrak{B}_2}[A_2]_{\mathfrak{B}_1}).$$
3. (composition) If $A_1 \in \mathcal{B}(\mathfrak{N}_1, \mathfrak{N}_2)$ and $A_2 \in \mathcal{B}(\mathfrak{N}_2, \mathfrak{N}_3)$, then
$${}_{\mathfrak{B}_3}[A_2A_1]_{\mathfrak{B}_1} = ({}_{\mathfrak{B}_3}[A_2]_{\mathfrak{B}_2})({}_{\mathfrak{B}_2}[A_1]_{\mathfrak{B}_1}).$$
4. (power) If $A \in \mathcal{B}(\mathfrak{N})$ is a self-adjoint and Ω is the Gram matrix of $\mathfrak{B}$: $\Omega = \{\langle b_i, b_j\rangle_{\mathfrak{B}} : i, j = 1, \ldots, m\}$, then, for any $c \in \mathbb{R}$ for which $(-1)^c \in \mathbb{R}$, we have
$${}_{\mathfrak{B}}[A^c]_{\mathfrak{B}} = \Omega^{\dagger 1/2}(\Omega^{1/2}\,{}_{\mathfrak{B}}[A]_{\mathfrak{B}}\,\Omega^{\dagger 1/2})^c\,\Omega^{1/2}.$$
5. (identity) If $I \in \mathcal{B}(\mathfrak{N})$ is the identity mapping, then ${}_{\mathfrak{B}}[I]_{\mathfrak{B}} = Q_{\mathfrak{B}}$, where $Q_{\mathfrak{B}}$ is the projection on to $\mathrm{span}\{[b_1]_{\mathfrak{B}}, \ldots, [b_m]_{\mathfrak{B}}\}$ in the Euclidean space $\mathbb{R}^m$.
6. (matrix of operators) If $\{A_{ij}\}_{i,j=1}^p \in \times_{i,j=1}^p \mathcal{B}(\mathfrak{M}_i, \mathfrak{M}_j)$, $\mathfrak{M}_i$ is spanned by $\mathfrak{B}_i$, and $\mathfrak{B} = \cup_{i=1}^p \mathfrak{B}_i$, then ${}_{\mathfrak{B}}[\{A_{i,j}\}_{i,j=1}^p]_{\mathfrak{B}} = \{{}_{\mathfrak{B}_i}[A_{ij}]_{\mathfrak{B}_j}\}_{i,j=1}^p$.

6.2. Construction of Nested Hilbert Spaces

Let $X_1, \ldots, X_n$ be an iid sample of X. For clarity, we use letters such as $i, j, \ldots$ to represent the components and letters such as $a, b, \ldots$ to represent subjects. Thus, X_a^i is the ith component of the ath vector X_a in the sample $\{X_a : a = 1, \ldots, n\}$. In practice, we can only observe the function $X_a(t) = (X_a^1(t), \ldots, X_a^p(t))^{\mathsf{T}}$

over a finite set of time points, say $S_a = \{t_{a1}, \ldots, t_{am_a}\} \subseteq T$. The functions $t \mapsto X_a(t)$ must then be estimated using the observed data. Commonly used methods for estimating $X_a(t)$ are smoothing spline or RKHS, both of which can be formulated as projection on to a finite-dimensional Hilbert space. In this article, we use a positive definite kernel to generate $\mathcal{H}_i$, but, instead of using the RKHS inner product, we use the L_2 inner product, which is approximated by the Simpson's rule. This inner product works better than the RKHS inner product in our simulations.

Specifically, let $S = \cup_{a=1}^n S_a$, and denote its members as $s_1, \ldots, s_N$ where $N \le \sum_{b=1}^n m_a$. Let J_a be the index set of S_a, so that $S_a = \{s_u : u \in J_a\}$. Let $\kappa_T : T \times T \to \mathbb{R}$ be a positive definite kernel. We take $\mathcal{H}_1, \ldots, \mathcal{H}_p$ to be the same linear subspace spanned by set of functions $\mathcal{L} = \{\kappa_T(\cdot, s_u) : u = 1, \ldots, N\}$ with κ_T as the reproducing kernel.

We use the subset $\mathcal{L}_a = \{\kappa_T(\cdot, s_u) : u \in J_a\}$ to construct an approximation of the function X_a^i. To avoid complicated notations we use the same notation X_a^i to represent the approximation. We set those components of $[X_a^i]_{\mathcal{L}}$ whose indices are not in J_a to 0. Let $[X_a^i]_{\mathcal{L}_a}$ denote the m_a-dimensional subvector of $[X_a^i]_{\mathcal{L}}$ with index $b \in J_a$. Let $X_a(S_a)$ be the observed part of X_a, that is, $X_a(S_a) = \{X_a(s_u) : u \in J_a\}$, and let $K_T^{(a,a)}$ be the submatrix $\{\kappa_T(s_u, s_v) : u, v \in J_a\}$. Then, for any $t \in T$, $X_a^i(t) = \sum_{u \in J_a} ([X_a^i]_{\mathcal{L}})_u \kappa_T(t, s_u)$, and consequently,

$$X_a^i(S_a) = K_T^{(a,a)}[X_a^i]_{\mathcal{L}_a}.$$

Solve this equation with Tychonoff regularization to obtain $[X_a^i]_{\mathcal{L}_a} = (K_T^{(a,a)} + \epsilon^{(T)})^{-1} X_a^i(S_a)$. To summarize, the coordinate $[X_i]_{\mathcal{L}}$ is constructed as

$$([X_a^i]_{\mathcal{L}})_u = \begin{cases} (K_T^{(a,a)} + \epsilon^{(T)})^{-1} X_a^i(S_a) & \text{if } u \in J_a \\ 0 & \text{if } u \notin J_a. \end{cases} \tag{10}$$

Having obtained the entire function X_a^i, we next approximate the inner product between X_a^i and X_b^i in $\mathcal{H}_i$ by the Simpson's rule, as follows. By the above construction, we have

$$X_a^i = [X_a^i]_{\mathcal{L}}^{\mathsf{T}} \kappa_T(\cdot, S), \quad X_b^i = [X_b^i]_{\mathcal{L}}^{\mathsf{T}} \kappa_T(\cdot, S), \tag{11}$$

where $\kappa_T(\cdot, S)$ is the vector of functions $\{\kappa_T(\cdot, s) : s \in S\}$. Evaluate X_a^i at an equally spaced set of points in J, say $U = \{u_0, \ldots, u_\ell\}$, where ℓ is even, and u_0 and u_ℓ are the left and right ends of the interval J. Then, by the Simpson's rule,

$$\int_{u_0}^{u_\ell} X_a^i(t) X_b^i(t) dt \approx X_a^i(U)^{\mathsf{T}} D X_b^i(U), \tag{12}$$

where $D = (h/3)\mathrm{diag}(1, 4, 2, 4, \ldots, 2, 4, 1)$, and $h = (u_\ell - u_0)/\ell$. Substituting (11) into the right-hand side of (12), we have

$$\langle X_a^i, X_b^i \rangle_{\mathcal{H}_i} \approx [X_a^i]_{\mathcal{L}}^{\mathsf{T}} \kappa_T(S, U) D \kappa_T(U, S) [X_b^i]_{\mathcal{L}},$$

where $\kappa_T(U, S)$ denotes the matrix $\{\kappa_T(u, s) : u \in U, s \in S\}$, and $[X_a^i]_{\mathcal{L}}$ and $[X_b^i]_{\mathcal{L}}$ are determined by (10).

This estimation framework accommodates both the balanced case, where $S_1, \ldots, S_n$ are the same, and the unbalanced case, where they are not. In particular, in the balanced case, $X_1^i, \ldots, X_n^i$ are observed on the same set of time points $\{t_1, \ldots, t_m\}$, so that $N = m_a = m$ for all a. In this case, each X_a^i is expressed as a linear combination of $\{\kappa_T(\cdot, t_1), \ldots, \kappa_T(\cdot, t_N)\}$, and no change of notation is needed.

Having constructed $\mathcal{H}_1, \ldots, \mathcal{H}_p$, $X_1, \ldots, X_n$, and the inner product in $\mathcal{H}_i$, we now define the second-level spaces $\mathfrak{M}_i$ as the RKHS spanned by

$$\mathfrak{B}_i = \{\phi_a^i = \kappa_i(\cdot, X_a^i) - E_n \kappa_i(\cdot, X^i) : a = 1, \ldots, n\},$$

where the kernel κ_i is determined by the inner product of $\mathcal{H}_i$, as described in Section 2.1. Note that $\mathfrak{B}_i$ is not a linearly independent set: it spans an $n-1$ dimensional space. Let 1_n be the n-dimension vector with its components being identically 1, I_n be the $n \times n$ identity matrix, and $Q = I_n - 1_n 1_n^{\mathsf{T}}/n$. Then it is easy to see that $Q = Q_{\mathfrak{B}_i}$. Consequently, for any $\phi \in \mathfrak{M}_i$, $[\phi]_{\mathfrak{B}_i} = Q[\phi]_{\mathfrak{B}_i}$. Thus, if we let $K_i = \{\kappa_i(X_\ell^i, X_m^i)\}_{\ell,m=1}^n$ and $G_i = QK_iQ$, then the inner product in $\mathfrak{M}_i$ can be expressed in coordinate system as

$$\langle \phi, \psi \rangle_{\mathfrak{B}_i} = [\phi]_{\mathfrak{B}_i}^{\mathsf{T}} K_i [\psi]_{\mathfrak{B}_i} = [\phi]_{\mathfrak{B}_i}^{\mathsf{T}} G_i [\psi]_{\mathfrak{B}_i}.$$

6.3. Estimation of FAPO and FASG

Our goal is to find the norms $\|\hat{\Theta}_{X^iX^j}\|$ and then threshold them to estimate the edge set E. The coordinate of $\hat{\Sigma}_{X^iX^j}$ with respect to $\mathfrak{B}_i$ and $\mathfrak{B}_j$ is given by the next proposition.

