Journal of Machine Learning Research 23 (2022) 1-82 Submitted 3/20; Published 3/22

FuDGE: A Method to Estimate a Functional Differential Graph in a High-Dimensional Setting

Boxin Zhao BOXINZ@UCHICAGO.EDU
Booth School of Business
The University of Chicago
Chicago, IL 60637, USA
Y. Samuel Wang YSW7@CORNELL.EDU
Department of Statistics and Data Science
Cornell University
Ithaca, NY 14853, USA
Mladen Kolar MKOLAR@CHICAGOBOOTH.EDU
Booth School of Business
The University of Chicago
Chicago, IL 60637, USA

Editor: Daniela Witten

Abstract

We consider the problem of estimating the difference between two undirected functional graphical models with shared structures. In many applications, data are naturally regarded as a vector of random functions rather than as a vector of scalars. For example, electroencephalography (EEG) data are treated more appropriately as functions of time. In such a problem, not only can the number of functions measured per sample be large, but each function is itself an infinite-dimensional object, making estimation of model parameters challenging. This is further complicated by the fact that curves are usually observed only at discrete time points. We first define a functional differential graph that captures the differences between two functional graphical models and formally characterize when the functional differential graph is well defined. We then propose a method, FuDGE, that directly estimates the functional differential graph without first estimating each individual graph. This is particularly beneficial in settings where the individual graphs are dense but the differential graph is sparse. We show that FuDGE consistently estimates the functional differential graph even in a high-dimensional setting for both fully observed and discretely observed function paths. We illustrate the finite sample properties of our method through simulation studies. We also propose a competing method, the Joint Functional Graphical Lasso, which generalizes the Joint Graphical Lasso to the functional setting. Finally, we apply our method to EEG data to uncover differences in functional brain connectivity between a group of individuals with alcohol use disorder and a control group.

Keywords: differential graph estimation, functional data analysis, multivariate functional data, probabilistic graphical models, structure learning

©2022 Boxin Zhao, Y. Samuel Wang, and Mladen Kolar.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v23/20-231.html.

1. Introduction

We consider a setting where we observe two samples of multivariate functional data, $X_i(t)$ for $i = 1, \ldots, n_X$ and $Y_i(t)$ for $i = 1, \ldots, n_Y$. The primary goal is to determine if and how the underlying populations—specifically their conditional dependency structures—differ. As a motivating example, consider electroencephalography (EEG) data, where the electrical activity of multiple regions of the brain can be measured simultaneously over a period of time. Given samples from the general population, fitting a graphical model to the observed measurements would allow a researcher to determine which regions of the brain are dependent after conditioning on all other regions. The EEG data analyzed in Section 6.2 consists of two samples: one from a control group and the other from a group of individuals with alcohol use disorder (AUD). Using these data, researchers may be interested in explicitly comparing the two groups and investigating the complex question of how brain functional connectivity patterns in the AUD group differ from those in the control group.

The conditional independence structure within multivariate data is commonly represented by a graphical model (Lauritzen, 1996). Let $G = \{V, E\}$ denote an undirected graph where V is the set of vertices with $|V| = p$ and $E \subset V^2$ is the set of edges. At times, we also denote V as $[p] = \{1, 2, \ldots, p\}$. When the data consist of random vectors $X = (X_1, \ldots, X_p)^\top$, we say that X satisfies the pairwise Markov property with respect to G if $X_v \not\perp\!\!\!\perp X_w \mid \{X_u\}_{u \in V \setminus \{v,w\}}$ holds if and only if $\{v, w\} \in E$. When X follows a multivariate Gaussian distribution with covariance $\Sigma = \Theta^{-1}$, then $\Theta_{vw} \neq 0$ if and only if $\{v, w\} \in E$. Thus, recovering the structure of an undirected graph from multivariate Gaussian data is equivalent to estimating the support of the precision matrix, Θ.

When the primary interest is in characterizing the difference between the conditional independence structure of two populations, the object of interest may be the *differential graph*, $G_\Delta = \{V, E_\Delta\}$. When X and Y follow multivariate normal distributions with covariance matrices Σ^X and Σ^Y, let $\Delta = \Theta^X - \Theta^Y$, where $\Theta^X = (\Sigma^X)^{-1}$ and $\Theta^Y = (\Sigma^Y)^{-1}$ are the precision matrices of X and Y, respectively. The differential graph is then defined by letting $E_\Delta = \{\{v, w\} : \Delta_{v,w} \neq 0\}$. This type of differential model for vector-valued data has been adopted in Zhao et al. (2014a), Xu and Gu (2016), and Cai (2017).

In the motivating example of EEG data, electrical activity is observed over a period of time. When the measurements smoothly vary over time, it may be more natural to consider the observations as arising from an underlying function. This is particularly true when data from different subjects are observed at different time points. Furthermore, when characterizing conditional independence, it is likely that the activity of each region depends not only on what is occurring simultaneously in the other regions but also on what has previously occurred in other regions; this suggests that a functional graphical model might be appropriate.

In this paper, we define a differential graph for functional data that we refer to as a functional differential graphical model. Similar to differential graphs for vector-valued data, functional differential graphical models characterize the differences in the conditional dependence structures of two distributions of multivariate curves. We build on the functional graphical model developed in Qiao et al. (2019). However, while Qiao et al. (2019) required that the observed functions lie in a finite-dimensional space in order for the functional graphical model to be well defined, the functional differential graphical models may be well

defined even in certain cases where the observed functions live in an infinite-dimensional space.

We propose an algorithm called FuDGE to estimate the differential graph and show that this procedure enjoys many benefits, similar to differential graph estimation in the vector-valued setting. Most notably, we show that under suitable conditions, the proposed method can consistently recover the differential graph even in the high-dimensional setting where p, the number of observed variables, may be larger than n, the number of observed samples.

A conference version of this paper was presented at the Conference on Neural Information Processing Systems (Zhao et al., 2019). Compared to the conference version, this paper includes the following new results.

- We give a new definition for a differential graph for functional data, which allows us to circumvent the unnatural assumption made in the previous version and take a truly functional approach. Specifically, instead of defining the differential graph based on the difference between conditional covariance functions, we use the limit of the norm of the difference between finite-dimensional precision matrices.

- We include new theoretical guarantees for discretely observed curves. In practice, we can only observe the functions at discrete time points, so this extends the theoretical guarantees to a practical estimation procedure. Discrete observations bring an additional source of error when the estimated curves are used in the functional PCA. In Theorem 13, we give an error bound for estimating the covariance matrix of the PCA score vectors under mild conditions.

- We introduce the Joint Functional Graphical Lasso, which is a generalization of the Joint Graphical Lasso (Danaher et al., 2014) to the functional data setting. Empirically, we show that the procedure performs competitively in some settings but is generally outperformed by the FuDGE procedure.

The software implementation can be found at `https://github.com/boxinz17/FuDGE`. The repository also contains the code to reproduce the simulation results.

1.1 Related Work

The work we develop lies at the intersection of two different lines of literature: graphical models for functional data and direct estimation of differential graphs.

Many previous works have studied the structure estimation of a static undirected graphical model (Chow and Liu, 1968; Yuan and Lin, 2007; Cai et al., 2011; Meinshausen and Bühlmann, 2006; Kolar and Xing, 2012a; Wang and Kolar, 2016; Vogel and Fried, 2011; Sun et al., 2015; Suggala et al., 2017). Previous methods have also been proposed to characterize conditional independence for multivariate observations recorded over time. For example, Talih and Hengartner (2005), Xuan and Murphy (2007), Ahmed and Xing (2009), Song et al. (2009a), Song et al. (2009b), Kolar et al. (2010b), Kolar et al. (2009), Kolar and Xing (2009), Zhou et al. (2010), Yin et al. (2010), Kolar et al. (2010a), Kolar and Xing (2011), Kolar and Xing (2012b), Wang and Kolar (2014), Lu et al. (2018), Geng et al. (2019a), Geng et al. (2019b), Tsai et al. (2020) studied methods for dynamic graphical models that

assume that data are sampled independently at different time points but generated by related distributions. In these works, the authors proposed procedures to estimate a series of graphs that represent the conditional independence structure at each time point; however, they assumed that the observed data do not encode "longitudinal" dependence. In contrast, Wang et al. (2020) focused on graphical models for time series data, while Qiao et al. (2019), Zhu et al. (2016), Li and Solea (2018), Zhang et al. (2021), Zhao et al. (2021) considered the setting where the data are multivariate random functions. Most similar to the setting we consider, Qiao et al. (2019) assumed that the data are distributed as a multivariate Gaussian process (MGP) and use a graphical lasso type procedure on the estimated functional principal component scores. Zhu et al. (2016) also assumed an MGP but proposed a Bayesian procedure. Crucially, however, both required that the covariance kernel can essentially be represented by a finite-dimensional object. Zapata et al. (2021) showed that under various notions of separability—roughly when the covariance kernel can be decomposed into covariance across time and covariance across nodes—the conditional independence of the MGP is well defined even when the functional data are truly infinite-dimensional and that the conditional independence graph can be recovered by the union of a (potentially infinitely) countable number of graphs over finite-dimensional objects. Zhao et al. (2021) adopted a neighborhood selection approach to learn the conditional independence structure of an MGP, which does not need to assume that functional data are finite-dimensional or that the MGP is separable to ensure consistency. In a different approach, Li and Solea (2018) did not assume that random functions are Gaussian and instead used the notion of additive conditional independence to define a graphical model for random functions. Qiao et al. (2020) also assumed that the data are random functions, but allowed the dependency structure to change smoothly over time—similar to a dynamic graphical model.

We also draw on recent literature that has shown that when the object of interest is the difference between two distributions, directly estimating the difference can provide improvements over first estimating each distribution and then taking the difference. Most notably, when estimating the difference in graphs in a high-dimensional setting, even if each individual graph does not satisfy the appropriate sparsity conditions, the differential graph may still be recovered consistently. Zhao et al. (2014a) considered data drawn from two Gaussian graphical models and showed that even if both underlying graphs are dense, if the difference between the precision matrices of each distribution is sparse, the differential graph can still be recovered in the high-dimensional setting. Liu et al. (2014) proposed procedure based on KLIEP (Sugiyama et al., 2007) that estimates the differential graph by directly modeling the ratio of two densities. They did not assume Gaussianity but required that both distributions lie in some exponential family. Fazayeli and Banerjee (2016) extended this idea to estimate the differences in Ising models. Wang et al. (2018) also proposed direct difference estimators for directed graphs when data are generated by linear structural equation models that share a common topological ordering.

1.2 Notation

Let $|\cdot|_p$ denote the vector p-norm and $\|\cdot\|_p$ denote the matrix/operator p-norm. For example, for a $p \times 1$ vector $a = (a_1, a_2, \ldots, a_p)^\top$, we have $|a|_1 = \sum_j |a_j|$, $|a|_2 = (\sum_j |a_j^2|)^{1/2}$ and $|a|_\infty = \max_j |a_j|$. For a $p \times q$ matrix A with entries a_{jk}, $|A|_1 = \sum_{j,k} |a_{jk}|$, $\|A\|_1 =$

$\max_k \sum_j |a_{jk}|$, $|A|_\infty = \max_{j,k} |a_{jk}|$, and $\|A\|_\infty = \max_j \sum_k |a_{jk}|$. Let $\|A\|_{\mathrm{F}} = (\sum_{j,k} a_{jk}^2)^{1/2}$ be the Frobenius norm of A. When A is symmetric, let $\mathrm{tr}(A) = \sum_j a_{jj}$ denote the trace of A. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues, respectively. Let $a_n \asymp b_n$ denote $0 < C_1 \leq \inf_n |a_n/b_n| \leq \sup_n |a_n/b_n| \leq C_2 < \infty$ for some positive constants C_1 and C_2.

We assume that all random functions belong to a separable Hilbert space $\mathbb{H}$. For any two functions $f_1, f_2 \in \mathbb{H}$, we define their inner product as $\langle f_1, f_2 \rangle = \int f_1(t) f_2(t) dt$. The induced norm is $\|f_1\| = \|f_1\|_{\mathcal{L}^2} = \{\int f_1^2(t) dt\}^{1/2}$.

For a function vector $f(t) = (f_1(t), f_2(t), \ldots, f_p(t))^\top$, we let $\|f\|_{\mathcal{L}^2,2} = (\sum_{j=1}^p \|f_j\|^2)^{1/2}$ denote its $\mathcal{L}^2, 2$-norm. For a bivariate function $g(s,t)$, we define the Hilbert-Schmidt norm of $g(s,t)$ as $\|g\|_{\mathrm{HS}} = \int\int \{g(s,t)\}^2 ds dt$. Typically, we will use $f(\cdot)$ (and similarly $g(\cdot, *)$) to denote the entire function f, while we use $f(t)$ (and similarly $g(s,t)$) to mean the value of f evaluated at t.

For a vector space $\mathbb{V}$, we use $\mathbb{V}^\perp$ to denote its orthogonal complement. For $v_1, \ldots, v_K \in \mathbb{V}$ and $v = (v_1, \ldots, v_K)^\top$, we use $\mathrm{Span}\{v_1, v_2, \ldots, v_K\} = \mathrm{Span}(v)$ to denote the vector subspace spanned by $v_1, \ldots, v_K$.

2. Functional Differential Graphical Models

In this section, we review functional graphical models and introduce the notion of a functional differential graphical model.

2.1 Functional Graphical Model

Suppose $X_i(\cdot) = (X_{i1}(\cdot), X_{i2}(\cdot), \ldots, X_{ip}(\cdot))^\top$ is a p-dimensional *multivariate Gaussian process (MGP)* with mean zero and common domain $\mathcal{T}$, where $\mathcal{T}$ is a closed interval of the real line with length $|\mathcal{T}|$.[1] Each observation, for $i = 1, 2, \ldots, n$, is i.i.d. In addition, assume that for $j \in V$, $X_{ij}(\cdot)$ is a random element of a separable Hilbert space $\mathbb{H}$. Qiao et al. (2019), define the conditional cross-covariance function for $X_i(\cdot)$ as

$$C_{jl}^X(s,t) \;=\; \mathrm{Cov}\left(X_{ij}(s), X_{il}(t) \;\middle|\; \{X_{ik}(\cdot)\}_{k \neq j,l}\right). \tag{1}$$

If $C_{jl}^X(s,t) = 0$ for all $s, t \in \mathcal{T}$, then the random functions $X_j(\cdot)$ and $X_l(\cdot)$ are conditionally independent given the other random functions, and the graph $G_X = \{V, E_X\}$ represents the pairwise Markov property of $X_i(\cdot)$ if

$$E_X = \left\{(j,l) \,:\, j < l \text{ and } \|C_{jl}^X\|_{\mathrm{HS}} \neq 0\right\}. \tag{2}$$

In general, we cannot directly estimate (2), since $X_i(\cdot)$ may be an infinite-dimensional object. Thus, before applying a statistical estimation procedure, dimension reduction is typically required. Qiao et al. (2019) used *functional principal component analysis* (FPCA) to project each observed function onto an orthonormal function basis defined by a finite number of eigenfunctions. Their procedure then estimates the conditional independence structure from the "projection scores" of this basis. We outline their approach in the following. However, in contrast to Qiao et al. (2019), we do not restrict ourselves to dimension

1. We assume mean zero and a common domain $\mathcal{T}$ to simplify the notation, but the methodology and theory generalize to non-zero means and different time domains.

reduction by projecting onto the FPCA basis, and in our discussion we instead consider a general function subspace.

Let $\mathbb{V}_j^{M_j} \subseteq \mathbb{H}$ be a subspace of a separable Hilbert space $\mathbb{H}$ with dimension $M_j \in \mathbb{N}^+$ for all $j = 1, 2, \ldots, p$. Our theory easily generalizes to the setting where M_j may differ, but to simplify the notation, we assume $M_j = M$ for all j and simply write $\mathbb{V}_j^M$ instead of $\mathbb{V}_j^{M_j}$. Let $\mathbb{V}_{[p]}^M \coloneqq \mathbb{V}_1^M \otimes \mathbb{V}_2^M \otimes \cdots \otimes \mathbb{V}_p^M$.

For any function $g(\cdot) \in \mathbb{H}$ and a subspace $\mathbb{F} \subseteq \mathbb{H}$, let $\pi(g(\cdot); \mathbb{F}) \in \mathbb{F}$ denote the projection of the function $g(\cdot)$ onto the subspace $\mathbb{F}$, and let

$$\pi(X_i(\cdot); \mathbb{V}_{[p]}^M) = \left(\pi(X_{i1}(\cdot); \mathbb{V}_1^M), \pi(X_{i2}(\cdot); \mathbb{V}_2^M), \ldots, \pi(X_{ip}(\cdot); \mathbb{V}_p^M)\right)^\top.$$

When the choice of subspace is clear from the context, we will use the following shorthand notation: $X_{ij}^\pi(\cdot) = \pi(X_{ij}(\cdot); \mathbb{V}_j^M)$, $j = 1, 2, \ldots, p$, and $X_i^\pi(\cdot) = \pi(X_i(\cdot); \mathbb{V}_{[p]}^M)$.

Similarly to the definitions in (1) and (2), we define the conditional independence graph of $X^\pi(\cdot)$ as

$$E_X^\pi = \left\{ \{j, l\} \,:\, j < l \text{ and } \|C_{jl}^{X,\pi}\|_{\text{HS}} \neq 0 \right\}, \tag{3}$$

where

$$C_{jl}^{X,\pi}(s,t) \;=\; \text{Cov}\left(X_{ij}^\pi(s), X_{il}^\pi(t) \;\middle|\; \{X_{ik}^\pi(\cdot)\}_{k \neq j,l}\right).$$

Note that E_X^π depends on the choice of $\mathbb{V}_{[p]}^M$ through the projection operator π, and, as we discuss below, E_X^π may be recovered from the observed samples.

When data arise from an MGP, we can estimate the projected graphical structure by studying the precision matrix of projection score vectors (defined below) with *any* orthonormal function basis of the subspace $\mathbb{V}_{[p]}^M$. Let $e_j^M = (e_{j1}(\cdot), e_{j2}(\cdot), \ldots, e_{jM}(\cdot))^\top$ be any orthonormal function basis of $\mathbb{V}_j^M$ and let $e^M(\cdot) = \{e_j^M\}_{j=1}^p$ be an orthonormal function basis of $\mathbb{V}_{[p]}^M$. Let

$$a_{ijk}^X = \int_{\mathcal{T}} X_{ij}(t) e_{jk}(t) dt$$

denote the projection score of $X_{ij}(\cdot)$ onto $e_{jk}(\cdot)$ and let

$$a_{ij}^{X,M} = (a_{ij1}^X, a_{ij2}^X, \ldots, a_{ijM}^X)^\top \;\text{ and }\; a_i^{X,M} = ((a_{i1}^{X,M})^\top, \ldots, (a_{ip}^{X,M})^\top)^\top \in \mathbb{R}^{pM}.$$

Since $X_i(\cdot)$ is a p-dimensional MGP, $a_i^{X,M}$ follows a multivariate Gaussian distribution and we denote the covariance matrix of that distribution as $\Sigma^{X,M} = (\Theta^{X,M})^{-1} \in \mathbb{R}^{pM \times pM}$. Each function $X_{ij}(\cdot)$ is associated with M rows and columns of $\Sigma^{X,M}$ corresponding to $a_{ij}^{X,M}$. We use $\Theta_{jl}^{X,M}$ to refer to the $M \times M$ submatrix of $\Theta^{X,M}$ that corresponds to the functions $X_{ij}(\cdot)$ and $X_{il}(\cdot)$. Lemma 1, from Qiao et al. (2019), shows that the conditional independence structure of the projected functional data can be obtained from the block sparsity of $\Theta^{X,M}$.

Lemma 1 *[Qiao et al. (2019)] Let $\Theta^{X,M}$ be the inverse covariance of the projection scores. Then, $X_{ij}^\pi(s) \perp\!\!\!\perp X_{il}^\pi(t) \mid \{X_{ik}^\pi(\cdot)\}_{k \neq j,l}$ for all[2] $s, t \in \mathcal{T}$ if and only if $\Theta_{jl}^{X,M} \equiv 0$. This implies*

2. More precisely, we only need the conditional independence to hold for all $s, t \in \mathcal{T}$ except for a subset of $\mathcal{T}^2$ with zero measure.

that E_X^π *—as defined in* (3)*—can be equivalently defined as*

$$E_X^\pi \;=\; \left\{\{j,l\} \,:\, j<l \text{ and } \|\Theta_{jl}^{X,M}\|_F \neq 0\right\}.$$

Although Qiao et al. (2019) only considered projections onto the span of the FPCA basis (that is, the eigenfunctions of $X_{ij}(\cdot)$ corresponding to M largest eigenvalues), the result trivially extends to the more general case of *any subspace* and *any orthonormal function basis* of that subspace.

Although $\Theta^{X,M}$ depends on the specific basis onto which $X_i(\cdot)$ is projected, the edge set E_X^π only depends on the subspace $\mathbb{V}_{[p]}^M$, that is, the span of the basis onto which $X_i(\cdot)$ is projected. Thus, Lemma 1 implies that although the entries of $\Theta^{X,M}$ can change when using different orthonormal function bases to represent $\mathbb{V}_{[p]}^M$, the block sparsity pattern of $\Theta^{X,M}$ only depends on the span of the selected basis.

When $X_i(\cdot) \neq X_i^\pi(\cdot)$, E_X^π may not be the same as E_X; furthermore, it may not be the case that $E_X^\pi \subseteq E_X$ or $E_X \subseteq E_X^\pi$. Thus, Condition 2 of Qiao et al. (2019) requires a finite $M^\star < \infty$ such that X_{ij} lies in $\mathbb{V}_{[p]}^{M^\star}$ almost surely. When $M = M^\star$, then $X_i(\cdot) = X_i^\pi(\cdot)$ and $E_X^\pi = E_X$. Under this assumption, to estimate $E_X^\pi = E_X$, Qiao et al. (2019) proposed the functional graphical lasso estimator (fglasso), which solves the following objective:

$$\hat{\Theta}^{X,M} = \underset{\Theta^{X,M}}{\arg\max} \left\{ \log\det\left(\Theta^{X,M}\right) - \operatorname{tr}\left(S^{X,M}\Theta^{X,M}\right) - \gamma_n \sum_{j\neq l} \left\|\Theta_{jl}^{X,M}\right\|_{\mathrm{F}} \right\}. \tag{4}$$

In (4), $\Theta^{X,M}$ is a symmetric positive definite matrix, $\Theta_{jl}^{X,M} \in \mathbb{R}^{M\times M}$ corresponds to the (j,l) submatrix of $\Theta^{X,M}$, γ_n is a non-negative tuning parameter, and $S^{X,M}$ is an estimator of $\Sigma^{X,M}$. The matrix $S^{X,M}$ is obtained by using FPCA on the empirical covariance functions (see Section 2.3 for details). The resulting estimated edge set for the functional graph is

$$\hat{E}_X^\pi = \left\{\{j,l\} \,:\, j<l \text{ and } \left\|\hat{\Theta}_{jl}^{X,M}\right\|_{\mathrm{F}} > 0\right\}.$$

We also note that the objective in (4) was previously used in Kolar et al. (2013) and Kolar et al. (2014) for the estimation of graphical models from multi-attribute data.

However, the requirement that $X_i(\cdot)$ lies in a subspace with finite-dimension may be violated in many practical applications and negates one of the primary benefits of considering the observations as functions. Unfortunately, the extension to infinite-dimensional data is nontrivial, and indeed Condition 2 in Qiao et al. (2019) requires that the observed functional data lie within a finite-dimensional span. To see why, we first note that $\Sigma^{X,M^\star}$ is always a compact operator on $\mathbb{R}^{pM^\star}$. Thus, as $M^\star \to \infty$, the smallest eigenvalue of $\Sigma^{X,M^\star}$ will go to zero. As a consequence, $\Sigma^{X,M^\star}$ becomes increasingly ill-conditioned, and $\Theta^{X,M^\star}$, the inverse of $\Sigma^{X,M^\star}$ will become ill-defined when $M^\star = \infty$. This behavior makes the estimation of a functional graphical model—at least through the basis expansion approach proposed by Qiao et al. (2019)—generally infeasible for truly infinite-dimensional functional data. When the data are truly infinite-dimensional, the best we can do is to estimate a finite-dimensional approximation and hope that it captures the relevant information.

2.2 Functional Differential Graphical Models: Finite-Dimensional Setting

In this paper, rather than estimating the conditional independence structure of a single MGP, we are interested in characterizing the difference between two MGPs, X and Y. For brevity, we will typically only explicitly define the notation for X; however, the reader should infer that all the notation for Y is defined analogously. As described in the introduction, Li et al. (2007) and Zhao et al. (2014a) consider the setting where X and Y are multivariate Gaussian vectors, and define the differential graph $G_\Delta = \{V, E_\Delta\}$ by letting

$$E_\Delta = \{(v, w) : v < w \text{ and } \Delta_{vw} \neq 0\}$$

where $\Delta = (\Sigma^X)^{-1} - (\Sigma^Y)^{-1}$ and Σ^X, Σ^Y are the covariance matrices of X and Y.

We extend this definition to the functional data setting and define functional differential graphical models. To develop intuition, we first start by defining the differential graph with respect to the finite-dimensional projections of functional data, that is, with respect to $X_i^\pi(t)$ and $Y_i^\pi(t)$ for some choice of $\mathbb{V}_{[p]}^M$. As implied by Lemma 1, in the functional graphical model setting, the $M \times M$ blocks of the precision matrix of the projection scores play a similar role to the individual entries of a precision matrix in the vector-valued Gaussian graphical model setting. Thus, we also define a functional differential graphical model by the difference of the precision matrices of the projection scores. Note that for each $j \in V$, we require that both a_{ij}^X and a_{ij}^Y be calculated using the same function basis of $\mathbb{V}_j^M$. Let $\Theta^{X,M} = \left(\Sigma^{X,M}\right)^{-1}$ and $\Theta^{Y,M} = \left(\Sigma^{Y,M}\right)^{-1}$ be the precision matrices for the projection scores for X and Y, respectively, where the inverse should be understood as the pseudo-inverse when $\Sigma^{X,M}$ or $\Sigma^{Y,M}$ are not invertible.

We now define the functional differential graphical model. Let $\Delta^M = \Theta^{X,M} - \Theta^{Y,M}$ and Δ_{jl}^M be the (j,l)-th $M \times M$ block of Δ^M. We define the edges of the functional differential graph of the projected data as:

$$E_\Delta^\pi = \left\{(j, l) : j < l \text{ and } \|\Delta_{jl}^M\|_F > 0\right\}. \tag{5}$$

While the entries of Δ^M depend on the choice of orthonormal function basis, the definition of E_Δ^π is invariant to the particular basis and only depends on the span. The following lemma formally states this result.

Lemma 2 *Suppose that* $\text{span}(e^M(\cdot)) = \text{span}(\tilde{e}^M(\cdot))$ *for two orthonormal bases* $e^M(\cdot)$ *and* $\tilde{e}^M(\cdot)$. *Let* E_Δ^π *and* $E_\Delta^{\tilde{\pi}}$ *be defined by* (5) *when projecting* X *and* Y *onto* $e^M(\cdot)$ *and* $\tilde{e}^M(\cdot)$, *respectively. Then,* $E_\Delta^\pi = E_\Delta^{\tilde{\pi}}$.

Proof See Appendix B.1. ■

We have several comments about E_Δ^π defined in (5).

Projecting X and Y onto different subspaces: While we project both X and Y onto the same subspace $\mathbb{V}_{[p]}^M$, our definition can be easily generalized to a setting where we project X onto $\mathbb{V}_{[p]}^{X,M}$ and Y onto $\mathbb{V}_{[p]}^{Y,M}$, with $\mathbb{V}_{[p]}^{X,M} \neq \mathbb{V}_{[p]}^{Y,M}$. For example, naively following the procedure of Qiao et al. (2019), we could perform FPCA on X and Y separately,

and subsequently we could use the difference between the precision matrices of the projection scores to define the functional differential graph. Although defining the functional differential graph using this alternative approach may be suitable for some applications, it may result in the undesirable case where $(j,l) \in E_\Delta^\pi$ even though $C_{jl}^{X,\pi}(\cdot,*) = C_{jl}^{Y,\pi}(\cdot,*)$, $C_{jj}^{X,\pi}(\cdot,*) = C_{ll}^{Y,\pi}(\cdot,*)$, and $C_{ll}^{\setminus j,X,\pi}(\cdot,*) = C_{ll}^{\setminus j,Y,\pi}(\cdot,*)$. Therefore, we restrict our discussion to the setting where X and Y are projected onto the same subspace.

Connection to Multi-Attribute Graphical Models: The selection of a specific functional subspace is connected to multi-attribute graphical models (Kolar et al., 2014). If we treat the random function $X_{ij}(\cdot)$ as representing an infinite number of attributes, then $X_{ij}^\pi(\cdot)$ will be an approximation using M attributes. The chosen attributes are given by the subspace $\mathbb{V}_j^M$. While we allow different nodes to choose different attributes by allowing $\mathbb{V}_j^M$ to vary across j, we require that the same attributes are used to represent both X and Y by restricting $\mathbb{V}_{[p]}^M$ to be the same for X and Y. The specific choice of $\mathbb{V}_{[p]}^M$, can extract different attributes from the data. For instance, using the subspace spanned by the Fourier basis can be viewed as extracting frequency information, while using the subspace spanned by the eigenfunctions—as introduced in the next section—can be viewed as extracting the dominant modes of variation.

Given the definition (5) and the Lemma 2, there are two main questions to answer: First, how do we choose $\mathbb{V}_{[p]}^M$? Second, what happens when X and Y are infinite-dimensional? We answer the first question in Section 2.3 and the second question in Section 2.4.

2.3 Choosing Functional Subspace via FPCA

As discussed in Section 2.2, the choice of $\mathbb{V}_{[p]}^M$ in Definition 5 decides—roughly speaking—the attributes or dimensions in which we compare the conditional independence structures of X and Y. In some applications, we may have very good prior knowledge about this choice. However, in many cases, we may not have a strong prior knowledge. In this section, we describe our recommended "default choice" that uses FPCA on the combined X and Y observations. In particular, suppose that there exist subspaces $\{\mathbb{V}_j^{M^\star}\}_{j\in V}$ such that $\mathbb{V}_j^{M^\star}$ has dimension $M^\star < \infty$ and $X_{ij}(t), Y_{ij}(t) \in \mathbb{V}_j^{M^\star}$ for all $j \in V$. Then, FPCA—when given population values—recovers this subspace.

Similarly to the way principal component analysis provides the L_2 optimal lower dimensional representation of vector-valued data, FPCA provides the L_2 optimal finite-dimensional representation of functional data. Let $K_{jj}^X(t,s) = \mathrm{Cov}(X_{ij}(t), X_{ij}(s))$ denote the covariance function for X_{ij} for $j \in V$. Then, there exist orthonormal eigenfunctions and eigenvalues $\{\phi_{jk}^X(t), \lambda_{jk}^X\}_{k\in\mathbb{N}}$ such that $\int_\mathcal{T} K_{jj}^X(s,t)\phi_{jk}^X(t)dt = \lambda_{jk}^X\phi_{jk}^X(s)$ for all $k \in \mathbb{N}$ (Hsing and Eubank, 2015). Since $K_{jj}^X(s,t)$ is symmetric and non-negative definite, we assume, without loss of generality, that $\{\lambda_{js}^X\}_{s\in\mathbb{N}^+}$ is non-negative and non-increasing. By the Karhunen-Loève expansion (Hsing and Eubank, 2015, Theorem7.3.5), $X_{ij}(t)$ can be expressed as $X_{ij}(t) = \sum_{k=1}^\infty a_{ijk}^X\phi_{jk}^X(t)$, where the principal component scores satisfy $a_{ijk}^X = \int_\mathcal{T} X_{ij}(t)\phi_{jk}^X(t)dt$ and $a_{ijk}^X \sim N(0,\lambda_{jk}^X)$ with $E(a_{ijk}^X a_{ijl}^X) = 0$ if $k \neq l$. Because the eigenfunctions are orthonormal, the L_2 projection of X_{ij} onto the span of the first M eigenfunctions is $X_{ij}^M(t) = \sum_{k=1}^M a_{ijk}^X\phi_{jk}^X(t)$. Similarly, we can define $K_{jj}^Y(t,s)$,

$\{\phi^Y_{jk}(t), \lambda^Y_{jk}\}_{k\in\mathbb{N}}$ and $Y^M_{ij}(t)$. Let $K_{jj}(s,t) = K^X_{jj}(s,t) + K^Y_{jj}(s,t)$ and let $\{\phi_{jk}(t), \lambda_{jk}\}_{k\in\mathbb{N}}$ be the eigenfunction-eigenvalue pairs of $K_{jj}(s,t)$.

Lemma 3 implies that $X_{ij}(\cdot)$ and $Y_{ij}(\cdot)$ lie within the span of the eigenfunctions corresponding to the non-zero eigenvalues of K_{jj}. Furthermore, this subspace is minimal in the sense that no subspace of a smaller dimension contains $X_{ij}(\cdot)$ and $Y_{ij}(\cdot)$ almost surely. Thus, the FPCA basis of K_{jj} provides a good default choice for dimension reduction.

Lemma 3 *Let $|\mathbb{V}|$ denote the dimension of a subspace $\mathbb{V}$ and suppose that*

$$M'_j = \inf\{|\mathbb{V}| : \mathbb{V} \subseteq \mathbb{H}, X_{ij}(\cdot), Y_{ij}(\cdot) \in \mathbb{V} \text{ almost surely}\}.$$

Let $\{\phi_{jk}(t), \lambda_{jk}\}_{k\in\mathbb{N}}$ be the eigenfunction-eigenvalue pairs of $K_{jj}(s,t)$ and

$$M^\star_j = \sup\{M \in \mathbb{N}^+ : \lambda_{jM} > 0\}.$$

Then $M'_j = M^\star_j$ and $X_{ij}, Y_{ij} \in \mathrm{Span}\{\phi_{j1}(\cdot), \phi_{j2}(\cdot), \ldots, \phi_{j,M^\star_j}(\cdot)\}$ almost surely.

Proof See Appendix B.2. ■

2.4 Infinite-Dimensional Functional Data

In Section 2.2, we defined a functional differential graph for functional data that have finite-dimensional representation. In this section, we present a more general definition that also allows for infinite-dimensional functional data.

As discussed in Section 2.1, when the data are infinite-dimensional, estimating a functional graphical model is not straightforward because the precision matrix of the scores does not have a well-defined limit as M, the dimension of the projected data, increases to ∞. When estimating the differential graph, however, although $\|\Theta^{X,M}\|_{\mathrm{F}} \to \infty$ and $\|\Theta^{Y,M}\|_{\mathrm{F}} \to \infty$ as $M \to \infty$, it is still possible for $\|\Theta^{X,M} - \Theta^{Y,M}\|_{\mathrm{F}}$ to be bounded as $M \to \infty$. For instance, $x_n, y_n \in \mathbb{R}$ may both tend to infinity, but $\lim_n x_n - y_n$ may still exist and be bounded. Furthermore, even when $\|\Theta^{X,M} - \Theta^{Y,M}\|_{\mathrm{F}} \to \infty$, it is still possible for the difference $\Theta^{X,M} - \Theta^{Y,M}$ to be informative. This observation leads to Definition 4 below. To simplify notation, in the rest of the paper, we assume that $X_{ij}(\cdot)$ and $Y_{ij}(\cdot)$ live in an $M^\star$ dimensional space where $M^\star \leq \infty$. Recall that $\{\phi^X_{jk}(\cdot), \lambda^X_{jk}\}_{k\in\mathbb{N}}$ and $\{\phi^Y_{jk}(\cdot), \lambda^Y_{jk}\}_{k\in\mathbb{N}}$ denote the eigenpairs of K^X_{jj} and K^Y_{jj} respectively.

Definition 4 (Differential Graph Matrix and Comparability) *The MGPs X and Y are **comparable** if the following two conditions hold:*

1. *For all $j \in [p]$, K^X_{jj} and K^Y_{jj} have $M^\star$ non-zero eigenvalues and*

$$\mathrm{span}\left(\{\phi^X_{jk}\}_{k=1}^{M^\star}\right) = \mathrm{span}\left(\{\phi^Y_{jk}\}_{k=1}^{M^\star}\right).$$

2. *For every $(j,l) \in V^2$ where $j \neq l$ and a projection subspace sequence $\left\{\mathbb{V}^M_{[p]}\right\}_{M\geq 1}$ satisfying $\lim_{M\to M^\star} \mathbb{V}^M_j = \mathrm{span}\left(\{\phi^X_{jk}\}_{k=1}^{M^\star}\right)$, we have either:*

$$\lim_{M\to M^\star} \|\Delta^M_{jl}\|_F = 0 \qquad \text{or} \qquad \lim \inf_{M\to M^\star} \|\Delta^M_{jl}\|_F > 0.$$

We say that X and Y are ***incomparable****, if for some j, K_{jj}^X and K_{jj}^Y have a different number of non-zero eigenvalues, or if* $\text{span}\left(\{\phi_{jk}^X\}_{k=1}^{M^\star}\right) \neq \text{span}\left(\{\phi_{jk}^Y\}_{k=1}^{M^\star}\right)$*, or if there exists some (j,l) such that given* $\left\{\mathbb{V}_{[p]}^M\right\}_{M\geq 1}$ *satisfying* $\lim_{M\to M^\star} \mathbb{V}_j^M = \text{span}\left(\{\phi_{jk}^X\}_{k=1}^{M^\star}\right)$*, we have*

$$\lim \inf_{M\to M^\star} \|\Delta_{jl}^M\|_F = 0, \qquad \textit{but} \qquad \lim \sup_{M\to M^\star} \|\Delta_{jl}^M\|_F > 0.$$

When X and Y are comparable, we define the ***differential graph matrix*** *(DGM) $D = (D_{jl})_{(j,l)\in V^2} \in \mathbb{R}^{p\times p}$, where*

$$D_{jl} = \lim \inf_{M\to M^\star} \|\Delta_{jl}^M\|_F.$$

In Definition 4 we say $\lim_{M\to M^\star} \mathbb{V}_j^M = \text{span}\left(\{\phi_{jk}^X\}_{k=1}^{M^\star}\right)$, to mean the following: For any $\epsilon > 0$ and all $g \in \text{span}\left(\{\phi_{jk}^X\}_{k=1}^{M^\star}\right)$, there exists $M' = M'(\epsilon) < \infty$ such that $\|g - g_P^M\| < \epsilon$ for all $M \geq M'$, where g_P^M denotes the projection of g onto the subspace of $\mathbb{V}_j^M$.