Proposition 1. For any $i, j = 1, \ldots, p$, ${}_{\mathfrak{B}_i}[\Sigma_{X^iX^j}]_{\mathfrak{B}_j} = n^{-1} G_j$.

This result is well known in the multivariate setting (see Fukumizu, Bach, and Jordan 2009; Li, Chun, and Zhao 2012), and the proof is omitted. By Proposition 1 and Theorem 8 (parts 1, 2, 3, 5), the coordinate of $\hat{C}_{X^jX^i}$ for $i \neq j$ is

$${}_{\mathfrak{B}_j}[\hat{C}_{X^jX^i}]_{\mathfrak{B}_i} = (n^{-1}G_j + \epsilon^{(X)}Q)^{\dagger 1/2}(n^{-1}G_i)(n^{-1}G_i + \epsilon^{(X)}Q)^{\dagger 1/2}. \tag{13}$$

For $i = j$, $\hat{C}_{X^iX^i}$ is defined to be $I \in \mathcal{B}(\mathfrak{M}_i)$, and hence, by Theorem 8 (part 5),

$${}_{\mathfrak{B}_i}[\hat{C}_{X^iX^i}]_{\mathfrak{B}_i} = {}_{\mathfrak{B}_i}[I]_{\mathfrak{B}_i} = Q_{\mathfrak{B}_i} = Q. \tag{14}$$

Let $\mathfrak{B} = \cup_{i=1}^p \mathfrak{B}_i$. By Theorem 8 (part 6), ${}_{\mathfrak{B}}[\hat{C}_{XX}]_{\mathfrak{B}}$ is an $np \times np$ matrix and its inversion, if done directly, can be computationally expensive when both n and p are large. The next Lemma gives an explicit form of ${}_{\mathfrak{B}}[\hat{\Theta}_{XX}]_{\mathfrak{B}}$ that only involves inversion of $n \times n$ matrices.

Lemma 9. If $H \in \mathbb{R}^{np \times n}$ and $\Lambda \in \mathbb{R}^{np \times np}$ satisfy

$$H = HQ, \quad H = (I_n \otimes Q)H, \quad \Lambda = (I_n \otimes Q)\Lambda = \Lambda(I_n \otimes Q),$$
$$\Lambda\Lambda^{\dagger} = I_p \otimes Q,$$

then $(HH^{\mathsf{T}} + \Lambda)^{\dagger} = \Lambda^{\dagger} - \Lambda^{\dagger}H(H^{\mathsf{T}}\Lambda^{\dagger}H + Q)^{\dagger}H^{\mathsf{T}}\Lambda^{\dagger}$.

We now present the coordinate of $\hat{\Theta}_{XX}$ with respect to $\mathfrak{B}$. For a set of matrices $D_1, \ldots, D_m$, let $\mathrm{diag}(D_1, \ldots, D_m)$ be the block diagonal matrix with $D_1, \ldots, D_m$ as the diagonal blocks.

Theorem 9. Let $A_i = (n^{-1}G_i)^{1/2}(n^{-1}G_i + \epsilon^{(X)}Q)^{\dagger 1/2}$, $\Lambda_i = Q - A_i^2$. Let

$$\Lambda = \mathrm{diag}(\Lambda_1, \ldots, \Lambda_p), \quad H = (A_1, \ldots, A_p)^{\mathsf{T}},$$
$$\Omega = \mathrm{diag}(G_1, \ldots, G_p).$$

Then

$${}_{\mathfrak{B}}[\hat{\Theta}_{XX}]_{\mathfrak{B}} = \Omega^{\dagger 1/2}(\Lambda^{\dagger} - \Lambda^{\dagger}H(H^{\mathsf{T}}\Lambda^{\dagger}H + Q)^{\dagger}H^{\mathsf{T}}\Lambda^{\dagger})\Omega^{1/2}. \tag{15}$$

Note that, since Λ and Ω are block diagonal, $\Lambda^\dagger$ and $\Omega^\dagger$ can be computed by $\Lambda_i^\dagger$ and $G_i^\dagger$. Hence the largest matrices to be inverted in (15) are $n \times n$ matrices. The next corollary gives an explicit form of $\|\hat{\Theta}_{X^iX^j}\|$.

Corollary 4. If $i \neq j$, then $\|\hat{\Theta}_{X^iX^j}\|$ is the largest singular value of the matrix

$$\Lambda_i^\dagger H_i(\textstyle\sum_{\ell=1}^p H_\ell \Lambda_\ell^\dagger H_\ell + Q)^\dagger H_j \Lambda_j^\dagger.$$

Alternatively, the above matrix can be written as $G_i^{\dagger 1/2}{}_{\mathfrak{B}_i}[\hat{\Theta}_{X^iX^j}]_{\mathfrak{B}_j} G_i^{\dagger 1/2}$.

Having derived $\|\hat{\Theta}_{X^iX^j}\|$, we estimate the edge set E by

$$\hat{\mathrm{E}}(\epsilon^{(X)}, \rho_n) = \{(i,j) \in \mathsf{C} : \|\hat{\Theta}_{X^iX^j}\| \le \rho_n\},$$

where ρ_n is a predetermined sequence satisfying the condition in Corollary 3.

6.4. Tuning

The tuning parameters include the kernel parameters in κ_T and $\kappa_1, \ldots, \kappa_p$, and the regularization parameters ϵ_T and ϵ_X. We now propose their tuning procedures. For an integer $a \in \{1, \ldots, n\}$, we let e_a denote the n-dimensional vector whose ath component is 1 and whose other components are all 0, and let $\|\cdot\|_F$ denote the Frobenius matrix norm.

Tuning the kernel parameters. Assume that the Gaussian radial basis functions (RBF) are used. That is,

$$\begin{aligned} \kappa_T : &\quad T \times T \to \mathbb{R}, \quad (t_1, t_2) \mapsto \exp\{-\gamma_T (t_1 - t_2)^2\}, \\ \kappa_i : &\quad \mathcal{H}_i \times \mathcal{H}_i \to \mathbb{R}, \quad (x_1^i, x_2^i) \mapsto \exp\{-\gamma_i \|x_1^i - x_2^i\|_{\mathcal{H}_i}^2\}. \end{aligned} \tag{16}$$

Mimicking the criterion in Lee, Li, and Chiaromonte (2013), we choose γ_T and γ_i as

$$\begin{aligned} 1/\sqrt{\gamma_T} &= \binom{N}{2}^{-1} \sum_{k=1}^{N-1} \sum_{\ell=k+1}^{N} |s_k - s_\ell|, \\ 1/\sqrt{\gamma_i} &= \binom{n}{2}^{-1} \sum_{a=1}^{n-1} \sum_{b=a+1}^{n} \left\| X_a^i - X_b^i \right\|_{\mathcal{H}_i}. \end{aligned} \tag{17}$$

Tuning regularization constant ϵ_T. By (10), the function $X_a^i(t)$ for any t is approximated by $\kappa_T(t, S_a)^\mathsf{T} (K_T^{(a,a)} + \epsilon_T I_{m_a})^{-1} X_a^i(S_a)$. So the approximation error of $X_a^i(S_a)$ is

$$\|X_a^i(S_a) - \kappa_T(t, S_a)^\mathsf{T} (K_T^{(a,a)} + \epsilon_T I_{m_a})^{-1} X_a^i(S_a)\|^2.$$

The total approximation error of $X_a^i(S_a)$ for $a = 1, \ldots, n$ is then

$$\left\| X_a^i(S_a) - K_T^{(a,a)} \left(K_T^{(a,a)} + \epsilon_T I_{m_a} \right)^{-1} X_a^i(S_a) \right\|_F^2.$$

To achieve appropriate scaling, we reset ϵ_T to $\lambda_{\min}(K_T^{(a,a)})\epsilon_T$, where $\lambda_{\min}(K_T^{(a,a)})$ is the largest eigenvalue of the matrix $K_T^{(a,a)}$. The reset ϵ_T is the proportion of the largest eigenvalue of the matrix to be regularized. We propose the following generalized cross-validation criterion:

$$\begin{aligned} &\mathrm{GCV}_T(\epsilon_T) \\ &= \sum_{i=1}^p \sum_{a=1}^n \frac{\|X_a^i(S_a) - K_T^{(a,a)} (K_T^{(a,a)} + \epsilon_T \lambda_{\max}(K_T^{(a,a)}) I_{m_a})^{-1} X_a^i(S_a)\|_F^2}{\{1 - \mathrm{trace}[K_T^{(a,a)} (K_T^{(a,a)} + \epsilon_T \lambda_{\max}(K_T^{(a,a)}) I_{m_a})^{-1}]/m_a\}^2}, \end{aligned}$$

where $m_a = \mathrm{card}(S_a)$. We minimize this criterion over a grid to determine the optimal ϵ_T.