When $M^\star < \infty$, the conditional independence structure in X_i and Y_i can be fully captured by a finite-dimensional representation. When $M^\star = \infty$, as $M \to \infty$, Δ_{jl}^M approaches the difference of two matrices with unbounded eigenvalues. However, when X and Y are comparable, the limits are still informative. This would suggest that by using a sufficiently large subspace, we can capture such a difference arbitrarily well. However, if the MGPs are not comparable, then using a larger subspace may not improve the approximation regardless of the sample size. For this reason, in the remainder of the article, we focus only on the setting where X and Y are comparable.

To our knowledge, there is no existing procedure to estimate a graphical model for functional data when the functions are infinite-dimensional. Thus, it is not straightforward to determine whether the comparability condition is stronger or weaker than what might be required for estimating the graphs separately and then comparing post hoc. However, we hope to provide some intuition to the reader.

Suppose that X and Y are of the same dimension, $M^\star$. If $M^\star < \infty$ and the functional graphical model for each sample could be estimated separately (that is, $\|\Theta^{X,M}\|_F < \infty$ and $\|\Theta^{Y,M}\|_F < \infty$), then X and Y are comparable when the minimal basis that spans X and Y is the same. Thus, the functional differential graph is also well defined. On the other hand, the conditions required by Qiao et al. (2019, Condition 2) for consistent estimation are not satisfied when $M^\star = \infty$, since $\lim_{M\to\infty} \|\Theta^{X,M}\|_F = \infty$ due to the compactness of the covariance operator. However, X and Y may still be comparable depending on the limiting behavior of $\Theta^{X,M}$ and $\Theta^{Y,M}$. Thus, there are settings where the differential graph may exist and can be consistently recovered even when each individual graph cannot be recovered (even when p is fixed).

However, when one MGP is finite-dimensional and the other is infinite-dimensional, then the MGPs are incomparable. To see this, without loss of generality, we assume that MGP X has infinite-dimension $M_j^X = M_X^\star = \infty$ for all $j \in V$ and MGP Y has finite-dimension $M_j^Y = M_Y^\star < \infty$ for all $j \in V$. Then $\Theta^{Y,M}$ is ill-defined when $M > M_Y^\star$ and recovering the differential graph is not straightforward.

We now define the notion of a functional differential graph.

Definition 5 *When two MGPs X and Y are comparable, we define their* ***functional differential graph*** *as an undirected graph $G_\Delta = \{V, E_\Delta\}$, where E_Δ is defined as*

$$E_\Delta = \{\{j,l\} \,:\, j < l \text{ and } D_{jl} > 0\}\,.$$

Remark 6 *The functional graphical model defined by Qiao et al. (2019) uses the conditional covariance function $C^X_{jl}(\cdot,*)$ given in (1). Thus, it would be quite natural to use the conditional covariance functions directly to define a differential graph, where*

$$E_\Delta = \left\{\{j,l\} \,:\, j < l \text{ and } C^X_{jl}(\cdot,*) \neq C^Y_{jl}(\cdot,*)\right\}. \tag{6}$$

Unfortunately, this definition does not always coincide with the one we propose in Definition 5. However, the functional differential graph given in Definition 5 has many nice statistical properties and retains important features of the graph defined in (6).

The primary statistical benefit of the graph defined in Definition 5 is that it can be directly estimated without estimating each conditional independence function: $C^X_{jl}(\cdot,\cdot)$ and $C^Y_{jl}(\cdot,\cdot)$. Similarly to the vector-valued case considered by (Zhao et al., 2014a), this allows for a much lower sample complexity when each individual graph is dense but the difference is sparse. In some settings, there may not be enough samples to accurately estimate each individual graph, but the difference may still be recovered. This result is demonstrated in Theorem 10.

The statistical advantages of our estimand unfortunately come at the cost of a slightly less precise characterization of the difference in the conditional covariance functions. However, many of the key characteristics are still preserved. Suppose X_i and Y_i are both $M^\star$-dimensional with $M^\star < \infty$ and further suppose that $\{\phi_{jm}(\cdot)\phi_{lm'}(*)\}_{m,m' \in [M^\star]\times[M^\star]}$ is a linearly independent set of functions. Suppose that the conditional covariance functions for $j,l \in V$ are unchanged so that $C^X_{jj}(\cdot,*) = C^Y_{jj}(\cdot,*)$ and $C^{\setminus j,X}_{ll}(\cdot,*) = C^{\setminus j,Y}_{ll}(\cdot,*)$, where

$$C^{\setminus j,X}_{ll}(\cdot,*) \coloneqq \mathrm{Cov}(X_l(\cdot), X_l(*) \,|\, X_k(\cdot), k \neq j,l)$$

and $C^{\setminus j,Y}_{ll}(\cdot,*)$ is defined similarly; then, $\Delta_{jl} = 0$ if and only if $C^X_{jl}(\cdot,*) = C^{Y,\pi}_{jl}(\cdot,*)$. When this holds for all pairs $j,l \in V$, then the definitions of a differential graph in Definition 5 and (6) are equivalent. When the conditional covariance functions change so that $C^X_{jj}(\cdot,*) \neq C^Y_{jj}(\cdot,*)$, then we still have $\Delta_{jl} \neq 0$ if $C^{X,\pi}_{jl}(\cdot,*) = 0$ and $C^{Y,\pi}_{jl}(\cdot,*) \neq 0$ (or vice versa). Thus, even in this more general setting, the functional differential graph given in Definition 5 captures all qualitative differences between conditional covariance functions $C^X_{jl}(\cdot,*)$ and $C^Y_{jl}(\cdot,*)$.

Our objective is to directly estimate E_Δ without first estimating E_X or E_Y. Since the functions we consider may be infinite-dimensional objects, in practice, what we can directly estimate is actually E^π_Δ defined in (5). We will use a sieve estimator to estimate Δ^M, where M increases with the sample size n. When $M^\star = M$, then $E^\pi_\Delta = E_\Delta$. When $M < M^\star \leq \infty$, then this is generally not true; however, we would expect the graphs to be similar when M is large enough compared to $M^\star$. Thus, by constructing a suitable estimator of Δ^M, we can still recover E_Δ.

2.5 Illustration of Comparability

We provide a few examples that illustrate the notion of comparability. In the first two examples, the graphs are comparable, whereas in the third example, the graphs are incomparable. First, we state a lemma that will be helpful in the following discussion. The lemma follows directly from the properties of the multivariate normal and the inverse of block matrices.

Lemma 7 *Let* $H_{jl}^{X,M} = \mathrm{Cov}(a_{ij}^{X,M}, a_{il}^{X,M} \mid a_{ik}^{X,M}, k \neq j, l)$ *and* $H_{jj}^{\setminus l,X,M} = \mathrm{Var}(a_{ij}^{X,M} \mid a_{ik}^{X,M}, k \neq j, l)$. *For any* $j \in V$, *we have* $\Theta_{jj}^{X,M} = (H_{jj}^{X,M})^{-1}$. *For any* $(j,l) \in V^2$ *and* $j \neq l$, *we have* $\Theta_{jl}^{X,M} = -(H_{jj}^{X,M})^{-1} H_{jl}^{X,M} (H_{ll}^{\setminus j,X,M})^{-1}$.

Proof See Appendix B.3. ■

The following proposition follows directly from Lemma 7.

Proposition 8 *Assume that for any* $(j,l) \in V^2$ *and* $j \neq l$, *we have*

$$a_{ijm}^{X} \perp\!\!\!\perp a_{ijm'}^{X} \mid a_{ik}^{X,M}, k \neq j \qquad \text{and} \qquad a_{ijm}^{X} \perp\!\!\!\perp a_{ijm'}^{X} \mid a_{ik}^{X,M}, k \neq j, l,$$

for any M *and* $1 \leq m \neq m' \leq M$. *We then have*

$$\Theta_{jj}^{X,M} = \mathrm{diag}\left(\frac{1}{\mathrm{Var}\left(a_{ij1}^{X} \mid a_{ik}^{X,M}, k \neq j\right)}, \ldots, \frac{1}{\mathrm{Var}\left(a_{ijM}^{X} \mid a_{ik}^{X,M}, k \neq j\right)} \right)$$

and

$$\Theta_{jl,mm'}^{X,M} = \frac{\mathrm{Cov}\left(a_{ijm}^{X}, a_{ilm'}^{X} \mid a_{ik}^{X,M}, k \neq j, l\right)}{\mathrm{Var}\left(a_{ijm}^{X} \mid a_{ik}^{X,M}, k \neq j\right) \mathrm{Var}\left(a_{ilm'}^{X} \mid a_{ik}^{X,M}, k \neq j\right)} \triangleq \bar{v}_{mm'}^{X,jl,M},$$

for any M *and* $1 \leq m \neq m' \leq M$. *In addition, if*

$$a_{ijm}^{Y} \perp\!\!\!\perp a_{ijm'}^{Y} \mid a_{ik}^{Y,M}, k \neq j \quad \text{and} \quad a_{ijm}^{Y} \perp\!\!\!\perp a_{ijm'}^{Y} \mid a_{ik}^{Y,M}, k \neq j, l,$$

for any M *and* $1 \leq m \neq m' \leq M$, *then*

$$\begin{aligned} \Theta_{jj}^{X,M} - \Theta_{jj}^{Y,M} &= \mathrm{diag}\left(\left\{ \frac{\mathrm{Var}\left(a_{ijm}^{Y} \mid a_{ik}^{Y,M}, k \neq j\right) - \mathrm{Var}\left(a_{ijm}^{X} \mid a_{ik}^{X,M}, k \neq j\right)}{\mathrm{Var}\left(a_{ijm}^{X} \mid a_{ik}^{X,M}, k \neq j\right) \mathrm{Var}\left(a_{ijm}^{Y} \mid a_{ik}^{Y,M}, k \neq j\right)} \right\}_{m=1}^{M} \right) \\ &\triangleq \mathrm{diag}\left(\bar{w}_1^{j,M}, \bar{w}_2^{j,M}, \ldots, \bar{w}_M^{j,M} \right) \end{aligned}$$

and

$$\Theta_{jl,mm'}^{X,M} - \Theta_{jl,mm'}^{Y,M} = \frac{\mathrm{Cov}\left(a_{ijm}^{X}, a_{ilm'}^{X} \mid a_{ik}^{X,M}, k \neq j, l\right)}{\mathrm{Var}\left(a_{ijm}^{X} \mid a_{ik}^{X,M}, k \neq j\right) \mathrm{Var}\left(a_{ilm'}^{X} \mid a_{ik}^{X,M}, k \neq j\right)}$$

$$- \frac{\operatorname{Cov}\left(a_{ijm}^{Y}, a_{ilm'}^{Y} \mid a_{ik}^{Y,M}, k \neq j, l\right)}{\operatorname{Var}\left(a_{ijm}^{Y} \mid a_{ik}^{Y,M}, k \neq j\right) \operatorname{Var}\left(a_{ilm'}^{Y} \mid a_{ik}^{Y,M}, k \neq j\right)}$$
$$= \bar{v}_{mm'}^{Y,jl,M} - \bar{v}_{mm'}^{X,jl,M} \triangleq \bar{z}_{mm'}^{jl,M},$$

for any M and $1 \leq m \neq m' \leq M$.

With the notation defined in Proposition 8, we have that

$$\|\Delta_{jj}^{M}\|_{\mathrm{HS}}^{2} = \sum_{m=1}^{M} \left(\bar{w}_{m}^{j,M}\right)^{2} \qquad \text{and} \qquad \|\Delta_{jl}^{M}\|_{\mathrm{HS}}^{2} = \sum_{m'=1}^{M} \sum_{m=1}^{M} \left(\bar{z}_{mm'}^{jl,M}\right)^{2}.$$

As a result, we have the following condition for comparability.

Proposition 9 *Under the assumptions in Proposition 8, assume that MGPs X and Y are $M^\star$-dimensional, with $1 \leq M^\star \leq \infty$, and lie in the same space. Then they are comparable if and only if for every $(j,l) \in V \times V$, we have either*

$$\liminf_{M \to M^\star} \sum_{m'=1}^{M} \sum_{m=1}^{M} \left(\bar{z}_{mm'}^{jl,M}\right)^{2} > 0 \qquad \text{or} \qquad \lim_{M \to M^\star} \sum_{m'=1}^{M} \sum_{m=1}^{M} \left(\bar{z}_{mm'}^{jl,M}\right)^{2} = 0,$$

where $\bar{z}_{mm'}^{jl,M}$ are defined in Proposition 8.

We now give an infinite-dimensional comparable example.

Example 1 *Assume that $\{\epsilon_{i1k}^{X}\}_{k\geq 1}$, $\{\epsilon_{i2k}^{X}\}_{k\geq 1}$, and $\{\epsilon_{i3k}^{X}\}_{k\geq 1}$ are all independent mean zero Gaussian variables with $\operatorname{Var}(\epsilon_{ijk}^{X}) = \sigma_{X,jk}^{2}$, $j = 1,2,3$, $k \geq 1$ for all i. For any $k \geq 1$, let*

$$a_{i1k}^{X} = a_{i2k}^{X} + \epsilon_{i1k}^{X}, \quad a_{i2k}^{X} = \epsilon_{i2k}^{X}, \quad a_{i3k}^{X} = a_{i2k}^{X} + \epsilon_{i3k}^{X}.$$

Let $a_{ij}^{X,M} = (a_{ij1}^{X}, \cdots, a_{ijM}^{X})^{\top}$, $j = 1,2,3$. We then define $X_{ij}(t) = \sum_{k=1}^{\infty} a_{ijk}^{X} b_k(t)$, $j = 1,2,3$, where $\{b_k(t)\}_{k=1}^{\infty}$ is some orthonormal function basis of $\mathbb{H}$. We define $\{\epsilon_{ijk}^{Y}\}_{k\geq 1}$, $\{a_{ijk}^{Y}\}_{k\geq 1}$, $a_{ij}^{Y,M}$, and $Y_{ij}(t)$, $j = 1,2,3$, similarly.

The graph structure of X and Y is shown in Figure 1. Since $a_{ij}^{X,M}$ follows a multivariate Gaussian distribution, for any $M \geq 2$, $1 \leq m, m' \leq M$ and $m \neq m'$:

$$\operatorname{Var}\left(a_{i1m}^{X} \mid a_{i2}^{X,M}, a_{i3}^{X,M}\right) = \sigma_{X,1m}^{2},$$
$$\operatorname{Var}\left(a_{i3m}^{X} \mid a_{i1}^{X,M}, a_{i2}^{X,M}\right) = \sigma_{X,3m}^{2},$$
$$\operatorname{Var}\left(a_{i2m}^{X} \mid a_{i1}^{X,M}, a_{i3}^{X,M}\right) = \frac{\sigma_{X,1m}^{2}\sigma_{X,2m}^{2}\sigma_{X,3m}^{2}}{\sigma_{X,1m}^{2}\sigma_{X,2m}^{2} + \sigma_{X,1m}^{2}\sigma_{X,3m}^{2} + \sigma_{X,2m}^{2}\sigma_{X,3m}^{2}},$$

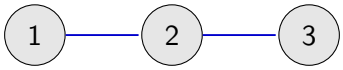


Figure 1: The conditional independence graph for both X and Y in Example 1. The differential graph between X and Y has the same structure.

and

$$\begin{aligned}
\mathrm{Var}\left(a^X_{i1m} \mid a^{X,M}_{i2}\right) &= \sigma^2_{X,1m},\\
\mathrm{Var}\left(a^X_{i1m} \mid a^{X,M}_{i3}\right) &= \frac{\sigma^2_{X,1m}\sigma^2_{X,2m} + \sigma^2_{X,1m}\sigma^2_{X,3m} + \sigma^2_{X,2m}\sigma^2_{X,3m}}{\sigma^2_{2m} + \sigma^2_{3m}},\\
\mathrm{Var}\left(a^X_{i3m} \mid a^{X,M}_{i2}\right) &= \sigma^2_{X,3m},\\
\mathrm{Var}\left(a^X_{i3m} \mid a^{X,M}_{i1}\right) &= \frac{\sigma^2_{X,1m}\sigma^2_{X,2m} + \sigma^2_{X,1m}\sigma^2_{X,3m} + \sigma^2_{X,2m}\sigma^2_{X,3m}}{\sigma^2_{2m} + \sigma^2_{1m}},\\
\mathrm{Var}\left(a^X_{i2m} \mid a^{X,M}_{i1}\right) &= \frac{\sigma^2_{X,1m}\sigma^2_{X,2m}}{\sigma^2_{X,1m} + \sigma^2_{X,2m}},\\
\mathrm{Var}\left(a^X_{i2m} \mid a^{X,M}_{i3}\right) &= \frac{\sigma^2_{X,3m}\sigma^2_{X,2m}}{\sigma^2_{X,3m} + \sigma^2_{X,2m}}.
\end{aligned}$$

In addition, we also have

$$\begin{aligned}
&\mathrm{Cov}(a^X_{i1m}, a^X_{3m'} \mid a^{X,M}_{i2}) = 0,\\
&\mathrm{Cov}(a^X_{i1m}, a^X_{i2m'} \mid a^{X,M}_{i3}) = \mathbb{1}(m = m') \cdot \frac{\sigma^2_{X,3m}\sigma^2_{X,2m}}{\sigma^2_{X,3m} + \sigma^2_{X,2m}},\\
&\mathrm{Cov}(a^X_{i2m}, a^X_{i3m'} \mid a^{X,M}_{i3}) = \mathbb{1}(m = m') \cdot \frac{\sigma^2_{X,1m}\sigma^2_{X,2m}}{\sigma^2_{X,1m} + \sigma^2_{X,2m}}.
\end{aligned}$$

Similar results hold for Y. Suppose that

$$\sigma^2_{X,jk}, \sigma^2_{Y,jk} \asymp k^{-\alpha} \quad \textit{and} \quad |\sigma^2_{X,jk} - \sigma^2_{Y,jk}| \asymp k^{-\beta}, \quad j = 1, 2, 3,$$

where $\alpha, \beta > 0$ and $\beta > \alpha$. Then

$$\begin{aligned}
\bar{z}^{13,M}_{mm'} &= 0,\\
\bar{z}^{12,M}_{mm'} &= \mathbb{1}(m = m')\frac{\sigma^2_{X,1m} - \sigma^2_{Y,1m}}{\sigma^2_{X,1m} \cdot \sigma^2_{Y,1m}} \asymp \mathbb{1}(m = m') \cdot m^{-(\beta-\alpha)},\\
\bar{z}^{23,M}_{mm'} &= \mathbb{1}(m = m')\frac{\sigma^2_{X,3m} - \sigma^2_{Y,3m}}{\sigma^2_{X,3m} \cdot \sigma^2_{Y,3m}} \asymp \mathbb{1}(m = m') \cdot m^{-(\beta-\alpha)}.
\end{aligned}$$

This implies that

$$\|\Delta_{13}^M\|_F^2 = \sum_{m'=1}^{M}\sum_{m=1}^{M}\left(\bar{z}_{mm'}^{13,M}\right)^2 = 0,$$
$$\|\Delta_{12}^M\|_F^2 = \sum_{m'=1}^{M}\sum_{m=1}^{M}\left(\bar{z}_{mm'}^{12,M}\right)^2 \asymp \sum_{m=1}^{M}\frac{1}{m^{\beta-\alpha}},$$
$$\|\Delta_{23}^M\|_F^2 = \sum_{m'=1}^{M}\sum_{m=1}^{M}\left(\bar{z}_{mm'}^{23,M}\right)^2 \asymp \sum_{m=1}^{M}\frac{1}{m^{\beta-\alpha}}.$$

When $\beta > \alpha+1$, we have $0 < \lim_{M\to\infty}\|\Delta_{12}^M\|_F = \lim_{M\to\infty}\|\Delta_{23}^M\|_F < \infty$. When $\beta \le \alpha+1$, we have $\lim_{M\to\infty}\|\Delta_{12}^M\|_F = \lim_{M\to\infty}\|\Delta_{23}^M\|_F = \infty$. In both cases, the two graphs are comparable.

Comparability describes population quantities and does not immediately imply that the differential graph is easy to estimate. The following example describes a sequence of MGPs that are comparable; however, the differential graph can be arbitrarily hard to estimate.

Example 2 *We define $\{\epsilon_{ijk}^X\}_{k\ge1}$, $\{a_{ijk}^X\}_{k\ge1}$, $\{\epsilon_{ijk}^Y\}_{k\ge1}$, and $\{a_{ijk}^Y\}_{k\ge1}$ as in Example 1. Let $X_{ij}(t) = \sum_{k=1}^{M^\star} a_{ijk}^X b_k(t)$ and $Y_{ij}(t) = \sum_{k=1}^{M^\star} a_{ijk}^Y b_k(t)$, $j = 1,2,3$, where $M^\star$ is a positive integer. Suppose that*

$$\sigma_{X,jk}^2, \sigma_{Y,jk}^2 \asymp k^{-\alpha} \quad \text{and} \quad |\sigma_{X,jk}^2 - \sigma_{Y,jk}^2| \asymp \mathbb{1}(k = M^\star)k^{-\beta}, \quad j = 1,2,3,$$

where $\alpha, \beta > 0$ and $\beta > \alpha$. Following the argument in Example 1, for any $1 \le M \le M^\star$, we have

$$\bar{z}_{mm'}^{13,M} = 0,$$
$$\bar{z}_{mm'}^{12,M} = \mathbb{1}(m = m')\mathbb{1}(m = M^\star)\cdot\frac{\sigma_{X,1m}^2 - \sigma_{Y,1m}^2}{\sigma_{X,1m}^2\cdot\sigma_{Y,1m}^2} \asymp \mathbb{1}(m = m')\mathbb{1}(m = M^\star)\cdot m^{-(\beta_1-2\alpha_1)},$$
$$\bar{z}_{mm'}^{23,M} = \mathbb{1}(m = m')\mathbb{1}(m = M^\star)\cdot\frac{\sigma_{X,3m}^2 - \sigma_{Y,3m}^2}{\sigma_{X,3m}^2\cdot\sigma_{Y,3m}^2} \asymp \mathbb{1}(m = m')\mathbb{1}(m = M^\star)\cdot m^{-(\beta_3-2\alpha_3)}.$$

This implies that

$$\|\Delta_{13}^M\|_F^2 = \sum_{m'=1}^{M}\sum_{m=1}^{M}\left(\bar{z}_{mm'}^{13,M}\right)^2 = 0,$$
$$\|\Delta_{12}^M\|_F^2 = \sum_{m'=1}^{M}\sum_{m=1}^{M}\left(\bar{z}_{mm'}^{12,M}\right)^2 \asymp M^{-2(\beta-2\alpha)}\mathbb{1}(M = M^\star),$$
$$\|\Delta_{23}^M\|_F^2 = \sum_{m'=1}^{M}\sum_{m=1}^{M}\left(\bar{z}_{mm'}^{23,M}\right)^2 \asymp M^{-2(\beta-2\alpha)}\mathbb{1}(M = M^\star).$$

On the basis of the calculation above, we observe that estimation of the differential graph here is intrinsically hard. For any $M < M^\star$, we have $\|\Delta_{12}^M\|_F = \|\Delta_{23}^M\|_F = 0$.

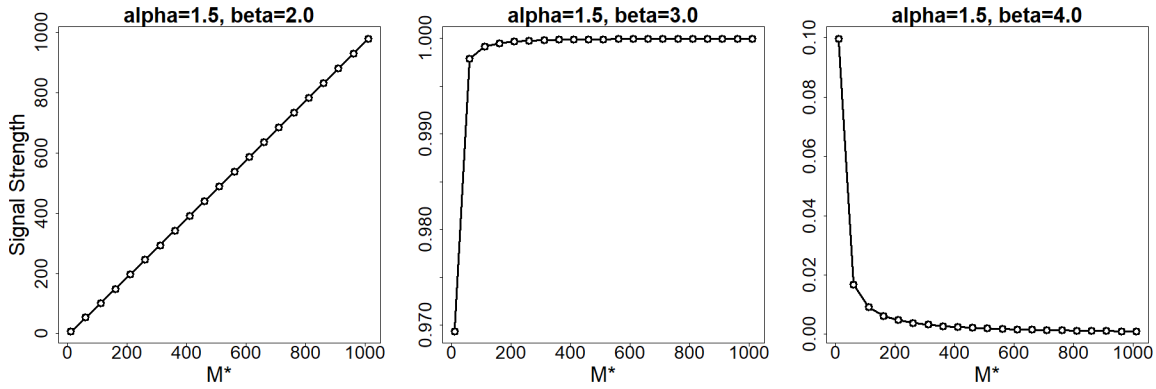


Figure 2: Signal Strength $D_{12} \asymp (M^\star)^{-2(\beta-2\alpha)}$ in Example 2.

Thus, when $M < M^\star$ is used for estimation, the resulting target graph E^π_Δ would be empty. However, by Definition 4 and Definition 5, we have $D_{12} = D_{23} \asymp (M^\star)^{-2(\beta-2\alpha)} > 0$ and $E_\Delta = \{(1,2),(2,3)\}$.

In practice, if $M^\star$ is very large and we do not have enough samples to accurately estimate Δ^M for a large M, then it is impossible for us to estimate the differential graph correctly. Furthermore, the situation is worse if $\beta > 2\alpha$, since D_{12} and D_{23}—the signal strength—vanishes as $M^\star$ increases. Figure 2 shows how the signal strength (defined as D_{12}) changes as $M^\star$ increases for three cases: $\beta < 2\alpha$, $\beta = 2\alpha$, and $\beta > 2\alpha$.

This problem is intrinsically hard because the difference between two graphs occurs only between components with the smallest positive eigenvalue. To capture this difference, we have to use a large number of bases M to approximate the functional data, which is statistically expensive. As we increase M, no useful information is captured until $M = M^\star$. Furthermore, if the difference between the eigenvalues decreases quickly relative to the decrease of the eigenvalues, the signal strength will be very weak when the intrinsic dimension is large.

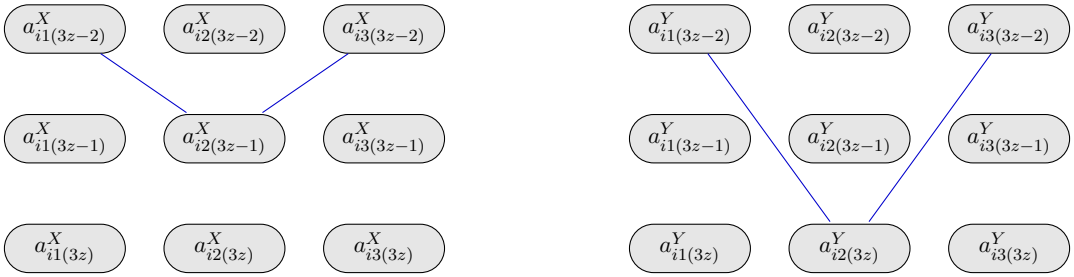
In Example 1, we characterized a pair of infinite-dimensional MGPs that are comparable, and in Example 2 we discussed a sequence of models that are all comparable, but increasingly difficult to recover. The following example demonstrates that there are infinite-dimensional MGPs that may share the same eigenspace but are still not comparable.

Example 3 *We construct two MGPs that are both infinite-dimensional and have the same eigenspace but are incomparable. As in the previous two examples, let $V = \{1,2,3\}$. We assume that X and Y share a common set of eigenfunctions: $\{\phi_m\}_{m=1}^\infty$ for $j = 1,2,3$.*

We construct the distribution of the scores of X and Y as follows. For any $m \in \mathbb{N}^+$, let $a^X_{i\cdot m}$ denote the vector of scores $(a^X_{i1m}, a^X_{i2m}, a^X_{i3m})$ and define $a^Y_{i\cdot m}$ analogously. For any natural number z, we first assume that

$$a^X_{i\cdot(3z-2)}, a^X_{i\cdot(3z-1)}, a^X_{i\cdot(3z)} \perp\!\!\!\perp \{a^X_{i\cdot k}\}_{k\neq 3z,3z-1,3z-2}.$$

Thus, the conditional independence graph for individual scores is a set of disconnected subgraphs corresponding to $\{a^X_{i\cdot(3z-2)}, a^X_{i\cdot,(3z-1)}, a^X_{i\cdot(3z)}\}$ for $z \in \mathbb{N}^+$. We make an analogous assumption for the scores of Y.



(a) CI graph for X scores (b) CI graph for Y scores

Figure 3: CI graph for the individual scores for two incomparable MGPs.

Within sets $\{a^X_{i\cdot(3z-2)}, a^X_{i\cdot(3z-1)}, a^X_{i\cdot(3z)}\}$ *and* $\{a^Y_{i\cdot(3z-2)}, a^Y_{i\cdot(3z-1)}, a^Y_{i\cdot(3z)}\}$, *we assume that the conditional independence graph has the structure shown in Figure 3. By construction, when projecting onto the span of the first* M *functions, the edge set of individual functional graphical models for* X^π *and* Y^π *is not stable as* $M \to \infty$. *In particular, for both* X *and* Y, *the edges* $(1,2)$ *and* $(2,3)$ *will persist; however, the edge* $(1,3)$ *will appear or disappear depending on* M.

If $M = 3z-2$ *for some* $z \in \mathbb{N}^+$, *which corresponds to the first row in Figure 3, then*

$$\{a^X_{i1k}\}_{k<M} \perp\!\!\!\perp \{a^X_{i3k}\}_{k<M} \mid \{a^X_{i2k}\}_{k\le M} \quad \text{and} \quad \{a^Y_{i1k}\}_{k<M} \perp\!\!\!\perp \{a^Y_{i3k}\}_{k<M} \mid \{a^Y_{i2k}\}_{k\le M}.$$

However, $a^X_{i1M} \not\perp\!\!\!\perp a^X_{i3M} \mid \{a^X_{i2k}\}_{k\le M}$ *since we do not condition on* $a^X_{i2(M+1)}$. *Similarly,* $a^Y_{i1M} \not\perp\!\!\!\perp a^Y_{i3M} \mid \{a^Y_{i2k}\}_{k\le M}$ *since we do not condition on* $a^Y_{i2(M+2)}$. *Thus, the edge* $(1,3)$ *is in the functional graphical model for both* X^π *and* Y^π; *however, the specific values of* $\Theta^{X,M}$ *and* $\Theta^{Y,M}$ *may differ.*

In contrast to the previous case, when $M = 3z-1$ *for some* $z \in \mathbb{N}^+$, *which corresponds to the second row of Figure 3, the functional graphical models for* X^π *and* Y^π *now differ. Note that* $\{a^X_{i1k}\}_{k\le M} \perp\!\!\!\perp \{a^X_{i3k}\}_{k\le M} \mid \{a^X_{i2k}\}_{k\le M}$. *Thus, the edge* $(1,3)$ *is absent from the functional graphical model for* X^π *and* $\Theta^{X,M}_{1,3} = 0$. *Considering* Y^π, *we have* $\{a^Y_{i1k}\}_{k<M-1} \perp\!\!\!\perp \{a^Y_{i3k}\}_{k<M-1} \mid \{a^Y_{i2k}\}_{k\le M}$. *However, because we do not condition on* $a^Y_{i2(M+1)}$ *(the node in the third row of Figure 3), the* $(1,3)$ *edge exists in the functional graphical model for* Y^π *since* $a^Y_{i1(M-1)} \not\perp\!\!\!\perp a^Y_{i3(M-1)} \mid \{a^Y_{i2k}\}_{k\le M}$.

Suppose that the covariance of the scores is

$$z^{-\beta} \times \begin{array}{c} \\ a^Y_{i1(3z-2)} \\ a^Y_{i1(3z-1)} \\ a^Y_{i1(3z)} \\ a^Y_{i2(3z-2)} \\ a^Y_{i2(3z-1)} \\ a^Y_{i2(3z)} \\ a^Y_{i3(3z-2)} \\ a^Y_{i3(3z-1)} \\ a^Y_{i3(3z)} \end{array} \begin{array}{c} \begin{array}{ccccccccc} a^Y_{i1(3z-2)} & a^Y_{i1(3z-1)} & a^Y_{i1(3z)} & a^Y_{i2(3z-2)} & a^Y_{i2(3z-1)} & a^Y_{i2(3z)} & a^Y_{i3(3z-2)} & a^Y_{i3(3z-1)} & a^Y_{i3(3z)} \end{array} \\ \begin{bmatrix} 3/2 & 0 & 0 & 0 & 0 & -1 & 1/2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 2 & -1 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & -1 & 3/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array},$$

where $\beta > 0$ is a parameter that determines the decay rate of the eigenvalues (see Assumption 3). We then set all other elements of the covariance as 0. The support of the inverse of this matrix corresponds to the edges of the graph in Figure 3. However, when we consider the marginal distribution of the first M scores and invert the corresponding covariance, $\Theta^{Y,M}_{1,3}$ is 0 everywhere except for the element corresponding to $a^Y_{i,1,M-1}$ and $a^Y_{i,3,M-1}$, that is, the nodes in the top row of Figure 3, which is equal to $-1/4 \times ((M+1)/3)^\beta$. Thus, $\|\Delta^M_{1,3}\|_F = 1/4 \times ((M+1)/3)^\beta$ and $\limsup_{M\to\infty} \|\Delta^M_{1,3}\|_F = \infty$.

Finally, when $M = 3z$ for some $z \in \mathbb{N}^+$, the $(1,3)$ edge is absent in both functional graphical models for X^π and Y^π because

$$\{a^X_{i1k}\}_{k\le M} \perp\!\!\!\perp \{a^X_{i3k}\}_{k\le M} \mid \{a^X_{i2k}\}_{k\le M} \quad \text{and} \quad \{a^Y_{i1k}\}_{k\le M} \perp\!\!\!\perp \{a^Y_{i3k}\}_{k\le M} \mid \{a^Y_{i2k}\}_{k\le M}.$$

Thus, $\Theta^{X,M}_{1,3} = \Theta^{Y,M}_{1,3} = \Delta^M_{1,3} = 0$. This implies that $\liminf_{M\to\infty} \|\Delta^M_{1,3}\|_F = 0$.

Because $\liminf_{M\to\infty} \|\Delta^M_{1,3}\|_F = 0$, but $\limsup_{M\to\infty} \|\Delta^M_{1,3}\|_F = \infty$, X and Y are incomparable.

The notion of comparability illustrates the intrinsic difficulty of dealing with functional data. However, it also illustrates when we can still hope to estimate the differential network consistently. We have formally stated when two infinite-dimensional functional graphical models will be comparable and have given conditions and examples of comparability. Unfortunately, these conditions cannot be checked using the observed data. For this reason, we mainly discuss the methodology and theoretical properties for estimation of E^π_Δ. Prior knowledge about the problem at hand should be used to decide whether two infinite-dimensional functional graphs are comparable. This is similar to other assumptions common in the graphical modeling literature, such as "faithfulness" (Spirtes et al., 2000), that are critical to graph recovery, but can not be verified.

3. Functional Differential Graph Estimation: FuDGE

In this section, we detail our methodology for estimating a functional differential graph. Unfortunately, in most situations, there may not be prior knowledge on which subspace to use to define the functional differential graph. In such situations, we suggest using the principle component scores of $K_{jj}(s,t) = K^X_{jj}(s,t) + K^Y_{jj}(s,t)$, $j \in V$ as a default choice. In addition, each observed function is only recorded (potentially with measurement error) at discrete time points. In Section 3.1 we consider this practical setting. Of course, if an appropriate basis for dimension reduction is known in advance or if the functions are fully observed at all time points, then the estimated objects can always be replaced with their known/observed counterparts.

3.1 Estimating the Covariance of the Scores

For each X_{ij}, suppose we have measurements at time points t_{ijk}, $k = 1, \ldots, T$,[3] and the recorded data, h_{ijk}, are the function values with random noise. That is,

$$h_{ijk} = g_{ij}(t_{ijk}) + \epsilon_{ijk}, \tag{7}$$

3. For simplicity, we assume that all functions have the same number of observations, however, our method and theory can be trivially extended to allow a different number of observations for each function.

where g_{ij} can denote either X_{ij} or Y_{ij} and the unobserved noise ϵ_{ijk} is i.i.d. Gaussian with mean 0 and variance σ_0^2. Without loss of generality, we assume that $t_{ij1} < \ldots < t_{ijT}$ for any $1 \le i \le n$ and $1 \le j \le p$. We do not assume that $t_{ijk} = t_{i'jk}$ for $i \ne i'$, so that each observation may be observed on a different grid.

We first use a basis expansion to estimate a least squares approximation of the whole curve $X_{ij}(t)$ (see Section 4.2 in Ramsay and Silverman (2005)). Specifically, given an initial basis function vector $b(t) = (b_1(t), \ldots, b_L(t))^\top$—for example, the B-spline or Fourier basis—our estimated approximation for $X_{ij}(t)$ is given by:

$$\hat{X}_{ij}(t) = \hat{\beta}_{ij}^\top b(t),$$
$$\hat{\beta}_{ij} = \left(B_{ij}^\top B_{ij}\right)^{-1} B_{ij}^\top h_{ij},$$

where $h_{ij} = (h_{ij1}, h_{ij2}, \ldots, h_{ijT})^\top$ and B_{ij} is the design matrix for g_{ij}:

$$B_{ij} = \begin{bmatrix} b_1(t_{ij1}) & \cdots & b_L(t_{ij1}) \\ \vdots & \ddots & \vdots \\ b_1(t_{ijT}) & \cdots & b_L(t_{ijT}) \end{bmatrix} \in \mathbb{R}^{T \times L}. \tag{8}$$

The computational complexity of the basis expansion procedure is $O(npT^3L^3)$, and in practice, there are many efficient package implementations of this step; for example, `fda` (Ramsay et al., 2020).