Tuning regularization constant ϵ_X. Let $\mathfrak{R}_i = \mathrm{ran}(\Sigma_{X^iX^i})$, that is,

$$\mathfrak{R}_i = \mathrm{span}\{\kappa(\cdot, X_a^i) - E_n \kappa(\cdot, X^i) : a = 1, \ldots, n\}, \quad i = 1, \ldots, p.$$

Our strategy is to predict the functions in $\mathfrak{R}_j$ by the functions in $\mathfrak{R}_i$ for all $i \neq j$. Let $b_a^i = \kappa(\cdot, X_a^i) - E_n \kappa(\cdot, X^i)$. By Lee, Li, and Zhao (2016b), the member of $\mathfrak{R}_i$ that is stochastically closest to b_a^j is $\Sigma_{X^iX^i}^{-1} \Sigma_{X^iX^j} b_a^j$. That is, the stochastic difference

$$\mathrm{var}_n \left\{ b_a^j(X^j) - \left(\Sigma_{X^iX^i}^{-1} \Sigma_{X^iX^j} b_a^j \right)(X^i) \right\}$$

is minimum among all members of $\mathfrak{R}_i$. The coordinate of $\Sigma_{X^iX^i}^{-1} \Sigma_{X^iX^j} b_a^j$ is

$$\left[\Sigma_{X^iX^i}^{-1} \Sigma_{X^iX^j} b_a^j \right] = \left[\Sigma_{X^iX^i}^{-1} \right] \left[\Sigma_{X^iX^j} \right] \left[b_a^j \right] = G_i^\dagger G_j e_a,$$

where $[b_a^j] = e_a$ because $[b_a^j]$ is the coordinate with respect to $\{b_1^j, \ldots, b_n^j\}$. We use the regularized version $\Sigma_{X^iX^i}^{-1} \Sigma_{X^iX^j} b_a^j$—that is,

$$[(G_i + \epsilon_X I_n)^{-1} G_j e_a]^\mathsf{T} \left(b_1^i, \ldots, b_n^i \right)^\mathsf{T} \equiv \hat{b}_a^{j|i}, \tag{18}$$

to estimate b_a^j. The next lemma gives the stochastic distance between b_a^j and $\hat{b}_a^{j|i}$.

Lemma 10. If b_a^j and $\hat{b}_a^{j|i}$ as defined the last paragraph, then

$$\mathrm{var}_n \left\{ b_a^j(X^j) - \hat{b}_a^{j|i}(X^i) \right\} = \|G_j e_a - G_i (G_i + \epsilon_X I_n)^{-1} G_j e_a\|^2. \tag{19}$$

The total stochastic error for estimating b_a^j for $a = 1, \ldots, n$ is, then,

$$\sum_{a=1}^n \|G_j e_a - G_i (G_i + \epsilon_X I_n)^{-1} G_j e_a\|^2 = \|G_j - G_i (G_i + \epsilon_X I_n)^{-1} G_j\|_F^2.$$

As before, we reset ϵ_X to $\epsilon_X \lambda_{\max}(G_i)$ to achieve appropriate scaling, and propose the following generalized cross-validation criterion for ϵ_X

$$\mathrm{GCV}_X(\epsilon_X) = \sum_{i<j} \frac{\|G_j - G_j (G_i + \epsilon_X \lambda_{\max}(G_i) I_n)^{-1} G_i\|^2}{\{1 - \mathrm{trace}[(G_i + \epsilon_X \lambda_{\max}(G_i) I_n)^{-1} G_i]/n\}^2}.$$

We minimize this criterion over a grid of numbers to determine the optimal ϵ_X.

6.5. Algorithm

We summarize the procedures developed in Sections 6.2–6.4 as the following algorithm.

1. Choose the kernels κ_T for $\mathcal{H}_1, \ldots, \mathcal{H}_p$. For example, if we choose κ_T as the RBF in (16), then choose the tuning constant γ_T according to (17); if we choose κ_T to be the Brownian motion covariance kernel $\min(s, t)$, then no tuning parameter is needed.

2. Choose the regularization parameter ϵ_T by minimizing $\mathrm{GCV}_T(\epsilon_T)$ over a grid of points.
3. Using the results of steps 1 and 2 to construct functions $X_1, \ldots, X_n$ according to (10).
4. Using the result of step 3 to compute K_i, G_i, Λ_i, and H_i to form Λ, H, Ω and eventually ${}_{\mathfrak{B}}[\hat{\Theta}_{XX}]_{\mathfrak{B}}$ as defined in Theorem 9, with ϵ_X in ${}_{\mathfrak{B}}[\hat{\Theta}_{XX}]_{\mathfrak{B}}$ chosen by minimizing $\mathrm{GCV}_X(\epsilon_X)$ over a grid of numbers.
5. For each $(i, j) \in \mathsf{C}$, read off the submatrices ${}_{\mathfrak{B}_i}[\hat{\Theta}_{X^iX^j}]_{\mathfrak{B}_j}$ from ${}_{\mathfrak{B}}[\hat{\Theta}_{XX}]_{\mathfrak{B}}$, and compute the largest singular value of $\hat{\sigma}_{ij} = G_i^{1/2}\,{}_{\mathfrak{B}_i}[\hat{\Theta}_{X^iX^j}]_{\mathfrak{B}_j} G_i^{\dagger 1/2}$.
6. For a chosen $\rho_n > 0$, let $\hat{\mathrm{E}} = \{(i, j) \in \mathsf{C} : \hat{\sigma}_{ij} > \rho_n\}$. For example, if we assume $\beta = 1$, then we can take $\rho_n \propto n^{-1/5}$.

7. Simulations

In this section, we compare our FAPO estimator with the FGGM estimator proposed by Qiao, James, and Lv (2014) and a heuristic copula method derived from the nonparanormal model proposed by Liu, Lafferty, and Wasserman (2009) in the classical setting.

The method of Liu, Lafferty, and Wasserman (2009) was developed for conventional graphical models with a scalar random variable observed on each vertex, and no extension has yet been available to the functional setting. While Solea and Li (2016) proposed functional copula graphical model, it is by no means a straightforward generalization. To avoid too much digression from the main theme, we use the following naive adaptation of the nonparanormal method to the current setting in the balanced case. Let $\{(X_a^i(t_1), \ldots, X_a^i(t_m)) : a = 1, \ldots, n, i = 1, \ldots, p\}$ be the raw data—observation on the random function X_a^i at time points $t_1, \ldots, t_m$. We first transform each sample $\{X_a^i(t_r) : a = 1, \ldots, n\}$ into normal scores using nonparametric copula Gaussian transformation based on the Windsorized empirical distribution described in Liu, Lafferty, and Wasserman (2009). We then apply the FGGM to the transformed data to construct the network. We refer to this method as the naive normal score method, abbreviated as NNS in the manuscript. NNS has been incorporated in all the simulation comparisons for the balanced cases and the fMRI data example.

As mentioned in the Introduction, nonlinearity and heteroscedasticity are two of the consequences of non-Gaussian interdependence, where a Gaussian-based method such as FGGM is expected to perform poorly. On the other hand, under the Gaussian assumption the nonparametric nature of FAPO estimator may incur loss of efficiency relative to FGGM. For these reasons, we include three scenarios in our comparison: nonlinearity dependence, heteroscedasticity dependence, and Gaussian dependence. Because in this section superscripts and powers will frequently appear within the same formulas, we use $X^{(i)}$ to indicate X^i, and $(X^{(i)})^c$ to indicate the power of $X^{(i)}$.

Comparison 1: nonlinear models. We use the following models with $p = 5$ and $p = 10$:

$$\begin{aligned}
\text{Model I:}\quad & X^{(i)}(t) = \epsilon^{(i)}(t),\ i = 2, 5;\ X^{(1)}(t) = (1 + |X^{(4)}(t)|)^2 \\
& \quad + \epsilon^{(1)}(t), \\
& X^{(3)}(t) = \sin(\pi X^{(1)}(t)) + \epsilon^{(3)}(t),\ X^{(4)}(t) = 3X^{(2)}(t)^2 \\
& \quad + \epsilon^{(4)}(t).
\end{aligned}$$

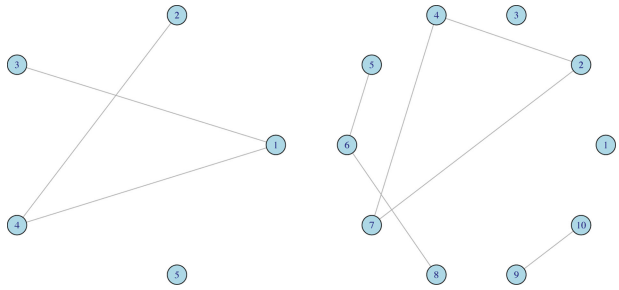

Figure 2. Graphs of models I and II.

$$\begin{aligned}
\text{Model II:}\quad & X^{(i)}(t) = \epsilon^{(i)}(t),\ i = 1, 2, 3, 5, 9; \\
& X^{(4)}(t) = X^{(2)}(t)^2 + \epsilon^{(4)}(t);\ X^{(6)}(t) = \sin(\pi X^{(5)}(t)) \\
& \quad + \epsilon^{(6)}(t); \\
& X^{(7)}(t) = X^{(2)}(t)^2 + (1 + |X^{(4)}(t)|)^3 + \epsilon^{(7)}(t); \\
& X^{(8)}(t) = X^{(6)}(t)^3 + \epsilon^{(8)}(t);\ X^{(10)}(t) = \exp[X^{(9)}(t)] \\
& \quad + \epsilon^{(10)}(t).
\end{aligned} \tag{20}$$

The random functions $\epsilon^{(i)}(t)$ are generated as $\sum_{k=1}^m \xi_k \kappa_T(t, t_k)$, where $\xi_1, \ldots, \xi_m$ are iid $N(0, 1)$, $t_1, \ldots t_m$ are iid $U[0, 1]$, $\kappa_T(t, s) = \min(t, s)$ is the Brownian motion covariance kernel, and $m = 50$. We include both balanced and unbalanced samples. For the balanced samples, $t_{i1}, \ldots, t_{im_i}$ are equally spaced 10 points in $[0, 1]$. For the unbalanced samples, $t_{i1}, \ldots, t_{im_i}$ are 10 points randomly chosen without replacement from 100 equally spaced points in $[0, 1]$.

Each set of structural equations in (20) generates a directed acyclic graph (DAG), which is designed so that its moral graph coincides with its skeleton. In this case, we can simply remove the arrows the DAG to obtain the corresponding undirected graphs, which are used as the targets of estimation. Figure 2 shows the graphs for Models I and II.

In the simulation, the sample sizes are taken to be $n = 100$ and $n = 200$. The simulation sample size is $n_{\mathrm{sim}} = 50$. The FAPO estimator is computed using the algorithm in Section 6.5 with $\kappa_T(s, t) = \min(s, t)$. To save computing time, for each batch of $n_{\mathrm{sim}} = 50$ simulation runs, we take the average of the GCV outcomes from the first five simulation runs and use it for the rest of the batch. For FGGM, the number of functional PCA components used is the smallest number that explains at least 90% of the total variation, which is typically 2.