We repeat this process for the observed Y functions. After obtaining $\{\hat{X}_{ij}(t)\}_{j \in V, i=1,\ldots,n_X}$ and $\{\hat{Y}_{ij}(t)\}_{j \in V, i=1,\ldots,n_Y}$, we use them as inputs for the FPCA procedure. Specifically, we first estimate the sum of the covariance functions by

$$\hat{K}_{jj}(s,t) = \hat{K}_{jj}^X(s,t) + \hat{K}_{jj}^Y(s,t) = \frac{1}{n_X}\sum_{i=1}^{n_X} \hat{X}_{ij}(s)\hat{X}_{ij}(t) + \frac{1}{n_Y}\sum_{i=1}^{n_Y} \hat{Y}_{ij}(s)\hat{Y}_{ij}(t). \tag{9}$$

Using $\hat{K}_{jj}(s,t)$ as input to FPCA, we can estimate the corresponding eigenfunctions $\hat{\phi}_{jk}(t)$, $k = 1, \ldots, M$, $j = 1, \ldots, p$. Given the estimated eigenfunctions, we compute the estimated projection scores

$$\hat{a}_{ijk}^X = \int_{\mathcal{T}} \hat{X}_{ij}(t)\hat{\phi}_{jk}(t)dt \qquad \text{and} \qquad \hat{a}_{ijk}^Y = \int_{\mathcal{T}} Y_{ij}(t)\hat{\phi}_{jk}(t)dt,$$

and collect them into vectors

$$\hat{a}_{ij}^{X,M} = (\hat{a}_{ij1}^X, \cdots, \hat{a}_{ijM}^X)^\top \in \mathbb{R}^M \qquad \text{and} \qquad \hat{a}_i^{X,M} = ((\hat{a}_{i1}^{X,M})^\top, \ldots, (\hat{a}_{ip}^{X,M})^\top)^\top \in \mathbb{R}^{pM},$$
$$\hat{a}_{ij}^{Y,M} = (\hat{a}_{ij1}^Y, \cdots, \hat{a}_{ijM}^Y)^\top \in \mathbb{R}^M \qquad \text{and} \qquad \hat{a}_i^{Y,M} = ((\hat{a}_{i1}^{Y,M})^\top, \ldots, (\hat{a}_{ip}^{Y,M})^\top)^\top \in \mathbb{R}^{pM}.$$

Finally, we estimate the covariance matrices of the score vectors, $\Sigma^{X,M}$ and $\Sigma^{Y,M}$, as

$$S^{X,M} = \frac{1}{n_X}\sum_{i=1}^{n_X} \hat{a}_i^{X,M}(\hat{a}_i^{X,M})^\top \qquad \text{and} \qquad S^{Y,M} = \frac{1}{n_Y}\sum_{i=1}^{n_Y} \hat{a}_i^{Y,M}(\hat{a}_i^{Y,M})^\top.$$

3.2 FuDGE: Functional Differential Graph Estimation

Now we describe the FuDGE algorithm for **Fu**nctional **D**ifferential **G**raph **E**stimation. To estimate Δ^M, we solve the following optimization program:

$$\hat{\Delta}^M \in \underset{\Delta \in \mathbb{R}^{pM \times pM}}{\arg\min}\ \mathcal{L}(\Delta) + \lambda_n \sum_{\{i,j\} \in V^2} \|\Delta_{ij}\|_F, \tag{10}$$

where

$$\mathcal{L}(\Delta) = \operatorname{tr}\left[\frac{1}{2} S^{Y,M} \Delta^\top S^{X,M} \Delta - \Delta^\top \left(S^{Y,M} - S^{X,M}\right)\right] \tag{11}$$

and $S^{X,M}$ and $S^{Y,M}$ are obtained as described in Section 3.1.

We construct the loss function, $\mathcal{L}(\Delta)$, so that the true parameter value, that is, $\Delta^M = \left(\Sigma^{X,M}\right)^{-1} - \left(\Sigma^{Y,M}\right)^{-1}$, minimizes the population loss $\mathbb{E}\left[\mathcal{L}(\Delta)\right]$, which for a differentiable and convex loss function is equivalent to selecting $\mathcal{L}$ so that $\mathbb{E}\left[\nabla \mathcal{L}(\Delta^M)\right] = 0$. Since Δ^M satisfies

$$\Sigma^{X,M} \Delta^M \Sigma^{Y,M} - (\Sigma^{Y,M} - \Sigma^{X,M}) = 0,$$

a choice for $\nabla \mathcal{L}(\Delta)$ is

$$\nabla \mathcal{L}(\Delta^M) = S^{X,M} \Delta^M S^{Y,M} - \left(S^{Y,M} - S^{X,M}\right) \tag{12}$$

so that

$$\mathbb{E}\left[\nabla \mathcal{L}(\Delta^M)\right] = \Sigma^{X,M} \Delta^M \Sigma^{Y,M} - (\Sigma^{Y,M} - \Sigma^{X,M}) = 0.$$

Given this choice of $\nabla \mathcal{L}(\Delta)$, $\mathcal{L}(\Delta)$ in (10) directly follows from properties of the differential of the trace function. The chosen loss is quadratic (see (B.6) in appendix) and leads to an efficient algorithm. Similar loss functions are used in Xu and Gu (2016), Yuan et al. (2017), Na et al. (2021), and Zhao et al. (2014a).

We also include the additional group lasso penalty (Yuan and Lin, 2006) to promote blockwise sparsity in $\hat{\Delta}^M$. The objective in (10) can be solved by a proximal gradient method detailed in Algorithm 1. Finally, we form $\hat{E}_\Delta$ by thresholding $\hat{\Delta}^M$ so that:

$$\hat{E}_\Delta = \left\{ \{j,l\} \,:\, \|\hat{\Delta}^M_{jl}\|_F > \epsilon_n \ \text{ or } \ \|\hat{\Delta}^M_{lj}\|_F > \epsilon_n \right\}. \tag{13}$$

The thresholding step in (13) is used for theoretical purposes. Specifically, it helps correct for bias induced by finite-dimensional truncation and relaxes commonly used assumptions for the recovery of the graph structure, such as the irrepresentability or incoherence condition (van de Geer and Bühlmann, 2009). In practice, one can simply set $\epsilon_n = 0$, as we do in the simulations.

3.3 Optimization Algorithm for FuDGE

The optimization program (10) can be solved using a proximal gradient method (Parikh and Boyd, 2014) summarized in Algorithm 1. Specifically, at each iteration, we update the current value of Δ, denoted as Δ^{old}, by solving the following problem:

$$\Delta^{\text{new}} = \underset{\Delta}{\arg\min} \left(\frac{1}{2} \left\| \Delta - \left(\Delta^{\text{old}} - \eta \nabla \mathcal{L}\left(\Delta^{\text{old}}\right) \right) \right\|_F^2 + \eta \cdot \lambda_n \sum_{j,l=1}^{p} \|\Delta_{jl}\|_F \right), \tag{14}$$

Algorithm 1 Functional differential graph estimation

Input: $S^{X,M}, S^{Y,M}, \lambda_n, \eta$.
Output: $\hat{\Delta}^M$.
Initialize $\Delta^{(0)} = 0_{pM}$.
repeat
$A = \Delta - \eta \nabla \mathcal{L}(\Delta) = \Delta - \eta \left(S^{X,M} \Delta S^{Y,M} - \left(S^{Y,M} - S^{X,M} \right) \right)$
for $1 \leq i, j \leq p$ **do**
$\Delta_{jl} \leftarrow \left(\frac{\|A_{jl}\|_F - \lambda_n \eta}{\|A_{jl}\|_F} \right)_+ \cdot A_{jl}$
end for
until Converge

where $\nabla \mathcal{L}(\Delta)$ is defined in (12) and η is a step size specified by the user. Note that $\nabla \mathcal{L}(\Delta)$ is Lipschitz continuous with Lipschitz constant $\lambda^S_{\max} = \|S^{Y,M} \otimes S^{X,M}\|_2 = \lambda_{\max}(S^{Y,M}) \lambda_{\max}(S^{X,M})$. Thus, for any step size η such that $0 < \eta \leq 1/\lambda^S_{\max}$, the proximal gradient method is guaranteed to converge (Beck and Teboulle, 2009).

The update in (14) has a closed-form solution:

$$\Delta^{\text{new}}_{jl} = \left[\left(\|A^{\text{old}}_{jl}\|_F - \lambda_n \eta \right) / \|A^{\text{old}}_{jl}\|_F \right]_+ \cdot A^{\text{old}}_{jl}, \qquad 1 \leq j, l \leq p, \tag{15}$$

where $A^{\text{old}} = \Delta^{\text{old}} - \eta \nabla \mathcal{L}(\Delta^{\text{old}})$ and $x_+ = \max\{0, x\}, x \in \mathbb{R}$, represents the positive part of x. Detailed derivations are given in the appendix. Note that although true Δ^M is symmetric, we do not explicitly enforce symmetry in $\hat{\Delta}^M$ in Algorithm 1.

After performing FPCA, the proximal gradient descent method converges in $O\left(\lambda^S_{\max}/\text{tol}\right)$ iterations, where tol is a user specified optimization error tolerance, and each iteration takes $O((pM)^3)$ operations; see Tibshirani (2010) for a convergence analysis of the general proximal gradient descent algorithm.

3.4 Selection of Tuning Parameters

There are four tuning parameters that must be chosen for the implementation of FuDGE: L (basis dimension used to estimate the curves from the discretely observed data), M (subspace dimension to estimate projection scores), λ_n (regularization parameter to tune the block sparsity of Δ^M), and ϵ_n (thresholding parameter for $\hat{E}_\Delta$). While we need the thresholding parameter ϵ_n in (13) to establish theoretical results, in practice, we simply set $\epsilon_n = 0$. To select M, we follow the procedure in Qiao et al. (2019). More specifically, for each discretely-observed curve, we first estimate the underlying functions by fitting an L-dimensional B-spline basis. Both M and L are chosen by 5-fold cross-validation as discussed in Qiao et al. (2019).

Finally, to choose λ_n, we recommend using selective cross-validation (SCV) (She, 2012). Given a value of λ_n, we use the entire data set to estimate a sparsity pattern. Then, fixing the sparsity pattern, we use a typical cross-validation procedure to calculate the CV error, where the CV error is measured by an unpenalized version of the loss function in (11) on the held-out data set. Ultimately, we choose the value of λ_n that results in the sparsity pattern that minimizes CV error. In addition to SCV, if we have prior knowledge about

the number of edges in the differential graph, we can also choose λ_n which results in the desired level of sparsity of the differential graph.

4. Theoretical Properties

In this section, we provide theoretical guarantees for FuDGE. We first give a deterministic result for $\hat{E}_\Delta$ defined in (13) when the max-norm of the difference between the estimates $S^{X,M}, S^{Y,M}$ and their corresponding parameters, $\Sigma^{X,M}, \Sigma^{Y,M}$, is bounded by δ_n. We then show that when projecting the data onto either a fixed basis or an estimated basis—under some mild conditions—δ_n can be controlled and the bias of the finite-dimensional projection decreases fast enough that E_Δ can be consistently recovered.

4.1 Deterministic Guarantees for $\hat{E}_\Delta$

In this section, we assume that $S^{X,M}, S^{Y,M}$ are good estimates of $\Sigma^{X,M}, \Sigma^{Y,M}$ and give a deterministic result in Theorem 10. Let $n = \min\{n_X, n_Y\}$. We assume that the following holds.

Assumption 1 *The matrices $S^{X,M}, S^{Y,M}$ are estimates of $\Sigma^{X,M}, \Sigma^{Y,M}$ that satisfy*

$$\max\left\{|S^{X,M} - \Sigma^{X,M}|_\infty, |S^{Y,M} - \Sigma^{Y,M}|_\infty\right\} \leq \delta_n. \tag{16}$$

We also require that E_Δ be sparse. This does not preclude the case where E_X and E_Y are dense, as long as there are not too many differences in the precision matrices. This assumption is also required when estimating a differential graph from vector-valued data; for example, see Condition 1 in Zhao et al. (2014a).

Assumption 2 *There are s edges in the differential graph; that is, $|E_\Delta| = s$ and $s \ll p$.*

We introduce the following three quantities that characterize the problem instance and will be used in Theorem 10 below:

$$\nu_1 = \nu_1(M) = \min_{(j,l)\in E_\Delta} \|\Delta^M_{jl}\|_F, \quad \nu_2 = \nu_2(M) = \max_{(j,l)\in E^C_\Delta} \|\Delta^M_{jl}\|_F,$$

and

$$\tau = \tau(M) = \nu_1(M) - \nu_2(M).$$

Roughly speaking, $\nu_1(M)$ indicates the "signal strength" present when we use the M-dimensional representation and $\nu_2(M)$ measures the bias. By Definition 4, when X and Y are comparable, we have $\liminf_{M\to M^\star} \nu_1(M) > 0$ and $\lim_{M\to M^\star} \nu_2(M) = 0$. Therefore, for a sufficiently large M, we have $\tau > 0$. However, a smaller τ implies that the differential graph is harder to recover.

Before we give the deterministic result in Theorem 10, we first define additional quantities that will be used in subsequent results. Let

$$\begin{aligned}
\sigma_{\max} &= \max\{|\Sigma^{X,M}|_\infty, |\Sigma^{Y,M}|_\infty\}, \\
\lambda^*_{\min} &= \lambda_{\min}\left(\Sigma^{X,M}\right) \times \lambda_{\min}\left(\Sigma^{Y,M}\right), \text{ and} \\
\Gamma^2_n &= \frac{9\lambda_n^2 s}{\kappa^2_{\mathcal{L}}} + \frac{2\lambda_n}{\kappa_{\mathcal{L}}}(\omega^2_{\mathcal{L}} + 2p^2\nu_2),
\end{aligned}$$

where

$$\begin{aligned}\lambda_n &= 2M\left[\left(\delta_n^2 + 2\delta_n\sigma_{\max}\right)\left|\Delta^M\right|_1 + 2\delta_n\right],\\ \kappa_{\mathcal{L}} &= (1/2)\lambda^*_{\min} - 8M^2 s\left(\delta_n^2 + 2\delta_n\sigma_{\max}\right),\\ \omega_{\mathcal{L}} &= 4Mp^2\nu_2\sqrt{\delta_n^2 + 2\delta_n\sigma_{\max}},\end{aligned}$$

and δ_n is defined in Assumption 1. Note that Γ_n—which measures the estimation error of $\|\hat{\Delta}^M - \Delta^M\|_{\mathrm{F}}$—implicitly depends on δ_n through λ_n, $\kappa_{\mathcal{L}}$, and $\omega_{\mathcal{L}}$. We observe that Γ_n decreases to zero as δ_n goes to zero. The quantity $\kappa_{\mathcal{L}}$ is the maximum restricted eigenvalue from the analysis framework of Negahban et al. (2012). Finally, $\omega_{\mathcal{L}}$ is the tolerance parameter that comes from the fact that ν_2 could be larger than zero, and will decrease to zero as ν_2 goes to zero.

Theorem 10 *Given Assumptions 1 and 2, when $\nu_1(M), \nu_2(M), \delta_n, \lambda_n, \sigma_{\max}, M$ and s satisfy*

$$0 < \Gamma_n < \tau/2 \qquad \text{and} \qquad \delta_n < (1/4)\sqrt{\left(\lambda^*_{\min} + 16M^2 s(\sigma_{\max})^2\right)/\left(M^2 s\right)} - \sigma_{\max},$$

then setting $\epsilon_n \in [\nu_2 + \Gamma_n, \nu_1 - \Gamma_n)$ ensures that $\hat{E}_\Delta = E_\Delta$.

As shown in Section 4.2, under some additional conditions, Assumption 1 holds for a sequence of δ_n that decreases to 0 as n goes to infinity. Thus, as M and n both increase to infinity, we have $\nu_2 + \Gamma_n \approx 0$ and $\nu_1 - \Gamma_n \approx \min_{(j,l)\in E_\Delta} D_{jl}$, and we only require $0 \le \epsilon_n < \min_{(j,l)\in E_\Delta} D_{jl}$.

4.2 Theoretical Guarantees for $S^{X,M}$ and $S^{Y,M}$

In this section, we prove that under some mild conditions, (16) will hold with a high probability for specific values of δ_n. We discuss the results in two cases: the case where the curves are fully observed and the case where the curves are only observed at discrete time points.

4.2.1 Fully Observed Curves

In this section, we discuss the case where each curve is fully observed. We first consider the case where the basis defining the differential graph is known in advance; that is, the exact form of $\{e_{jk}\}_{k\ge1}$ for all $j \in V$ is known. In this case, the projection score vectors $a_i^{X,M}$ and $a_i^{Y,M}$ can be recovered exactly for all $i = 1, 2, \ldots, n$. By the assumption that $X_i(t)$ and $Y_i(t)$ are p-dimensional multivariate Gaussian processes with mean zero, we then have $a_i^{X,M} \sim N(0, \Sigma^{X,M})$ and $a_i^{Y,M} \sim N(0, \Sigma^{Y,M})$. The following result follows directly from the standard results on the sample covariance of multivariate Gaussian variables.

Theorem 11 *Assume that $S^{X,M}$ and $S^{Y,M}$ are computed as in Section 3.1, except that the basis functions $\{e_{jk}\}_{k\ge1}$, $j \in V$, are fixed and known in advance. Recall that*

$$n = \min\{n_X, n_Y\} \quad \text{and} \quad \sigma_{\max} = \max\{|\Sigma^{X,M}|_\infty, |\Sigma^{Y,M}|_\infty\}.$$

Fix $\iota \in (0, 1]$. Suppose that n is large enough so that

$$\delta_n = \sigma_{\max}\sqrt{\frac{C_1}{n}\log\left(\frac{8p^2M^2}{\iota}\right)} \le C_2,$$

for some universal constants $C_1, C_2 > 0$. Then (16) *holds with probability at least $1 - \iota$.*

Proof The proof follows directly from Lemma 1 of Ravikumar et al. (2011) and a union bound. ■

With fully observed curves and known basis functions, it follows from Theorem 11 that (1) is satisfied for $\delta_n \asymp \sqrt{\log(p^2M^2)/n}$ with high probability. As assumed in Section 2.2 (and also in Qiao et al. (2019)), when $\lambda^X_{jm'} = \lambda^Y_{jm'} = 0$ for all j and $m' > M$ (where M is allowed to grow with n), then $\nu_2(M) = 0$, $\tau(M) = \nu_1(M) = \min_{(j,l)\in E_\Delta} D_{jl} > 0$, and $E_\Delta = E^\pi_\Delta$. We can recover E_Δ with high probability even in the high-dimensional setting, as long as

$$\max\left\{\frac{sM^2\log(p^2M^2)|\Delta^M|_1^2/((\lambda^\star_{\min})^2\tau^2)}{n}, \frac{sM^2\log(p^2M^2)/\lambda^\star_{\min}}{n}\right\} \to 0.$$

Even with an infinite number of positive eigenvalues, high-dimensional consistency is still possible for quickly increasing ν_1 and quickly decaying ν_2.

We then consider the case where the curves are fully observed, but we do not have any prior knowledge on which orthonormal function basis should be used. In this case, as discussed in Section 2.3, we recommend using the eigenfunctions of $K_{jj}(\cdot,*) = K^X_{jj}(\cdot,*) + K^Y_{jj}(\cdot,*)$ as basis functions. We use FPCA to estimate the eigenfuctions of $K_{jj}(\cdot,*)$ and make the following assumption.

Assumption 3 *Let $\{\lambda_{jk}, \phi_{jk}(\cdot)\}$ be the eigenpairs of $K_{jj}(\cdot,*) = K^X_{jj}(\cdot,*)+K^Y_{jj}(\cdot,*)$, $j \in V$, and suppose that λ_{jk} are non-increasing in k.*

(i) Suppose $\max_{j\in V}\sum_{k=1}^\infty \lambda_{jk} < \infty$ and assume that there exists a constant $\beta > 1$ such that, for each $k \in \mathbb{N}$, $\lambda_{jk} \asymp k^{-\beta}$ and $d_{jk}\lambda_{jk} = O(k)$ uniformly in $j \in V$, where $d_{jk} = 2\sqrt{2}\max\{(\lambda_{j(k-1)} - \lambda_{jk})^{-1}, (\lambda_{jk} - \lambda_{j(k+1)})^{-1}\}$, $k \geq 2$, and $d_{j1} = 2\sqrt{2}(\lambda_{j1} - \lambda_{j2})^{-1}$.

(ii) For all k, $\phi_{jk}(\cdot)$'s are continuous on the compact set $\mathcal{T}$ and satisfy

$$\max_{j\in V}\sup_{s\in\mathcal{T}}\sup_{k\geq 1}|\phi_{jk}(s)|_\infty = O(1).$$

This assumption was used in Qiao et al. (2019, Condition 1). We have the following result.

Theorem 12 *Suppose Assumption 3 holds and the basis functions are estimated using the FPCA of $K_{jj}(\cdot,*)$ with fully observed curves. Fix $\iota \in (0,1]$. Suppose that n is large enough so that*

$$\delta_n = M^{1+\beta}\sqrt{\frac{\log\left(2C_2p^2M^2/\iota\right)}{n}} \leq C_1,$$

for some universal constants $C_1, C_2 > 0$. Then (16) *holds with probability at least $1 - \iota$. Note that C_1, C_2 may be different constants from those of Theorem 11.*

Proof The proof follows directly from Theorem 1 of Qiao et al. (2019) and the fact that $\|\hat{K}_{jj}(\cdot,*) - K_{jj}(\cdot,*)\|_{\text{HS}} \leq \|\hat{K}^X_{jj}(\cdot,*) - K^X_{jj}(\cdot,*)\|_{\text{HS}} + \|\hat{K}^Y_{jj}(\cdot,*) - K^Y_{jj}(\cdot,*)\|_{\text{HS}}$. ■

It follows from Theorem 12 that (1) holds for $\delta_n \asymp M^{1+\beta}\sqrt{\log(p^2M^2/)/n}$ with high probability. Compared to Theorem 11, there is an additional $M^{1+\beta}$ term that arises from FPCA estimation error. Similarly, when $\lambda^X_{jm'} = \lambda^Y_{jm'} = 0$ for all j and $m' > M$, we can recover E_Δ with high probability as long as

$$\max\left\{\frac{sM^{(4+2\beta)}\log(p^2M^2)|\Delta^M|_1^2/((\lambda^\star_{\min})^2\tau^2)}{n}, \frac{sM^{(4+2\beta)}\log(p^2M^2)/\lambda^\star_{\min}}{n}\right\} \to 0.$$

4.2.2 Discretely-Observed Curves

Finally, we discuss the case where the curves are only observed at discrete time points—possibly with measurement error. Following Chapter 1 of Kokoszka and Reimherr (2017), we first estimate each curve from the available observations by basis expansion; then, we use the estimated curves to form empirical covariance functions from which we estimate the eigenfunctions using FPCA. The estimated eigenfunctions are then used to calculate the scores.

Recall the model for discretely observed functions given in (7):

$$h_{ijk} = g_{ij}(t_{ijk}) + \epsilon_{ijk},$$

where g_{ij} denotes either X_{ij} or Y_{ij}, ϵ_{ijk} are i.i.d. Gaussian noise with mean 0 and variance σ_0^2. Assume that $t_{ij1} < \cdots < t_{ijT}$ for any $1 \le i \le n$ and $1 \le j \le p$. Note that we do not need X and Y to be observed at the same time points and we use t_{ijk} to represent either t^X_{ijk} or t^Y_{ijk}. Furthermore, recall that we first compute a least squares estimator of $X_{ij}(\cdot)$ and $Y_{ij}(\cdot)$ by projecting it onto the basis $b(\cdot) = (b_1(\cdot), \ldots, b_L(\cdot))$.

First, we assume that as we increase the number of basis functions, we can approximate any function in $\mathbb{H}$ arbitrarily well.

Assumption 4 *We assume that $\{b_l\}_{l=1}^\infty$ is a complete orthonormal system (CONS) (See Definition 2.4.11 of Hsing and Eubank, 2015) of $\mathbb{H}$, that is, $\overline{\text{Span}\left(\{b_l\}_{l=1}^\infty\right)} = \mathbb{H}$.*

Assumption 4 requires that the basis functions are orthonormal. When this assumption is violated—for example, when using the B-splines basis—we can always first use an orthonormalization process, such as Gram-Schmidt, to convert the basis to an orthonormal one. For B-splines, there are many algorithms that can efficiently provide orthonormalization (Liu et al., 2022).

To establish theoretical guarantees for the least squares estimator, we require smoothness in both the curves we are trying to estimate as well as the basis functions we use.

Assumption 5 *We assume that the basis functions $\{b_l(\cdot)\}_{l=1}^\infty$ satisfy the following conditions:*

$$D_{0,b} \coloneqq \sup_{l\ge1}\sup_{t\in\mathcal{T}}|b_l(t)| < \infty, \qquad D_{1,b}(l) \coloneqq \sup_{t\in\mathcal{T}}|b_l'(t)| < \infty, \qquad D_{1,b,L} \coloneqq \max_{1\le l\le L} D_{1,b}(l).$$

We also require that the curves g_{ij} satisfy the following smoothness condition:

$$\max_{1\le j\le p}\sum_{m=1}^\infty \mathbb{E}\left[\left(\langle g_{ij}, b_m\rangle\right)^2\right] D^2_{1,b}(m) < \infty. \tag{17}$$

To better understand Assumption 5, we use the Fourier basis as an example. Let $\mathcal{T} = [0,1]$ and $b_m(t) = \sqrt{2}\cos(2\pi m t)$, $0 \le t \le 1$ and $m \in \mathbb{N}$. Thus, $\{b_m(t)\}_{m=0}^{\infty}$ then constitutes an orthonormal basis of $\mathbb{H} = \mathcal{L}^2[0,1]$. We then have $b'(t) = -2\sqrt{2}\pi m \sin(2\pi m t)$, $D_{0,b} = \sqrt{2}$, $D_{1,b}(m) = 2\sqrt{2}\pi m$ and $D_{1,b,L} = 2\sqrt{2}\pi L$. In this case, (17) is equivalent to

$$\max_{1\le j\le p} \sum_{m=1}^{\infty} \mathbb{E}\left[\left(\langle g_{ij}, b_m\rangle\right)^2\right] m^2 < \infty.$$

On the other hand, $g_{ij}(t) = \sum_{m=1}^{\infty}\langle g_{ij}, b_m\rangle b_m(t)$ and $g'_{ij}(t) = \sum_{m=1}^{\infty}\langle g_{ij}, b_m\rangle b'_m(t)$. Suppose that $\mathbb{E}\left[\|g'_{ij}\|^2\right] < \infty$. Then

$$\mathbb{E}\left[\|g'_{ij}\|^2\right] = \sum_{m=1}^{\infty} \mathbb{E}\left[\left(\langle g_{ij}, b_m\rangle\right)^2\right] \|b'_m\|^2 \asymp \sum_{m=1}^{\infty} \mathbb{E}\left[\left(\langle g_{ij}, b_m\rangle\right)^2\right] m^2.$$

Therefore, $\max_{1\le j\le p}\mathbb{E}\left[\|g'_{ij}\|^2\right] < \infty$, which is a commonly used assumption in nonparameteric statistics (e.g., Section 7.2 of Wasserman (2006)), implies (17).

Finally, we require that each function be observed at time points that are "evenly spaced." Formally, we require the following assumption.

Assumption 6 *The observation time points* $\{t_{ijk} : 1 \le i \le n, 1 \le j \le p, 1 \le k \le T\}$ *satisfy*

$$\max_{1\le i\le n}\max_{1\le j\le p}\max_{1\le k\le T+1}\left|\frac{t_{ijk} - t_{ij(k-1)}}{|\mathcal{T}|} - \frac{1}{T}\right| \le \frac{\zeta_0}{T^2},$$

where t_{ij0} *and* $t_{ij(T+1)}$ *are endpoints of* $\mathcal{T}$ *for any* $1 \le i \le n$, $1 \le j \le p$, *and* ζ_0 *is a positive constant that does not depend on* i *or* j.

Any g_{ij} can be decomposed into $g_{ij} = g_{ij}^{\shortparallel} + g_{ij}^{\perp}$, where $g_{ij}^{\shortparallel} \in \mathrm{Span}(b)$ and $g_{ij}^{\perp} \in \mathrm{Span}(b)^{\perp}$. We denote the eigenvalues of the covariance operator of g_{ij} as $\{\lambda_{jk}\}_{k\ge1}$ and $\lambda_{j0} = \sum_{k=1}^{\infty}\lambda_{jk}$; and we denote the eigenvalues of the covariance operator of $g_{ij}^{\perp}$ as $\{\lambda_{jk}^{\perp}\}_{k\ge1}$ and $\lambda_{j0}^{\perp} = \sum_{k=1}^{\infty}\lambda_{jk}^{\perp}$. Note that under Assumption 3, we have $\max_{1\le j\le p}\lambda_{j0} < \infty$. Let $1 < \lambda_{0,\max} < \infty$ be a constant such that $\max_{1\le j\le p}\lambda_{j0} \le \lambda_{0,\max}$. Let B_{ij} be the design matrix of g_{ij} as defined in (8) and let $\lambda_{\min}^{B} = \min_{1\le i\le n, 1\le j\le p}\left\{\lambda_{\min}(B_{ij}^{\top}B_{ij})\right\}$. We define

$$\tilde{\psi}_1(T,L) = \frac{\sigma_0 L}{\sqrt{\lambda_{\min}^{B}}}, \quad \tilde{\psi}_2(T,L) = \frac{L^2}{(\lambda_{\min}^{B})^2}\left(\lambda_0\left(\tilde{c}_1 D_{1,b,L}^2 + \tilde{c}_2\right)\tilde{\psi}_3(L) + \tilde{c}_1\tilde{\psi}_4(L)\right),$$

$$\tilde{\psi}_3(L) = \max_{1\le j\le p}\left(\lambda_{j0}^{\perp}/\lambda_{j0}\right), \quad \tilde{\psi}_4(L) = \max_{1\le j\le p}\sum_{m>L}\mathbb{E}\left[\left(\langle g_{ij}, b_m\rangle\right)^2\right] D_{1,b}^2(m),$$

$$\Phi(T,L) = \min\left\{1/\tilde{\psi}_1(T,L), 1/\sqrt{\tilde{\psi}_3(L)}\right\},$$

where $\tilde{c}_1 = 18 D_{0,b}^2(\zeta_0+1)^4|\mathcal{T}|^2$ and $\tilde{c}_2 = 36 D_{0,b}^4(2\zeta_0+1)^2$.

We now use superscripts or subscripts to indicate the specific quantities for X and Y. In this way, we define L_X, L_Y, T_X, T_Y, $\tilde{\psi}_1^X$-$\tilde{\psi}_4^X$, $\tilde{\psi}_1^Y$-$\tilde{\psi}_4^Y$, and Φ^X, Φ^Y. Furthermore, let $T = \min\{T_X, T_Y\}$, $L = \min\{L_X, L_Y\}$, $\bar{\psi}_k = \max\{\tilde{\psi}_k^X, \tilde{\psi}_k^Y\}$, $k = 1, \cdots, 4$, $\bar{\Phi} = \min\{\Phi^X, \Phi^Y\}$, and let n, β be defined as in Section 4.1.

Theorem 13 *Assume the observation model given in* (7)*. Suppose Assumption 3 holds and Assumptions 4-6 hold for both X and Y. Suppose T and L are large enough so that*

$$\bar{\psi}_1(T,L) \leq \gamma_1 \frac{\delta_n}{M^{1+\beta}}, \quad \bar{\psi}_3(L) \leq \gamma_3 \frac{\delta_n^2}{M^{2+2\beta}}$$

where

$$\delta_n = \max\left\{ \frac{M^{1+\beta}\log\left(4\bar{C}_1 np/\iota\right)}{\bar{C}_2 \bar{\Phi}(T,L)}, M^{1+\beta}\sqrt{\frac{1}{C_6}\bar{\psi}_2(T,L)\log\left(\frac{C_5 npL}{\iota}\right)}, M^{1+\beta}\sqrt{\frac{\log\left(4\bar{C}_3 p^2M^2/\iota\right)}{\bar{C}_4 n}} \right\}, \quad (18)$$

$\bar{C}_1 = \max\{C_1^X, C_1^Y\}$, $\bar{C}_2 = \min\{C_2^X, C_2^Y\}$, $\bar{C}_3 = \max\{C_3^X, C_3^Y\}$, $\bar{C}_4 = \min\{C_4^X, C_4^Y\}$, $\bar{C}_5 = \max\{C_5^X, C_6^Y\}$, $\bar{C}_6 = \min\{C_6^X, C_6^Y\}$. γ_k^X, γ_k^Y, $k = 1, 3$, and C_k^X, C_k^Y, $k = 1, \cdots, 6$ are constants that do not depend on n, p, and M. Then

$$\max\left\{|S^{X,M} - \Sigma^{X,M}|_\infty, |S^{Y,M} - \Sigma^{Y,M}|_\infty\right\} \leq \delta_n$$

holds with probability at least $1 - \iota$.

Proof See Appendix B.5. ■

The rate δ_n in Theorem 13 is composed of three terms. The first two terms correspond to the error incurred by measuring the curves at discrete locations and are approximation errors. The third term, which also appears in Theorem 12, is the sampling error.

We provide some insight into how $\tilde{\psi}_1$, $\tilde{\psi}_2$, $\tilde{\psi}_3$, and $\tilde{\psi}_4$ depend on T and L. Note that we choose an orthonormal basis. Then, as $T \to \infty$, we have

$$\begin{aligned}
\frac{1}{T}B_{ij}^\top B_{ij} &= \frac{1}{T}\sum_{k=1}^{T}\begin{bmatrix} b_1^2(t_{ijk}) & b_1(t_{ijk})b_2(t_{ijk}) & \cdots & b_1(t_{ijk})b_L(t_{ijk}) \\ \vdots & \vdots & \ddots & \vdots \\ b_L(t_{ijk})b_1(t_{ijk}) & b_L(t_{ijk})b_2(t_{ijk}) & \cdots & b_L^2(t_{ijk}) \end{bmatrix} \\
&\approx \begin{bmatrix} \|b_1\|^2 & \langle b_1, b_2\rangle & \cdots & \langle b_1, b_L\rangle \\ \vdots & \vdots & & \vdots \\ \langle b_L, b_1\rangle & \langle b_L, b_2\rangle & \cdots & \|b_L\|^2 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.
\end{aligned}$$

Thus, as T grows, we expect $\lambda_{\min}(B_{ij}^\top B_{ij}) \approx T$ for any $1 \leq j \leq p$ and $1 \leq i \leq n$. This implies that $\tilde{\psi}_1(T,L) \approx L/\sqrt{T}$ and $\tilde{\psi}_2(T,L) \approx \left(D_{1,b,L}^2 \tilde{\psi}_3(L) + \tilde{\psi}_4(L)\right) L^2/T^2$. Furthermore, $D_{1,b,L}^2 \asymp L^2$ when we use the Fourier basis.

To understand $\tilde{\psi}_3(L)$ and $\tilde{\psi}_4(L)$, note that $\lambda_{j0}^\perp = \mathbb{E}[\|g_{ij}^\perp\|^2] = \mathbb{E}_{g_{ij}}[\mathbb{E}_\epsilon[\|g_{ij}^\perp\|^2 \mid g_{ij}]]$. Under Assumption 4, $\lambda_{j0}^\perp \to 0$ as $L \to \infty$; however, the speed at which $\lambda_{j0}^\perp$ goes to zero

will depend on $\mathbb{H}$ and the choice of the basis functions. For example, for a fixed g_{ij}, by well-known approximation results (see, for example, Barron and Sheu (1991)), if g_{ij} has r-th continuous and square integrable derivatives, $\|g_{ij}^{\perp}\|^2 \approx 1/L^r$ for frequently used bases such as the Legendre polynomials, B-splines, and Fourier basis. Thus, roughly speaking, we should have $\tilde{\psi}_3(L) \approx 1/L^r$ when $\mathbb{H}$ is a Sobolev space of order r. When g_{ij} is an infinitely differentiable function and all derivatives can be uniformly bounded, then $\|g_{ij}^{\perp}\|^2 \approx \exp(-L)$ and thus $\tilde{\psi}_3(L) \approx \exp(-L)$. Similarly, we have $\tilde{\psi}_4(L) \approx 1/L^{r-1}$ if g_{ij} has r-th continuous and square integrable derivatives; and $\tilde{\psi}_4(L) \approx \exp(-L)$ if g_{ij} is an infinitely differentiable function and all derivatives can be uniformly bounded.

To roughly show how M, T, L and n may co-vary, we assume that p and s are fixed and all elements of $\mathbb{H}$ have r-th continuous and square integrable derivatives. Then FuDGE will recover the differential graph with high probability if $M \ll n^{1/(2+2\beta)}$, $\sqrt{T}/L \gg M^{1+\beta}$, $T \gg L^{2-r/2}$, and $L \gg M^{(1+\beta)/r}$.

As a reviewer pointed out, the noise term in (7) will create a nugget effect in the covariance, meaning that $\mathrm{Var}(h_{ijk}) = \mathrm{Var}(g_{ij}(t_{ijk})) + \sigma_0^2$. This nugget effect leads to bias in the estimated eigenvalues (variances of the scores). In our theorem, the nugget effect is reflected by σ_0 in $\tilde{\psi}_1$. When σ_0 is large, adding a regularization term when estimating eigenvalues can improve the estimation of FPCA scores and their covariance matrices (see Chapter 6 of Hsing and Eubank (2015)). However, adding a regularization term increases the number of tuning parameters that need to be chosen. An alternative approach to estimating the covariance matrix is through local polynomial regression (Zhang and Wang, 2016). Since the focus of the paper is on the estimation of differential functional graphical models, we do not explore ways to improve the estimation of FPCA scores. However, we recognize that there are alternative approaches that can perform better in some cases.