In this and all the other simulation experiments in this section, we use the Brownian motion covariance kernel to construct the first-level space $\mathcal{H}_i$, coupled with the L_2-inner product described in Section 6.2. For the second-level space $\mathfrak{M}_i$, we used the Gaussian radial basis function (second line in (16)) with γ_i therein defined by (17) as the kernel for the RKHS. The regularization constants ϵ_T and ϵ_X are chosen by the GCV criteria described in Section 6.4, with the grid for ϵ_T being $\{5 \times 10^{-\ell} : \ell = 0, \ldots, 6\}$, and the grid for ϵ_X being $\{5 \times 10^{-\ell} : \ell = 0, \ldots, 5\}$. Because the tuning constants are very stable from sample to sample, for each simulation scenario we just compute the GCV's for the first five samples, and use their average value for the entire simulation scenario.

Figures 3 shows the ROC curves for the combinations of Models (I, II) and sample sizes $n = (100, 200)$ for the balanced

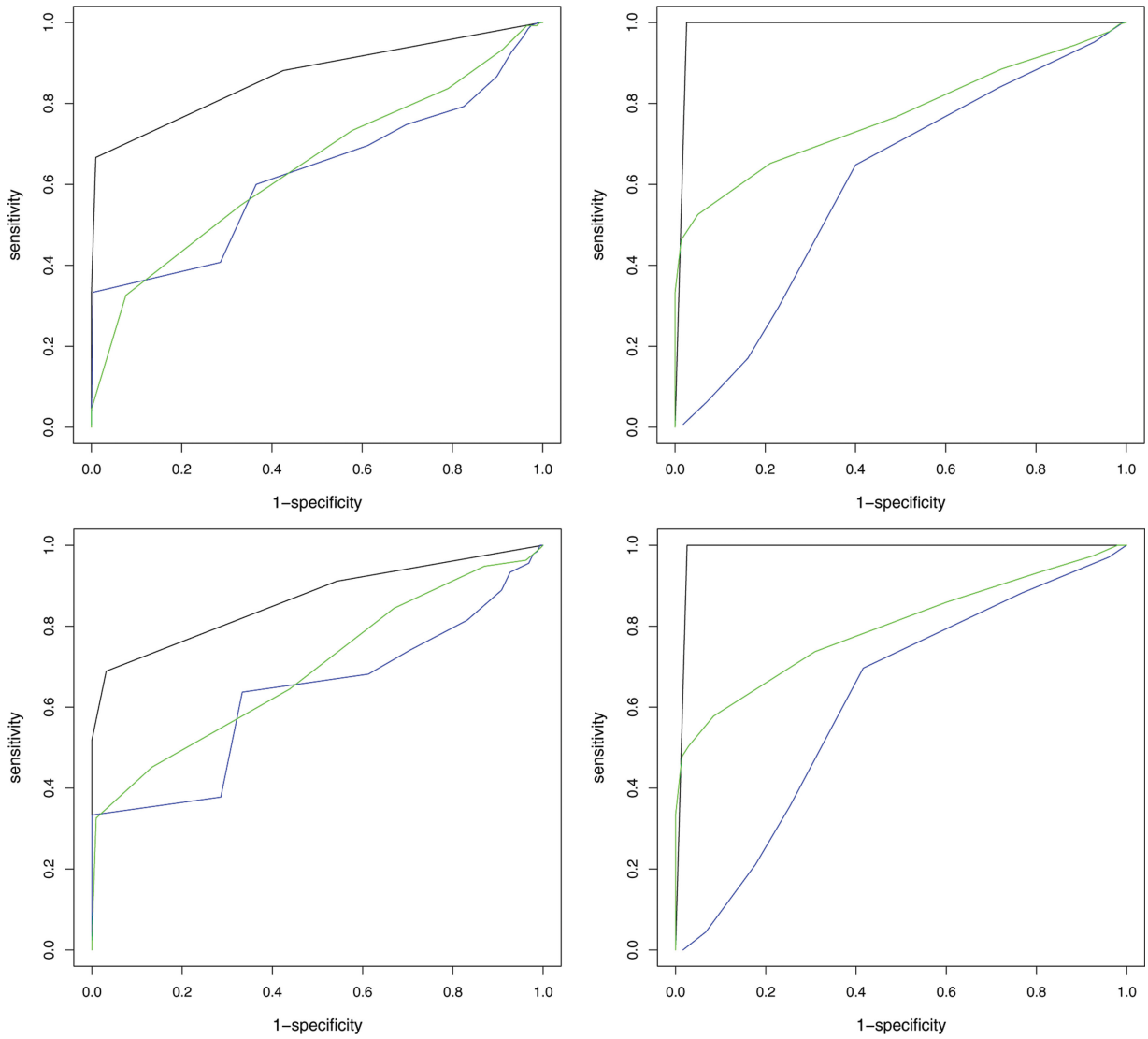

Figure 3. ROC curves (black for FAPO, blue for FGGM, and green for NNS) for Models I (left panels) and II (right panels), and for $n = 100$ (upper panels), and $n = 200$ (lower panels) in balanced case.

case. Each panel contains two ROC curves, with black indicating FAPO and blue indicating FGGM, and each curve being the average of the ROC curves across the $n_{\text{sim}} = 50$ simulation runs. We can see that the areas under curve (AUC) for the FAPO estimator are substantially larger than FGGM. We then repeat the simulation for the unbalanced samples (Figure 4), which again exhibits the superior performance of the FAPO estimator.

The reason underlying the better performance of FAPO is the presence of nonlinear interactions, particularly those quadratic relations with no intercepts, such as equations for $X^{(1)}$ and $X^{(4)}$ in Model I, and equations for $X^{(4)}$ and $X^{(7)}$ in Model II. Essentially, Gaussian graphical models rely on mutual linear regressions among random elements on the vertices—this is implicit in Yuan and Lin (2007) but explicit in some other Gaussian-assumption-based approaches such as Meinshausen and Bühlmann (2006), Friedman, Hastie, and Tibshirani (2008), and Peng et al. (2009). The same can be said of the FGGM. Because a linear regression model cannot pick up any information from a relation that is symmetric about the origin, such as $X^{(4)} = 3X^{(2)}(t)^2 + \epsilon^{(4)}$, the edges with this type of dependence are totally missed by FGGM. The FAPO estimator can detect this type of dependence because the kernel offers a rich family of functions to accommodate such symmetric relations.

In Table 1, we report the AUC for the above simulation. Entries of the table are means and standard deviations (in parentheses) of AUC over $n_{\text{sim}} = 50$ simulation samples.

Comparison 2: heteroscedastic model. Another advantage of the FAPO estimator is it regresses, in effect, the family of functions at one vertix on the family of functions at another vertex. Thus, implicitly, it not only regresses $X^{(1)}$ on $X^{(2)}$, but also $(X^{(1)})^2$ on $X^{(2)}$, $X^{(1)}$ on $(X^{(2)})^2$, $(X^{(1)})^2$ on $(X^{(2)})^2$, and so on, with all the regression coefficients implicitly contained in the covariance operator. This feature makes it possible to capture the dependency that eludes any mean-based regression models such as $X^{(1)} = f(X^{(2)}) + \epsilon^{(2)}$, of which FGGM is a special case. We use a simple model to demonstrate this point:

$$\text{Model III: } X^{(1)}(t) = 5\exp(X^{(3)}(t))\epsilon^{(1)}(t),$$
$$X^{(2)}(t) = (1 + 0.5|X^{(3)}(t)|)^3\epsilon^{(2)}(t),\ X^{(3)}(t) = \epsilon^{(3)}(t).$$

Obviously, the edge set is $\mathrm{E} = \{(1, 3), (2, 3)\}$. In this model, $X^{(2)}$ depends on $X^{(3)}$ through the conditional variance