5. Joint Functional Graphical Lasso

In this section, we introduce two variants of a *Joint Functional Graphical Lasso (JFGL)* estimator, which we compare empirically with our proposed FuDGE procedure in Section 6.1. Danaher et al. (2014) proposed the *Joint Graphical Lasso (JGL)* to estimate multiple related Gaussian graphical models from different classes simultaneously. Given $Q \geq 2$ data sets, where the q-th data set consists of n_q independent random vectors drawn from $N(\mu_q, \Sigma_q)$, JGL simultaneously estimates $\{\Theta\} = \{\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(Q)}\}$, where $\Theta^{(q)} = \Sigma_q^{-1}$ is the precision matrix of the q-th data set. Specifically, JGL constructs an estimator $\{\hat{\Theta}\} = \{\hat{\Theta}^{(1)}, \hat{\Theta}^{(2)}, \ldots, \hat{\Theta}^{(Q)}\}$ by solving the penalized log-likelihood:

$$\{\hat{\Theta}\} = \underset{\{\Theta\}}{\arg\min} \left\{ -\sum_{q=1}^{Q} n_q \left(\log \det \Theta^{(q)} - \mathrm{trace}\left(S^{(q)} \Theta^{(q)} \right) \right) + P(\{\Theta\}) \right\}, \tag{19}$$

where $S^{(q)}$ is the sample covariance of the q-th data set and $P(\{\Theta\})$ is a penalty function. The *fused graphical lasso (FGL)* is obtained by setting

$$P(\{\Theta\}) = \lambda_1 \sum_{q=1}^{Q} \sum_{i \neq j} |\Theta_{ij}^{(q)}| + \lambda_2 \sum_{q<q'} \sum_{i \neq j} |\Theta_{ij}^{(q)} - \Theta_{ij}^{(q')}|,$$

while the *group graphical lasso (GGL)* is obtained by setting

$$P(\{\Theta\}) = \lambda_1 \sum_{q=1}^{Q} \sum_{i \neq j} |\Theta_{ij}^{(q)}| + \lambda_2 \sum_{i \neq j} \sqrt{\sum_{q=1}^{Q} \left(\Theta_{ij}^{(q)}\right)^2}.$$

The terms λ_1 and λ_2 are non-negative tuning parameters, while $\Theta_{ij}^{(q)}$ denotes the (i,j)-th entry of $\Theta^{(q)}$. For both penalties, the first term is the lasso penalty, which encourages sparsity for the off-diagonal entries of all precision matrices; however, FGL and GGL differ in the second term. For FGL, the second term encourages the off-diagonal entries of precision matrices among all classes to be similar, which means that it encourages not only a similar network structure but also similar edge values. For GGL, the second term is a group lasso penalty, which encourages the support of the precision matrices to be similar but allows the specific values to differ. See Tsai et al. (2021) for a recent survey of joint estimation procedures for joint Gaussian graphical models.

A similar approach can be used to estimate the precision matrix of the score vectors. Unlike the direct estimation procedure proposed in Section 3, we could first estimate $\hat{\Theta}^{X,M}$ and $\hat{\Theta}^{Y,M}$ using a joint graphical lasso objective, and then take the difference to estimate Δ.

In the functional graphical model setting, we are interested in the block sparsity, so we modify the entry-wise penalties to a block-wise penalty. Specifically, we propose to solve the objective function in (19), where $S^{(q)}$ and $\Theta^{(q)}$ denote the sample covariance and the estimated precision of the projection scores for the q-th group. Note that now $S^{(q)}$, $\Theta^{(q)}$ and $\hat{\Theta}^{(q)}$, $q = 1, \ldots, Q$ are all $pM \times pM$ matrices. Similarly to the GGL and FGL procedures, we define the *Grouped Functional Graphical Lasso (GFGL)* and *Fused Functional Graphical Lasso (FFGL)* penalties for functional graphs. Specifically, letting $\Theta_{jl}^{(q)}$ denote the (j,l)-th $M \times M$ block matrix, the GFGL penalty is

$$P(\{\Theta\}) = \lambda_1 \sum_{q=1}^{Q} \sum_{j \neq l} \|\Theta_{jl}^{(q)}\|_{\mathrm{F}} + \lambda_2 \sum_{j \neq l} \sqrt{\sum_{q=1}^{Q} \|\Theta_{jl}^{(q)}\|_{\mathrm{F}}^2}, \tag{20}$$

where λ_1 and λ_2 are non-negative tuning parameters. The FFGL penalty can be defined in two ways. The first way is to use the Frobenius norm for the second term:

$$P(\{\Theta\}) = \lambda_1 \sum_{q=1}^{Q} \sum_{j \neq l} \|\Theta_{jl}^{(q)}\|_{\mathrm{F}} + \lambda_2 \sum_{q<q'} \sum_{j,l} \|\Theta_{jl}^{(q)} - \Theta_{jl}^{(q')}\|_{\mathrm{F}}. \tag{21}$$

The second way is to keep the element-wise L_1 norm as in FGL:

$$P(\{\Theta\}) = \lambda_1 \sum_{q=1}^{Q} \sum_{j \neq l} \|\Theta_{jl}^{(q)}\|_{\mathrm{F}} + \lambda_2 \sum_{q<q'} \sum_{j,l} |\Theta_{jl}^{(q)} - \Theta_{jl}^{(q')}|_1, \tag{22}$$

where λ_1 and λ_2 are non-negative tuning parameters.

The Joint Functional Graphical Lasso accommodates an arbitrary Q. However, when estimating the functional differential graph, we set $Q = 2$. We will refer to (21) as FFGL and to (22) as FFGL2. The algorithms to solve GFGL, FFGL, and FFGL2 are given in the Appendix A.

6. Experiments

We examine the performance of FuDGE using both simulations and a real data set.[4]

6.1 Simulations

Given a graph G_X, we generate samples of X such that $X_{ij}(t) = b'(t)^\top \delta_{ij}^X$. The coefficients $\delta_i^X = ((\delta_{i1}^X)^\top, \ldots, (\delta_{ip}^X)^\top)^\top \in \mathbb{R}^{mp}$ are drawn from $N\left(0, (\Omega^X)^{-1}\right)$ where Ω_X is described below. In all cases, $b'(t)$ is an m-dimensional basis with disjoint support over $[0,1]$ such that for $k = 1, \ldots m$:

$$b'_k(t) = \begin{cases} \cos\left(10\pi\left(x - (2k-1)/10\right)\right) + 1 & \text{if } (k-1)/m \le x < k/m; \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

To generate noisy observations at discrete time points, we sample data

$$h_{ijk}^X = X_{ij}(t_k) + e_{ijk}, \quad e_{ijk} \sim N(0, 0.5^2),$$

for 200 evenly spaced time points $0 = t_1 \le \ldots \le t_{200} = 1$. $Y_{ij}(t)$ and h_{ijk}^Y are sampled in an analogous procedure. We use $m = 5$ for the experiments below, except for the simulation, where we explore the effect of m on empirical performance.

We consider three different simulation settings for the construction of G_X and G_Y. In each setting, we let $n_X = n_Y = 100$ and $p = 30, 60, 90, 120$, and replicate the procedure 30 times for each p and the model setting.

Model 1: This model is similar to the setting considered in Zhao et al. (2014a), but modified for the functional case. We generate the support of Ω^X according to a graph with $p(p-1)/10$ edges and a power law degree distribution with an expected power parameter of 2. Although the graph is sparse with only 20% of all possible edges present, the power-law structure mimics certain real-world graphs by creating hub nodes with a large degree (Newman, 2003). For each non-zero block, we set $\Omega_{jl}^X = \delta' I_5$, where δ' is sampled uniformly from $\pm[0.2, 0.5]$. To ensure positive definiteness, we further scale each off-diagonal block by $1/2, 1/3, 1/4, 1/5$ for $p = 30, 60, 90, 120$ respectively. Each diagonal element of Ω^X is set to 1 and the matrix is symmetrized by averaging it with its transpose. To get Ω^Y, we first select the top 2 hub nodes in G_X (i.e., the nodes with top 2 largest degree), and for each hub node we select the top (by magnitude) 20% of edges. For each selected edge, we set $\Omega_{jl}^Y = \Omega_{jl}^X + W$ where $W_{kk'} = 0$ for $|k - k'| \le 2$, and $W_{kk'} = c$ otherwise, where c is generated the same way as δ'. For all other blocks, $\Omega_{jl}^Y = \Omega_{jl}^X$.

Model 2: We first generate a tridiagonal block matrix Ω_X^* with $\Omega_{X,jj}^* = I_5$, $\Omega_{X,j,j+1}^* = \Omega_{X,j+1,j}^* = 0.6 I_5$, and $\Omega_{X,j,j+2}^* = \Omega_{X,j+2,j}^* = 0.4 I_5$ for $j = 1, \ldots, p$. All other blocks are set to 0. We form G_Y by adding four edges to G_X. Specifically, we first let $\Omega_{Y,jl}^* = \Omega_{X,jl}^*$ for all blocks, then for $j = 1, 2, 3, 4$, we set $\Omega_{Y,j,j+3}^* = \Omega_{Y,j+3,j}^* = W$, where $W_{kk'} = 0.1$ for all $1 \le k, k' \le M$. Finally, we set $\Omega^X = \Omega_X^* + \delta I$, $\Omega^Y = \Omega_Y^* + \delta I$, where $\delta = \max\left\{|\min(\lambda_{\min}(\Omega_X^*), 0)|, |\min(\lambda_{\min}(\Omega_Y^*), 0)|\right\} + 0.05$.

Model 3: We generate Ω_X^* according to an Erdös-Rényi graph. We first set $\Omega_{X,jj}^* = I_5$. With probability .8, we set $\Omega_{X,jl}^* = \Omega_{X,lj}^* = 0.1 I_5$, and set it to 0 otherwise. Thus, we

4. Code to replicate the simulations is available at `https://github.com/boxinz17/FuDGE`.

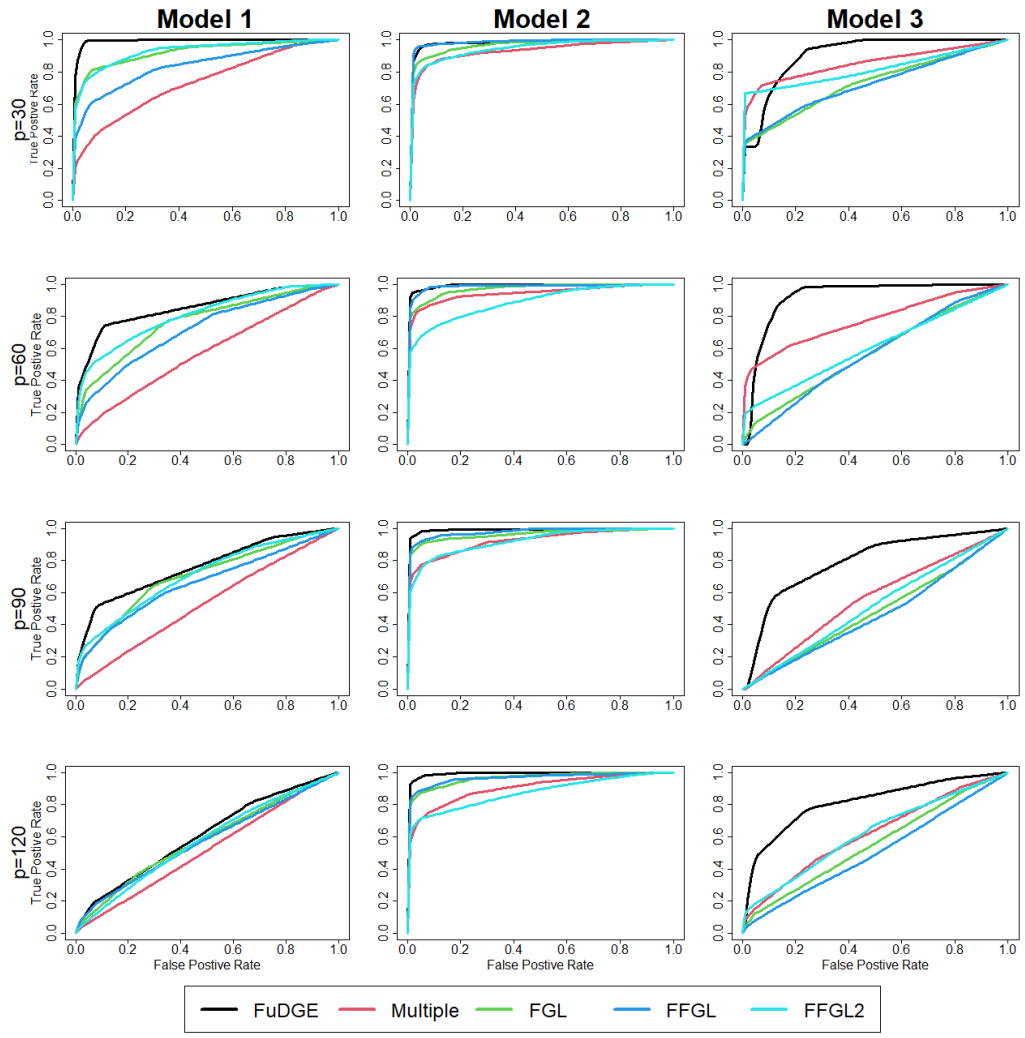


Figure 4: Average ROC curves across 30 simulations. Different columns correspond to different models, different rows correspond to different dimensions.

expect 80% of all possible edges to be present. Then we form G_Y by randomly adding s new edges to G_X, where $s = 3$ for $p = 30$, $s = 4$ for $p = 60$, $s = 5$ for $p = 90$, and $s = 6$ for $p = 120$. We set each corresponding block as $\Omega^*_{Y,jl} = W$, where $W_{kk'} = 0$ when $|k - k'| \leq 1$ and $W_{kk'} = c$ otherwise. We let $c = 2/5$ for $p = 30$, $c = 4/15$ for $p = 60$, $c = 1/5$ for $p = 90$, and $c = 4/25$ for $p = 120$. Finally, we set $\Omega^X = \Omega^*_X + \delta I$, $\Omega^Y = \Omega^*_Y + \delta I$, where $\delta = \max\{|\min(\lambda_{\min}(\Omega^*_X), 0)|, |\min(\lambda_{\min}(\Omega^*_Y), 0)|\} + 0.05$.

We compare FuDGE with four competing methods. The first competing method (denoted *multiple* in Figure 4) ignores the functional nature of the data. We select 15 equally spaced time points and at each time point implement a direct difference estimation procedure (Zhao et al., 2014a) to estimate the graph at that time point. Specifically, for each t, $X_i(t)$ and $Y_i(t)$ are simply p-dimensional random vectors, and we use their sample covari-

ances in (10) to obtain a $p \times p$ matrix $\hat{\Delta}$. This produces 15 differential graphs, and we use a majority vote to form a single differential graph. The ROC curve is obtained by changing the L_1 penalty, λ_n, used for all time points.

The other three competing methods all estimate two functional graphical models using either the Joint Graphical Lasso or the Functional Joint Graphical Lasso introduced in Section 5. For each method, we first estimate the sample covariances of the FPCA scores for X and Y. The second competing method (denoted *FGL*) ignores the block structure in precision matrices and applies the fused graphical lasso method directly. The third and fourth competing methods take into account the block structure and apply FFGL and FFGL2 defined in Section 5. To draw an ROC curve, we follow the same approach as in Zhao et al. (2014a). We first fix $\lambda_1 = 0.1$, which controls the overall sparsity in each graph; then we form an ROC curve by varying λ_2, which controls the similarity between two graphs.

For each setting and method, the ROC curve averaged across the 30 replications is shown in Figure 4. We see that FuDGE clearly has the best overall performance in recovering the support of the differential graph for all cases. We also note that explicit consideration of block structure in the joint graphical methods does not seem to make a substantial difference as the performance of FGL is comparable to FFGL and FFGL2.

The effect of the number of basis functions: To examine how the accuracy of the estimation is associated with the dimension of the functional data, we repeat the experiment under Model 1 with $p = 30$ and vary the number of basis functions used to generate the data in (23). In each case, the number of principal components selected by cross-validation is $M = 4$. In Figure 5, we see that as the gap between the true dimension m and the number of dimensions used M increases, the performance of FuDGE degrades slightly, but remains relatively robust. This is because the FPCA procedure is data adaptive and produces an eigenfunction basis that approximates the true functions well with a relatively small number of basis functions.

6.2 Neuroscience Application

We apply our method to electroencephalogram (EEG) data obtained from a study (Zhang et al., 1995; Ingber, 1997), which included 122 total subjects; 77 individuals with alcohol use disorder (AUD) and 45 in the control group. Specifically, the EEG data was measured by placing $p = 64$ electrodes at various locations on the subject's scalp and measuring voltage values over time. We follow the preprocessing procedure in Knyazev (2007) and Zhu et al. (2016), which filters the EEG signals at α frequency bands between 8 and 12.5 Hz.

Qiao et al. (2019) estimate separate functional graphs for each group, but we directly estimate the differential graph using FuDGE. We choose λ_n so that the estimated differential graph has approximately 1% of possible edges. The estimated edges of the differential graph are shown in Figure 6.

In this setting, an edge in the differential graph suggests that the communication pattern between two different regions of the brain may be affected by alcohol use disorder. However, the differential graph does not exactly indicate how the communication pattern has changed. For example, the edge between P4 and P6 suggests that AUD affects the communication pattern between those two regions; however, it could be that these two regions

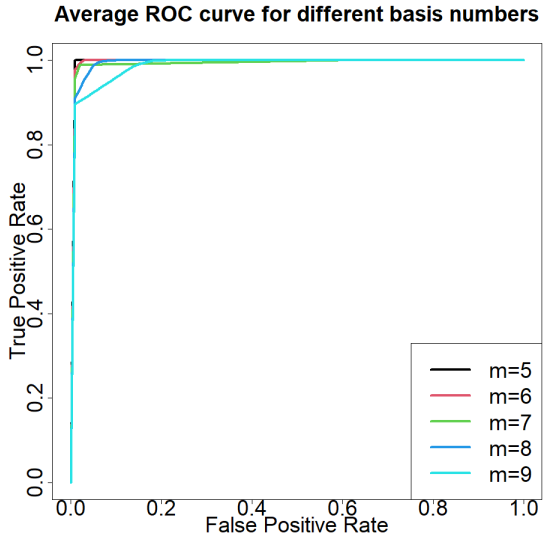


Figure 5: ROC curves for Model 1 with $p = 30$ and changing number of basis functions m. Each curve is drawn by averaging across 30 simulations. The number of eigenfunctions, M, selected by cross-validation is 4 in each replication.

are (conditionally) associated with the control group, but not with the AUD group or vice versa. It could also be that the two regions are (conditionally) associated in both groups, but the conditional covariance is different. However, many interesting observations can be gleaned from the results and may generate interesting hypotheses that could be investigated more thoroughly in an experimental setting.

We give two specific observations. First, edges are generally between nodes located in the same region—either the anterior region or the posterior region—and there is no edge that crosses between regions. This observation is consistent with the result in Qiao et al. (2019) where there are no connections between the anterior and posterior regions for both groups. We also note that electrode X, lying in the middle left region, has a high degree in the estimated differential graph. Although there is no direct connection between the anterior and posterior regions, this region may play a role in helping the two parts communicate and may be greatly affected by AUD. Similarly, P08 in the anterior region also has a high degree and is connected to other nodes in the anterior region, which may indicate that this region can be an information exchange center for the anterior regions, which, at the same time, may be heavily affected by AUD.

7. Discussion

We proposed a method to directly estimate the differential graph for functional graphical models. In certain settings, direct estimation allows the differential graph to be recovered consistently, even if each underlying graph cannot be consistently recovered. Experiments with simulated data also show that preserving the functional nature of the data rather

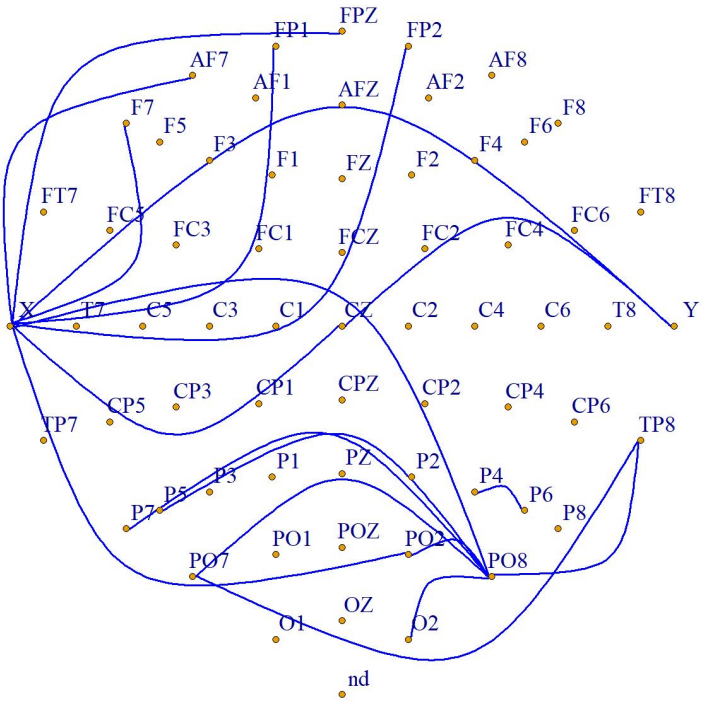


Figure 6: Estimated differential graph for EEG data. The anterior region is the top of the figure and the posterior region is the bottom of the figure.

than treating the data as multivariate scalars can also result in better estimation of the differential graph.

A key step in the procedure is to first represent the functions with an M-dimensional basis using FPCA. Definition 4 ensures that there exists some M large enough so that the signal, $\nu_1(M)$, is larger than the bias, $\nu_2(M)$, due to the use of a finite-dimensional representation. Intuitively, $\tau = \nu_1(M) - \nu_2(M)$ is tied to the eigenvalue decay rate; however, we defer the derivation of the explicit connection for future work.

We have provided a method for direct estimation of the differential graph, but the development of methods that allow for inference and hypothesis testing in functional differential graphs is a fruitful avenue for future work. In recent years, a number of studies have focused on inference in high-dimensional linear models (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014; Zhao et al., 2014b; Bradic and Kolar, 2017; Wang et al., 2021). Subsequently, these approaches were extended for statistical inference of low-dimensional parameters in graphical models (Ren et al., 2015; Wasserman et al., 2014; Janková and van de Geer, 2015, 2017; Barber and Kolar, 2018; Yu et al., 2016, 2020) and differential graphical models (Xia et al., 2015; Liu, 2017; Kim et al., 2021). Future work may extend these results to the functional graph setting. A promising approach would be to

extend the inference procedures developed for semi- and non-parametric models (see, e.g., Lu et al., 2020; Dai and Kolar, 2021).

Acknowledgments

We thank the associate editor and reviewers for their helpful feedback which has greatly improved the manuscript. This work is partially supported by the William S. Fishman Faculty Research Fund at the University of Chicago Booth School of Business. This work was completed in part with resources provided by the University of Chicago Research Computing Center.

A. Derivation of Optimization Algorithm

In this section, we derive the key steps for optimization algorithms.

A.1 Optimization Algorithm for FuDGE

We derive closed-form updates for the proximal method stated in (15). In particular, recall that for all $1 \leq j, l \leq p$, we have

$$\Delta_{jl}^{\text{new}} \;=\; \left[\left(\|A_{jl}^{\text{old}}\|_F - \lambda_n\eta\right)/\|A_{jl}^{\text{old}}\|_F\right]_+ \times A_{jl}^{\text{old}},$$

where $A^{\text{old}} = \Delta^{\text{old}} - \eta\nabla\mathcal{L}(\Delta^{\text{old}})$ and $x_+ = \max\{0, x\}$ represents the positive part of $x \in \mathbb{R}$.
Proof [Proof of (15)] Let $A^{\text{old}} = \Delta^{\text{old}} - \eta\nabla\mathcal{L}(\Delta^{\text{old}})$ and let f_{jl} denote the loss decomposed over each j, l block so that

$$f_{jl}(\Delta_{jl}) \;=\; \frac{1}{2\lambda_n\eta}\|\Delta_{jl} - A_{jl}^{\text{old}}\|_F^2 + \|\Delta_{jl}\|_F$$

and

$$\Delta_{jl}^{\text{new}} \;=\; \underset{\Delta_{jl}\in\mathbb{R}^{M\times M}}{\arg\min}\; f_{jl}(\Delta_{jl}).$$

The loss $f_{jl}(\Delta_{jl})$ is convex, so the first-order optimality condition implies that:

$$0 \in \partial f_{jl}\left(\Delta_{jl}^{\text{new}}\right), \tag{A.1}$$

where $\partial f_{jl}\left(\Delta_{jl}\right)$ is the subdifferential of f_{jl} at Δ_{jl}:

$$\partial f_{jl}(\Delta_{jl}) \;=\; \frac{1}{\lambda_n\eta}\left(\Delta_{jl} - A_{jl}^{\text{old}}\right) + Z_{jl},$$

where

$$Z_{jl} \;=\; \begin{cases} \frac{\Delta_{jl}}{\|\Delta_{jl}\|_F} & \text{if } \Delta_{jl} \neq 0 \\ \left\{Z_{jl} \in \mathbb{R}^{M\times M} \colon \|Z_{jl}\|_F \leq 1\right\} & \text{if } \Delta_{jl} = 0. \end{cases} \tag{A.2}$$

Claim 1 If $\|A_{jl}^{\text{old}}\|_F > \lambda_n\eta > 0$, then $\Delta_{jl}^{\text{new}} \neq 0$.

We verify this claim by proving the contrapositive. Suppose $\Delta_{jl}^{\text{new}} = 0$. Then by (A.1) and (A.2), there exists a $Z_{jl} \in \mathbb{R}^{M\times M}$ such that $\|Z_{jl}\|_F \leq 1$ and

$$0 = -\frac{1}{\lambda_n\eta}A_{jl}^{\text{old}} + Z_{jl}.$$

Thus, $\|A_{jl}^{\text{old}}\|_F = \|\lambda_n\eta \cdot Z_{jl}\|_F \leq \lambda_n\eta$, so that Claim 1 holds.

Combining Claim 1 with (A.1) and (A.2), for any j, l such that $\|A_{jl}^{\text{old}}\|_F > \lambda_n\eta$, we have

$$0 = \frac{1}{\lambda_n\eta}\left(\Delta_{jl}^{\text{new}} - A_{jl}^{\text{old}}\right) + \frac{\Delta_{jl}^{\text{new}}}{\|\Delta_{jl}^{\text{new}}\|_F},$$

which is solved by

$$\Delta_{jl}^{\text{new}} = \frac{\|A_{jl}^{\text{old}}\|_F - \lambda_n\eta}{\|A_{jl}^{\text{old}}\|_F}A_{jl}^{\text{old}}. \tag{A.3}$$

Claim 2 If $\|A_{jl}^{\text{old}}\|_F \leq \lambda_n \eta$, then $\Delta_{jl}^{\text{new}} = 0$.

Again, we verify the claim by proving the contrapositive. Suppose $\Delta_{jl}^{\text{new}} \neq 0$. Then the first-order optimality implies the updates in (A.3). However, taking the Frobenius norm on both sides of the equation gives $\|\Delta_{jl}^{\text{new}}\|_F = \|A_{jl}^{\text{old}}\|_F - \lambda_n \eta$, which implies that $\|A_{jl}^{\text{old}}\|_F - \lambda_n \eta \geq 0$.

Updates in (15) follow immediately by combining Claim 2 and (A.3). ■

A.2 Solving the Joint Functional Graphical Lasso

As in Danaher et al. (2014), we use the alternating directions method of multipliers (ADMM) algorithm to solve (19); see Boyd et al. (2011) for a detailed exposition of ADMM.

To solve (19), we first rewrite the problem as:

$$\max_{\{\Theta\},\{Z\}} \left\{ -\sum_{q=1}^{Q} n_q \left(\log\det\Theta^{(q)} - \text{trace}\left(S^{(q)}\Theta^{(q)} \right) \right) + P(\{Z\}) \right\},$$

subject to $\Theta^{(q)} \succ 0$ and $Z^{(q)} = \Theta^{(q)}$, where $\{Z\} = \{Z^{(1)}, Z^{(2)}, \ldots, Z^{(Q)}\}$. The scaled augmented Lagrangian (Boyd et al., 2011) is given by

$$L_\rho\left(\{\Theta\},\{Z\},\{U\}\right) = -\sum_{q=1}^{Q} n_q \left(\log\det\Theta^{(q)} - \text{trace}\left(S^{(q)}\Theta^{(q)} \right) \right) + P(\{Z\}) + \frac{\rho}{2}\sum_{q=1}^{Q} \|\Theta^{(q)} - Z^{(q)} + U^{(q)}\|_{\text{F}}^2, \quad \text{(A.4)}$$

where $\rho > 0$ is a tuning parameter and $\{U\} = \{U^{(1)}, U^{(2)}, \ldots, U^{(Q)}\}$ are dual variables. The ADMM algorithm will then solve (A.4) by iterating the following three steps. At the i-th iteration, they are as follows:

1. $\{\Theta_{(i)}\} \leftarrow \arg\min_{\{\Theta\}} L_\rho\left(\{\Theta\}, \{Z_{(i-1)}\}, \{U_{(i-1)}\}\right)$.
2. $\{Z_{(i)}\} \leftarrow \arg\min_{\{Z\}} L_\rho\left(\{\Theta_{(i)}\}, \{Z\}, \{U_{(i-1)}\}\right)$.
3. $\{U_{(i)}\} \leftarrow \{U_{(i-1)}\} + (\{\Theta_{(i)}\} - \{Z_{(i)}\})$.

We now give more details on the above three steps.

ADMM algorithm for solving the joint functional graphical lasso problem

Input: $\{S^{(q)}\}_{q=1}^{Q}$, $\{n_q\}_{q=1}^{Q}$, and the penalty term $P(\cdot)$.

Output: $\{\hat{\Theta}^{(q)}\}_{q=1}^{Q}$.

(a) Initialize the variables: $\Theta_{(0)}^{(q)} = I_{pM}$, $U_{(0)}^{(q)} = 0_{pM}$, and $Z_{(0)}^{(q)} = 0_{pM}$, $q = 1, \ldots, Q$.

(b) Select a scalar $\rho > 0$.

(c) For $i = 1, 2, 3, \ldots$ until convergence

(i) For $q = 1, \ldots, Q$, update $\Theta_{(i)}^{(q)}$ as the minimizer (with respect to $\Theta^{(q)}$) of

$$-n_q \left(\log\det\Theta^{(q)} - \text{trace}\left(S^{(q)}\Theta^{(q)} \right) \right) + \frac{\rho}{2}\|\Theta^{(q)} - Z_{(i-1)}^{(q)} + U_{(i-1)}^{(q)}\|_{\text{F}}^2$$

Let $VDV^\top$ denote the eigendecomposition of $S^{(q)} - \rho Z^{(q)}_{(i-1)}/n_q + \rho U^{(q)}_{(i-1)}/n_q$. The solution is given by $V\tilde{D}V^\top$, where $\tilde{D}$ is the diagonal matrix with the j-th diagonal element being

$$\frac{n_q}{2\rho}\left(-D_{jj} + \sqrt{D_{jj}^2 + 4\rho/n_q}\right),$$

where D_{jj} is the (j,j)-th entry of D.

(ii) Update $\{Z_{(i)}\}$ as minimizer (with respect to $\{Z\}$) of

$$\min_{\{Z\}} \frac{\rho}{2}\sum_{q=1}^{Q}\|Z^{(q)} - A^{(q)}\|_{\mathrm{F}}^2 + P(\{Z\}), \tag{A.5}$$

where $A^{(q)} = \Theta^{(q)}_{(i)} + U^{(q)}_{(i-1)}$, $q = 1, \ldots, Q$.

(iii) $U^{(q)}_{(i)} \leftarrow U^{(q)}_{(i-1)} + (\Theta^{(q)}_{(i)} - Z^{(q)}_{(i)})$, $q = 1, \ldots, Q$.

(d) Output $\hat{\Theta}^{(q)}$ as $\Theta^{(q)}_{(i)}$, $q = 1, \ldots, Q$, from the final round.

There are three things that are worth noting. **1.** The key step is to solve (A.5), which depends on the form of penalty term $P(\cdot)$; **2.** This algorithm is guaranteed to converge to the global optimum when $P(\cdot)$ is convex (Boyd et al., 2011); **3.** The positive-definiteness constraint on $\{\hat{\Theta}\}$ is naturally enforced by step (c) (i).

A.3 Solving (A.5) for different penalty functions

We provide solutions to (A.5) for three problems (GFGL, FFGL, FFGL2) defined by (20), (21), and (22).

A.3.1 Solution to (A.5) for GFGL

Let the solution for

$$\min_{\{Z\}} \frac{\rho}{2}\sum_{q=1}^{Q}\|Z^{(q)} - A^{(q)}\|_{\mathrm{F}}^2 + \lambda_1\sum_{q=1}^{Q}\sum_{j\neq l}\|Z^{(q)}_{jl}\|_{\mathrm{F}} + \lambda_2\sum_{j\neq l}\left(\sum_{q=1}^{Q}\|Z^{(q)}_{jl}\|_{\mathrm{F}}^2\right)^{1/2}$$

be denoted as $\{\hat{Z}\} = \{\hat{Z}^{(1)}, \hat{Z}^{(2)}, \ldots, \hat{Z}^{(Q)}\}$. Let $Z^{(q)}_{jl}$, $\hat{Z}^{(q)}_{jl}$ be the (j,l)-th $M \times M$ block of $Z^{(q)}$ and $\hat{Z}^{(q)}$, $q = 1, \ldots, Q$. Then, for $j = 1, \ldots, p$, we have

$$\hat{Z}^{(q)}_{jj} = A^{(q)}_{jj}, \qquad q = 1, \ldots, Q, \tag{A.6}$$

and, for $j \neq l$, we have

$$\hat{Z}^{(q)}_{jl} = \left(\frac{\|A^{(q)}_{jl}\|_{\mathrm{F}} - \lambda_1/\rho}{\|A^{(q)}_{jl}\|_{\mathrm{F}}}\right)_+ \left(1 - \frac{\lambda_2}{\rho\sqrt{\sum_{q=1}^{Q}\left(\|A^{(q)}_{jl}\|_{\mathrm{F}} - \lambda_1/\rho\right)_+^2}}\right)_+ A^{(q)}_{jl}, \tag{A.7}$$

where $q = 1, \ldots, Q$. Details of the proof of (A.6) and (A.7) are given in Appendix A.4.

A.3.2 Solution to (A.5) for FFGL

For FFGL, there is no simple closed-form solution. When $Q = 2$, (A.5) becomes

$$\min_{\{Z\}} \frac{\rho}{2}\sum_{q=1}^{2}\|Z^{(q)} - A^{(q)}\|_{\mathrm{F}}^2 + \lambda_1\left(\sum_{q=1}^{2}\sum_{j\neq l}\|Z_{jl}^{(q)}\|_{\mathrm{F}}\right) + \lambda_2\sum_{j,l}\|Z_{jl}^{(1)} - Z_{jl}^{(2)}\|_{\mathrm{F}}.$$

For each $1 \leq j, l \leq p$, we compute $\hat{Z}_{jl}^{(1)}$, $\hat{Z}_{jl}^{(2)}$ by solving

$$\min_{\{Z_{jl}^{(1)}, Z_{jl}^{(2)}\}} \frac{1}{2}\sum_{q=1}^{2}\|Z_{jl}^{(q)} - A_{jl}^{(q)}\|_{\mathrm{F}}^2 + \frac{\lambda_1}{\rho}\mathbb{1}_{j\neq l}\sum_{q=1}^{2}\|Z_{jl}^{(q)}\|_{\mathrm{F}} + \frac{\lambda_2}{\rho}\|Z_{jl}^{(1)} - Z_{jl}^{(2)}\|_{\mathrm{F}}, \tag{A.8}$$

where $\mathbb{1}_{j\neq l} = 1$ when $j \neq l$ and 0 otherwise.

When $j = l$, by Lemma 16, we have the following closed-form updates for $\{\hat{Z}_{jj}^{(1)}, \hat{Z}_{jj}^{(2)}\}$, $j = 1, \ldots, p$. If $\|A_{jj}^{(1)} - A_{jj}^{(2)}\|_{\mathrm{F}} \leq 2\lambda_2/\rho$, then

$$\hat{Z}_{jj}^{(1)} = \hat{Z}_{jj}^{(2)} = \frac{1}{2}\left(A_{jj}^{(1)} + A_{jj}^{(2)}\right).$$

If $\|A_{jj}^{(1)} - A_{jj}^{(2)}\|_{\mathrm{F}} > 2\lambda_2/\rho$, then

$$\hat{Z}_{jj}^{(1)} = A_{jj}^{(1)} - \frac{\lambda_2/\rho}{\|A_{jj}^{(1)} - A_{jj}^{(2)}\|_{\mathrm{F}}}\left(A_{jj}^{(1)} - A_{jj}^{(2)}\right),$$
$$\hat{Z}_{jj}^{(2)} = A_{jj}^{(2)} + \frac{\lambda_2/\rho}{\|A_{jj}^{(1)} - A_{jj}^{(2)}\|_{\mathrm{F}}}\left(A_{jj}^{(1)} - A_{jj}^{(2)}\right).$$

For $j \neq l$, we get $\{\hat{Z}_{jl}^{(1)}, \hat{Z}_{jl}^{(2)}\}$ using the ADMM algorithm again. We construct the scaled augmented Lagrangian as:

$$L'_{\rho'}(\{W\}, \{R\}, \{V\}) = \frac{1}{2}\sum_{q=1}^{2}\|W^{(q)} - B^{(q)}\|_{\mathrm{F}} + \frac{\lambda_1}{\rho}\sum_{q=1}^{2}\|W^{(q)}\|_{\mathrm{F}}$$
$$+ \frac{\lambda_2}{\rho}\|R^{(1)} - R^{(2)}\|_{\mathrm{F}} + \frac{\rho'}{2}\sum_{q=1}^{2}\|W^{(q)} - R^{(q)} + V^{(q)}\|_{\mathrm{F}}^2,$$

where $\rho' > 0$ is a tuning parameter, $B^{(q)} = A_{jl}^{(q)}$, $q = 1, 2$, and $W^{(q)}, R^{(q)}, V^{(q)} \in \mathbb{R}^{M\times M}$, $q = 1, 2$. $\{W\} = \{W^{(1)}, W^{(2)}\}$, $\{R\} = \{R^{(1)}, R^{(2)}\}$, and $\{V\} = \{V^{(1)}, V^{(2)}\}$. The detailed ADMM algorithm is described as below:

ADMM algorithm for solving (A.8) **for** $j \neq l$

Input: $A_{jl}^{(q)}$, $q = 1, 2$; $\lambda_1, \lambda_2 \geq 0$.