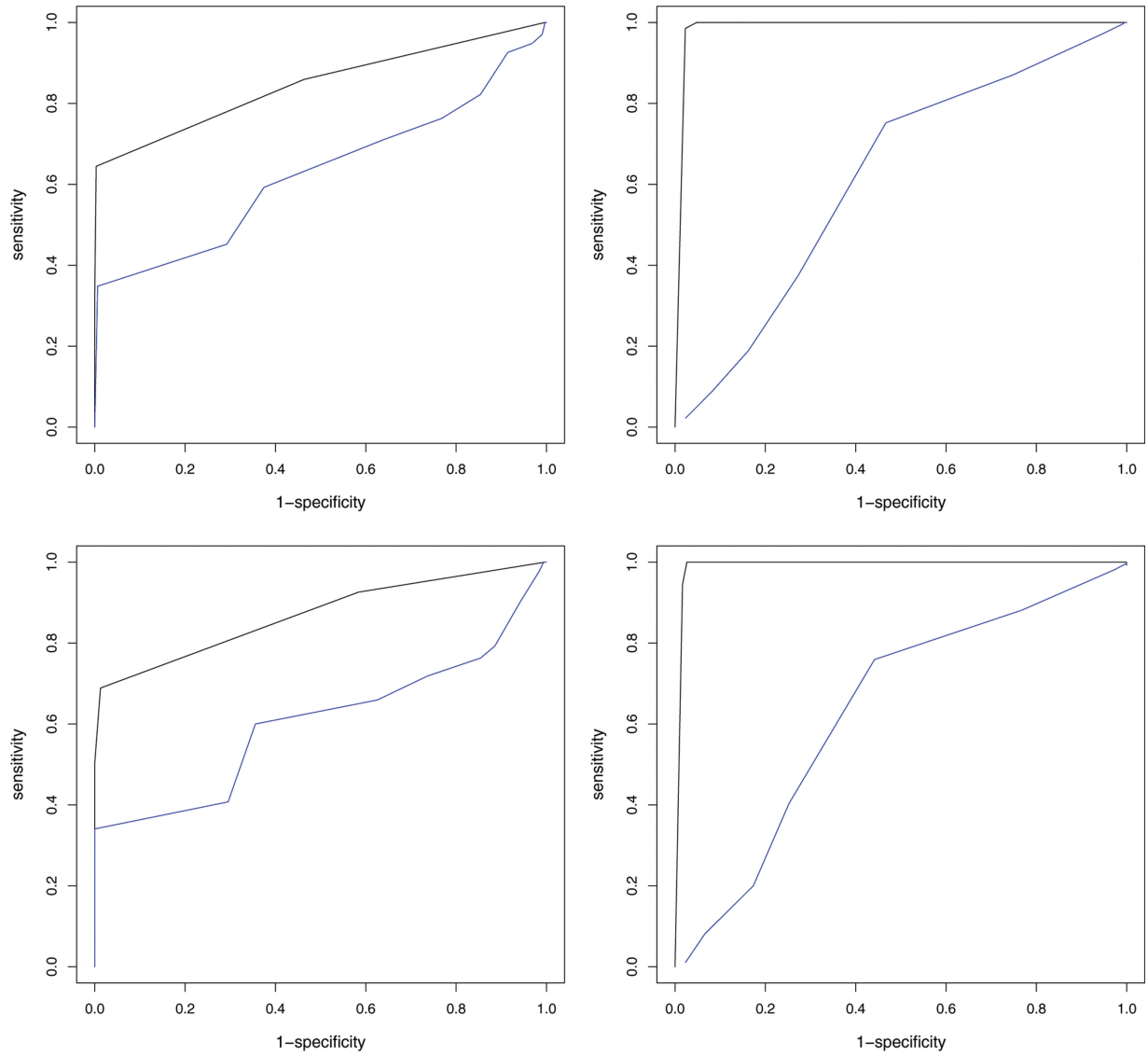

Figure 4. FAPO (black) and FGGM (blue) applied to the same comparison scenarios as in Figure 3 for the unbalanced case.

$\text{var}(X^{(2)}|X^{(3)})$ rather than the conditional mean $E(X^{(2)}|X^{(3)})$. The same can be said of the dependence of $X^{(1)}$ on $X^{(3)}$.

The random functions $\epsilon^{(1)}$, $\epsilon^{(2)}$, and $\epsilon^{(3)}$ are generated in the same way as in Comparison 1. Figure 5 shows the average ROC curves for the two estimators, for four combinations of scenarios: {balance, unbalanced} $\times$ $\{n = 100,\ n = 200\}$, which indicate that FAPO performs much better than FGGM.

In Table 2, we report the averages and standard errors of the AUC's based on $n_{\text{sim}} = 50$ for different comparison scenarios.

Comparison 3: Gaussian model. We now compare the performances of FAPO and FGGM under the Gaussian assumption to see how much information we lose as compared with a parametric model under model assumption. This model is taken from Qiao, James, and Lv (2014), as follows:

$$\text{Model IV:} \quad X_i(t) = \sum_{k=1}^{m} \xi_{ik} \upsilon_{ik}(t), \quad i = 1, \ldots p,$$

Table 2. Averages and standard errors of AUC for Model III.

	$n = 100$			$n = 200$		
Time	FAPO	FGGM	NNS	FAPO	FGGM	NNS
B	0.87 (0.24)	0.45 (0.11)	0.52 (0.19)	0.84 (0.24)	0.67 (0.12)	0.45 (0.26)
U	0.94 (0.22)	0.5 (0.00)	NA	0.98 (0.01)	0.25 (0.00)	NA

Table 1. Averages and standard errors (in parentheses) of AUC for Models I and II. The left column indicates sample type: B for balanced and U for unbalanced.

		$n = 100$			$n = 200$		
Time	Model	FAPO	FGGM	NNS	FAPO	FGGM	NNS
B	I	0.87 (0.10)	0.62 (0.08)	0.64 (0.20)	0.87 (0.08)	0.63 (0.05)	0.69 (0.18)
	II	0.99 (0.00)	0.61 (0.03)	0.77 (0.10)	0.98 (0.00)	0.63 (0.04)	0.80 (0.10)
U	I	0.85 (0.08)	0.73 (0.12)	NA	0.87 (0.08)	0.61 (0.06)	NA
	II	0.99 (0.00)	0.62 (0.05)	NA	0.99 (0.005)	0.65 (0.04)	NA

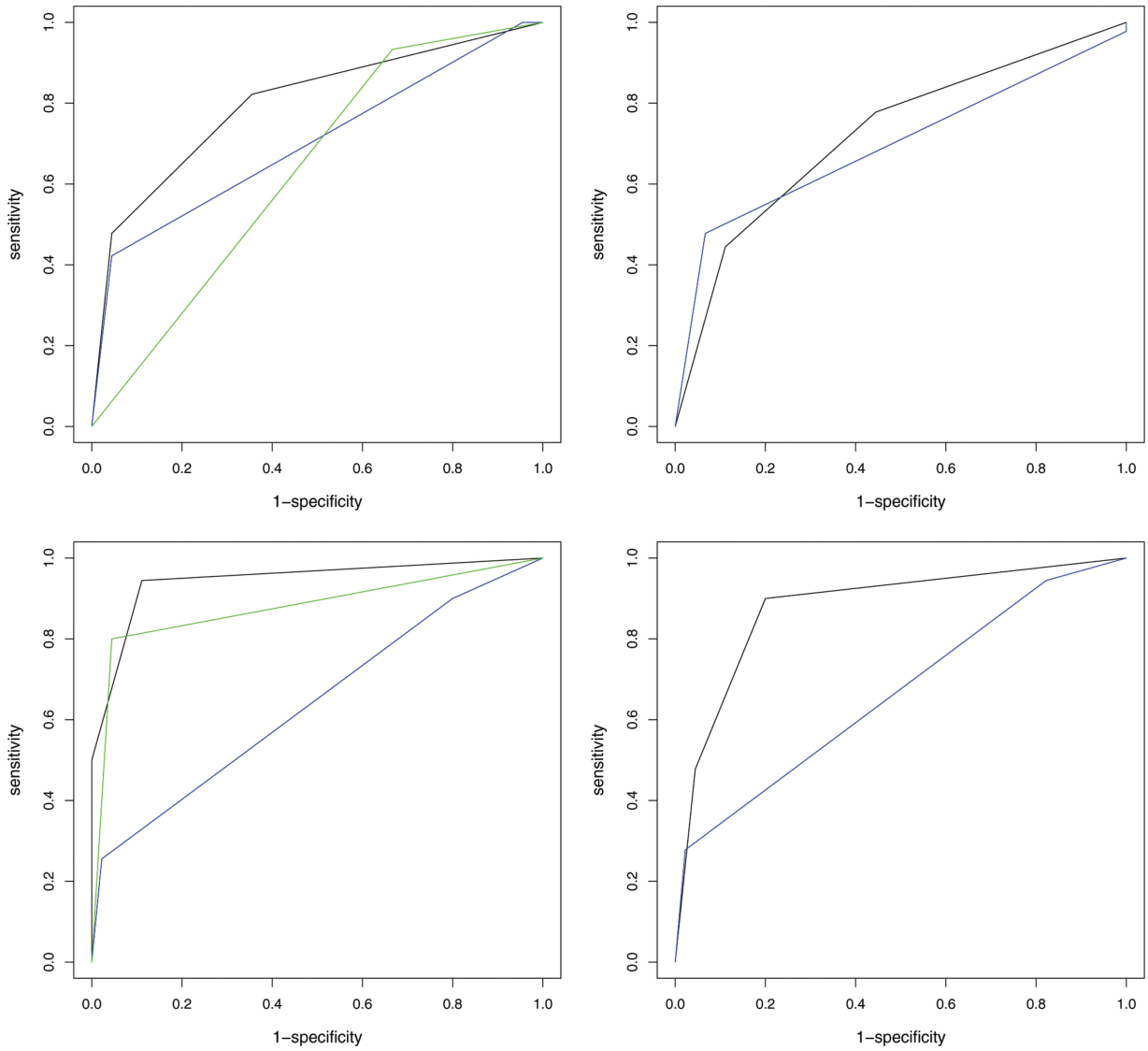

Figure 5. ROC curves (black for FAPO, blue for FGGM, and green for NNS) for Model III with $n = 100$ (upper panels) and $n = 200$ (lower panels), in the balanced case (left panels) and unbalanced case (right panels).

where $m = 5$, $\{v_{ik},\ k = 1, 2, 3, 4, 5\}$ are the first five functions in the Fourier basis:

$$1,\ \sqrt{2}\sin(2\pi t),\ \sqrt{2}\cos(2\pi t),\ \sqrt{2}\sin(4\pi t),\ \sqrt{2}\cos(4\pi t),$$

and $\xi = (\xi_{11}, \ldots, \xi_{1m}, \ldots, \xi_{p1}, \ldots, \xi_{pm})^{\mathsf{T}}$ is multivariate Gaussian with mean 0 and block precision matrix $\Theta \in \mathbb{R}^{pm \times pm}$

$$\Theta_{ij} = \begin{cases} I_m & i = j \\ 0.4I_m & |i - j| = 1 \\ 0.2I_m & |i - j| = 2 \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

The graph determined by Model IV is shown in Figure 6.

The FAPO and FGGM estimators are computed as before. The averaged ROC curves and the averaged AUC and their standard deviations for the two estimators are shown in Figure 7 and Table 3, which indicate that, under the Gaussian assumption, the nonparametric FAPO does not perform as well as the parametric FGGM estimator and the NNS method.