Output: $\{\hat{Z}_{jl}^{(1)}, \hat{Z}_{jl}^{(2)}\}$.

(a) Initialize the variables: $W_{(0)}^{(q)} = I_M$, $R_{(0)}^{(q)} = 0_M$, $V_{(0)}^{(q)} = 0_M$, $B^{(q)} = A_{jl}^{(q)}$, $q = 1, 2$.

(b) Select a scalar $\rho' > 0$.

(c) For $i = 1, 2, 3, \ldots$ until convergence

(i) $\{W_{(i)}\} \leftarrow \arg\min_{\{W\}} L'_{\rho'}\left(\{W\}, \{R_{(i-1)}\}, \{V_{(i-1)}\}\right)$.

This is equivalent to

$$\{W_{(i)}\} \leftarrow \underset{\{W\}}{\arg\min} \frac{1}{2}\sum_{q=1}^{2}\|W^{(q)} - C^{(q)}\|_{\mathrm{F}}^2 + \frac{\lambda_1}{\rho(1+\rho')}\sum_{q=1}^{2}\|W^{(q)}\|_{\mathrm{F}},$$

where

$$C^{(q)} = \frac{1}{1+\rho'}\left[B^{(q)} + \rho'\left(R_{(i-1)}^{(q)} - V_{(i-1)}^{(q)}\right)\right].$$

Similar to (14), we have

$$W_{(i)}^{(q)} \leftarrow \left(\frac{\|C^{(q)}\|_{\mathrm{F}} - \lambda_1/(\rho(1+\rho'))}{\|C^{(q)}\|_{\mathrm{F}}}\right)_{+} \cdot C^{(q)}, \qquad q = 1, 2.$$

(ii) $\{R_{(i)}\} \leftarrow \arg\min_{\{R\}} L'_{\rho'}\left(\{W_{(i)}\}, \{R\}, \{V_{(i-1)}\}\right)$.
This is equivalent to

$$\{R_{(i)}\} \leftarrow \underset{\{R\}}{\arg\min} \frac{1}{2}\sum_{q=1}^{2}\|R^{(q)} - D^{(q)}\|_{\mathrm{F}}^2 + \frac{\lambda_2}{\rho\rho'}\|R^{(1)} - R^{(2)}\|_{\mathrm{F}},$$

where $D^{(q)} = W_{(i)}^{(q)} + V_{(i-1)}^{(q)}$. By Lemma 16, if $\|D^{(1)} - D^{(2)}\|_{\mathrm{F}} \le 2\lambda_2/(\rho\rho')$, then

$$R_{(i)}^{(1)} = R_{(i)}^{(2)} \leftarrow \frac{1}{2}\left(D^{(1)} + D^{(2)}\right),$$

and if $\|D^{(1)} - D^{(2)}\|_{\mathrm{F}} > 2\lambda_2/(\rho\rho')$, then

$$R^{(1)} \leftarrow D^{(1)} - \frac{\lambda_2/(\rho\rho')}{\|D^{(1)} - D^{(2)}\|_{\mathrm{F}}}\left(D^{(1)} - D^{(2)}\right),$$
$$R^{(2)} \leftarrow D^{(2)} + \frac{\lambda_2/(\rho\rho')}{\|D^{(1)} - D^{(2)}\|_{\mathrm{F}}}\left(D^{(1)} - D^{(2)}\right).$$

(iii) $V_{(i)}^{(q)} \leftarrow V_{(i-1)}^{(q)} + W_{(i)}^{(q)} - R_{(i)}^{(q)}$, $q = 1, 2$.

(d) Output $\{\hat{Z}_{jl}^{(1)}, \hat{Z}_{jl}^{(2)}\}$ as $\{W_{(i)}^{(1)}, W_{(i)}^{(2)}\}$ from the final round.

A.3.3 Solution to (A.5) for FFGL2

For FFGL2, there is also no closed-form solution. Similarly to Section A.3.2, we compute a closed-form solution for $\{\hat{Z}_{jj}^{(1)}, \hat{Z}_{jj}^{(2)}\}$, $j = 1, \ldots, p$, and use an ADMM algorithm to compute $\{\hat{Z}_{jl}^{(1)}, \hat{Z}_{jl}^{(2)}\}$, $1 \le j \ne l \le p$.

For any $1 \le j, l \le p$, we solve:

$$\min_{\{Z_{jl}^{(1)}, Z_{jl}^{(2)}\}} \frac{1}{2}\sum_{q=1}^{2}\|Z_{jl}^{(q)} - A_{jl}^{(q)}\|_{\mathrm{F}}^2 + \frac{\lambda_1}{\rho}\mathbb{1}_{j\ne l}\sum_{q=1}^{2}\|Z_{jl}^{(q)}\|_{\mathrm{F}} + \frac{\lambda_2}{\rho}\sum_{1\le a,b\le M}|Z_{jl,ab}^{(1)} - Z_{jl,ab}^{(2)}|, \quad \text{(A.9)}$$

where $\mathbb{1}_{j\ne l} = 1$ when $j \ne l$ and 0 otherwise.

By Lemma 16, when $j = l$ we have

$$\begin{aligned}&\left(\hat{Z}^{(1)}_{jj,ab}, \hat{Z}^{(2)}_{jj,ab}\right)\\ &= \begin{cases} \left(A^{(1)}_{jl,ab} - \lambda_2/\rho, A^{(2)}_{jl,ab} + \lambda_2/\rho\right) & \text{if } A^{(1)}_{jl,ab} > A^{(2)}_{jl,ab} + 2\lambda_2/\rho \\ \left(A^{(1)}_{jl,ab} + \lambda_2/\rho, A^{(2)}_{jl,ab} - \lambda_2/\rho\right) & \text{if } A^{(1)}_{jl,ab} < A^{(2)}_{jl,ab} - 2\lambda_2/\rho \\ \left(\left(A^{(1)}_{jl,ab} + A^{(2)}_{jl,ab}\right)/2, \left(A^{(1)}_{jl,ab} + A^{(2)}_{jl,ab}\right)/2\right) & \text{if } \left|A^{(1)}_{jl,ab} - A^{(2)}_{jl,ab}\right| \le 2\lambda_2/\rho \end{cases},\end{aligned}$$

where the subscript denotes the (a, b)-th entry, $1 \le a, b \le M$ and $j = 1, \ldots, p$.

For $j \neq l$, we get $\{\hat{Z}^{(1)}_{jl}, \hat{Z}^{(2)}_{jl}\}$, $1 \le j \neq l \le p$ using an ADMM algorithm. Let $B^{(q)} = A^{(q)}_{jl}$, $q = 1, 2$. We first construct the scaled augmented Lagrangian:

$$\begin{aligned}L'_{\rho'}(\{W\}, \{R\}, \{V\}) = \frac{1}{2}\sum_{q=1}^{2}\|W^{(q)} - B^{(q)}\|_{\mathrm{F}} + \frac{\lambda_1}{\rho}\sum_{q=1}^{2}\|W^{(q)}\|_{\mathrm{F}}\\ + \frac{\lambda_2}{\rho}\sum_{a,b}|R^{(1)}_{a,b} - R^{(2)}_{a,b}| + \frac{\rho'}{2}\sum_{q=1}^{2}\|W^{(q)} - R^{(q)} + V^{(q)}\|^2_{\mathrm{F}},\end{aligned}$$

where $\rho' > 0$ is a tuning parameter, $W^q, R^{(q)}, V^{(q)} \in \mathbb{R}^{M\times M}$, $q = 1, 2$, $\{W\} = \{W^{(1)}, W^{(2)}\}$, $\{R\} = \{R^{(1)}, R^{(2)}\}$, and $\{V\} = \{V^{(1)}, V^{(2)}\}$. The detailed ADMM algorithm is described below.

ADMM algorithm for solving (A.9) **for** $j \neq l$

Input: $A^{(q)}_{jl}$, $q = 1, 2$; $\lambda_1, \lambda_2 \ge 0$.

Output: $\{\hat{Z}^{(1)}_{jl}, \hat{Z}^{(2)}_{jl}\}$.

(a) Initialize the variables: $W^{(q)}_{(0)} = I_M$, $R^{(q)}_{(0)} = 0_M$, $V^{(q)}_{(0)} = 0_M$, $B^{(q)} = A^{(q)}_{jl}$, $q = 1, 2$.

(b) Select a scalar $\rho' > 0$.

(c) For $i = 1, 2, 3, \ldots$ until convergence

(i) $\{W_{(i)}\} \leftarrow \arg\min_{\{W\}} .L'_{\rho'}\left(\{W\}, \{R_{(i-1)}\}, \{V_{(i-1)}\}\right)$
This is equivalent to

$$\{W_{(i)}\} \leftarrow \arg\min_{\{W\}} \frac{1}{2}\sum_{q=1}^{2}\|W^{(q)} - C^{(q)}\|^2_{\mathrm{F}} + \frac{\lambda_1}{\rho(1+\rho')}\sum_{q=1}^{2}\|W^{(q)}\|_{\mathrm{F}},$$

where

$$C^{(q)} = \frac{1}{1+\rho'}\left[B^{(q)} + \rho'\left(R^{(q)}_{(i-1)} - V^{(q)}_{(i-1)}\right)\right].$$

Similarly to (14), we have

$$W^{(q)}_{(i)} \leftarrow \left(\frac{\|C^{(q)}\|_{\mathrm{F}} - \lambda_1/(\rho(1+\rho'))}{\|C^{(q)}\|_{\mathrm{F}}}\right)_+ \cdot C^{(q)}, \qquad q = 1, 2.$$

(ii) $\{R_{(i)}\} \leftarrow \arg\min_{\{R\}} L'_{\rho'}\left(\{W_{(i)}\}, \{R\}, \{V_{(i-1)}\}\right)$

This is equivalent to

$$\{R_{(i)}\} \leftarrow \underset{\{R\}}{\arg\min} \frac{1}{2}\sum_{q=1}^{2}\|R^{(q)} - D^{(q)}\|_{\mathrm{F}}^2 + \frac{\lambda_2}{\rho\rho'}\sum_{a,b}\left|R_{ab}^{(1)} - R_{ab}^{(2)}\right|,$$

where $D^{(q)} = W_{(i)}^{(q)} + V_{(i-1)}^{(q)}$. Then, by Lemma 16, we have

$$\begin{aligned}&\left(R_{(i),ab}^{(1)}, R_{(i),ab}^{(2)}\right)\\ &= \begin{cases} \left(D_{ab}^{(1)} - \lambda_2/(\rho\rho'), D_{ab}^{(2)} + \lambda_2/(\rho\rho')\right) & \text{if } D_{ab}^{(1)} > D_{ab}^{(2)} + 2\lambda_2/(\rho\rho') \\ \left(D_{ab}^{(1)} + \lambda_2/(\rho\rho'), D_{ab}^{(2)} - \lambda_2/(\rho\rho')\right) & \text{if } D_{ab}^{(1)} < D_{ab}^{(2)} - 2\lambda_2/(\rho\rho') \\ \left(\left(D_{ab}^{(1)} + D_{ab}^{(2)}\right)/2, \left(D_{ab}^{(1)} + D_{ab}^{(2)}\right)/2\right) & \text{if } \left|D_{ab}^{(1)} - D_{ab}^{(1)}\right| \le 2\lambda_2/(\rho\rho') \end{cases},\end{aligned}$$

where the subscript denotes the (a,b)-th entry, $1 \le a, b \le M$ and $1 \le j, l \le p$.

(iii) $V_{(i)}^{(q)} \leftarrow V_{(i-1)}^{(q)} + W_{(i)}^{(q)} - R_{(i)}^{(q)}$, $q = 1, 2$.

(d) Output $\{\hat{Z}_{jl}^{(1)}, \hat{Z}_{jl}^{(2)}\}$ as $\{W_{(i)}^{(1)}, W_{(i)}^{(2)}\}$ from the final round.

A.4 Derivation of (A.6) and (A.7)

Note that for any $1 \le j, l \le p$, we can obtain $\hat{Z}_{jl}^{(1)}, \hat{Z}_{jl}^{(2)}, \ldots, \hat{Z}_{jl}^{(Q)}$ by solving

$$\underset{Z_{jl}^{(1)}, Z_{jl}^{(2)}, \ldots, Z_{jl}^{(Q)}}{\arg\min} \frac{\rho}{2}\sum_{q=1}^{Q}\|Z_{jl}^{(q)} - A_{jl}^{(q)}\|_{\mathrm{F}}^2 + \lambda_1 \mathbb{1}_{j\neq l}\sum_{q=1}^{Q}\|Z_{jl}^{(q)}\|_{\mathrm{F}} + \lambda_2 \mathbb{1}_{j\neq l}\left(\sum_{q=1}^{Q}\|Z_{jl}^{(q)}\|_{\mathrm{F}}^2\right)^{1/2}, \quad \text{(A.10)}$$

where $\mathbb{1}_{j\neq l} = 1$ when $j \neq l$ and 0 otherwise. By (A.10), we have $\hat{Z}_{jj}^{(q)} = A_{jj}^{(q)}$ for any $j = 1, \ldots, p$ and $q = 1, \ldots, Q$, which is (A.6). We then prove (A.7). Denote the objective function in (A.10) by $\tilde{L}_{jl}$. Then, for $j \neq l$, the subdifferential of $\tilde{L}_{jl}$ with respect to $Z_{jl}^{(q)}$ is

$$\partial_{Z_{jl}^{(q)}}\tilde{L}_{jl} = \rho(Z_{jl}^{(q)} - A_{jl}^{(q)}) + \lambda_1 G_{jl}^{(q)} + \lambda_2 D_{jl}^{(q)},$$

where

$$G_{jl}^{(q)} = \begin{cases} \frac{Z_{jl}^{(q)}}{\|Z_{jl}^{(q)}\|_{\mathrm{F}}} & \text{when } Z_{jl}^{(q)} \neq 0 \\ \{G_{jl}^{(q)} \in \mathbb{R}^{M\times M} : \|G_{jl}^{(q)}\|_{\mathrm{F}} \le 1\} & \text{otherwise} \end{cases},$$

and

$$D_{jl}^{(q)} = \begin{cases} \frac{Z_{jl}^{(q)}}{\left(\sum_{q=1}^{Q}\|Z_{jl}^{(q)}\|_{\mathrm{F}}^2\right)^{1/2}} & \text{when } \sum_{q=1}^{Q}\|Z_{jl}^{(q)}\|_{\mathrm{F}}^2 > 0 \\ \{D_{jl}^{(q)} \in \mathbb{R}^{M\times M} : \sum_{q=1}^{Q}\|D_{jl}^{(q)}\|_{\mathrm{F}}^2 \le 1\} & \text{otherwise} \end{cases}.$$

To obtain the optimum, we need

$$0 \in \partial_{Z_{jl}^{(q)}}\tilde{L}_{jl}(\hat{Z}_{jl}^{(q)})$$

for all $q = 1, \ldots, Q$. Now we split our discussion into two cases.

(a) Suppose $\sum_{q=1}^{Q} \|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}^2 = 0$ or equivalently $\hat{Z}_{jl}^{(q)} = 0$ for all $q = 1, \ldots, Q$.

In this case, there exist $G_{jl}^{(q)}$, where $\|G_{jl}^{(q)}\|_{\mathrm{F}} \leq 1$, $q = 1, \ldots, Q$; and also $D_{jl}^{(q)}$, where $\sum_{q=1}^{Q} \|D_{jl}^{(q)}\|_{\mathrm{F}}^2 \leq 1$, such that

$$0 = -\rho \cdot A_{jl}^{(q)} + \lambda_1 G_{jl}^{(q)} + \lambda_2 D_{jl}^{(q)}.$$

This implies that

$$D_{jl}^{(q)} = \frac{\rho}{\lambda_2} \left(A_{jl}^{(q)} - \frac{\lambda_1}{\rho} G_{jl}^{(q)} \right).$$

Thus, we have

$$\begin{aligned} \|D_{jl}^{(q)}\|_{\mathrm{F}} = \frac{\rho}{\lambda_2} \left\| A_{jl}^{(q)} - \frac{\lambda_1}{\rho} G_{jl}^{(q)} \right\|_{\mathrm{F}} &\geq \frac{\rho}{\lambda_2} \left(\|A_{jl}^{(q)}\|_{\mathrm{F}} - \frac{\lambda_1}{\rho} \|G_{jl}^{(q)}\|_{\mathrm{F}} \right)_+ \\ &\geq \frac{\rho}{\lambda_2} \left(\|A_{jl}^{(q)}\|_{\mathrm{F}} - \frac{\lambda_1}{\rho} \right)_+, \end{aligned}$$

which implies that

$$\frac{\rho^2}{\lambda_2^2} \sum_{q=1}^{Q} \left(\|A_{jl}^{(q)}\|_{\mathrm{F}} - \frac{\lambda_1}{\rho} \right)_+^2 \leq \sum_{q=1}^{Q} \|D_{jl}^{(q)}\|_{\mathrm{F}}^2 \leq 1.$$

Therefore,

$$\sqrt{\sum_{q=1}^{Q} \left(\|A_{jl}^{(q)}\|_{\mathrm{F}} - \lambda_1/\rho \right)_+^2} \leq \lambda_2/\rho. \tag{A.11}$$

(b) Suppose $\sum_{q=1}^{Q} \|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}^2 > 0$.

For those q's such that $\hat{Z}_{jl}^{(q)} = 0$, there exists $G_{jl}^{(q)}$, where $\|G_{jl}^{(q)}\|_{\mathrm{F}} = 1$, such that

$$0 = -\rho A_{jl}^{(q)} + \lambda_1 G_{jl}^{(q)}.$$

Thus, we have

$$\|A_{jl}^{(q)}\|_{\mathrm{F}} = \frac{\lambda_1}{\rho} \|G_{jl}^{(q)}\|_{\mathrm{F}} \leq \frac{\lambda_1}{\rho},$$

which implies that

$$\left(\|A_{jl}^{(q)}\|_{\mathrm{F}} - \lambda_1/\rho \right)_+ = 0. \tag{A.12}$$

On the other hand, for those q's such that $\hat{Z}_{jl}^{(q)} \neq 0$, we have

$$0 = \rho \left(\hat{Z}_{jl}^{(q)} - A_{jl}^{(q)} \right) + \lambda_1 \frac{\hat{Z}_{jl}^{(q)}}{\|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}} + \lambda_2 \frac{\hat{Z}_{jl}^{(q)}}{\left(\sum_{q=1}^{Q} \|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}^2 \right)^{1/2}},$$

which implies that

$$A_{jl}^{(q)} = \hat{Z}_{jl}^{(q)} \left(1 + \frac{\lambda_1}{\rho \|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}} + \frac{\lambda_2}{\rho \left(\sum_{q=1}^{Q} \|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}^2 \right)^{1/2}} \right), \tag{A.13}$$

and

$$\|A_{jl}^{(q)}\|_{\mathrm{F}} = \|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}} + \lambda_1/\rho + (\lambda_2/\rho) \cdot \frac{\|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}}{\left(\sum_{q=1}^{Q} \|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}^2 \right)^{1/2}}. \tag{A.14}$$

By (A.14), we have

$$\left(\|A_{jl}^{(q)}\|_{\mathrm{F}} - \lambda_1/\rho \right)_+ > \frac{\lambda_2}{\rho} \cdot \frac{\|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}}{\sqrt{\sum_{q=1}^{Q} \|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}^2}} > 0. \tag{A.15}$$

By (A.12) and (A.15), we have

$$\begin{aligned} \sum_{q=}^{Q} \left(\|A_{jl}^{(q)}\|_{\mathrm{F}} - \lambda_1/\rho \right)_+^2 &= \sum_{q:\|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}} \neq 0} \left(\|A_{jl}^{(q)}\|_{\mathrm{F}} - \lambda_1/\rho \right)_+^2 \\ &> \frac{\lambda_2^2}{\rho^2} \sum_{q:\|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}} \neq 0} \frac{\|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}^2}{\sum_{q=1}^{Q} \|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}^2} \\ &> \lambda_2^2/\rho^2. \end{aligned} \tag{A.16}$$

Now we make the following claims.

Claim 1. $\sum_{q=1}^{Q} \|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}^2 = 0 \Longleftrightarrow \sqrt{\sum_{q=}^{Q} \left(\|A_{jl}^{(q)}\|_{\mathrm{F}} - \lambda_1/\rho \right)_+^2} \leq \lambda_2/\rho.$
This claim is easily shown by (A.11) and (A.16).

Claim 2. When $\sum_{q=1}^{Q} \|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}^2 > 0$, we have $\|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}} = 0 \Longleftrightarrow \|A_{jl}^{(q)}\|_{\mathrm{F}} \leq \lambda_1/\rho.$
This claim is easily shown by (A.12) and (A.15).

Claim 3. When $\|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}} \neq 0$, then we have

$$\hat{Z}_{jl}^{(q)} = \left(\frac{\|A_{jl}^{(q)}\|_{\mathrm{F}} - \lambda_1/\rho}{\|A_{jl}^{(q)}\|_{\mathrm{F}}} \right) \left(1 - \frac{\lambda_2}{\rho \sqrt{\sum_{q=}^{Q} \left(\|A_{jl}^{(q)}\|_{\mathrm{F}} - \lambda_1/\rho \right)_+^2}} \right) A_{jl}^{(q)}.$$

To prove this claim, note that by Claim 2 and (A.14), we have

$$\left(\|A_{jl}^{(q)}\|_{\mathrm{F}} - \lambda_1/\rho \right)_+ = \|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}} \left(1 + \frac{\lambda_2}{\rho \left(\sum_{q=1}^{Q} \|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}^2 \right)^{1/2}} \right), \qquad q = 1, \ldots, Q.$$

Thus,

$$\sqrt{\sum_{q=1}^{Q}\left(\|A_{jl}^{(q)}\|_{\mathrm{F}}-\lambda_1/\rho\right)_+^2}=\sqrt{\sum_{q=1}^{Q}\|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}^2}+\lambda_2/\rho,$$

which implies that

$$\sqrt{\sum_{q=1}^{Q}\|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}^2}=\sqrt{\sum_{q=1}^{Q}\left(\|A_{jl}^{(q)}\|_{\mathrm{F}}-\lambda_1/\rho\right)_+^2}-\lambda_2/\rho.$$

Thus, by (A.14), we have

$$\begin{aligned}\|\hat{Z}_{jl}^{(q)}\|_{\mathrm{F}}&=\frac{\|A_{jl}^{(q)}\|_{\mathrm{F}}-\lambda_1/\rho}{1+\frac{\lambda_2/\rho}{\sqrt{\sum_{q'=1}^{Q}\left(\|A_{jl}^{(q')}\|_{\mathrm{F}}-\lambda_1/\rho\right)_+^2}-\lambda_2/\rho}}\\&=\left(1-\frac{\lambda_2}{\rho\sqrt{\sum_{q'=1}^{Q}\left(\|A_{jl}^{(q')}\|_{\mathrm{F}}-\lambda_1/\rho\right)_+^2}}\right)\left(\|A_{jl}^{(q)}\|_{\mathrm{F}}-\lambda_1/\rho\right).\end{aligned}$$

Claim 3 follows by combining the above display with (A.13).

Finally, combining Claims 1-3, we obtain (A.7).

B. Main Technical Proofs

We give proofs of the results given in the main text.

B.1 Proof of Lemma 2

We only need to prove that when we use two sets of orthonormal function basis $e^M(t) = \{e_j^M(t)\}_{j=1}^p$ and $\tilde{e}^M(t) = \{\tilde{e}_j^M(t)\}_{j=1}^p$ to expand the same subspace $\mathbb{V}_{[p]}^M$, the definition of E_Δ^π will not change. Since both $e_j^M(t) = (e_{j1}^M(t), e_{j2}^M(t), \ldots, e_{jM}^M(t))^\top$ and $\tilde{e}_j^M(t) = (\tilde{e}_{j1}^M(t), \tilde{e}_{j2}^M(t), \ldots, \tilde{e}_{jM}^M(t))^\top$ are orthonormal function basis of $\mathbb{V}_j^M$, there must exist an orthonormal matrix $U_j \in \mathbb{R}^{M\times M}$ satisfying $U_j^\top U_j = U_j U_j^\top = I_M$, such that $\tilde{e}_j^M(t) = U_j e_j^M(t)$. Let $a_{ij}^{X,M}$ be the projection score vectors of $X_{ij}(t)$ onto $e_j^M(t)$ and $\tilde{a}_{ij}^{X,M}$ be the projection score vectors of $X_{ij}(t)$ onto $\tilde{e}_j^M(t)$. Then $\tilde{a}_{ij}^{X,M} = U_j a_{ij}^{X,M}$. Denote

$$U = \mathrm{diag}\{U_1, U_2, \ldots, U_p\} \in \mathbb{R}^{pM\times pM}.$$

We then have

$$\begin{aligned}\tilde{a}_i^{X,M} &= ((\tilde{a}_{i1}^{X,M})^\top, (\tilde{a}_{i2}^{X,M})^\top, \ldots, (\tilde{a}_{ip}^{X,M})^\top)^\top \\ &= ((a_{i1}^{X,M})^\top U_1^\top, (a_{i2}^{X,M})^\top U_2^\top, \ldots, (a_{ip}^{X,M})^\top U_p^\top)^\top = U a_i^{X,M}\end{aligned}$$

and

$$\tilde{\Sigma}^{X,M} = \mathrm{Cov}\left(\tilde{a}^{X,M}\right) = U\mathrm{Cov}\left(\tilde{a}^{X,M}\right)U^\top = U\Sigma^{X,M}U^\top.$$

Thus

$$\tilde{\Theta}^{X,M} = \left(\tilde{\Sigma}^{X,M}\right)^{-1} = U\left(\Sigma^{X,M}\right)^{-1}U^\top = U\Theta^{X,M}U^\top.$$

Therefore, $\tilde{\Theta}_{jl}^{X,M} = U_j\Theta_{jl}^{X,M}U_l^\top$ for all $j,l \in V^2$ and, therefore, $\|\tilde{\Theta}_{jl}^{X,M}\|_{\mathrm{F}} = \|\Theta_{jl}^{X,M}\|_{\mathrm{F}}$ for all $j,l \in V^2$. This implies the final result.

B.2 Proof of Lemma 3

We first show that $X_{ij}, Y_{ij} \in \mathrm{Span}\left\{\phi_{j1}, \ldots, \phi_{jM_j^\star}\right\}$ almost surely. Let

$$M_j^X = \sup\{M \in \mathbb{N}^+ : \lambda_{jM}^X > 0\}.$$

By Karhunen–Loève theorem, we have $X_{ij} = \sum_{k=1}^{M_j^X}\langle X_{ij}, \phi_{jk}^X\rangle\phi_{jk}^X$ almost surely. Thus, we have $X_{ij} \in \mathrm{Span}\left\{\phi_{j1}^X, \ldots, \phi_{j,M_j^X}^X\right\}$ almost surely. For any $1 \le k \le M_j^X$, we have that

$$\int_{\mathcal{T}} K_{jj}(s,t)\phi_k^X(s)\phi_k^X(t)dsdt \ge \int_{\mathcal{T}} K_{jj}^X(s,t)\phi_k^X(s)\phi_k^X(t)dsdt = \lambda_{jk}^X > 0,$$

which implies that $\phi_k^X \in \mathrm{Span}\left\{\phi_{j1}, \ldots, \phi_{jM_j^\star}\right\}$. Thus, we have $\mathrm{Span}\left\{\phi_{j1}^X, \ldots, \phi_{j,M_j^X}^X\right\} \subseteq \mathrm{Span}\left\{\phi_{j1}, \ldots, \phi_{jM_j^\star}\right\}$ and $X_{ij} \in \mathrm{Span}\left\{\phi_{j1}, \ldots, \phi_{jM_j^\star}\right\}$ almost surely. Similarly, we have that $Y_{ij} \in \mathrm{Span}\left\{\phi_{j1}, \ldots, \phi_{jM_j^\star}\right\}$ almost surely.

Next, we show that $M_j' = M_j^\star$ by contradiction. By the definition of M_j', we have that $M_j' \leq M_j^\star$. If $M_j' \neq M_j^\star$, then we have $\mathbb{V}_j^{M_j'} \subseteq \mathbb{H}$ such that $M_j' < M_j^\star$ and $X_{ij}, Y_{ij} \in \mathbb{V}_j^{M_j'}$ almost surely. This implies that there exists $\phi \in \text{Span}\left\{\phi_{j1}, \ldots, \phi_{jM_j^\star}\right\} \setminus \mathbb{V}_j^{M_j'}$ such that

$$\begin{aligned}
&\mathbb{E}\left[(\langle \phi_{jk}(t), X_{ij}(t)\rangle)^2\right] = 0 \quad \text{and} \quad \mathbb{E}\left[(\langle \phi_{jk}(t), Y_{ij}(t)\rangle)^2\right] = 0 \\
\Rightarrow &\int_{\mathcal{T}} K_{jj}^X(s,t)\phi_{jk}(s)\phi_{jk}(t)dsdt = 0 \quad \text{and} \quad \int_{\mathcal{T}} K_{jj}^Y(s,t)\phi_{jk}(s)\phi_{jk}(t)dsdt = 0 \\
\Rightarrow &\int_{\mathcal{T}} K_{jj}(s,t)\phi_{jk}(s)\phi_{jk}(t)dsdt = 0, \\
\Rightarrow &\lambda_{jk} = 0,
\end{aligned}$$

which contradicts the definition of $M_j^\star$. Thus, we must have $M_j' = M_j^\star$.

B.3 Proof of Lemma 7

Let $U = V \backslash \{j, l\}$, and $a_U^{X,M} = \left((a_j^{X,M})^\top, j \in U\right)^\top$. Without loss of generality, assume that $\Sigma^{X,M}$ and $\Theta^{X,M}$ take the following block structure:

$$\Sigma^{X,M} = \begin{bmatrix} \Sigma_{jj}^{X,M} & \Sigma_{jl}^{X,M} & \Sigma_{jU}^{X,M} \\ \Sigma_{lj}^{X,M} & \Sigma_{ll}^{X,M} & \Sigma_{lU}^{X,M} \\ \Sigma_{Uj}^{X,M} & \Sigma_{Ul}^{X,M} & \Sigma_{UU}^{X,M} \end{bmatrix}, \quad \Theta^{X,M} = \begin{bmatrix} \Theta_{jj}^{X,M} & \Theta_{jl}^{X,M} & \Theta_{jU}^{X,M} \\ \Theta_{lj}^{X,M} & \Theta_{ll}^{X,M} & \Theta_{lU}^{X,M} \\ \Theta_{Uj}^{X,M} & \Theta_{Ul}^{X,M} & \Theta_{UU}^{X,M} \end{bmatrix}.$$

Let P denote the submatrix:

$$P = \begin{bmatrix} \Theta_{jj}^{X,M} & \Theta_{jl}^{X,M} \\ \Theta_{lj}^{X,M} & \Theta_{ll}^{X,M} \end{bmatrix}.$$

By standard results for the multivariate Gaussian (Johnson and Wichern, 2014), we have

$$\begin{aligned}
&\text{Var}\left(a_j^{X,M} \mid a_k^{X,M}, k \neq j\right) = H_{jj}^{X,M} = (\Theta_{jj}^{X,M})^{-1}, \\
&\text{Var}\left(\begin{bmatrix} a_j^{X,M} \\ a_l^{X,M} \end{bmatrix} \mid a_U^{X,M}\right) = P^{-1} = \begin{bmatrix} (P^{-1})_{11} & (P^{-1})_{12} \\ (P^{-1})_{21} & (P^{-1})_{22} \end{bmatrix}.
\end{aligned}$$

Thus, the first statement directly follows from the first equation. To prove the second statement, we only need to note that

$$\begin{aligned}
H_{jl}^{X,M} &= \text{Cov}\left(a_j^{X,M}, a_l^{X,M} \mid a_U^{X,M}\right) \\
&= (P^{-1})_{12} \\
&= -(\Theta_{jj}^{X,M})^{-1}\Theta_{jl}^{X,M}(P^{-1})_{22} \\
&= -H_{jj}^{X,M}\Theta_{jl}^{X,M}H_{ll}^{\backslash j,X,M},
\end{aligned}$$

where the second to last equation follows from the 2×2 block matrix inverse and the last equation follows from the property of multivariate Gaussian. This completes the proof.

B.4 Proof of Theorem 10

We provide the proof of Theorem 10, following the framework introduced in Negahban et al. (2012). We start by introducing some notation.

We use $\otimes$ to denote the Kronecker product. For $\Delta \in \mathbb{R}^{pM \times pM}$, let $\theta = \text{vec}(\Delta) \in \mathbb{R}^{p^2M^2}$ and $\theta^* = \text{vec}(\Delta^M)$, where Δ^M is defined in Section 2.2. Let $\mathcal{G} = \{G_t\}_{t=1,\dots,N_\mathcal{G}}$ be a set of indices, where $N_\mathcal{G} = p^2$ and $G_t \subset \{1, 2, \cdots, p^2M^2\}$ is the set of indices for θ that correspond to the t-th $M \times M$ submatrix of Δ^M. Thus, if $t = (j-1)p + l$, then $\theta_{G_t} = \text{vec}\,(\Delta_{jl}) \in \mathbb{R}^{M^2}$, where Δ_{jl} is the (j,l)-th $M \times M$ submatrix of Δ. Denote the group indices of θ^* that belong to blocks corresponding to E_Δ as $S_\mathcal{G} \subseteq \{1, 2, \cdots, N_\mathcal{G}\}$. Note that we define $S_\mathcal{G}$ using E_Δ and not E_{Δ^M}. Therefore, as stated in Assumption 2, $|S_\mathcal{G}| = s$. We further define the subspace $\mathcal{M}$ as

$$\mathcal{M} \coloneqq \{\theta \in \mathbb{R}^{p^2M^2} \mid \theta_{G_t} = 0 \text{ for all } t \notin S_\mathcal{G}\}. \tag{B.1}$$

Its orthogonal complement with respect to the Euclidean inner product is

$$\mathcal{M}^\perp \coloneqq \{\theta \in \mathbb{R}^{p^2M^2} \mid \theta_{G_t} = 0 \text{ for all } t \in S_\mathcal{G}\}.$$

For a vector θ, let $\theta_\mathcal{M}$ and $\theta_{\mathcal{M}^\perp}$ be the projection of θ on the subspaces $\mathcal{M}$ and $\mathcal{M}^\perp$, respectively. Let $\langle \cdot, \cdot \rangle$ represent the Euclidean inner product. Let

$$\mathcal{R}(\theta) \coloneqq \sum_{t=1}^{N_\mathcal{G}} |\theta_{G_t}|_2 \triangleq |\theta|_{1,2}. \tag{B.2}$$

For any $v \in \mathbb{R}^{p^2M^2}$, the dual norm of $\mathcal{R}$ is given by

$$\mathcal{R}^*(v) \coloneqq \sup_{u \in \mathbb{R}^{p^2M^2} \setminus \{0\}} \frac{\langle u, v \rangle}{\mathcal{R}(u)} = \sup_{\mathcal{R}(u) \leq 1} \langle u, v \rangle. \tag{B.3}$$

The subspace compatibility constant of $\mathcal{M}$ with respect to $\mathcal{R}$ is defined as

$$\Psi(\mathcal{M}) \coloneqq \sup_{u \in \mathcal{M} \setminus \{0\}} \frac{\mathcal{R}(u)}{|u|_2}. \tag{B.4}$$

Proof By Lemma 15 and Assumption 1, we have

$$|(S^{Y,M} \otimes S^{X,M}) - (\Sigma^{Y,M} \otimes \Sigma^{X,M})|_\infty \leq \delta_n^2 + 2\delta_n \sigma_{\max} \tag{B.5}$$

and

$$|\,\text{vec}\,(S^{Y,M} - S^{X,M}) - \text{vec}\,(\Sigma^{Y,M} - \Sigma^{X,M})|_\infty \leq 2\delta_n.$$

The problem (10) can be written in the following form:

$$\hat{\theta}_{\lambda_n} \in \underset{\theta \in \mathbb{R}^{p^2M^2}}{\arg\min}\, \mathcal{L}(\theta) + \lambda_n \mathcal{R}(\theta),$$

where

$$\mathcal{L}(\theta) = \frac{1}{2}\theta^\top (S^{Y,M} \otimes S^{X,M})\theta - \theta^\top \text{vec}(S^{Y,M} - S^{X,M}). \tag{B.6}$$

Here, we slightly abuse the notation and use $\mathcal{L}(\cdot)$ to denote the function of θ rather than Δ. The loss $\mathcal{L}(\theta)$ is convex and differentiable with respect to θ, and it can easily be verified that $\mathcal{R}(\cdot)$ defines a vector norm. For $h \in \mathbb{R}^{p^2M^2}$, the error of the first-order Taylor series expansion of $\mathcal{L}$ is:

$$\delta\mathcal{L}(h, \theta^*) \coloneqq \mathcal{L}(\theta^* + h) - \mathcal{L}(\theta^*) - \langle \nabla\mathcal{L}(\theta^*), h\rangle = \frac{1}{2} h^\top (S^{Y,M} \otimes S^{X,M}) h. \tag{B.7}$$

From (B.6), we see that $\nabla\mathcal{L}(\theta) = (S^{Y,M} \otimes S^{X,M})\theta - \mathrm{vec}(S^{Y,M} - S^{X,M})$. By Lemma 19, we have

$$\mathcal{R}^*(\nabla\mathcal{L}(\theta^*)) = \max_{t=1,2,\cdots,N_{\mathcal{G}}} \left| \left[(S^{Y,M} \otimes S^{X,M})\theta^* - \mathrm{vec}(S^{Y,M} - S^{X,M}) \right]_{G_t} \right|_2 .$$

Now we establish an upper bound for $\mathcal{R}^*(\nabla\mathcal{L}(\theta^*))$. First, note that

$$(\Sigma^{Y,M} \otimes \Sigma^{X,M})\theta^* - \mathrm{vec}(\Sigma^{Y,M} - \Sigma^{X,M}) = \mathrm{vec}(\Sigma^{X,M}\Delta^M\Sigma^{Y,M} - (\Sigma^{Y,M} - \Sigma^{X,M})) = 0.$$

Letting $(\cdot)_{jl}$ denote the (j,l)-th submatrix, we have

$$\begin{aligned}
&\left| \left[(S^{Y,M} \otimes S^{X,M})\theta^* - \mathrm{vec}(S^{Y,M} - S^{X,M}) \right]_{G_t} \right|_2 \\
&= \left| \left[(S^{Y,M} \otimes S^{X,M} - \Sigma^{Y,M} \otimes \Sigma^{X,M})\theta^* - \mathrm{vec}\left((S^{Y,M} - \Sigma^{Y,M}) - (S^{X,M} - \Sigma^{X,M}) \right) \right]_{G_t} \right|_2 \\
&= \| (S^{X,M}\Delta^M S^{Y,M} - \Sigma^{X,M}\Delta^M\Sigma^{Y,M})_{jl} - (S^{Y,M} - \Sigma^{Y,M})_{jl} - (S^{X,M} - \Sigma^{X,M})_{jl} \|_F \\
&\le \| (S^{X,M}\Delta^M S^{Y,M} - \Sigma^{X,M}\Delta^M\Sigma^{Y,M})_{jl} \|_F + \| (S^{Y,M} - \Sigma^{Y,M})_{jl} \|_F + \| (S^{X,M} - \Sigma^{X,M})_{jl} \|_F .
\end{aligned}$$

For any $M \times M$ matrix A, $\|A\|_F \le M|A|_\infty$, so

$$\begin{aligned}
&\left| \left[(S^{Y,M} \otimes S^{X,M})\theta^* - \mathrm{vec}(S^{Y,M} - S^{X,M}) \right]_{G_t} \right|_2 \\
&\le M \left[\left| (S^{X,M}\Delta^M S^{Y,M} - \Sigma^{X,M}\Delta^M\Sigma^{Y,M})_{jl} \right|_\infty \right. \\
&\qquad \left. + \left| (S^{Y,M} - \Sigma^{Y,M})_{jl} \right|_\infty + \left| (S^{X,M} - \Sigma^{X,M})_{jl} \right|_\infty \right] \\
&\le M \left[\left| S^{X,M}\Delta^M S^{Y,M} - \Sigma^{X,M}\Delta^M\Sigma^{Y,M} \right|_\infty + |S^{Y,M} - \Sigma^{Y,M}|_\infty + |S^{X,M} - \Sigma^{X,M}|_\infty \right] .
\end{aligned}$$

For any $A \in \mathbb{R}^{k\times k}$ and $v \in \mathbb{R}^k$, we have $|Av|_\infty \le |A|_\infty |v|_1$. Thus, we also have

$$\begin{aligned}
|S^{X,M}\Delta^M S^{Y,M} - \Sigma^{X,M}\Delta^M\Sigma^{Y,M}|_\infty &= |[(S^{Y,M} \otimes S^{X,M}) - (\Sigma^{X,M} \otimes \Sigma^{Y,M})] \,\mathrm{vec}\,(\Delta^M)|_\infty \\
&\le |(S^{Y,M} \otimes S^{X,M}) - (\Sigma^{X,M} \otimes \Sigma^{Y,M})|_\infty |\,\mathrm{vec}\,(\Delta^M)|_1 \\
&= |(S^{Y,M} \otimes S^{X,M}) - (\Sigma^{X,M} \otimes \Sigma^{Y,M})|_\infty |\Delta^M|_1 .
\end{aligned}$$

Combining the inequalities gives an upper bound uniform over $\mathcal{G}$ (i.e., for all G_t):

$$\begin{aligned}
&\left| \left[(S^{Y,M} \otimes S^{X,M})\theta^* - \mathrm{vec}(S^{Y,M} - S^{X,M}) \right]_{G_t} \right|_2 \\
&\qquad \le M[|(S^{Y,M} \otimes S^{X,M}) - (\Sigma^{X,M} \otimes \Sigma^{Y,M})|_\infty |\Delta^M|_1 \\
&\qquad\qquad + |S^{Y,M} - \Sigma^{Y,M}|_\infty + |S^{X,M} - \Sigma^{X,M}|_\infty],
\end{aligned}$$

which implies

$$\mathcal{R}^*\left(\nabla\mathcal{L}(\theta^*)\right) \leq M[|(S^{Y,M}\otimes S^{X,M}) - (\Sigma^{X,M}\otimes\Sigma^{Y,M})|_\infty|\Delta^M|_1 + |S^{Y,M}-\Sigma^{Y,M}|_\infty + |S^{X,M}-\Sigma^{X,M}|_\infty].$$

Assuming $|S^{X,M}-\Sigma^{X,M}|_\infty \leq \delta_n$ and $|S^{Y,M}-\Sigma^{Y,M}|_\infty \leq \delta_n$ implies

$$\mathcal{R}^*\left(\nabla\mathcal{L}(\theta^*)\right) \leq M[(\delta_n^2+2\delta_n\sigma_{\max})|\Delta^M|_1 + 2\delta_n].$$

Setting

$$\lambda_n = 2M\left[\left(\delta_n^2+2\delta_n\sigma_{\max}\right)\left|\Delta^M\right|_1 + 2\delta_n\right], \tag{B.8}$$

then implies that $\lambda_n \geq 2\mathcal{R}^*\left(\nabla\mathcal{L}(\theta^*)\right)$. Thus, invoking Lemma 1 in Negahban et al. (2012), $h = \hat{\theta}_{\lambda_n} - \theta^*$ must satisfy

$$\mathcal{R}(h_{\mathcal{M}^\perp}) \leq 3\mathcal{R}(h_{\mathcal{M}}) + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp}),$$

where $\mathcal{M}$ is defined in (B.1). Equivalently,

$$|h_{\mathcal{M}^\perp}|_{1,2} \leq 3|h_{\mathcal{M}}|_{1,2} + 4|\theta^*_{\mathcal{M}^\perp}|_{1,2}. \tag{B.9}$$

By the definition of ν_2, we have

$$|\theta^*_{\mathcal{M}^\perp}|_{1,2} = \sum_{t\notin\mathcal{S}_\mathcal{G}}|\theta^*_{G_t}|_2 \leq \left(p(p+1)/2-s\right)\nu_2 \leq p^2\nu_2.$$

Next, we show that $\delta\mathcal{L}(h,\theta^*)$, as defined in (B.7), satisfies the Restricted Strong Convexity property: $\delta\mathcal{L}(h,\theta^*) \geq \kappa_\mathcal{L}|h|_2^2 - \omega_\mathcal{L}^2(\theta^*)$ whenever h satisfies (B.9). We have

$$\begin{aligned}\theta^\top(S^{Y,M}\otimes S^{X,M})\theta &= \theta^\top(\Sigma^{Y,M}\otimes\Sigma^{X,M})\theta + \theta^\top(S^{Y,M}\otimes S^{X,M}-\Sigma^{Y,M}\otimes\Sigma^{X,M})\theta\\ &\geq \theta^\top(\Sigma^{Y,M}\otimes\Sigma^{X,M})\theta - |\theta^\top(S^{Y,M}\otimes S^{X,M}-\Sigma^{Y,M}\otimes\Sigma^{X,M})\theta|\\ &\geq \lambda^*_{\min}|\theta|_2^2 - M^2|S^{Y,M}\otimes S^{X,M}-\Sigma^{Y,M}\otimes\Sigma^{X,M}|_\infty|\theta|_{1,2}^2,\end{aligned}$$

where the last inequality follows from Lemma 17 and $\lambda^*_{\min} = \lambda_{\min}(\Sigma^{X,M})\times\lambda_{\min}(\Sigma^{Y,M}) = \lambda_{\min}(\Sigma^{Y,M}\otimes\Sigma^{X,M}) > 0$. Thus,

$$\begin{aligned}\delta\mathcal{L}(h,\theta^*) &= \frac{1}{2}h^\top(S^{Y,M}\otimes S^{X,M})h\\ &\geq \frac{1}{2}\lambda^*_{\min}|h|_2^2 - \frac{1}{2}M^2|S^{Y,M}\otimes S^{X,M}-\Sigma^{Y,M}\otimes\Sigma^{X,M}|_\infty|h|_{1,2}^2.\end{aligned}$$

By Lemma 18 and (B.9), we have

$$\begin{aligned}|h|_{1,2}^2 &= (|h_{\mathcal{M}}|_{1,2}+|h_{\mathcal{M}^\perp}|_{1,2})^2 \leq 16(|h_{\mathcal{M}}|_{1,2}+|\theta^*_{\mathcal{M}^\perp}|_{1,2})^2\\ &\leq 16(\sqrt{s}|h|_2 + p^2\nu_2)^2 \leq 32s|h|_2^2 + 32p^4\nu_2^2.\end{aligned}$$

Combining with the above equation, we get

$$\begin{aligned}\delta\mathcal{L}(h,\theta^*) &\geq \left[\frac{1}{2}\lambda^*_{\min} - 16M^2s|S^{Y,M}\otimes S^{X,M}-\Sigma^{Y,M}\otimes\Sigma^{X,M}|_\infty\right]|h|_2^2\\ &\qquad - 16M^2p^4\nu_2^2|S^{Y,M}\otimes S^{X,M}-\Sigma^{Y,M}\otimes\Sigma^{X,M}|_\infty\\ &\geq \left[\frac{1}{2}\lambda^*_{\min} - 8M^2s\left(\delta_n^2+2\delta_n^2\sigma_{\max}\right)\right]|h|_2^2\\ &\qquad - 16M^2p^4\nu_2^2\left(\delta_n^2+2\delta_n\sigma_{\max}\right).\end{aligned}$$

Thus, appealing to (B.5), the Restricted Strong Convexity property holds with

$$\begin{aligned}\kappa_{\mathcal{L}} &= \frac{1}{2}\lambda^*_{\min} - 8M^2 s\left(\delta^2 + 2\delta_n\sigma_{\max}\right),\\ \omega_{\mathcal{L}} &= 4Mp^2\nu_2\sqrt{\delta_n^2 + 2\delta_n\sigma_{\max}}.\end{aligned}$$

When $\delta_n < \frac{1}{4}\sqrt{\frac{\lambda^*_{\min}+16M^2s(\sigma_{\max})^2}{M^2 s}} - \sigma_{\max}$ as we assumed in the theorem, then $\kappa_{\mathcal{L}} > 0$. By Theorem 1 of Negahban et al. (2012) and Lemma 18, letting

$$\lambda_n = 2M\left[\left(\delta_n^2 + 2\delta_n\sigma_{\max}\right)|\Delta^M|_1 + 2\delta_n\right],$$

as in (B.8), ensures that

$$\begin{aligned}\|\hat{\Delta}^M - \Delta^M\|_F^2 &= |\hat{\theta}_{\lambda_n} - \theta^*|_2^2\\ &\le 9\frac{\lambda_n^2}{\kappa_{\mathcal{L}}^2}\Psi^2(\mathcal{M}) + \frac{\lambda_n}{\kappa_{\mathcal{L}}}\left(2\omega_{\mathcal{L}}^2 + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp})\right)\\ &= \frac{9\lambda_n^2 s}{\kappa_{\mathcal{L}}^2} + \frac{2\lambda_n}{\kappa_{\mathcal{L}}}(\omega_{\mathcal{L}}^2 + 2p^2\nu_2)\\ &= \Gamma_n^2.\end{aligned}$$

We then prove that $\hat{E}_\Delta = E_\Delta$. Recall that we have assumed that $0 < \Gamma_n < \tau/2 = (\nu_1 - \nu_2)/2$ and $\nu_2 + \Gamma_n \le \epsilon_n < \nu_1 - \Gamma_n$. Note that we have $\|\hat{\Delta}^M_{jl} - \Delta^M_{jl}\|_F \le \|\hat{\Delta}^M - \Delta^M\|_F \le \Gamma_n$ for any $(j,l) \in V^2$. Recall that

$$E_\Delta = \{(j,l) \in V^2 :\ j \ne l, D_{jl} > 0\}.$$

First, we prove that $E_\Delta \subseteq \hat{E}_\Delta$. For any $(j,l) \in E_\Delta$, by the definition of ν_1 in Section 4.1, we have

$$\begin{aligned}\|\hat{\Delta}^M_{jl}\|_F &\ge \|\Delta^M_{jl}\|_F - \|\hat{\Delta}^M_{jl} - \Delta^M_{jl}\|_F\\ &\ge \nu_1 - \Gamma_n\\ &> \epsilon_n.\end{aligned}$$

The last inequality holds because we have assumed that $\epsilon_n < \nu_1 - \Gamma_n$. Thus, by the definition of $\hat{E}_\Delta$ in (13), we have $(j,l) \in \hat{E}_\Delta$, which further implies that $E_\Delta \subseteq \hat{E}_\Delta$.

We then show $\hat{E}_\Delta \subseteq E_\Delta$. Let $\hat{E}^c_\Delta$ and E^c_Δ denote the complement of $\hat{E}_\Delta$ and E_Δ. For any $(j,l) \in E^c_\Delta$, which also means that $(l,j) \in E^c_\Delta$, by the definition of ν_2, we have that

$$\begin{aligned}\|\hat{\Delta}^M_{jl}\|_F &\le \|\Delta^M_{jl}\|_F + \|\hat{\Delta}^M_{jl} - \Delta^M_{jl}\|_F\\ &\le \nu_2 + \Gamma_n\\ &\le \epsilon_n.\end{aligned}$$

Again, the last inequality is true because we have assumed $\epsilon_n \ge \nu_2 + \Gamma_n$. Thus, by the definition of $\hat{E}_\Delta$, we have $(j,l) \notin \hat{E}_\Delta$ or $(j,l) \in \hat{E}^c_\Delta$. This implies that $E^c_\Delta \subseteq \hat{E}^c_\Delta$, or $\hat{E}_\Delta \subseteq E_\Delta$. Combining with the previous conclusion that $E_\Delta \subseteq \hat{E}_\Delta$, the proof is complete. ■

B.5 Proof of Theorem 13

We only need to prove that

$$P\left(|S^M - \Sigma^M|_\infty > \delta\right) \le C_1 np \exp\{-C_2 \Phi(T,L) M^{-(1+\beta)}\delta\} + C_3 (pM)^2 \exp\{-C_4 n M^{-2(1+\beta)}\delta^2\} + C_5 npL \exp\left\{-\frac{C_6 M^{-2(1+\beta)}\delta^2}{\tilde{\psi}_2(T,L)}\right\}, \quad \text{(B.10)}$$

where S^M can be understood as $S^{X,M}$ or $S^{Y,M}$ and Σ^M can be understood as $\Sigma^{X,M}$ or $\Sigma^{Y,M}$, with $C_k = C_k^X$ or $C_k = C_k^Y$ for $k = 1,2,3,4$. To see that (B.10) implies (18), we first note that (B.10) implies that

$$\begin{aligned} &P\left(|S^{X,M} - \Sigma^{X,M}|_\infty \le \delta \text{ and } |S^{Y,M} - \Sigma^{Y,M}|_\infty \le \delta\right) \\ &\quad \ge 1 - P\left(|S^{X,M} - \Sigma^{X,M}|_\infty > \delta\right) - P\left(|S^{Y,M} - \Sigma^{Y,M}|_\infty > \delta\right) \\ &\quad \ge 1 - 2\bar{C}_1 pM \exp\{-\bar{C}_2 \Phi(T,L) M^{-(1+\beta)}\delta\} - 2\bar{C}_3 (pM)^2 \exp\{-\bar{C}_4 n M^{-2(1+\beta)}\delta^2\}, \end{aligned}$$

where $\bar{C}_k$ for $k = 1,2,3,4$ are defined in Theorem 13. Thus, letting the last two terms in the last line of the above equation be $\iota/2$, we then have (18). In this way, the rest of the proof will focus on proving (B.10).

Denote the (j,l)-th submatrix of S^M as S^M_{jl}, and the (k,m)-th entry of S^M_{jl} as $\hat{\sigma}_{jl,km}$. We have $S^M = (\hat{\sigma}_{jl,km})_{1\le j,l\le p, \le k,m \le M}$ and $\Sigma^M = (\sigma_{jl,km})_{1\le j,l\le p, \le k,m \le M}$. Then, by the definition of S^M and Σ^M, we have

$$\hat{\sigma}_{jl,km} = \frac{1}{n}\sum_{i=1}^n \hat{a}_{ijk}\hat{a}_{ilm} \qquad \text{and} \qquad \sigma_{jl,km} = \mathbb{E}\left[a_{ijk}a_{ilm}\right].$$

Note that

$$\begin{aligned} \hat{a}_{ijk} &= \langle \hat{g}_{ij}, \hat{\phi}_{jk}\rangle \\ &= \langle g_{ij} + \hat{g}_{ij} - g_{ij}, \phi_{jk} + \hat{\phi}_{jk} - \phi_{jk}\rangle \\ &= \langle g_{ij}, \phi_{jk}\rangle + \langle g_{ij}, \hat{\phi}_{jk} - \phi_{jk}\rangle + \langle \hat{g}_{ij} - g_{ij}, \phi_{jk}\rangle + \langle \hat{g}_{ij} - g_{ij}, \hat{\phi}_{jk} - \phi_{jk}\rangle \\ &= a_{ijk} + \langle g_{ij}, \hat{\phi}_{jk} - \phi_{jk}\rangle + \langle \hat{g}_{ij} - g_{ij}, \phi_{jk}\rangle + \langle \hat{g}_{ij} - g_{ij}, \hat{\phi}_{jk} - \phi_{jk}\rangle. \end{aligned}$$

Thus, we have

$$\hat{\sigma}_{jl,km} - \sigma_{jl,km} = \frac{1}{n}\sum_{i=1}^n (\hat{a}_{ijk}\hat{a}_{ilm} - \sigma_{jl,km}) = \sum_{u=1}^{16} I_u,$$

where

$$\begin{aligned} I_1 &= \frac{1}{n}\sum_{i=1}^n \left(a_{ijk}a_{ilm} - \mathbb{E}(a_{ijk}a_{ilm})\right), \\ I_2 &= \frac{1}{n}\sum_{i=1}^n a_{ijk}\langle \hat{g}_{il} - g_{il}, \phi_{lm}\rangle, \\ I_3 &= \frac{1}{n}\sum_{i=1}^n a_{ijk}\langle g_{il}, \hat{\phi}_{lm} - \phi_{lm}\rangle, \end{aligned}$$

$$
\begin{aligned}
I_4 &= \frac{1}{n}\sum_{i=1}^{n} a_{ijk}\langle \hat{g}_{il} - g_{il}, \hat{\phi}_{lm} - \phi_{lm}\rangle,\\
I_5 &= \frac{1}{n}\sum_{i=1}^{n} a_{ilm}\langle \hat{g}_{ij} - g_{ij}, \phi_{jk}\rangle,\\
I_6 &= \frac{1}{n}\sum_{i=1}^{n} \langle \hat{g}_{ij} - g_{ij}, \phi_{jk}\rangle\langle \hat{g}_{il} - g_{il}, \phi_{lm}\rangle,\\
I_7 &= \frac{1}{n}\sum_{i=1}^{n} \langle \hat{g}_{ij} - g_{ij}, \phi_{jk}\rangle\langle g_{il}, \hat{\phi}_{lm} - \phi_{lm}\rangle,\\
I_8 &= \frac{1}{n}\sum_{i=1}^{n} \langle \hat{g}_{ij} - g_{ij}, \phi_{jk}\rangle\langle \hat{g}_{il} - g_{il}, \hat{\phi}_{lm} - \phi_{lm}\rangle,\\
I_9 &= \frac{1}{n}\sum_{i=1}^{n} \langle g_{ij}, \hat{\phi}_{jk} - \phi_{jk}\rangle a_{ilm},\\
I_{10} &= \frac{1}{n}\sum_{i=1}^{n} \langle g_{ij}, \hat{\phi}_{jk} - \phi_{jk}\rangle\langle \hat{g}_{il} - g_{il}, \phi_{lm}\rangle,\\
I_{11} &= \frac{1}{n}\sum_{i=1}^{n} \langle g_{ij}, \hat{\phi}_{jk} - \phi_{jk}\rangle\langle g_{il}, \hat{\phi}_{lm} - \phi_{lm}\rangle,\\
I_{12} &= \frac{1}{n}\sum_{i=1}^{n} \langle g_{ij}, \hat{\phi}_{jk} - \phi_{jk}\rangle\langle \hat{g}_{il} - g_{il}, \hat{\phi}_{lm} - \phi_{lm}\rangle,\\
I_{13} &= \frac{1}{n}\sum_{i=1}^{n} \langle \hat{g}_{ij} - g_{ij}, \hat{\phi}_{jk} - \phi_{jk}\rangle a_{ilm},\\
I_{14} &= \frac{1}{n}\sum_{i=1}^{n} \langle \hat{g}_{ij} - g_{ij}, \hat{\phi}_{jk} - \phi_{jk}\rangle\langle \hat{g}_{il} - g_{il}, \phi_{lm}\rangle,\\
I_{15} &= \frac{1}{n}\sum_{i=1}^{n} \langle \hat{g}_{ij} - g_{ij}, \hat{\phi}_{jk} - \phi_{jk}\rangle\langle g_{il}, \hat{\phi}_{lm} - \phi_{lm}\rangle,\\
I_{16} &= \frac{1}{n}\sum_{i=1}^{n} \langle \hat{g}_{ij} - g_{ij}, \hat{\phi}_{jk} - \phi_{jk}\rangle\langle \hat{g}_{il} - g_{il}, \hat{\phi}_{lm} - \phi_{lm}\rangle.
\end{aligned}
$$

Note that I_u, $u = 1, \ldots, 16$ depend on j, l, k, m. To simplify the notation, we do not explicitly denote this fact. Thus, for any $0 < \delta \leq 1$, when for any $1 \leq j, l \leq p$ and $1 \leq k, m \leq M$, if $|I_u| \leq \delta/16$, $u = 1, \ldots, 16$, we have $|S^M - \Sigma^M|_\infty \leq \delta$. We now calculate the probability of $|I_u| \leq \delta/16$, $u = 1, \ldots, 16$, $1 \leq j, l \leq p$ and $1 \leq k, m \leq M$.

By Assumption 3 (i), we have constants $d_1, d_2 > 0$, such that $\lambda_{jk} \leq d_1 k^{-\beta}$, $d_{jk} \leq d_2 k^{1+\beta}$ for any $j = 1, \ldots, p$ and $k \geq 1$. Let $d_0 = \max\{1, \sqrt{d_1}, d_2\}$ and $\xi_{ijk} = \lambda_{jk}^{-1/2} a_{ijk}$ so that $\xi_{ijk} \sim N(0, 1)$ are i.i.d. for $i = 1, \ldots, n$. Let

$$
\delta_1 = \frac{\delta}{144 d_0^2 M^{1+\beta}\sqrt{3\lambda_{0,\max}}} \quad \text{and} \quad \delta_2 = 9\lambda_{0,\max}\delta_1 = \frac{\delta}{16 d_0^2 M^{1+\beta}\sqrt{3\lambda_{0,\max}}},
$$

where $\lambda_{0,\max} = \max_{j\in V}\sum_{k=1}^{\infty}\lambda_{jk}$. Recall that $\hat{K}_{jj}$, $j = 1,\ldots,p$, are defined in (9). We define five events A_1-A_5 as follows:

$$\begin{aligned}
A_1:&\ \|\hat{g}_{ij} - g_{ij}\| \le \delta_1, \quad \forall i = 1,\ldots,n\ \forall j = 1,\ldots,p,\\
A_2:&\ \|\hat{K}_{jj} - K_{jj}\|_{\mathrm{HS}} \le \delta_2 \quad \forall j = 1,\ldots,p,\\
A_3:&\ \frac{1}{n}\sum_{i=1}^{n}\xi_{ijk}^2 \le \frac{3}{2} \quad \forall j = 1,\ldots,p\ \forall k = 1,\ldots,M,\\
A_4:&\ \frac{1}{n}\sum_{i=1}^{n}\|g_{ij}\|^2 \le 2\lambda_{0,\max} \quad \forall j = 1,\ldots,p,\\
A_5:&\ |\frac{1}{n}\sum_{i=1}^{n}a_{ijk}a_{ilm} - \sigma_{jl,km}| \le \frac{\delta}{16} \quad \forall 1 \le j,l \le\ 1 \le k,m \le M.
\end{aligned}$$

Without loss of generality, we assume that $\langle\hat{\phi}_{jl},\phi_{jl}\rangle \ge 0$ for any $1 \le j \le p$ and $1 \le k \le M$ (if this is not true, we only need to use $-\phi_{jl}$ to substitute ϕ_{jl}). Then, by Lemma 20-Lemma 35, when A_1-A_5 hold simultaneously, we have $|I_u| \le \delta/16$ for all $u = 1,\ldots,16$, $1 \le j,l \le p$ and $1 \le k,m \le M$. Therefore,

$$\begin{aligned}
&P\left(|S^M - \Sigma^M|_\infty \le \delta\right)\\
&\quad\ge P\left(|I_u| \le \delta/16,\ \text{for all } 1 \le u \le 16, 1 \le j,l \le\ 1 \le k,m \le M\right)\\
&\quad\ge P\left(\bigcap_{w=1}^{5}A_w\right),
\end{aligned}$$

which implies

$$P\left(|S^M - \Sigma^M|_\infty > \delta\right) \le P\left(\bigcup_{w=1}^{5}\bar{A}_w\right) \le \sum_{w=1}^{5}P\left(\bar{A}_w\right),$$

where the last inequality follows Boole's inequality and $\bar{A}$ denotes the complement of A. Then we only need to give an upper bound for $P(\bar{A}_w)$, $w = 1,\ldots,5$.

By Theorem 14 and the definition of $\tilde{\psi}_1$-$\tilde{\psi}_4$, we have

$$\begin{aligned}
P(\bar{A}_1) &= P\left(\|\hat{g}_{ij} - g_{ij}\| > \delta_1\ \exists 1 \le i \le n, 1 \le j \le p\right)\\
&\le 2(np)\left\{\exp\left(-\frac{\delta_1^2}{72\tilde{\psi}_1^2(T,L) + 6\sqrt{2}\tilde{\psi}_1(T,L)\delta_1}\right) + L\exp\left(-\frac{\delta_1^2}{\tilde{\psi}_2(T,L)}\right)\right.\\
&\qquad\left. + \exp\left(-\frac{\delta_1^2}{72\lambda_{0,\max}\tilde{\psi}_3(L) + 6\sqrt{2\lambda_{0,\max}\tilde{\psi}_3(L)}\delta_1}\right)\right\}.
\end{aligned}$$

Let $\gamma_1 = \sqrt{2}/(12\times 144 d_0^2 3\sqrt{3\lambda_{0,\max}})$ and $\gamma_3 = 1/(72\lambda_{0,\max}\times(144 d_0^2\sqrt{3\lambda_{0,\max}})^2)$. If $\tilde{\psi}_1 < \gamma_1\cdot\delta/M^{1+\beta}$ and $\tilde{\psi}_3 < \gamma_3\cdot\delta^2/M^{2+2\beta}$, then $72\tilde{\psi}_1^2 < 6\sqrt{2}\tilde{\psi}_1\delta_1$ and $72\lambda_{0,\max}\tilde{\psi}_3 < 6\sqrt{2\lambda_{0,\max}\tilde{\psi}_3}\delta_1$,

which implies that

$$
\begin{aligned}
&P(\bar{A}_1) \\
&\leq 2np\left\{\exp\left(-\frac{\delta_1}{12\sqrt{2}\tilde{\psi}_1(T,L)}\right)+\exp\left(-\frac{\delta_1}{12\sqrt{2\lambda_{0,\max}}\sqrt{\tilde{\psi}_3(L)}}\right)+L\exp\left(-\frac{\delta_1^2}{\tilde{\psi}_2(T,L)}\right)\right\} \\
&\overset{(i)}{\leq} 2np\left\{\exp\left(-\frac{\delta_1}{12\sqrt{2}}\Phi(T,L)\right)+\exp\left(-\frac{\delta_1}{12\sqrt{2\lambda_{0,\max}}}\Phi(T,L)\right)+L\exp\left(-\frac{\delta_1^2}{\tilde{\psi}_2(T,L)}\right)\right\} \\
&\overset{(ii)}{\leq} 4np\exp\left(-\frac{\delta_1}{12\sqrt{2\lambda_{0,\max}}}\Phi(T,L)\right)+2npL\exp\left(-\frac{\delta_1^2}{\tilde{\psi}_2(T,L)}\right) \\
&= 4np\exp\left(-\frac{1}{1728\sqrt{6}\lambda_{0,\max}d_0^2}\cdot\frac{\delta}{M^{1+\beta}}\cdot\Phi(T,L)\right) \\
&\quad+2npL\exp\left(-\frac{\delta^2}{6228d_0^4\lambda_{0,\max}M^{2+2\beta}\tilde{\psi}_2(T,L)}\right),
\end{aligned}
\tag{B.11}
$$

where (i) follows the definition of $\Phi(T,L)$ and (ii) follows the fact that $\lambda_{0,\max}>1$.

Next, we bound $P(\bar{A}_4)$. For any two real values z_1, z_2 and any positive integer k, we have

$$
(z_1+z_2)^k \leq (|z_1|+|z_2|)^k = 2^k\left(\frac{1}{2}|z_1|+\frac{1}{2}|z_2|\right)^k \leq 2^{k-1}\left(|z_1|+|z_2|\right),
$$

where the last line follows from Jensen's inequality. Since $\mathbb{E}[\|g_{ij}\|^2]=\lambda_{j0}$, $i=1,\ldots,n$, $j=1,2\ldots,p$, then, by Jensen's inequality and Lemma 41, for any $k\geq 2$, we have

$$
\mathbb{E}\left[\left(\|g_{ij}\|^2-\lambda_{j0}\right)^k\right] \leq 2^{k-1}\left(\mathbb{E}\left[\|g_{ij}\|^{2k}+\lambda_{j0}^k\right]\right) \leq 2^{k-1}\left((2\lambda_{j0})^k k!+\lambda_{j0}^k\right) \leq (4\lambda_{j0})^k k!\,.
$$

Thus,

$$
\sum_{i=1}^n \mathbb{E}\left[\left(\|g_{ij}\|^2-\lambda_{j0}\right)^k\right] \leq \frac{k!}{2}n\times(32\lambda_{j0}^2)\times(4\lambda_{j0})^{k-2}.
$$

Then by Lemma 39, for any $\epsilon>0$, we have

$$
P\left(\left|\frac{1}{n}\sum_{i=1}^n\|g_{ij}\|^2-\lambda_{j0}\right|>\epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{64\lambda_{j0}^2+8\lambda_{j0}\epsilon}\right).
$$

Finally,

$$
\begin{aligned}
P\left(\frac{1}{n}\sum_{i=1}^n\|g_{ij}\|^2>2\lambda_{0,\max}\right) &\leq P\left(\frac{1}{n}\sum_{i=1}^n\|g_{ij}\|^2>2\lambda_{j0}\right) \\
&\leq P\left(\left|\frac{1}{n}\sum_{i=1}^n\|g_{ij}\|^2-\lambda_{j0}\right|>\lambda_{j0}\right) \\
&\leq 2\exp\left(-\frac{n}{72}\right)
\end{aligned}
$$

and

$$P(\bar{A}_4) = P\left(\frac{1}{n}\sum_{i=1}^{n}\|g_{ij}\|^2 > 2\lambda_{0,\max},\ \exists j = 1,\dots,p\right) \le 2p\exp\left(-\frac{n}{72}\right). \tag{B.12}$$

Next, we bound $P(\bar{A}_2)$. Let $\hat{K}^g_{jj}(s,t) = \frac{1}{n}\sum_{i=1}^{n} g_{ij}(s)g_{ij}(t)$ and $K_{jj}(s,t) = \mathbb{E}[g_{ij}(s)g_{ij}(t)]$, $j \in V$, and define

$$A_2' :\ \|\hat{K}^g_{jj} - K^g_{jj}\|_{\mathrm{HS}} \le \delta_2 \quad \forall j = 1,\dots,p.$$

Note that

$$\begin{aligned}
&\|\hat{K}^g_{jj}(s,t) - K^g_{jj}(s,t)\|_{\mathrm{HS}} \\
&= \left\|\frac{1}{n}\sum_{i=1}^{n}\left[\hat{g}_{ij}(s) - g_{ij}(s) + g_{ij}(s)\right]\left[\hat{g}_{ij}(t) - g_{ij}(t) + g_{ij}(t)\right] - K^g_{jj}(s,t)\right\|_{\mathrm{HS}} \\
&\le \frac{1}{n}\sum_{i=1}^{n}\|\hat{g}_{ij} - g_{ij}\|^2 + \frac{2}{n}\sum_{i=1}^{n}\|\hat{g}_{ij} - g_{ij}\|\cdot\|g_{ij}\| + \left\|\frac{1}{n}\sum_{i=1}^{n}\left[g_{ij}(s)g_{ij}(t) - K^g_{jj}(s,t)\right]\right\|_{\mathrm{HS}}.
\end{aligned}$$

Let

$$A_6 :\ \left\|\frac{1}{n}\sum_{i=1}^{n}\left[g_{ij}(s)g_{ij}(t) - K^g_{jj}(s,t)\right]\right\|_{\mathrm{HS}} \le 4\lambda_{0,\max}\delta_1,\ \forall j = 1,\dots,p.$$

We show that $A_1 \cap A_4 \cap A_6 \Longrightarrow A_2'$. By Jensen's inequality, we have

$$\frac{1}{n}\sum_{i=1}^{n}\|g_{ij}\| \le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|g_{ij}\|^2}.$$

On the event A_4, we have $(1/n)\sum_{i=1}^{n}\|g_{ij}\| \le \sqrt{2\lambda_{0,\max}}$ for any $j = 1,\dots,p$. When A_1, A_4, and A_6 hold simultaneously, we have

$$\|\hat{K}^g_{jj}(s,t) - K^g_{jj}(s,t)\|_{\mathrm{HS}} \le \delta_1^2 + 2\sqrt{2\lambda_{0,\max}}\delta_1 + 4\lambda_{0,\max}\delta_1 \le 9\lambda_{0,\max}\delta_1,$$

which is A_2. Therefore, $A_1 \cap A_4 \cap A_6 \Longrightarrow A_2'$, which implies that $\bar{A}'_2 \Longrightarrow \bar{A}_1 \cup \bar{A}_4 \cup \bar{A}_6$ and $P(\bar{A}'_2) \le P(\bar{A}_1) + P(\bar{A}_4) + P(\bar{A}_6)$. We upper bound $P(\bar{A}_6)$ next.

By Lemma 42, for any $j = 1,\dots,p$, we have

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\left[g_{ij}(s)g_{ij}(t) - K^g(s,t)\right]\right\|_{\mathrm{HS}} > 4\lambda_{0,\max}\delta_1\right) \le 2\exp\left(-\frac{n\delta_1^2}{6}\right).$$

Thus,

$$P(\bar{A}_6) \le 2p\exp\left(-\frac{n\delta_1^2}{6}\right) = 2p\exp\left(-\frac{1}{373248 d_0^4\lambda_{0,\max}^2} \times n\frac{\delta^2}{M^{2+2\beta}}\right). \tag{B.13}$$

Combining (B.11), (B.12), and (B.13), we have

$$P(\bar{A}'_2) \leq 4pM \exp\left(-\frac{1}{1728\sqrt{6}\lambda_{0,\max}d_0^2} \cdot \frac{\delta}{M^{1+\beta}} \cdot \Phi(T,L)\right) + 2p\exp\left(-\frac{n}{72}\right) + 2p\exp\left(-\frac{1}{373248 d_0^4 \lambda_{0,\max}^2} \times n\frac{\delta^2}{M^{2+2\beta}}\right).$$

Finally, $P(\bar{A}_2) \leq P(\bar{A}'_{X,2}) + P(\bar{A}'_{Y,2})$, where $A'_{X,2}$ and $A'_{Y,2}$ are defined similarly to A'_2 with g being X and Y, since

$$\|\hat{K}_j j(s,t) - K_{jj}(s,t)\|_{\text{HS}} \leq \|\hat{K}_j^X j(s,t) - K_{jj}^X(s,t)\|_{\text{HS}} + \|\hat{K}_j^Y j(s,t) - K_{jj}^Y(s,t)\|_{\text{HS}}.$$

Thus, we have

$$P(\bar{A}_2) \leq 8pM \exp\left(-\frac{1}{1728\sqrt{6}\lambda_{0,\max}d_0^2} \cdot \frac{\delta}{M^{1+\beta}} \cdot \Phi(T,L)\right) + 4p\exp\left(-\frac{n}{72}\right) + 4p\exp\left(-\frac{1}{373248 d_0^4 \lambda_{0,\max}^2} \times n\frac{\delta^2}{M^{2+2\beta}}\right).$$

For $P(\bar{A}_3)$, note that $\sum_{i=1}^n \xi_{ijk}^2 \sim \chi_n^2$ for any $j = 1, \ldots, p$ and $k = 1, \ldots, M$. By Pages 28-29 of Boucheron et al. (2013), for any $\epsilon > 0$, we have

$$P\left(\frac{1}{n}\sum_{i=1}^n \xi_{ijk}^2 - 1 > \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{4+4\epsilon}\right).$$

Letting $\epsilon = 1/2$, we have

$$P(\bar{A}_3) \leq pM \exp\left(-\frac{n}{24}\right).$$

Finally, we upper bound $P(\bar{A}_5)$. Note that

$$\begin{aligned}\mathbb{E}\left[(a_{ijk}a_{ilm} - \mathbb{E}(a_{ijk}a_{ilm}))^k\right] &= \lambda_{jk}^{k/2}\lambda_{lm}^{k/2}\mathbb{E}\left[(\xi_{ijk}\xi_{ilm} - \mathbb{E}(\xi_{ijk}\xi_{ilm}))^k\right] \\ &\leq d_0^k \mathbb{E}\left[(\xi_{ijk}\xi_{ilm} - \mathbb{E}(\xi_{ijk}\xi_{ilm}))^k\right],\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[(\xi_{ijk}\xi_{ilm} - \mathbb{E}(\xi_{ijk}\xi_{ilm}))^k\right] &\leq 2^{k-1}\left(\mathbb{E}\left[|\xi_{ijk}\xi_{ilm}|^k\right] + |\mathbb{E}(\xi_{ijk}\xi_{ilm})|^k\right) \\ &\leq 2^{k-1}\left(\mathbb{E}[\xi_{ij1}^{2k}] + 1\right) \\ &\leq 2^{k-1}(2^k k! + 1) \\ &\leq 4^k k!.\end{aligned}$$

Thus

$$\mathbb{E}\left[(a_{ijk}a_{ilm} - \mathbb{E}(a_{ijk}a_{ilm}))^k\right] \leq (4d_0)^k k!$$

and Lemma 39 tells us that for any $1 \leq j, l \leq p$ and $1 \leq k, m \leq M$, we have

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n a_{ijk}a_{ilm} - \sigma_{jl,km}\right| > \frac{\delta}{16}\right) \leq 2\exp\left(-\frac{n\delta^2}{16512 d_0^2}\right).$$

Therefore,

$$P\left(\bar{A}_5\right) \leq 2(pM)^2 \exp\left(-\frac{n\delta^2}{16512 d_0^2}\right).$$

Let $C_1 = 12$, $C_2 = 1/(1728\sqrt{6}\lambda_{0,\max})$, $C_3 = 9$, $C_4 = 1/(373248 d_0^4 \lambda_{0,\max}^2)$, $C_5 = 2$, and $C_6 = 1/(6228 d_0^4 \lambda_{0,\max})$. The final result follows by combining the upper bounds on $P(\bar{A}_w)$, $w = 1, \ldots, 5$.