Comparison 4: higher dimensions and larger sample sizes. We now compare FAPO with other methods for higher dimensions

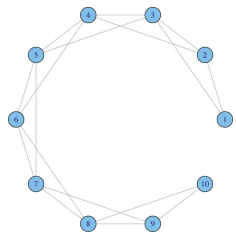

Figure 6. Graph of model IV.

and larger sample sizes: $p = 20, 30, 40$ and $n = 100, 200, 300$. All three models, the nonlinear model, the heteroscedastic model, and the Gaussian model are included in this comparison. Increasing the size of the network p requires us to create

Table 3. Averages and standard errors for AUC for model IV.

$n = 100$			$n = 200$		
FAPO	FGGM	NNS	FAPO	FGGM	NNS
0.68 (0.05)	0.82 (0.04)	0.82 (0.04)	0.65 (0.05)	0.89 (0.04)	0.88 (0.04)

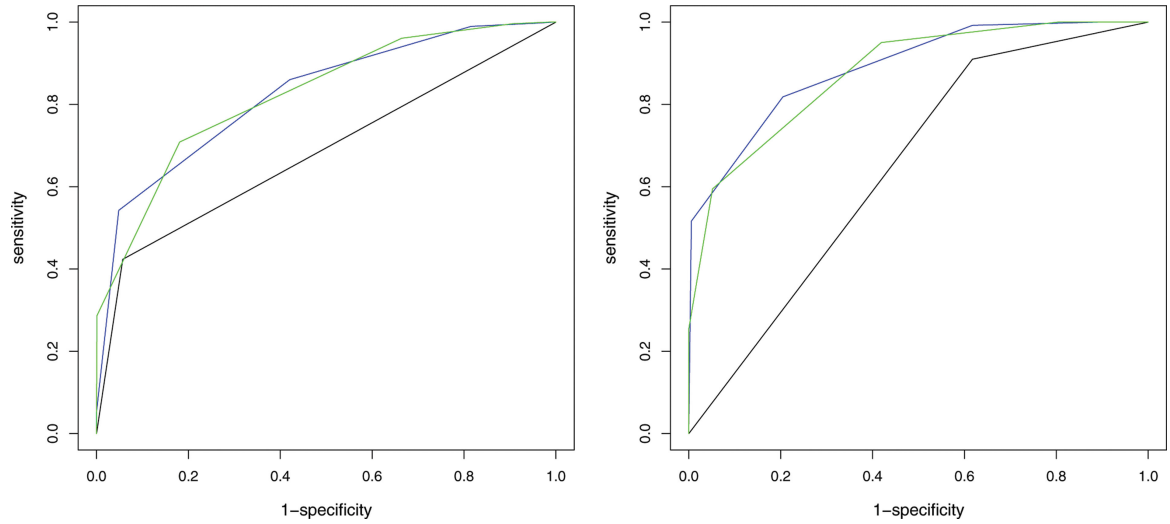

Figure 7. ROC curves (black for FAPO, blue for FGGM, and green for NNS) for Model IV with $n = 100$ (left) and $n = 200$ (right) in the balanced case.

new edges in the graph, which we do by augmenting Model II (nonlinear), III (heteroscedastic), and IV (Gaussian).

For the nonlinear model with $p = 20$, we augment model II by adding the nodes $11, \ldots, 20$ and the following edges:

$$\text{Model II}' : \begin{array}{l} X^{(14)}(t) = (1 + |X^{(17)}(t)|)^3 + \epsilon^{(14)}(t); \\ X^{(15)}(t) = (1 + |X^{(14)}(t)|)^2 + \epsilon^{(15)}(t); \\ X^{(18)}(t) = 3X^{(14)}(t)^2 + \epsilon^{(18)}(t); \\ X^{(20)}(t) = \exp[X^{(19)}(t)] + \epsilon^{(20)}(t). \end{array}$$

For $p = 30$, we augment Model II′ by adding nodes $21, \ldots, 30$ and the edges:

$$\text{Model II}'' : \begin{array}{l} X^{(22)}(t) = (1 + |X^{(23)}(t)|)^2 + \epsilon^{(22)}(t); \\ X^{(26)}(t) = 3X^{(22)}(t)^2 + \epsilon^{(26)}(t); \\ X^{(27)}(t) = \sin(\pi X^{(29)}(t)) + \epsilon^{(27)}(t). \end{array}$$

For $p = 40$, we further augment Model II″ by adding nodes $31, \ldots, 40$ and the edges:

$$\text{Model II}''' : \begin{array}{l} X^{(32)}(t) = 3X^{(30)}(t)^2 + \epsilon^{(32)}(t); \\ X^{(39)}(t) = (1 + |X^{(32)}(t)|)^2 + \epsilon^{(39)}(t). \end{array}$$

We create the heteroscedastic model with $p = 20$ as follows:

$$\text{Model III}' : \begin{array}{ll} X^{(i)}(t) = \epsilon^{(i)}(t),\ i \in M; & X^{(4)}(t) = X^{(2)}(t)^2\epsilon^{(4)}(t); \\ X^{(6)}(t) = \sin(\pi X^{(5)}(t))\epsilon^{(6)}(t); & X^{(7)}(t) = (X^{(2)}(t)^2 + (1 + |X^{(4)}(t)|)^3)\epsilon^{(7)}(t); \\ X^{(8)}(t) = X^{(6)}(t)^3\epsilon^{(8)}(t); & X^{(10)}(t) = \exp[X^{(9)}(t)]\epsilon^{(10)}(t); \\ X^{(14)}(t) = (1 + |X^{(17)}(t)|)^3\epsilon^{(14)}(t); & X^{(15)}(t) = (1 + |X^{(14)}(t)|)^2\epsilon^{(15)}(t); \\ X^{(18)}(t) = 3X^{(14)}(t)^2\epsilon^{(18)}(t); & X^{(20)}(t) = \exp[X^{(19)}(t)]\epsilon^{(20)}(t), \end{array}$$

where $M = \{1, 2, 3, 5, 9, 11, 12, 13, 17, 19\}$. For $p = 30$, we augment Model III′ by adding the nodes $21, \ldots, 30$ and the edges

$$\text{Model III}'' : \begin{array}{l} X^{(22)}(t) = (1 + |X^{(23)}(t)|)^2\epsilon^{(22)}(t); \\ X^{(26)}(t) = 3X^{(22)}(t)^2\epsilon^{(26)}(t); \\ X^{(27)}(t) = \sin(\pi X^{(29)}(t))\epsilon^{(27)}(t). \end{array}$$

For $p = 40$, we augment Model III″ by adding the nodes $31, \ldots, 40$ and the edges

$$\text{Model III}''' : \begin{array}{l} X^{(32)}(t) = 3X^{(30)}(t)^2\epsilon^{(32)}(t); \\ X^{(39)}(t) = (1 + |X^{(32)}(t)|)^2\epsilon^{(39)}(t). \end{array}$$

For the Gaussian with $p = 20, 30, 40$, we simply augment the block precision matrix (21) for $i, j = 1, \ldots, p$. We label these augmented models as IV′, IV″, and IV‴. Table 4 shows the AUC values of the estimates for these models for balanced cases. We can see the advantage of FAPO holds up very well against the increase in the network size.

8. Application to fMRI Data

We now apply FAPO and FGGM to the fMRI dataset mentioned in the Introduction, which is taken from Consortium (2012). The goal is to infer the brain network structures for healthy children and for children with ADHD. The data consist of 79 subjects, with 42 healthy subjects and 37 ADHD subjects. The ADHD group is further divided into 23 in the ADHD Combined group, 12 in the ADHD Inattentive group, and 2 in the ADHD Hyperactive group. In our analysis, we used the 42 healthy subjects and the 22 subjects in the ADHD Combined group (one subject in the ADHD group is removed because it contains a significant amount of missing observations). Technical details regarding the sample and the scanning parameters can be found at the ADHD-200 Consortium.

The dataset was preprocessed by the NeuroBureau community using the Athena pipeline. One hundred sixteen brain regions-of-interest were constructed for the preprocessed resting-state fMRI using the anatomical labeling atlas (AAL) developed by Craddock et al. (2012). fMRI time series were extracted for each of the 116 regions by averaging all voxels time series within each region at each time point, resulting in 74 time points for each of the 116 regions for each subject. Hence, for each subject we have 116 different regional fMRI time series, observed at 74 time points. The AAL atlas and the regional fMRI time series are publicly available at NITRC

Table 4. Averages and standard errors of AUC for higher p and n.

			Sample size n		
p	Model	Method	100	200	300
20	I′	FAPO	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)
		FGGM	0.72 (0.04)	0.73 (0.04)	0.73 (0.03)
		NNS	0.70 (0.04)	0.74 (0.04)	0.74 (0.04)
	III′	FAPO	0.91 (0.03)	0.93 (0.02)	0.94 (0.02)
		FGGM	0.57 (0.04)	0.57 (0.04)	0.58 (0.03)
		NNS	0.63 (0.06)	0.64 (0.08)	0.65 (0.07)
	IV′	FAPO	0.62 (0.03)	0.73 (0.02)	0.78 (0.02)
		FGGM	0.85 (0.03)	0.95 (0.01)	0.98 (0.01)
		NNS	0.85 (0.03)	0.94 (0.01)	0.98 (0.01)
30	I″	FAPO	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)
		FGGM	0.72 (0.04)	0.72 (0.04)	0.74 (0.04)
		NNS	0.69 (0.03)	0.70 (0.04)	0.72 (0.03)
	III″	FAPO	0.89 (0.04)	0.92 (0.03)	0.93 (0.02)
		FGGM	0.57 (0.03)	0.57 (0.03)	0.58 (0.03)
		NNS	0.64 (0.07)	0.66 (0.07)	0.66 (0.05)
	IV″	FAPO	0.66 (0.03)	0.73 (0.02)	0.79 (0.01)
		FGGM	0.77 (0.03)	0.93 (0.02)	0.97 (0.01)
		NNS	0.77 (0.03)	0.93 (0.01)	0.97 (0.01)
40	I‴	FAPO	0.99 (0.01)	0.99 (0.00)	0.99 (0.00)
		FGGM	0.72 (0.05)	0.73 (0.04)	0.73 (0.03)
		NNS	0.70 (0.04)	0.73 (0.03)	0.72 (0.03)
	III‴	FAPO	0.91 (0.03)	0.96 (0.01)	0.98 (0.01)
		FGGM	0.58 (0.04)	0.58 (0.04)	0.59 (0.04)
		NNS	0.61 (0.06)	0.63 (0.03)	0.65 (0.06)
	IV‴	FAPO	0.68 (0.02)	0.73 (0.01)	0.79 (0.01)
		FGGM	0.77 (0.02)	0.85 (0.02)	0.80 (0.02)
		NNS	0.72 (0.04)	0.75 (0.01)	0.73 (0.02)

Because of the size of the data, it takes considerable amount time to apply FGGM in its original form. We therefore use a thresholding version of FGGM, that is, if the operator norm of the (i, j)th block of the precision matrix of the functional principal components is smaller than a threshold, then declare there is no edge between vertices i and j. We applied FAPO and the threshold-FGGM to the data to estimate the network for the control and the ADHD Combined group. For each group, we computed brain networks by taking the thresholds for both estimators to be such that 5% of the pairs of vertices are edges. This is a somewhat arbitrary threshold which we chose so that the edges do not look too crowded. Ideally chosen by a significance test of ACI, but this is beyond the scope of the present article and we leave it to future research.

Again, we used the Brownian motion covariance kernel for the first-level spaces, and the Gaussian radial basis kernel for the second-level spaces. The tuning constants ϵ_T and ϵ_T are selected in the same way as in the simulation experiments. We also used the Gaussian radial basis kernel for the first-level space, with similar result (not reported here).

Figure 8 shows the constructed networks for scenarios in the combinations

$$\{\text{FAPO, FGGM}\} \times \{\text{control, ADHD}\}.$$

Red vertices correspond to those with more than 20 edges. We observe that the brain networks have different patterns for the two groups. For example, from the FAPO network, we observe dense connectivity in the inferior occipital gyrus regions (nodes 53 and 54) for the ADHD group relative to the control group. Moreover, compared to the ADHD group, we see increased functional connectivity for the control group in the middle frontal gyri nodes (nodes 9 and 10), in the gyrus rectus node (node 27), and in the vermis nodes (nodes 110, 113, and 114).

From Figure 8, we see that the graphs for the ADHD and the control groups are quite different. It is then natural to ask whether this difference is due to group variation or random variation among individuals. We now investigate this question for the FAPO networks. Since we are unaware of any formal test that can determine this in our nonparametric and functional data setting, we use the following heuristic approach. Since the graphs are determined by the covariance operator Σ_{XX}, the question boils down to whether the difference in the covariance operators is due to group variation or the variation among individuals. More formally, letting $\Sigma_{XX}^{(1)}$ and $\Sigma_{XX}^{(0)}$ represent the covariance operators for the ADHD and the control group, respectively, we would like to see whether there is a significant difference between them. To do so, we randomly split the sample into two parts, of sample sizes 23 and 42, respectively, which correspond to the actual sample sizes for the ADHD and the control groups. We perform the random splitting 100 times. For the sth split sample, we apply (6) to each group to estimate $\Sigma_{XX}^{(0)}$ and $\Sigma_{XX}^{(1)}$, and denote them by $\hat{\Sigma}_{XX}^{(0)}(s)$ and $\hat{\Sigma}_{XX}^{(1)}(s)$. We then compute the operator norms

$$R_s = \left\| \hat{\Sigma}_{XX}^{(1)}(s) - \hat{\Sigma}_{XX}^{(0)}(s) \right\|, \quad s = 1, \ldots, 100.$$

Let $s = 0$ represent the true the sample, and let $R_0 = \|\hat{\Sigma}_{XX}^{(1)}(0) - \hat{\Sigma}_{XX}^{(0)}(0)\|$. Figure 9 shows the histogram of $R_1, \ldots, R_{100}$ with the position of R_0 marked by a red vertical line. It shows that the R_0 is significantly larger than those produced by random splitting.

9. Discussion

In this article, we introduce a nonparametric functional graphical model based on additive conditional independence. The generalization hinges on the additively nested Hilbert spaces, where the first space contains functions of time, which is used to represent functional data, and the second space consists of functions defined on the first space, which is used to characterize the potentially nonlinear and heteroscedastic relations among the random functions. The two spaces are additively nested—additive in the sense that functions from different vertices are put together by linear combination; nested in the sense that the inner product of the first space determines the kernel of the second space. The additive nature of this approach allows us to avoid high-dimensional kernels, which is a source of the curse of dimensionality.

Our simulation studies indicate that the new method works better than the FGGM when the interdependencies among vertices are nonlinear or heteroscedastic. We have developed an efficient algorithms to implement the estimation and tuning for FAPO. The method is very easy to use: the largest inverses involved in the algorithm are $n \times n$ matrices regardless of the size of the network; so it can be applied relative large networks.

In several ways, this article goes far beyond a formal generalization of the ASG model to the functional case. First, we developed the consistency of the estimator, which was not developed in Li, Chun, and Zhao (2014) in the multivariate setting. Second, we derived the convergence rate for FAPO. This

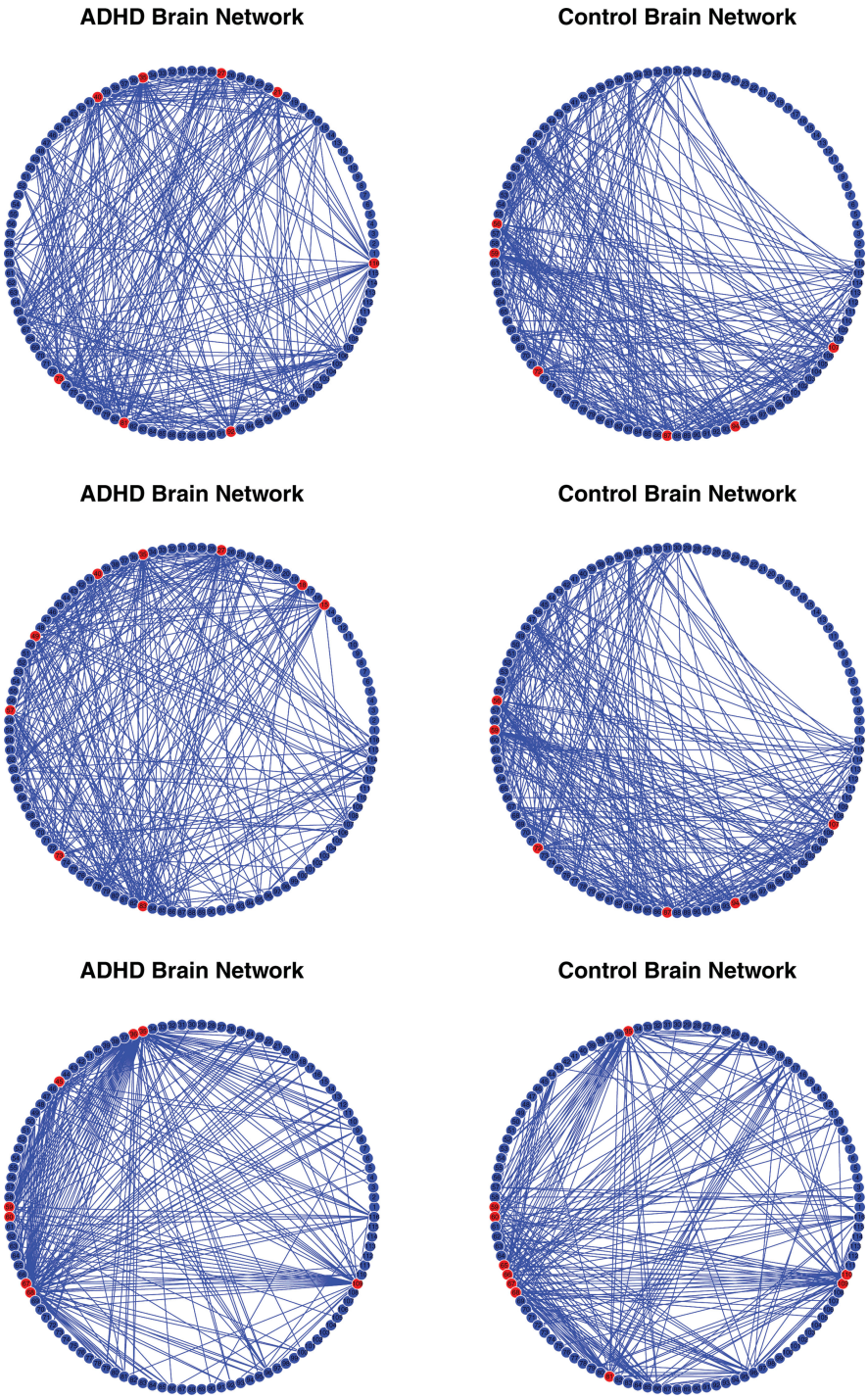

Figure 8. Brain networks by FGGM (upper panels), NNS (middle panels), and FAPO (lower panels) from the fMRI data for the ADHD Combined group (left panels) and control group (right panels).

is significant because such convergence rates are completely novel. In particular, our FAPO estimator is derived from an extension of the correlation operator of Fukumizu, Bach, and Gretton (2007), and to this point no convergence rates were available for this operator. Since our convergence rate is easily modified to cover the multivariate setting, our work also fills a gap in that theory. Third, our asymptotic derivation leads us to construct an FAPO estimator that is different from what a straightforward extension of the estimator of Li, Chun, and Zhao (2014) would be, in that the Tychonoff regularization is placed on different operators. This modification assists us to prove consistency and derive the convergence rates. It also simplifies computation because it only requires one Tychonoff regularization parameter; whereas the procedure in Li, Chun, and Zhao (2014) requires an additional regularization parameter. Fourth, in developing the asymptotic theory for the FAPO estimator, we gather together a list of six important properties of the coordinate mapping, which is conducive to systematically developing operator-based estimation procedures. Such procedures are increasingly important for big and complex data. Finally, our application to the fMRI network data reveals the great potentials for nonparametric functional graphical models, and the need for further developing related theories and methodologies. Our results indicate that ADHD and