C. Additional Results

In this section, we establish additional results that are needed to prove the main results.

C.1 Theorem 14 and Its Proof

We give a non-asymptotic error bound on the function estimated using the basis expansion, which is subsequently used to establish Theorem 13.

For a random function $g(t) \in \mathbb{H}$, where $t \in \mathcal{T}$, $\mathcal{T}$ is a closed interval of the real line, and $\mathbb{H}$ is a separable Hilbert space, we have noisy discrete observations at time points $t_1, t_2, \ldots, t_T$ generated from the model below:

$$h_k = g(t_k) + \epsilon_k,$$

where $\epsilon_k \overset{\text{i.i.d.}}{\sim} N(0, \sigma_0^2)$, $k = 1, \ldots, T$. Let $b(t) = (b_1(t), b_2(t), \ldots, b_L(t))^\top$ be the vector of the basis functions. Let $\hat{g}(t) = \hat{\beta}^\top b(t)$ be the estimator of $g(t)$, where $\hat{\beta} \in \mathbb{R}^L$ is obtained by minimizing the least square loss:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^L}{\arg\min} \sum_{k=1}^{T} \left(\beta^\top b(t_k) - h_k \right)^2.$$

We define the design matrix B as

$$B = \begin{bmatrix} b_1(t_1) & \cdots & b_L(t_1) \\ \vdots & \ddots & \vdots \\ b_1(t_T) & \cdots & b_L(t_T) \end{bmatrix} \in \mathbb{R}^{T \times L},$$

so that

$$\hat{\beta} = \left(B^\top B \right)^{-1} B^\top h,$$

where $h = (h_1, h_2, \ldots, h_T)^\top \in \mathbb{R}^T$.

We assume that $g(t) = \sum_{m=1}^{\infty} \beta_m^* b_m(t)$, and we can decompose $g(t)$ as $g = g^{\shortparallel} + g^{\perp}$, where $g^{\shortparallel} \in \text{Span}(b)$ and $g^{\perp} \in \text{Span}(b)^{\perp}$. Let $\lambda_0 \coloneqq \mathbb{E}[\|g\|^2]$ and $\lambda_0^{\perp} \coloneqq \mathbb{E}[\|g^{\perp}\|^2]$. It is then easy to check that $\lambda_0 = \sum_{m=1}^{\infty} \mathbb{E}[(\beta_m^*)^2]$ and $\lambda_0^{\perp} = \sum_{m>L}^{\infty} \mathbb{E}[(\beta_m^*)^2]$. We assume that the basis functions $\{b_l(t)\}_{l=1}^{\infty}$ make up a complete orthonormal system (CONS) of $\mathbb{H}$, that is, $\overline{\text{Span}\left(\{b_l\}_{l=1}^{\infty}\right)} = \mathbb{H}$ (see Definition 2.4.11 of Hsing and Eubank (2015)), and have continuous derivative functions with

$$D_{0,b} \coloneqq \sup_{l \geq 1} \sup_{t \in \mathcal{T}} |b_l(t)| < \infty, \qquad D_{1,b}(l) \coloneqq \sup_{t \in \mathcal{T}} |b_l'(t)| < \infty, \qquad D_{1,b,L} \coloneqq \max_{1 \leq l \leq L} D_{1,b}(l).$$

We further assume that the observation time points $\{t_k : 1 \leq k \leq T\}$ satisfy

$$\max_{1 \leq k \leq T+1} \left| \frac{t_k - t_{(k-1)}}{|\mathcal{T}|} - \frac{1}{T} \right| \leq \frac{\zeta_0}{T^2},$$

where t_0 and $t_{(T+1)}$ are endpoints of $\mathcal{T}$ and ζ_0 is a positive constant. We further assume that $\sum_{m=1}^{\infty} \mathbb{E}\left[(\beta_m^*)^2\right] D_{1,b}^2(m) < \infty$, and we define

$$\psi_4(L) = \sum_{m>L} \mathbb{E}\left[(\beta_m^*)^2\right] D_{1,b}^2(m).$$

Let

$$\psi_1(T,L)=\frac{\sigma_0 L}{\sqrt{\lambda_{\min}\left(B^\top B\right)}},\qquad \psi_3(L)=\lambda_0^{\perp}/\lambda_0,$$

and

$$\psi_2(T,L)=\frac{1}{(\lambda_{\min}^B)^2}\left(18\lambda_0\left[D_{0,b}^2(\zeta_0+1)^4|\mathcal{T}|^2D_{1,b,L}^2+2D_{0,b}^4(2\zeta_0+1)^2\right]L^2\psi_3(L)\right.$$
$$\left.+D_{0,b}^2(\zeta_0+1)^4|\mathcal{T}|^2L^2\psi_4(L)\right),$$

Then we have the following theorem.

Theorem 14 *For any $\delta>0$, we have*

$$P\left(\|g-\hat{g}\|>\delta\right)\le 2\exp\left(-\frac{\delta^2}{72\psi_1^2(T,L)+6\sqrt{2}\psi_1(T,L)\delta}\right)+L\exp\left(-\frac{\delta^2}{\psi_2(T,L)}\right)$$
$$+2\exp\left(-\frac{\delta^2}{72\lambda_0\psi_3(L)+6\sqrt{2\lambda_0}\sqrt{\psi_3(L)}\delta}\right).$$

Proof For a fixed g, since $\overline{\text{Span}\left(\{b_l\}_{l=1}^{\infty}\right)}=\mathbb{H}$, we can assume that $g(t)=\sum_{l=1}^{\infty}\beta_l^* b_l(t)$ where $\beta_l^*=\langle g,b_l\rangle=\int_{\mathcal{T}}g(t)b_l(t)dt$. Let $\beta^*=(\beta_1^*,\cdots,\beta_L^*)^\top\in\mathbb{R}^L$. We have $g^{\shortparallel}(t)=(\beta^*)^\top b(t)=\sum_{l=1}^{L}\beta_l^*b_l(t)$ and $g^{\perp}(t)=\sum_{l>L}\beta_l^*b_l(t)$. Thus, we have

$$h_k=g(t_k)+\epsilon_k=(\beta^*)^\top b(t_k)+g^{\perp}(t_k)+\epsilon_k.$$

Let $h^{\perp}=\left(g^{\perp}(t_1),g^{\perp}(t_2),\ldots,g^{\perp}(t_T)\right)^\top$ and $\epsilon=(\epsilon_1,\epsilon_2,\ldots,\epsilon_T)^\top$, so that $h=B\beta^*+h^{\perp}+\epsilon$. Then, $\mathbb{E}(\hat{\beta})=\beta^*+\left(B^\top B\right)^{-1}B^\top h^{\perp}$ and

$$\hat{g}(t)-g(t)=\hat{g}(t)-g^{\shortparallel}(t)-g^{\perp}(t)=\hat{g}(t)-(\beta^*)^\top b(t)-g^{\perp}(t)$$
$$=\left(\hat{\beta}-\mathbb{E}(\hat{\beta})\right)^\top b(t)+\left(\left(B^\top B\right)^{-1}B^\top h^{\perp}\right)^\top b(t)-g^{\perp}(t).$$

By Lemma 36, we then have

$$\begin{aligned}\|\hat{g}-g\|&\le\left\|\left(\hat{\beta}-\mathbb{E}(\hat{\beta})\right)^\top b(t)\right\|+\left\|\left(\left(B^\top B\right)^{-1}B^\top h^{\perp}\right)^\top b(t)\right\|+\left\|g^{\perp}\right\|\\&\le\left|\hat{\beta}-\mathbb{E}(\hat{\beta})\right|_2\times\|b\|_{\mathcal{L}^2,2}+\left|\left(B^\top B\right)^{-1}B^\top h^{\perp}\right|_2\times\|b\|_{\mathcal{L}^2,2}+\left\|g^{\perp}\right\|\\&\le\left|\hat{\beta}-\mathbb{E}(\hat{\beta})\right|_2\times\|b\|_{\mathcal{L}^2,2}+\frac{1}{\lambda_{\min}(B^\top B)}\times\left|B^\top h^{\perp}\right|_2\times\|b\|_{\mathcal{L}^2,2}+\left\|g^{\perp}\right\|.\end{aligned}$$

Let

$$J_1=\left|\hat{\beta}-\mathbb{E}(\hat{\beta})\right|_2\times\|b\|_{\mathcal{L}^2,2},\quad J_2=\frac{1}{\lambda_{\min}(B^\top B)}\times|B^\top h^{\perp}|_2\times\|b\|_{\mathcal{L}^2,2},\quad J_3=\|g^{\perp}\|,$$

where $|\mathcal{T}|$ denotes the length of the interval, then $\|\hat{g} - g\| \leq J_1 + J_2 + J_3$. This equation holds with probability one, since it holds for any $g \in \mathbb{H}$. We then bound J_1, J_2, and J_3 individually.

We bound J_1. Recall that $\|b\|_{\mathcal{L}^2,2} = \sqrt{L}$ and $\psi_1(T,L) = \sigma_0\|b\|_{\mathcal{L}^2,2}\sqrt{L}/\sqrt{\lambda_{\min}\left(B^\top B\right)}$. Treating g as fixed, we have $\hat{\beta} \sim N_L\left(\mathbb{E}(\hat{\beta}), \sigma_0^2\left(B^\top B\right)^{-1}\right)$ and

$$\frac{1}{\sigma_0}\left(B^\top B\right)^{1/2}\left(\hat{\beta} - \mathbb{E}(\hat{\beta})\right) \sim N_L\left(0, I_L\right).$$

Since

$$\begin{aligned} J_1 = \left|\hat{\beta} - \mathbb{E}(\hat{\beta})\right|_2 \times \|b\|_{\mathcal{L}^2,2} &= \left|\left(B^\top B\right)^{-1/2}\left(B^\top B\right)^{1/2}\left(\hat{\beta} - \mathbb{E}(\hat{\beta})\right)\right|_2 \times \|b\|_{\mathcal{L}^2,2} \\ &\leq \frac{\sigma_0\|b\|_{\mathcal{L}^2,2}}{\sqrt{\lambda_{\min}\left(B^\top B\right)}}\left|\frac{1}{\sigma_0}\left(B^\top B\right)^{1/2}\left(\hat{\beta} - \mathbb{E}(\hat{\beta})\right)\right|_2, \end{aligned}$$

we have

$$\begin{aligned} P(J_1 > \delta) &\leq P\left(\left|\frac{1}{\sigma_0}\left(B^\top B\right)^{1/2}\left(\hat{\beta} - \mathbb{E}(\hat{\beta})\right)\right|_2 > \frac{\delta}{\sigma_0\|b\|_{\mathcal{L}^2,2}/\sqrt{\lambda_{\min}\left(B^\top B\right)}}\right) \\ &\overset{(i)}{\leq} 2\exp\left(-\frac{\delta^2}{8\psi_1^2(T,L) + 2\sqrt{2}\psi_1(T,L)\delta}\right), \end{aligned} \tag{C.1}$$

where (i) follows Lemma 38. The bound does not depend on g, so it holds when g is also random.

Next, we bound J_2. Let $(B^\top h^\perp)_l$ denote the l-th element of the vector $B^\top h^\perp$. Then

$$(B^\top h^\perp)_l = \sum_{k=1}^{T} b_l(t_k)g^\perp(t_k) = \sum_{m>L}\beta_m^* \sum_{k=1}^{T} b_l(t_k)b_m(t_k)$$

and $(B^\top h^\perp)_l$ follows a mean zero Gaussian distribution, since g is a Gaussian random function with mean zero. Furthermore,

$$\mathbb{E}\left[(B^\top h^\perp)_l^2\right] = \sum_{m>L}\mathbb{E}\left[\beta_m^{*2}\right]\left(\sum_{k=1}^{T} b_l(t_k)b_m(t_k)\right)^2 \tag{C.2}$$

From the definition of $D_{0,b}$, $D_{1,b}(\cdot)$, for any $l < m$, we have

$$\sup_{t\in\mathcal{T}}(b_l(t)b_m(t)) \leq D_{0,b}^2,$$

$$\sup_{t\in\mathcal{T}}(b_l(t)b_m(t))' = \sup_{t\in\mathcal{T}}\{b_l'(t)b_m(t) + b_l(t)b_m'(t)\} \leq D_{0,b}(D_{1,b}(l) + D_{1,b}(m)).$$

Note that $\int_{\mathcal{T}} b_l(t)b_m(t)dt = 0$, $l < m$. Then, by Lemma 40, for all $1 \leq l < m < \infty$, we have

$$\begin{aligned} \left|\frac{1}{T}\sum_{k=1}^{T} b_l(t_k)b_m(t_k)\right| &= \left|\frac{1}{T}\sum_{k=1}^{T} b_l(t_k)b_m(t_k) - \frac{1}{|\mathcal{T}|}\int_{\mathcal{T}} b_l(t)b_m(t)dt\right| \\ &\leq \frac{D_{0,b}(D_{1,b}(l) + D_{1,b}(m))(\zeta_0+1)^2|\mathcal{T}|/2 + D_{0,b}^2(2\zeta_0+1)}{T}, \end{aligned}$$

which implies that

$$\left|\sum_{k=1}^{T} b_l(t_k)b_m(t_k)\right| \leq \frac{1}{2}D_{0,b}(\zeta_0+1)^2|\mathcal{T}|(D_{1,b}(l)+D_{1,b}(m)) + D_{0,b}^2(2\zeta_0+1).$$

Then,

$$\begin{aligned}\left(\sum_{k=1}^{T} b_l(t_k)b_m(t_k)\right)^2 &\leq \left(\frac{1}{2}D_{0,b}(\zeta_0+1)^2|\mathcal{T}|(D_{1,b}(l)+D_{1,b}(m)) + D_{0,b}^2(2\zeta_0+1)\right)^2 \\ &\leq \frac{1}{2}D_{0,b}^2(\zeta_0+1)^4|\mathcal{T}|^2(D_{1,b}(l)+D_{1,b}(m))^2 + 2D_{0,b}^4(2\zeta_0+1)^2 \\ &\leq D_{0,b}^2(\zeta_0+1)^4|\mathcal{T}|^2(D_{1,b}^2(l)+D_{1,b}^2(m)) + 2D_{0,b}^4(2\zeta_0+1)^2\end{aligned}$$

and, by (C.2),

$$\begin{aligned}\mathbb{E}\left[(B^\top h^\perp)_l^2\right] &\leq \left[D_{0,b}^2(\zeta_0+1)^4|\mathcal{T}|^2D_{1,b}^2(l) + 2D_{0,b}^4(2\zeta_0+1)^2\right]\sum_{m>L}\mathbb{E}\left[\beta_m^{*2}\right] \\ &\quad + D_{0,b}^2(\zeta_0+1)^4|\mathcal{T}|^2\sum_{m>L}\mathbb{E}\left[\beta_m^{*2}\right]D_{1,b}^2(m) \\ &\leq \left[D_{0,b}^2(\zeta_0+1)^4|\mathcal{T}|^2D_{1,b}^2(l) + 2D_{0,b}^4(2\zeta_0+1)^2\right]\lambda_0^\perp + D_{0,b}^2(\zeta_0+1)^4|\mathcal{T}|^2\psi_4(L) \\ &\leq \left[D_{0,b}^2(\zeta_0+1)^4|\mathcal{T}|^2D_{1,b,L}^2 + 2D_{0,b}^4(2\zeta_0+1)^2\right]\lambda_0^\perp + D_{0,b}^2(\zeta_0+1)^4|\mathcal{T}|^2\psi_4(L) \\ &= \lambda_0\left[D_{0,b}^2(\zeta_0+1)^4|\mathcal{T}|^2D_{1,b,L}^2 + 2D_{0,b}^4(2\zeta_0+1)^2\right]\psi_3(L) \\ &\quad + D_{0,b}^2(\zeta_0+1)^4|\mathcal{T}|^2\psi_4(L).\end{aligned}$$

Using a tail bound for Gaussian random variable (e.g., Section 2.1.2 of Wainwright (2019)), we have

$$\begin{aligned}\mathbb{P}\left(J_2>\delta\right) &\leq \mathbb{P}\left(|B^\top h^\perp|_2 > \frac{\lambda_{\min}^B\delta}{\sqrt{L}}\right) \leq \mathbb{P}\left(\max_{1\leq l\leq L}(B^\top h^\perp)_l > \frac{\lambda_{\min}^B\delta}{L}\right) \\ &\leq L\exp\left(-\frac{9\delta^2}{\psi_2(T,L)}\right).\end{aligned} \tag{C.3}$$

Finally, we bound J_3. By Lemma 41 and the definition of $\psi_3(L)$, we have $\mathbb{E}\left[\|g^\perp\|^{2k}\right] \leq (2\lambda_0\psi_3(L))^k k!$. By Jensen's inequality, we have

$$\mathbb{E}\left[\|g^\perp\|^k\right] = \mathbb{E}\left[\sqrt{\|g^\perp\|^{2k}}\right] \leq \sqrt{\mathbb{E}\left[\|g^\perp\|^{2k}\right]} \leq \left(\sqrt{2\lambda_0\psi_3(L)}\right)^k k!.$$

Thus, by Lemma 39, we have

$$P\left(J_3>\delta\right) = P\left(\|g^\perp\|>\delta\right) \leq 2\exp\left(-\frac{\delta^2}{8\lambda_0\psi_3(L)+2\sqrt{2\lambda_0}\sqrt{\psi_3(L)}\delta}\right). \tag{C.4}$$

The final result follows from (C.1), (C.3), and (C.4). ■

D. Lemmas and Their Proofs

In this section, we introduce some useful lemmas along with their proofs.

Lemma 15 *Let* $\sigma_{\max} = \max\{|\Sigma^{X,M}|_\infty,\ |\Sigma^{Y,M}|_\infty\}$. *Suppose that*

$$|S^{X,M} - \Sigma^{X,M}|_\infty \le \delta, \qquad |S^{Y,M} - \Sigma^{Y,M}|_\infty \le \delta, \tag{D.1}$$

for some $\delta \ge 0$. *Then*

$$|(S^{Y,M} \otimes S^{X,M}) - (\Sigma^{Y,M} \otimes \Sigma^{X,M})|_\infty \le \delta^2 + 2\delta\sigma_{\max},$$

and

$$|\operatorname{vec}(S^{Y,M} - S^{X,M}) - \operatorname{vec}(\Sigma^{Y,M} - \Sigma^{X,M})|_\infty \le 2\delta. \tag{D.2}$$

Proof Note that for any $(j,l),(j',l') \in V^2$ and $1 \le k,k',m,m' \le M$, by (D.1), we have

$$\begin{aligned}
&\left|S^{X,M}_{jl,km}S^{Y,M}_{j'l',k'm'} - \Sigma^{X,M}_{jl,km}\Sigma^{Y,M}_{j'l',k'm'}\right| \\
&\le \left|S^{X,M}_{jl,km} - \Sigma^{X,M}_{jl,km}\right| \cdot \left|S^{Y,M}_{j'l',k'm'} - \Sigma^{Y,M}_{j'l',k'm'}\right| + \left|\Sigma^{X,M}_{jl,km}\right| \cdot \left|S^{Y,M}_{j'l',k'm'} - \Sigma^{Y,M}_{j'l',k'm'}\right| \\
&\quad + \left|\Sigma^{Y,M}_{j'l',k'm'}\right| \cdot \left|S^{X,M}_{jl,km} - \Sigma^{X,M}_{jl,km}\right| \\
&\le \left|S^{X,M} - \Sigma^{X,M}\right|_\infty \left|S^{Y,M} - \Sigma^{Y,M}\right|_\infty + \sigma_{\max}\left|S^{Y,M} - \Sigma^{Y,M}\right|_\infty + \sigma_{\max}\left|S^{X,M} - \Sigma^{X,M}\right|_\infty \\
&\le \delta^2 + 2\delta\sigma_{\max}.
\end{aligned}$$

For (D.2), note that

$$\begin{aligned}
\left|\operatorname{vec}(S^{Y,M} - S^{X,M}) - \operatorname{vec}(\Sigma^{Y,M} - \Sigma^{X,M})\right|_\infty &= \left|(S^{X,M} - \Sigma^{X,M}) - (S^{Y,M} - \Sigma^{Y,M})\right|_\infty \\
&\le |S^{X,M} - \Sigma^{X,M}|_\infty + |S^{Y,M} - \Sigma^{Y,M}|_\infty \\
&\le 2\delta.
\end{aligned}$$

This completes the proof. ■

Lemma 16 *Given* $Z^{(1)}, Z^{(2)}, A^{(1)}, A^{(2)} \in \mathbb{R}^{M\times M}$ *and* $\lambda > 0$, *let* $\{\hat{Z}^{(1)}, \hat{Z}^{(2)}\}$ *denote the solution of*

$$\underset{\{Z^{(1)},Z^{(2)}\}}{\arg\min}\ \frac{1}{2}\sum_{q=1}^{2}\|Z^{(q)} - A^{(q)}\|_F^2 + \lambda\|Z^{(1)} - Z^{(2)}\|_F. \tag{D.3}$$

If $\|A^{(1)} - A^{(2)}\|_F \le 2\lambda$, *then*

$$\hat{Z}^{(1)} = \hat{Z}^{(2)} = \frac{1}{2}\left(A^{(1)} + A^{(2)}\right). \tag{D.4}$$

If $\|A^{(1)} - A^{(2)}\|_F > 2\lambda$, *then*

$$\begin{aligned}
\hat{Z}^{(1)} &= A^{(1)} - \frac{\lambda}{\|A^{(1)} - A^{(2)}\|_F}\left(A^{(1)} - A^{(2)}\right), \\
\hat{Z}^{(2)} &= A^{(2)} + \frac{\lambda}{\|A^{(1)} - A^{(2)}\|_F}\left(A^{(1)} - A^{(2)}\right).
\end{aligned} \tag{D.5}$$

Proof The subdifferential of the objective function in (D.3) is

$$G^{(1)}(Z^{(1)}, Z^{(2)}) \coloneqq Z^{(1)} - A^{(1)} + \lambda T(Z^{(1)}, Z^{(2)}),$$
$$G^{(2)}(Z^{(1)}, Z^{(2)}) \coloneqq Z^{(2)} - A^{(2)} - \lambda T(Z^{(1)}, Z^{(2)}),$$

where

$$T(Z^{(1)}, Z^{(2)}) = \begin{cases} \frac{Z^{(1)} - Z^{(2)}}{\|Z^{(1)} - Z^{(2)}\|_{\mathrm{F}}} & \text{if } Z^{(1)} \neq Z^{(2)} \\ \left\{T \in \mathbb{R}^{M \times M} : \|T\|_{\mathrm{F}} \leq 1\right\} & \text{if } Z^{(1)} = Z^{(2)} \end{cases}.$$

The optimality condition is $0 \in G^{(q)}(Z^{(1)}, Z^{(2)})$.

Claim We have $\hat{Z}^{(1)} \neq \hat{Z}^{(2)}$ if and only if $\|A^{(1)} - A^{(2)}\|_{\mathrm{F}} > 2\lambda$.

We first prove the necessity, that is, when $\hat{Z}^{(1)} \neq \hat{Z}^{(2)}$, then $\|A^{(1)} - A^{(2)}\|_{\mathrm{F}} > 2\lambda$. By the optimality condition, we have

$$\hat{Z}^{(1)} - \hat{Z}^{(2)} - \left(A^{(1)} - A^{(2)}\right) - 2\lambda \frac{\hat{Z}^{(1)} - \hat{Z}^{(2)}}{\|\hat{Z}^{(1)} - \hat{Z}^{(2)}\|_{\mathrm{F}}} = 0,$$

which implies that

$$\|A^{(1)} - A^{(2)}\|_{\mathrm{F}} = 2\lambda + \|\hat{Z}^{(1)} - \hat{Z}^{(2)}\|_{\mathrm{F}} > 2\lambda.$$

We then prove the sufficiency, that is, when $\|A^{(1)} - A^{(2)}\|_{\mathrm{F}} > 2\lambda$, then $\hat{Z}^{(1)} \neq \hat{Z}^{(2)}$. By the optimality condition, we have $\hat{Z}^{(1)} + \hat{Z}^{(2)} = A^{(1)} + A^{(2)}$. If $\hat{Z}^{(1)} = \hat{Z}^{(2)}$, then $\hat{Z}^{(1)} = \hat{Z}^{(2)} = (A^{(1)} + A^{(2)})/2$. Furthermore, $\|\hat{Z}^{(1)} - A^{(1)}\|_{\mathrm{F}} = \|A^{(1)} - A^{(2)}\|_{\mathrm{F}}/2 = \lambda \|T(\hat{Z}^{(1)}, \hat{Z}^{(2)})\|_{\mathrm{F}} \leq \lambda$, which implies that $\|A^{(1)} - A^{(2)}\|_{\mathrm{F}} \leq 2\lambda$. This contradicts the assumption that $\|A^{(1)} - A^{(2)}\|_{\mathrm{F}} > 2\lambda$. Thus, we must have $\hat{Z}^{(1)} \neq \hat{Z}^{(2)}$.

By proving the claim, we have already established (D.4). We now prove (D.5). When $\|A^{(1)} - A^{(2)}\|_{\mathrm{F}} > 2\lambda$, according to the claim above, we must have $\hat{Z}^{(1)} \neq \hat{Z}^{(2)}$. Then

$$\hat{Z}^{(1)} - A^{(1)} + \frac{\lambda}{\|\hat{Z}^{(1)} - \hat{Z}^{(2)}\|_{\mathrm{F}}} \left(\hat{Z}^{(1)} - \hat{Z}^{(2)}\right) = 0,$$
$$\hat{Z}^{(2)} - A^{(2)} - \frac{\lambda}{\|\hat{Z}^{(1)} - \hat{Z}^{(2)}\|_{\mathrm{F}}} \left(\hat{Z}^{(1)} - \hat{Z}^{(2)}\right) = 0,$$

which implies that $\hat{Z}^{(1)} - \hat{Z}^{(2)} = \alpha \cdot \left(A^{(1)} - A^{(2)}\right)$, where α is a constant. The result in (D.5) follows by substituting the relationship back into the above display. ■

Lemma 17 *Let $|\cdot|_{1,2}$ be defined as in* (B.2)*, where $\mathcal{G} = \{G_t\}_{t=1,\ldots,N_{\mathcal{G}}}$ is a set of indices. For any matrix $A \in \mathbb{R}^{p^2M^2 \times p^2M^2}$ and $\theta \in \mathbb{R}^{p^2M^2}$, we have $|\theta^\top A \theta| \leq M^2 |A|_\infty |\theta|_{1,2}^2$.*

Proof By direct calculation, we have

$$|\theta^\top A \theta| \leq \sum_i \sum_j |A_{ij}\theta_i\theta_j| \leq |A|_\infty \left(\sum_i |\theta_i|\right)^2 = |A|_\infty \left(\sum_{t=1}^{N_{\mathcal{G}}} \sum_{k \in G_t} |\theta_k|\right)^2$$
$$= |A|_\infty \left(\sum_{t=1}^{N_{\mathcal{G}}} |\theta_{G_t}|_1\right)^2 \leq |A|_\infty \left(\sum_{t=1}^{N_{\mathcal{G}}} M |\theta_{G_t}|_2\right)^2 = M^2 |A|_\infty |\theta|_{1,2}^2,$$

where we use that for any vector $v \in \mathbb{R}^n$, $|v|_1 \leq \sqrt{n}|v|_2$. ■

Lemma 18 *Suppose that $\mathcal{M}$ is defined as in* (B.1)*. For any $\theta \in \mathcal{M}$, we have $|\theta|_{1,2} \leq \sqrt{s}|\theta|_2$. Furthermore, for $\Psi(\mathcal{M})$ as defined in* (B.4)*, we have $\Psi(\mathcal{M}) = \sqrt{s}$.*

Proof By the definitions of $\mathcal{M}$ and $|\cdot|_{1,2}$, we have

$$|\theta|_{1,2} = \sum_{t \in S_{\mathcal{G}}} |\theta_{G_t}|_2 + \sum_{t \notin S_{\mathcal{G}}} |\theta_{G_t}|_2 = \sum_{t \in S_{\mathcal{G}}} |\theta_{G_t}|_2 \leq \sqrt{s} \left(\sum_{t \in S_{\mathcal{G}}} |\theta_{G_t}|_2^2 \right)^{\frac{1}{2}} = \sqrt{s}|\theta|_2.$$

To show $\Psi(\mathcal{M}) = \sqrt{s}$, it suffices to show that the upper bound above can be achieved. Select $\theta \in \mathbb{R}^{p^2M^2}$ so that $|\theta_{G_t}|_2 = c$, $\forall t \in S_{\mathcal{G}}$, where c is some positive constant. This implies that $|\theta|_{1,2} = sc$ and $|\theta|_2 = \sqrt{s}c$ so that $|\theta|_{1,2} = \sqrt{s}|\theta|_2$. Thus, $\Psi(\mathcal{M}) = \sqrt{s}$. ■

Lemma 19 *Let $\mathcal{R}(\cdot)$ be the norm defined in* (B.2)*. Its dual norm $\mathcal{R}^*(\cdot)$, defined in* (B.3)*, is*

$$\mathcal{R}^*(v) \;=\; \max_{t=1,\ldots,N_{\mathcal{G}}} |v_{G_t}|_2.$$

Proof For any u satisfying $|u|_{1,2} \leq 1$ and $v \in \mathbb{R}^{p^2M^2}$, we have

$$\begin{aligned}\langle v, u \rangle = \sum_{t=1}^{N_{\mathcal{G}}} \langle v_{G_t}, u_{G_t} \rangle \leq \sum_{t=1}^{N_{\mathcal{G}}} |v_{G_t}|_2 |u_{G_t}|_2 &\leq \left(\max_{t=1,2,\cdots,N_{\mathcal{G}}} |v_{G_t}|_2 \right) \sum_{t=1}^{N_{\mathcal{G}}} |u_{G_t}|_2 \\ &= \left(\max_{t=1,2,\cdots,N_{\mathcal{G}}} |v_{G_t}|_2 \right) |u|_{1,2} \leq \max_{t=1,2,\cdots,N_{\mathcal{G}}} |v_{G_t}|_2.\end{aligned}$$

We show that this upper bound can be obtained. Let $t^* = \arg\max_{t=1,2,\cdots,N_{\mathcal{G}}} |v_{G_t}|$ and set u such that

$$u_{G_t} = \begin{cases} 0 & t \neq t^* \\ \frac{v_{G_{t^*}}}{|v_{G_{t^*}}|_2} & t = t^* \end{cases}.$$

Then $|u|_{1,2} = 1$ and $\langle v, u \rangle = |v_{G_{t^*}}|_2 = \max_{t=1,\ldots,N_{\mathcal{G}}} |v_{G_t}|_2$. ■

Lemma 20 *Given that A1-A5 hold, we have $|I_1| \leq \delta/16$ for all $1 \leq j, l \leq p,\ 1 \leq k, m \leq M$.*

Proof This follows directly from the assumption that A_5 is true. ■

Lemma 21 *Given that A1-A5 hold, we have $|I_2| \leq \delta/16$ for all $1 \leq j, l \leq p,\ 1 \leq k, m \leq M$.*

Proof We have

$$\begin{aligned}|I_2| &= \left|\langle \frac{1}{n}\sum_{i=1}^{n} a_{ijk}(\hat{g}_{il} - g_{il}), \phi_{lm}\rangle\right| \leq \left\|\frac{1}{n}\sum_{i=1}^{n} a_{ijk}(\hat{g}_{il} - g_{il})\right\| \\ &\overset{(i)}{\leq} \sqrt{\frac{1}{n}\sum_{i=1}^{n} a_{ijk}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\hat{g}_{il} - g_{il}\|^2} \overset{(ii)}{\leq} \delta_1\sqrt{\frac{1}{n}\sum_{i=1}^{n} a_{ijk}^2} = \delta_1\lambda_{jk}^{1/2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\xi_{ijk}^2} \\ &\overset{(iii)}{\leq} \sqrt{\frac{3}{2}}\delta_1\lambda_{jk}^{1/2} \leq \sqrt{\frac{3}{2}}\sqrt{d_1}\delta_1 k^{-\beta/2} \leq \sqrt{\frac{3}{2}}\sqrt{d_1}\delta_1,\end{aligned}$$

where (i) follows Lemma 36, (ii) follows A_1, (iii) follows A_3. From the definition of d_0, we have $|I_2| \leq \sqrt{3/2}d_0\delta_1$. Since

$$\delta_1 = \delta/\left(144d_0^2M^{1+\beta}\sqrt{3\lambda_{0,\max}}\right) \leq \delta/(8\sqrt{6}d_0), \tag{D.6}$$

we have

$$\sqrt{\frac{3}{2}}d_0\delta_1 \leq \sqrt{\frac{3}{2}}d_0 \cdot \frac{\delta}{8\sqrt{6}d_0} = \frac{\delta}{16}. \tag{D.7}$$

Thus, $|I_2| \leq \delta/16$. ■

Lemma 22 *Given that A1-A5 hold, we have* $|I_3| \leq \delta/16$ *for all* $1 \leq j, l \leq p,\ 1 \leq k, m \leq M$.

Proof We have

$$\begin{aligned}|I_3| &= \left|\langle \frac{1}{n}\sum_{i=1}^{n} a_{ijk}g_{il}, \hat{\phi}_{lm} - \phi_{lm}\rangle\right| \leq \lambda_{jk}^{1/2}\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_{ijk}g_{il}\right\|\left\|\hat{\phi}_{lm} - \phi_{lm}\right\| \\ &\overset{(i)}{\leq} \lambda_{jk}^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_{ijk}^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\|g_{il}\|^2\right)^{1/2}\|\hat{\phi}_{lm} - \phi_{lm}\| \\ &\overset{(ii)}{\leq} \lambda_{jk}^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_{ijk}^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\|g_{il}\|^2\right)^{1/2} d_{lm}\|\hat{K}_{ll} - K_{ll}\|_{\mathrm{HS}},\end{aligned}$$

where (i) follows Lemma 36, and (ii) follows Lemma 37. Since $\lambda_{jk}^{1/2} \leq \sqrt{d_1}k^{-\beta/2}$, $d_{lm} \leq d_2m^{1+\beta}$, and A_2-A_4 hold, we have

$$|I_3| \leq \sqrt{d_1}d_2k^{-\beta/2}m^{1+\beta}\sqrt{\frac{3}{2}}\sqrt{2\lambda_{0,\max}}\delta_2 \leq d_0^2M^{1+\beta}\sqrt{3\lambda_{0,\max}}\delta_2.$$

By the definition of δ_2, we have

$$d_0^2M^{1+\beta}\sqrt{3\lambda_{0,\max}}\delta_2 \leq d_0^2M^{1+\beta}\sqrt{3\lambda_{0,\max}} \times \frac{\delta}{16d_0^2M^{1+\beta}\sqrt{3\lambda_{0,\max}}} = \frac{\delta}{16}. \tag{D.8}$$

Thus, $|I_3| \leq \delta/16$. ■

Lemma 23 *Given that A1-A5 hold, we have* $|I_4| \leq \delta/16$ *for all* $1 \leq j, l \leq p,\ 1 \leq k, m \leq M$.

Proof We have

$$\begin{aligned}
|I_4| &\leq \lambda_{jk}^{1/2} \frac{1}{n} \|\sum_{i=1}^{n} \xi_{ijk} \left(\hat{g}_{il} - g_{il}\right)\| \|\hat{\phi}_{lm} - \phi_{lm}\| \\
&\overset{(i)}{\leq} \lambda_{jk}^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} \xi_{ijk}^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} \|\hat{g}_{il} - g_{il}\|^2\right)^{1/2} \|\hat{\phi}_{lm} - \phi_{lm}\| \\
&\overset{(ii)}{\leq} \lambda_{jk}^{1/2} d_{lm} \left(\frac{1}{n}\sum_{i=1}^{n} \xi_{ijk}^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} \|\hat{g}_{il} - g_{il}\|^2\right)^{1/2} \|\hat{K}_{ll} - K_{ll}\|_{\mathrm{HS}},
\end{aligned}$$

where (i) follows Lemma 36, and (ii) follows Lemma 37. Since $\lambda_{jk}^{1/2} \leq \sqrt{d_1} k^{-\beta/2}$, $d_{lm} \leq d_2 m^{1+\beta}$, and A_1-A_3 hold, we have

$$\begin{aligned}
|I_4| &\leq \sqrt{\frac{3}{2}} \sqrt{d_1} d_2 k^{-\beta/2} m^{1+\beta} \delta_1 \delta_2 \leq \sqrt{\frac{3}{2}} d_0^2 M^{1+\beta} \delta_1 \delta_2 \\
&\overset{(iii)}{\leq} \frac{\delta}{16} \times \frac{\sqrt{\frac{3}{2}} d_0^2 M^{1+\beta} \delta_1 \delta_2}{\sqrt{\frac{3}{2}} d_0 \delta_1} \leq \frac{\delta}{16} \times \frac{\delta}{16 d_0 \sqrt{3\lambda_{0,\max}}} \leq \frac{\delta}{16},
\end{aligned}$$

where (iii) follows (D.7). ■

Lemma 24 *Given that A1-A5 hold, we have* $|I_5| \leq \delta/16$ *for all* $1 \leq j, l \leq p,\ 1 \leq k, m \leq M$.

Proof This proof is similar to the proof of Lemma 21, thus is omitted. ■

Lemma 25 *Given that A1-A5 hold, we have* $|I_6| \leq \delta/16$ *for all* $1 \leq j, l \leq p,\ 1 \leq k, m \leq M$.

Proof We have

$$\begin{aligned}
|I_6| &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\langle \hat{g}_{ij} - g_{ij}, \phi_{jk} \rangle|^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} |\langle \hat{g}_{il} - g_{il}, \phi_{lm} \rangle|^2} \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\hat{g}_{ij} - g_{ij}\|^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\hat{g}_{il} - g_{il}\|^2}.
\end{aligned}$$

By the assumption that A_1 holds, we have $|I_6| \leq \delta_1^2$. By (D.6),(D.7) and Lemma 21, we have

$$\delta_1^2 \leq \frac{\delta}{16} \times \frac{\delta_1^2}{\sqrt{\frac{3}{2}} d_0 \delta_1} \leq \frac{\delta}{16}, \tag{D.9}$$

which completes the proof. ■

Lemma 26 *Given that A1-A5 hold, we have* $|I_7| \leq \delta/16$ *for all* $1 \leq j, l \leq p,\ 1 \leq k, m \leq M$.

Proof We have

$$\begin{aligned}
|I_7| &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}|\langle \hat{g}_{ij} - g_{ij}, \phi_{jk}\rangle|^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}|\langle g_{il}, \hat{\phi}_{lm} - \phi_{lm}\rangle|^2} \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\hat{g}_{ij} - g_{ij}\|^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|g_{il}\|^2\|\hat{\phi}_{lm} - \phi_{lm}\|^2} \\
&\overset{(i)}{\leq} \delta_1\|\hat{\phi}_{lm} - \phi_{lm}\| \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|g_{il}\|^2} \overset{(ii)}{\leq} \delta_1\sqrt{2\lambda_{0,\max}} \cdot \|\hat{\phi}_{lm} - \phi_{lm}\| \\
&\overset{(iii)}{\leq} \delta_1\sqrt{2\lambda_{0,\max}}d_{lm}\|\hat{K}_{ll} - K_{ll}\|_{\mathrm{HS}} \overset{(iv)}{\leq} \delta_1\delta_2\sqrt{2\lambda_{0,\max}}d_{lm} \leq d_0\sqrt{2\lambda_{0,\max}}M^{1+\beta}\delta_1\delta_2,
\end{aligned}$$

where (i) follows since A_1 holds, (ii) follows since A_4 holds, (iii) follows from Lemma 37, and (iv) follows since A_2 holds. By (D.6) and (D.8), we have

$$|I_7| \leq \frac{\delta}{16} \times \frac{d_0\sqrt{2\lambda_{0,\max}}M^{1+\beta}\delta_1\delta_2}{d_0^2M^{1+\beta}\sqrt{3\lambda_{0,\max}}\delta_2} \leq \frac{\delta}{16} \times \sqrt{\frac{2}{3}} \times \frac{\delta}{8\sqrt{6}d_0^2} \leq \frac{\delta}{16},$$

which completes the proof. ■

Lemma 27 *Given that A1-A5 hold, we have* $|I_8| \leq \delta/16$ *for all* $1 \leq j, l \leq p,\ 1 \leq k, m \leq M$.

Proof We have

$$\begin{aligned}
|I_8| &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}|\langle \hat{g}_{ij} - g_{ij}, \phi_{jk}\rangle|^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}|\langle \hat{g}_{il} - g_{il}, \hat{\phi}_{lm} - \phi_{lm}\rangle|^2} \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\hat{g}_{ij} - g_{ij}\|^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\hat{g}_{il} - g_{il}\|^2\|\hat{\phi}_{lm} - \phi_{lm}\|^2} \\
&\overset{(i)}{\leq} \delta_1^2\|\hat{\phi}_{lm} - \phi_{lm}\| \overset{(ii)}{\leq} \delta_1^2 d_{lm}\|\hat{K}_{ll} - K_{ll}\|_{\mathrm{HS}} \leq \delta_1^2 d_2 m^{1+\beta}\|\hat{K}_{ll} - K_{ll}\|_{\mathrm{HS}} \\
&\leq \delta_1^2 d_0 M^{1+\beta}\|\hat{K}_{ll} - K_{ll}\|_{\mathrm{HS}} \overset{(iii)}{\leq} d_0 M^{1+\beta}\delta_1^2\delta_2,
\end{aligned}$$

where (i) follows since A_1 holds, (ii) follows from Lemma 37, and (iii) follows since A_2 holds. By (D.9), we have

$$|I_8| \leq \frac{\delta}{16} \times \frac{d_0 M^{1+\beta}\delta_1^2\delta_2}{\delta_1^2} \leq \frac{\delta}{16},$$

which completes the proof. ■

Lemma 28 *Given that A1-A5 hold, we have* $|I_9| \leq \delta/16$ *for all* $1 \leq j, l \leq p,\ 1 \leq k, m \leq M$.

Proof This proof is similar to the proof of Lemma 22, and is therefore omitted. ■

Lemma 29 *Given that A1-A5 hold, we have* $|I_{10}| \leq \delta/16$ *for all* $1 \leq j, l \leq p, 1 \leq k, m \leq M$.

Proof This proof is similar to the proof of Lemma 26, and is therefore omitted. ■

Lemma 30 *Given that A1-A5 hold, we have* $|I_{11}| \leq \delta/16$ *for all* $1 \leq j, l \leq p, 1 \leq k, m \leq M$.

Proof We have

$$\begin{aligned}
|I_{11}| &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}|\langle g_{ij}, \hat{\phi}_{jk} - \phi_{jk}\rangle|^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}|\langle g_{il}, \hat{\phi}_{lm} - \phi_{lm}\rangle|^2} \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|g_{ij}\|^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|g_{il}\|^2} \cdot \|\hat{\phi}_{jk} - \phi_{jk}\| \cdot \|\hat{\phi}_{lm} - \phi_{lm}\| \\
&\overset{(i)}{\leq} 2\lambda_{0,\max}\|\hat{\phi}_{jk} - \phi_{jk}\|\|\hat{\phi}_{lm} - \phi_{lm}\| \overset{(ii)}{\leq} 2\lambda_{0,\max}\delta_2^2 d_{jk} d_{lm} \leq 2\lambda_{0,\max}\delta_2^2 d_2^2 k^{1+\beta} m^{1+\beta},
\end{aligned}$$

where (i) follows since A_4 holds and (ii) follows from Lemma 37. Then, we have

$$|I_{11}| \leq 2d_0^2\lambda_{0,\max}M^{2+2\beta}\delta_2^2.$$

By (D.8), we have

$$2d_0^2\lambda_{0,\max}M^{2+2\beta}\delta_2^2 \leq \frac{\delta}{16} \times \frac{2d_0^2\lambda_{0,\max}M^{2+2\beta}\delta_2^2}{d_0^2 M^{1+\beta}\sqrt{3\lambda_{0,\max}}\delta_2} \leq \frac{\delta}{16}, \tag{D.10}$$

which completes the proof. ■

Lemma 31 *Given that A1-A5 hold, we have* $|I_{12}| \leq \delta/16$ *for all* $1 \leq j, l \leq p, 1 \leq k, m \leq M$.

Proof We have

$$\begin{aligned}
|I_{12}| &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}|\langle g_{ij}, \hat{\phi}_{jk} - \phi_{jk}\rangle|^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}|\langle \hat{g}_{il} - g_{il}, \hat{\phi}_{lm} - \phi_{lm}\rangle|^2} \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|g_{ij}\|^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\hat{g}_{il} - g_{il}\|^2} \cdot \|\hat{\phi}_{jk} - \phi_{jk}\| \cdot \|\hat{\phi}_{lm} - \phi_{lm}\| \\
&\overset{(i)}{\leq} \sqrt{2\lambda_{0,\max}}\delta_1\delta_2^2 d_{jk} d_{lm} \leq d_2^2\sqrt{2\lambda_{0,\max}}k^{1+\beta}m^{1+\beta}\delta_1\delta_2^2,
\end{aligned}$$

where (i) follows since A_1-A_3 hold and Lemma 37. Then, we have

$$|I_{12}| \leq d_0^2 \sqrt{2\lambda_{0,\max}} M^{2+2\beta} \delta_1 \delta_2^2.$$

By (D.6) and (D.10), we have

$$d_0^2 \sqrt{2\lambda_{0,\max}} M^{2+2\beta} \delta_1 \delta_2^2 \leq \frac{\delta}{16} \times \frac{d_0^2 \sqrt{2\lambda_{0,\max}} M^{2+2\beta} \delta_1 \delta_2^2}{2 d_0^2 \lambda_{0,\max} M^{2+2\beta} \delta_2^2} \leq \frac{\delta}{16}, \tag{D.11}$$

which completes the proof. ■

Lemma 32 *Given that A1-A5 hold, we have* $|I_{13}| \leq \delta/16$ *for all* $1 \leq j, l \leq p, 1 \leq k, m \leq M$.

Proof This proof is similar to the proof of Lemma 23, and is therefore omitted. ■

Lemma 33 *Given that A1-A5 hold, we have* $|I_{14}| \leq \delta/16$ *for all* $1 \leq j, l \leq p, 1 \leq k, m \leq M$.

Proof This proof is similar to the proof of Lemma 27, and is therefore omitted. ■

Lemma 34 *Given that A1-A5 hold, we have* $|I_{15}| \leq \delta/16$ *for all* $1 \leq j, l \leq p,\ \ 1 \leq k, m \leq M$.

Proof This proof is similar to the proof of Lemma 21, thus is omitted. ■

Lemma 35 *Given that A1-A5 hold, we have* $|I_{16}| \leq \delta/16$ *for all* $1 \leq j, l \leq p, 1 \leq k, m \leq M$.

Proof We have

$$\begin{aligned} |I_{16}| &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\hat{g}_{ij} - g_{ij}\|^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\hat{g}_{il} - g_{il}\|^2} \cdot \|\hat{\phi}_{jk} - \phi_{jk}\| \cdot \|\hat{\phi}_{lm} - \phi_{lm}\| \\ &\overset{(i)}{\leq} \delta_1^2 d_{jk} d_{lm} \delta_2^2 \leq d_2^2 k^{1+\beta} m^{1+\beta} \delta_1^2 \delta_2^2 \leq d_0^2 M^{2+2\beta} \delta_1^2 \delta_2^2, \end{aligned}$$

where (i) follows since A_1 and A_2 hold, and Lemma 37. Thus, by (D.7) and (D.11), we have

$$|I_{16}| \leq \frac{\delta}{16} \times \frac{d_0^2 M^{2+2\beta} \delta_1^2 \delta_2^2}{d_0^2 \sqrt{2\lambda_{0,\max}} M^{2+2\beta} \delta_1 \delta_2^2} \leq \frac{\delta}{16},$$

which completes the proof. ■

Lemma 36 *Suppose $f_1, f_2, \ldots, f_n \in \mathbb{H}$ and $v_1, v_2, \ldots, v_n \in \mathbb{R}$. Then*

$$\left\| \sum_{i=1}^{n} v_i f_i \right\| \leq \sqrt{\sum_{i=1}^{n} v_i^2} \cdot \sqrt{\sum_{i=1}^{n} \|f_i\|^2}.$$

Proof Note that

$$\left\| \sum_{i=1}^{n} v_i f_i \right\|^2 = \int \left(\sum_{i=1}^{n} v_i f_i(t) \right)^2 dt \overset{(i)}{\leq} \int \left(\sum_{i=1}^{n} v_i^2 \right) \left(\sum_{i=1}^{n} f_i^2(t) \right) dt = \left(\sum_{i=1}^{n} v_i^2 \right) \left(\sum_{i=1}^{n} \|f_i\|^2 \right),$$

where (i) follows the Cauchy-Schwarz inequality. This directly implies the result. ■

Lemma 37 (Lemma 4.3 of Bosq (2000)) *Suppose that Assumption 3 holds. Denote $\tilde{\phi}_{jk} = sgn\left(\langle \hat{\phi}_{jk}, \phi_{jk} \rangle\right) \phi_{jk}$, where $sgn(t) = 1$ if $t \geq 0$ and $sgn(t) = -1$ if $t < 0$. Then*

$$\|\hat{\phi}_{jk} - \tilde{\phi}_{jk}\| \leq d_{jk} \|\hat{K}_{jj} - K_{jj}\|_{HS},$$

where $d_{j1} = 2\sqrt{2}(\lambda_{j1} - \lambda_{j2})^{-1}$ and $d_{jk} = 2\sqrt{2} \max\{(\lambda_{j(k-1)} - \lambda_{jk})^{-1}, (\lambda_{jk} - \lambda_{j(k+1)})^{-1}\}$, $k \geq 2$.

Lemma 38 *Suppose $z \sim N_L(0, I_L)$. Then*

$$P\left(\|z\|_2 > \delta\right) \leq 2 \exp\left(-\frac{\delta^2}{8L + 2\sqrt{2L}\delta} \right), \qquad \delta > 0.$$

Proof Since

$$\mathbb{E}\left[\|z\|_2^{2k} \right] = \frac{\Gamma(\frac{L}{2} + k)}{\Gamma(\frac{L}{2})} \times 2^k \leq k!(2L)^k,$$

we have

$$\mathbb{E}\left[\|z\|_2^{k} \right] \leq \sqrt{\mathbb{E}\left[\|z\|_2^{2k} \right]} \leq \sqrt{k!} \left(\sqrt{2L} \right)^k \leq \frac{k!}{2} \cdot 4L \cdot (\sqrt{2L})^{k-2}$$

for $k \geq 2$. The result follows from Lemma 39. ■

Lemma 39 (Theorem 2.5 (2) of Bosq (2000)) *Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables in a separable Hilbert space with norm $\|\cdot\|$. If $\mathbb{E}[Z_i] = 0$, $i = 1, \ldots, n$, and*

$$\sum_{i=1}^{n} \mathbb{E}\left[\|Z_i\|^k \right] \leq \frac{k!}{2} n L_1 L_2^{k-2}, \qquad k \geq 2,$$

for two positive constants L_1 and L_2, then

$$P\left(\|\sum_{i=1}^{n} Z_i\| \geq n\delta \right) \leq 2 \exp\left(-\frac{n\delta^2}{2L_1 + 2L_2\delta} \right), \qquad \delta > 0.$$

Lemma 40 *Let $f(t)$ be a function defined on $\mathcal{T}$ and suppose that f has a continuous derivative. Let $D_{0,f} \coloneqq \sup_{t\in\mathcal{T}}|f(t)|$ and $D_{1,f} \coloneqq \sup_{t\in\mathcal{T}}|f'(t)|$. Assume that $D_{0,f}, D_{1,f} < \infty$. Let $|\mathcal{T}|$ denote the length of the interval $\mathcal{T}$, and let $u_1 < u_2 < \cdots < u_T \in \mathcal{T}$. We denote the endpoints of $\mathcal{T}$ as u_0 and u_{T+1}. Assume that there is a positive constant ζ_0 such that*

$$\max_{1\le k\le T+1}\left|\frac{u_k - u_{k-1}}{|\mathcal{T}|} - \frac{1}{T}\right| \le \frac{\zeta_0}{T^2}. \tag{D.12}$$

Let $\zeta_1 = \zeta_0 + 1$. Then

$$\left|\frac{1}{T}\sum_{k=1}^{T} f(u_k) - \frac{1}{|\mathcal{T}|}\int_{\mathcal{T}} f(t)dt\right| \le \frac{D_{1,f}\zeta_1^2|\mathcal{T}|/2 + D_{0,f}(\zeta_1+\zeta_0)}{T}.$$

Proof Since

$$\begin{aligned}\left|\frac{1}{T}\sum_{k=1}^{T} f(u_k) - \frac{1}{|\mathcal{T}|}\int_{\mathcal{T}} f(t)dt\right| \le{}& \left|\frac{1}{T}\sum_{k=1}^{T} f(u_k) - \frac{1}{|\mathcal{T}|}\sum_{k=1}^{T} f(u_k)(u_k - u_{k-1})\right| \\ &+ \left|\frac{1}{|\mathcal{T}|}\sum_{k=1}^{T} f(u_k)(u_k - u_{k-1}) - \frac{1}{|\mathcal{T}|}\int_{\mathcal{T}} f(t)dt\right|,\end{aligned}$$

we proceed to show that the first part is smaller than $D_{0,f}\zeta_0/T$ and that the second part is smaller than $(D_{1,f}\zeta_1^2|\mathcal{T}|/2 + D_{0,f}\zeta_1)/T$. For the first part, we have

$$\begin{aligned}\left|\frac{1}{T}\sum_{k=1}^{T} f(u_k) - \frac{1}{|\mathcal{T}|}\sum_{k=1}^{T} f(u_k)(u_k - u_{k-1})\right| &\le \sum_{k=1}^{T}|f(u_k)|\left|\frac{1}{T} - \frac{u_k - u_{k-1}}{|\mathcal{T}|}\right| \\ &\le \max_{1\le k\le T}\left|\frac{u_k - u_{k-1}}{|\mathcal{T}|} - \frac{1}{T}\right|\sum_{k=1}^{T}|f(u_k)| \le \frac{\zeta_0}{T^2}\times T\times D_{0,f} = \frac{\zeta_0 D_{0,f}}{T}.\end{aligned}$$

To prove the second part, we first note that based on (D.12), we have

$$\max_{1\le k\le T+1}|u_k - u_{k-1}| \le \frac{\zeta_1|\mathcal{T}|}{T}.$$

Then, for any $t \in (u_k, u_{k+1})$, by Taylor's expansion, we have $f(t) = f(u_k) + f'(\bar{t})(t - u_k)$, where $\bar{t} \in (u_k, t)$, and $|f(t) - f(u_k)| = |f'(\bar{t})|(t - u_k) \le D_{1,f}(t - u_k)$. Therefore,

$$\begin{aligned}&\left|\frac{1}{|\mathcal{T}|}\sum_{k=1}^{T} f(u_k)(u_k - u_{k-1}) - \frac{1}{|\mathcal{T}|}\int_{\mathcal{T}} f(t)dt\right| \\ &\le \frac{1}{|\mathcal{T}|}\sum_{k=1}^{T}\int_{u_{k-1}}^{u_k}|f(u_k) - f(t)|dt + \frac{1}{|\mathcal{T}|}\int_{u_T}^{u_{T+1}}|f(t)|dt \\ &\le \frac{1}{|\mathcal{T}|}\times T\times D_{1,f}\times\int_{u_{k-1}}^{u_k}(t - u_k)dt + \frac{1}{|\mathcal{T}|}\times D_{0,f}\times\frac{\zeta_1|\mathcal{T}|}{T}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{|\mathcal{T}|} \times T \times D_{1,f} \times \frac{(u_{k+1}-u_k)^2}{2} + \frac{1}{|\mathcal{T}|} \times D_{0,f} \times \frac{\zeta_1 |\mathcal{T}|}{T} \\
&\leq \frac{1}{|\mathcal{T}|} \times T \times \frac{D_{1,f}}{2} \times \left(\max_{1 \leq k \leq T+1} |u_{k+1}-u_k| \right)^2 + \frac{1}{|\mathcal{T}|} \times D_{0,f} \times \frac{\zeta_1 |\mathcal{T}|}{T} \\
&\leq \frac{1}{|\mathcal{T}|} \times T \times \frac{D_{1,f}}{2} \times \left(\frac{\zeta_1 |\mathcal{T}|}{T} \right)^2 + \frac{1}{|\mathcal{T}|} \times D_{0,f} \times \frac{\zeta_1 |\mathcal{T}|}{T} \\
&= \frac{D_{1,f} \zeta_1^2 |\mathcal{T}|/2 + D_{0,f} \zeta_1}{T}.
\end{aligned}
$$

The result follows by combining the two bounds. ■

Lemma 41 *Let g be a mean zero Gaussian random function in a Hilbert space $\mathbb{H}$. We have $\mathbb{E}\left[\|g\|^{2k}\right] \leq (2\lambda_0)^k \cdot k!$ where $\lambda_0 = \mathbb{E}\left[\|g\|^2\right]$.*

Proof Let $\{\phi_m\}_{m \geq 1}$ be the orthonormal eigenfunctions of g and $a_m = \langle g, \phi_m \rangle$. Then $a_m \sim N(0, \lambda_m)$ and $\lambda_0 = \sum_{m \geq 1} \lambda_m$. Let $\xi_m = \lambda_m^{-1/2} a_m$. By the Karhunen-Loève theorem, we have $g = \sum_{m=1}^{\infty} \lambda_m^{1/2} \xi_m \phi_m$. Thus, $\|g\| = \left(\sum_{m \geq 1} \lambda_m \xi_m^2 \right)^{1/2}$ and $\|g\|^{2k} = \left(\sum_{m \geq 1} \lambda_m \xi_m^2 \right)^k$. By Jensen's inequality, we have

$$
\begin{aligned}
\|g\|^{2k} &= \left(\sum_{m \geq 1} \lambda_m \right)^k \cdot \left(\frac{\sum_{m \geq 1} \lambda_m \xi_m^2}{\sum_{m \geq 1} \lambda_m} \right)^k \\
&\leq \left(\sum_{m \geq 1} \lambda_m \right)^k \cdot \frac{\sum_{m \geq 1} \lambda_m \xi_m^{2k}}{\sum_{m \geq 1} \lambda_m} = \left(\sum_{m \geq 1} \lambda_m \right)^{k-1} \cdot \left(\sum_{m \geq 1} \lambda_m \xi_m^{2k} \right).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathbb{E}\left[\|g\|^{2k}\right] &\leq \left(\sum_{m \geq 1} \lambda_m \right)^{k-1} \cdot \left(\sum_{m \geq 1} \lambda_m \mathbb{E}\left[\xi_m^{2k}\right] \right) = \left(\sum_{m \geq 1} \lambda_m \right)^k \mathbb{E}\left[\xi_1^{2k}\right] \\
&= \left(\sum_{m \geq 1} \lambda_m \right)^k \cdot \pi^{-1/2} \cdot 2^k \cdot \Gamma(k+1/2) \leq \left(\sum_{m \geq 1} \lambda_m \right)^k \cdot 2^k \cdot k! = (2\lambda_0)^k k!,
\end{aligned}
$$

which completes the proof. ■

Lemma 42 *For any $\delta > 0$ and any $j = 1, \ldots, p$, we have*

$$
P\left(\left\| \frac{1}{n} \sum_{i=1}^{n} \left[g_{ij}(t) g_{ij}(s) - K_{jj}(s,t) \right] \right\|_{HS} > \delta \right) \leq 2 \exp\left(- \frac{n\delta^2}{64 \lambda_{0,\max}^2 + 8 \lambda_{0,\max} \delta} \right).
$$

Proof Since $g_{ij}(t) = \sum_{m\geq 1} \lambda_{jm}^{1/2} \xi_{ijm} \phi_{jm}(t)$ and $\xi_{ijm} \sim N(0,1)$, we have

$$g_{ij}(s)g_{ij}(t) = \sum_{m,m'\geq 1} \lambda_{jm}^{1/2} \lambda_{jm'}^{1/2} \xi_{ijm}\xi_{ijm'} \phi_{jm}(s)\phi_{jm'}(t),$$

and

$$K_{jj}(s,t) = \mathbb{E}[g_{ij}(s)g_{ij}(t)] = \sum_{m,m'\geq 1} \lambda_{jm}^{1/2} \lambda_{jm'}^{1/2} \phi_{jm}(s)\phi_{jm'}(t) \mathbb{1}_{mm'},$$

where $\mathbb{1}_{mm'} = \mathbb{1}(m = m') = 1$ if $m = m'$ and 0 if $m \neq m'$. Thus,

$$\|g_{ij}(s)g_{ij}(t) - K_{jj}(s,t)\|_{\mathrm{HS}}^2 = \sum_{m,m'\geq 1} \lambda_{jm}\lambda_{jm'}(\xi_{ijm}\xi_{ijm'} - \mathbb{1}_{mm'})^2,$$

and, for any $k \geq 2$, we have

$$\begin{aligned}
&\mathbb{E}\left[\|g_{ij}(s)g_{ij}(t) - K_{jj}(s,t)\|_{\mathrm{HS}}^k\right] \\
&= \mathbb{E}\left[\left\{\sum_{m,m'\geq 1} \lambda_{jm}\lambda_{jm'}(\xi_{ijm}\xi_{ijm'} - \mathbb{1}_{mm'})^2\right\}^{k/2}\right] \\
&\overset{(i)}{\leq} \left(\sum_{m,m'\geq 1} \lambda_{jm}\lambda_{jm'}\right)^{k/2-1} \sum_{m,m'\geq 1} \lambda_{jm}\lambda_{jm'} \mathbb{E}\left[\left(\xi_{ijm}\xi_{ijm'} - \mathbb{1}_{mm'}\right)^k\right],
\end{aligned}$$

where (i) follows from Jensen's inequality. Since

$$\begin{aligned}
\mathbb{E}\left[\left(\xi_{ijm}\xi_{ijm'} - \mathbb{1}_{mm'}\right)^k\right] &\leq 2^{k-1}\left(\mathbb{E}\left[(\xi_{ijm}\xi_{ijm'})^k\right] + 1\right) \\
&\leq 2^{k-1}\left(\mathbb{E}[\xi_{ij1}^{2k}] + 1\right) \leq 2^{k-1}(2^k k! + 1) \leq 4^k k!,
\end{aligned}$$

we have $\mathbb{E}\left[\|g_{ij}(s)g_{ij}(t) - K_{jj}(s,t)\|_{\mathrm{HS}}^k\right] \leq (4\lambda_{j0})^k k! \leq (4\lambda_{0,\max})^k k!$. The result follows from Lemma 39. ■

References

A. Ahmed and E. P. Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106(29):11878–11883, 2009.

R. F. Barber and M. Kolar. ROCKET: Robust confidence intervals via kendall's tau for transelliptical graphical models. *Annals of Statistics*, 46(6B):3422–3450, 2018.

A. R. Barron and C.-H. Sheu. Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, 19(3):1347–1369, 1991.

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

D. Bosq. *Linear processes in function spaces*, volume 149 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2000. Theory and applications.

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

S. P. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

J. Bradic and M. Kolar. Uniform inference for high-dimensional quantile regression: linear functionals and regression rank scores. *arXiv preprint arXiv:1702.06209*, 2017.

T. T. Cai. Global testing and large-scale multiple testing for high-dimensional covariance structures. *Annual Review of Statistics and Its Application*, 4(1):423–446, 2017.

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

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

R. Dai and M. Kolar. Inference for high-dimensional varying-coefficient quantile regression. *Electronic Journal of Statistics*, 15(2):5696–5757, 2021, `arXiv:2002.07370v1`.

P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 76(2):373–397, 2014.

F. Fazayeli and A. Banerjee. Generalized direct change estimation in ising model structure. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016*, 2016.

S. Geng, M. Kolar, and O. Koyejo. Joint nonparametric precision matrix estimation with confounding. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence, UAI 2019*, 2019a.

S. Geng, M. Yan, M. Kolar, and S. Koyejo. Partially linear additive gaussian graphical models. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, 2019b.

T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2015.

L. Ingber. Statistical mechanics of neocortical interactions: Canonical momenta indicators of electroencephalography. *Physical Review E*, 55:4578–4593, 1997.

J. Janková and S. A. van de Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229, 2015.

J. Janková and S. A. van de Geer. Honest confidence regions and optimality in high-dimensional precision matrix estimation. *TEST*, 26(1):143–162, 2017.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research (JMLR)*, 15:2869–2909, 2014.

R. A. Johnson and D. W. Wichern. *Applied multivariate statistical analysis*, volume 6. Pearson London, UK:, 2014.

B. Kim, S. Liu, and M. Kolar. Two-sample inference for high-dimensional Markov networks. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 83(5):939–962, 2021.

G. G. Knyazev. Motivation, emotion, and their inhibitory control mirrored in brain oscillations. *Neuroscience and Biobehavioral Reviews*, 31(3):377–395, 2007.

P. Kokoszka and M. Reimherr. *Introduction to Functional Data Analysis*. Chapman and Hall/CRC, 2017.

M. Kolar and E. P. Xing. Sparsistent estimation of time-varying discrete markov random fields. *arXiv preprint, arXiv:0907.2337*, 2009.

M. Kolar and E. P. Xing. On time varying undirected graphs. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, AISTATS 2011*, 2011.

M. Kolar and E. P. Xing. Consistent covariance selection from data with missing values. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, 2012a.

M. Kolar and E. P. Xing. Estimating networks with jumps. *Electronic Journal of Statistics*, 6:2069–2106, 2012b.

M. Kolar, L. Song, and E. P. Xing. Sparsistent learning of varying-coefficient models with structural changes. In *Advances in Neural Information Processing Systems, NIPS 2009*, 2009.

M. Kolar, A. P. Parikh, and E. P. Xing. On sparse nonparametric conditional covariance selection. In *Proceedings of the 27th International Conference on Machine Learning, ICML 2010*, 2010a.

M. Kolar, L. Song, A. Ahmed, and E. P. Xing. Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1):94–123, 2010b.

M. Kolar, H. Liu, and E. P. Xing. Markov network estimation from multi-attribute data. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*, 2013.

M. Kolar, H. Liu, and E. P. Xing. Graph estimation from multi-attribute data. *Journal of Machine Learning Research (JMLR)*, 15(1):1713–1750, 2014.

S. Lauritzen. *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996. Oxford Science Publications.

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

K.-C. Li, A. Palotie, S. Yuan, D. Bronnikov, D. Chen, X. Wei, O.-W. Choi, J. Saarela, and L. Peltonen. Finding disease candidate genes by liquid association. *Genome Biology*, 8 (10):R205, 2007.

S. Liu, J. A. Quinn, M. U. Gutmann, T. Suzuki, and M. Sugiyama. Direct learning of sparse changes in Markov networks by density ratio estimation. *Neural Computation*, 26 (6):1169–1197, 2014.

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

X. Liu, H. Nassar, and K. Podgorski. Splinets – efficient orthonormalization of the B-splines. *BIT Numerical Mathematics*, 2022.

J. Lu, M. Kolar, and H. Liu. Post-regularization inference for time-varying nonparanormal graphical models. *Journal of Machine Learning Research (JMLR)*, 18(203):1–78, 2018.

J. Lu, M. Kolar, and H. Liu. Kernel meets sieve: post-regularization confidence bands for sparse additive model. *Journal of the American Statistical Association*, 115(532): 2084–2099, 2020, `arXiv:1503.02978`.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

S. Na, M. Kolar, and O. Koyejo. Estimating differential latent variable graphical models with applications to brain connectivity. *Biometrika*, 108(2):425–442, 2021.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4), 2012.

M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2): 167–256, 2003.

N. Parikh and S. P. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

X. Qiao, S. Guo, and G. M. James. Functional Graphical Models. *Journal of the American Statistical Association*, 114(525):211–222, 2019.

X. Qiao, C. Qian, G. M. James, and S. Guo. Doubly functional graphical models in high dimensions. *Biometrika*, 107(2):415–431, 2020.

J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.

J. O. Ramsay, H. Wickham, S. Graves, and G. Hooker. *fda: Functional Data Analysis*, 2020. R package version 2.4.8.1.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Z. Ren, T. Sun, C.-H. Zhang, and H. H. Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.

Y. She. An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics and Data Analysis*, 56(10):2976–2990, 2012.

L. Song, M. Kolar, and E. P. Xing. KELLER: estimating time-varying interactions between genes. *Bioinformatics*, 25(12):i128–i136, 2009a.

L. Song, M. Kolar, and E. P. Xing. Time-varying dynamic bayesian networks. In *Advances in Neural Information Processing Systems, NIPS 2009*, 2009b.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, And Search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2000.

A. S. Suggala, M. Kolar, and P. Ravikumar. The Expxorcist: Nonparametric graphical models via conditional exponential densities. In *Advances in Neural Information Processing Systems, NIPS 2017*, 2017.

M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems, NIPS 2007*, 2007.

S. Sun, M. Kolar, and J. Xu. Learning structured densities via infinite dimensional exponential families. In *Advances in Neural Information Processing Systems, NIPS 2015*, 2015.

M. Talih and N. Hengartner. Structural learning with time-varying components: tracking the cross-section of the financial time series. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 67(3):321–341, 2005.

R. Tibshirani. Proximal gradient descent and acceleration. *Lecture Notes*, 2010.

K. Tsai, M. Kolar, and O. Koyejo. A nonconvex framework for structured dynamic covariance recovery. *arXiv preprint, arXiv:2011.05601*, 2020.

K. Tsai, O. Koyejo, and M. Kolar. Joint gaussian graphical model estimation: A survey. *arXiv preprint, arXiv:2110.10281*, 2021.

S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

S. A. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42 (3):1166–1202, 2014.

D. Vogel and R. Fried. Elliptical graphical modelling. *Biometrika*, 98(4):935–951, 2011.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

J. Wang and M. Kolar. Inference for sparse conditional precision matrices. *arXiv preprint, arXiv:1412.7638*, 2014.

J. Wang and M. Kolar. Inference for high-dimensional exponential family graphical models. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, 2016.

X. Wang, M. Kolar, and A. Shojaie. Statistical inference for networks of high-dimensional point processes. *arXiv preprint, arXiv:2007.07448*, 2020.

Y. S. Wang, S. K. Lee, P. Toulis, and M. Kolar. Robust inference for high-dimensional linear models via residual randomization. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, 2021.

Y. Wang, C. Squires, A. Belyaeva, and C. Uhler. Direct estimation of differences in causal graphs. In *Advances in Neural Information Processing Systems, NeurIPS 2018*, 2018.

L. Wasserman. *All of nonparametric statistics*. Springer Texts in Statistics. Springer, New York, 2006.

L. Wasserman, M. Kolar, and A. Rinaldo. Berry-Esseen bounds for estimating undirected graphs. *Electronic Journal of Statistics*, 8(1):1188–1224, 2014.

Y. Xia, T. Cai, and T. T. Cai. Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika*, 102(2):247–266, 2015.

P. Xu and Q. Gu. Semiparametric differential graph models. In *Advances in Neural Information Processing Systems, NIPS 2016*, 2016.

X. Xuan and K. P. Murphy. Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th International Conference on Machine Learning, ICML 2007*, 2007.

J. Yin, Z. Geng, R. Li, and H. Wang. Nonparametric covariance model. *Statistica Sinica*, 20(1):469–479, 2010.

M. Yu, V. Gupta, and M. Kolar. Statistical inference for pairwise graphical models using score matching. In *Advances in Neural Information Processing Systems, NIPS 2016*, 2016.

M. Yu, V. Gupta, and M. Kolar. Simultaneous inference for pairwise graphical models with generalized score matching. *Journal of Machine Learning Research (JMLR)*, 21(91):1–51, 2020.

H. Yuan, R. Xi, C. Chen, and M. Deng. Differential network analysis via lasso penalized D-trace loss. *Biometrika*, 104(4):755–770, 2017.

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

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 68(1):49–67, 2006.

J. Zapata, S. Y. Oh, and A. Petersen. Partial separability and functional graphical models for multivariate Gaussian processes. *Biometrika*, 2021.

C. Zhang, H. Yan, S. Lee, and J. Shi. Dynamic multivariate functional data modeling via sparse subspace learning. *Technometrics*, 63(3):370–383, 2021.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 76(1):217–242, 2014.

X. L. Zhang, H. Begleiter, B. Porjesz, W. Wang, and A. Litke. Event related potentials during object recognition tasks. *Brain Research Bulletin*, 38(6):531–538, 1995.

X. Zhang and J.-L. Wang. From sparse to dense functional data and beyond. *The Annals of Statistics*, 44(5):2281–2321, 2016.

B. Zhao, Y. S. Wang, and M. Kolar. Direct estimation of differential functional graphical models. In *Advances in Neural Information Processing Systems, NeurIPS 2019*, 2019.

B. Zhao, S. Zhai, Y. S. Wang, and M. Kolar. High-dimensional functional graphical model structure learning via neighborhood selection approach. *arXiv preprint, arXiv:2105.02487*, 2021.

S. D. Zhao, T. T. Cai, and H. Li. Direct estimation of differential networks. *Biometrika*, 101(2):253–268, 2014a.

T. Zhao, M. Kolar, and H. Liu. A general framework for robust testing and confidence regions in high-dimensional quantile regression. *arXiv preprint arXiv:1412.8724*, 2014b.

S. Zhou, J. Lafferty, and L. Wasserman. Time varying undirected graphs. *Machine Learning*, 80(2-3):295–319, 2010.

H. Zhu, N. Strawn, and D. B. Dunson. Bayesian graphical models for multivariate functional data. *Journal of Machine Learning Research (JMLR)*, 17(204):1–27, 2016.