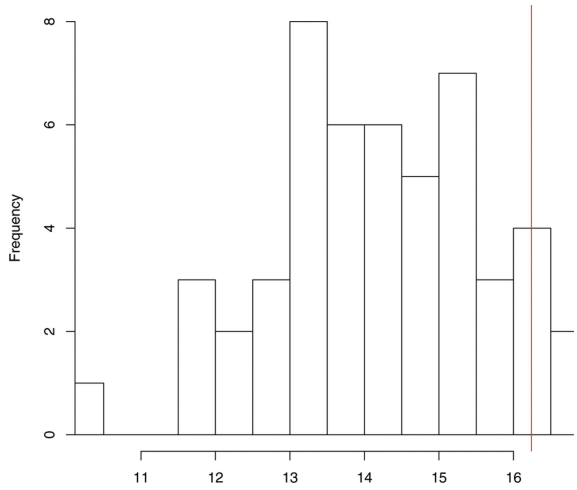

Figure 9. Histograms for R_s and position of R_0 (red line).

non-ADHD subjects do exhibit rather different brain network patterns—at least based on the fMRI data we have analyzed.

Kernel representation of conditional independence has been used previously in a variety of contexts. For example, Fukumizu, Bach, and Gretton (2007) used it in canonical correlation analysis, Fukumizu, Bach, and Jordan (2009) used it in sufficient dimension reduction, and Song, Fukumizu, and Gretton (2013) used it in kernel belief propagation and kernel Bayesian filtering. The difference between our approach and theirs is our use of ACI, which need not be driven by any conditional distribution, but nevertheless possesses the semigraphoid properties that are key to constructing a graphical model. Since ACI only relies on inner product, it is more flexible than conditional independence. Moreover, it allows us to avoid multi-dimensional kernels and mitigate the curse of dimensionality. Another manifestation of ACI is the Gaussian-like equivalence (5), which does not hold generally for conditional independence. Since its introduction by Li, Chun, and Zhao (2014), ACI has also been used and further developed in other contexts. For example, Lee, Li, and Zhao (2016b) used it in for variable selection; Lee, Li, and Zhao (2016a) introduced an additive partial correlation operator for evaluating ACI, and Liu, Lee, and Zhao (2016) used ACI for variable screening in the high-dimensional setting. The current article carries this idea further to construct nonparametric graphical models for functional observations.

The idea advanced in this article opens up a number of possibilities for further development. For example, one could apply, at the operator level, more sophisticated sparse estimation techniques than thresholding, such as LASSO (Tibshirani 1996), SCAD (Fan and Li 2001), and adaptive LASSO (Hui 2006). In this connection, Lee, Li, and Zhao (2016b) developed LASSO-based nonparametric variable selection procedure using the ACI principal in the nonparametric regression setting, by imposing sparsity at the operator level, which may be adopted to further refine our thresholding estimator for the FASG.

Supplementary Materials

The online supplementary materials contain the proofs of all the Theorems, Lemmas, and Corollaries in the main article.

Acknowledgments

The authors thank two referees and an associate editor for their many useful comments and suggestions that helped them greatly in improving an earlier manuscript. The authors are indebted to Professors X. Qiao, G. James, and J. Lv for sharing with them their computer codes, and to Professor N. Lazar for introducing them to the Human Connectome Project.

Funding

The research of Bing Li is supported in part by the National Science Foundation Grant DMS-1407557.

References

Bach, F. R. (2008), "Consistency of the Group Lasso and Multiple Kernel Learning," *The Journal of Machine Learning Research*, 9, 1179–1225. [1641]

Baker, C. R. (1973), "Joint Measures and Cross-Covariance Operators," *Transactions of The American Mathematical Society*, 186, 273–289. [1641]

Bickel, P., and Levia, E. (2008), "Covariance Regularization and Thresholding," *The Annals of Statistics*, 36, 2577–2604. [1637]

Bosq, D. (2000), *Linear Process in Function Spaces: Theory and Application* (*Lecture Notes in Statistics*, Vol. 149), New York: Springer. [1640]

Cheng, R., and Herskovits, E. H. (2007), "Graphical-Model-Based Multivariate Analysis of Functional Megnetic Resonance Data," *NeuroImage*, 35, 635–647. [1637]

Consortium (2012), "The ADHD-200 Consortium: A Model to Advance the Translational Potential of Neuroimaging in Clinical Neuroscience," *Frontiers in Systems Neuroscience*, 6. [1651]

Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P., and Mayberg, H. S. (2012), "A whole Brain fMRI Atlas Generated via Spatially Constrained Spectral Clustering," *Human Brain Mapping*, 33, 1914–1928. [1651]

Dawid, A. P. (1979), "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society*, Series B, 41, 1–31. [1640]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1654]

Ferraty, F., and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, New York: Springer Science & Business Media. [1637]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [1637,1648]

Fukumizu, K., Bach, F. R., and Gretton, A. (2007), "Statistical Consistency of Kernel Canonical Correlation Analysis," *The Journal of Machine Learning Research*, 8, 361–383. [1642,1643,1653,1654]

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2009), "Kernel Dimension Reduction in Regression," *The Annals of Statistics*, 37, 1871–1905. [1641,1645,1654]

Horn, R. A., and Johnson, C. R. (1985), *Matrix Analysis*, Cambridge: Cambridge University Press. [1644]

Horváth, L., and Kokoszka, P. (2012), *Inference for Functional Data With Applications* (Vol. 200), New York: Springer Science & Business Media. [1637]

Hsing, T., and Eubank, R. (2015), *Theoretical Foundations of Functional Data Analysis, With an Introduction to Linear Operators*, New York: Wiley. [1637,1644]

Hui, Z. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1654]

Lazar, N. A., Lura, B., Sweeney, J. A., and Eddy, W. F. (2002), "Combining Brains: A Survey of Methods for Statistical Pooling of Information," *NeuroImage*, 16, 538–550. [1637]

Lee, K.-Y., Li, B., and Chiaromonte, F. (2013), "A General Theory for Nonlinear Sufficient Dimension Reduction: Formulation and Estimation," *The Annals of Statistics*, 41, 221–249. [1646]

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. [1654]

——— (2016b), "Variable Selection via Additive Conditional Independence," *Journal of the Royal Statistical Society,* Series B, 78, 1037–1055. [1639,1641,1646,1654]

Li, B., Chun, H., and Zhao, H. (2012), "Sparse Estimation of Conditional Graphical Models With Application to Gene Networks," *Journal of the American Statistical Association*, 107, 152–167. [1645]

Li, B., Chun, H., and Zhao, H. (2014), "On an Additive Semigraphoid Model for Statistical Networks With Application to Pathway Analysis," *Journal of the American Statistical Association*, 109, 1188–1204. [1639,1640,1641,1652,1654]

Li, B., Kim, M. K., and Altman, N. (2010), "On Dimension Folding of Matrix-or Array-Valued Statistical Objects," *The Annals of Statistics*, 38, 1094–1121. [1637]

Li, B., and Song, J. (2017), "Nonlinear Sufficient Dimension Reduction for Functional Data," *The Annals of Statistics*, 45, 1059–1095. [1639]

Liu, H., Lafferty, J., and Wasserman, L. (2009), "The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs," *The Journal of Machine Learning Research*, 10, 2295–2328. [1647]

Liu, T., Lee, K.-Y., and Zhao, H. (2016), "Ultrahigh Dimensional Feature Selection via Kernel Canonical Correlation Analysis," available at *https://arxiv.org/abs/1604.07354*. [1654]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs With the Lasso," *The Annals of Statistics*, 34, 1436–1462. [1637,1648]

Pearl, J., Geiger, D., and Verma, T. (1989), "Conditional Independence and Its Representations," *Kybernetika*, 25, 33–44. [1639,1640]

Pearl, J., and Verma, T. (1987), *The Logic of Representing Dependencies by Directed Graphs*, University of California, Los Angeles, Computer Science Department. [1640]

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial Correlation Estimation by Joint Sparse Regression Models," *Journal of American Statistical Association*, 104, 735–749. [1637,1648]

Qiao, X., James, G., and Lv, J. (2014), "Functional Graphical Models," Technical Report, University of South California. [1637,1638,1639,1647,1649]

Ramsay, J. O., and Silverman, B. W. (2002), *Applied Functional Data Analysis: Methods and Case Studies* (Vol. 77), New York: Springer. [1637]

Silverman, B., and Ramsay, J. (2005), *Functional Data Analysis*, New York: Springer. [1637]

Solea, E., and Li, B. (2016), "Copula Gaussian Graphical Model for Functional Data," Manuscript submitted for publication. [1640,1647]

Song, L., Fukumizu, K., and Gretton, A. (2013), "Kernel Embeddings of Conditional Distributions: A Unified Kernel Framework for Nonparametric Inference in Graphical Models," *IEEE Signal Processing Magazine*, 30, 98–111. [1654]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [1654]

Wang, J.-L., Chiou, J.-M., and Muller, H.-G. (2016), "Functional Data Analysis," *Annual Review of Statistics and Its Application*, 3, 257–295. [1643]

Yao, F., Müller, H.-G., and Wang, J.-L. (2005), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590. [1637]

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in Gaussian Graphical Model," *Biometrika*, 94, 19–35. [1637,1648]

Zhu, H., Strawn, N., and Dunson, D. B. (2016), "Bayesian Graphical Models for Multivariate Functional Data," *Journal of Machine Learning Research*, 17, 1–27. [1637]