Functional Graphical Models

Xinghao Qiao[1], Shaojun Guo*[2], and Gareth M. James[3]

[1]*Department of Statistics, London School of Economics, U.K.*

[2]*Institute of Statistics and Big Data, Renmin University of China, P.R. China*

[3]*Department of Data Sciences and Operations, University of Southern California, USA*

Abstract

Graphical models have attracted increasing attention in recent years, especially in settings involving high dimensional data. In particular Gaussian graphical models are used to model the conditional dependence structure among multiple Gaussian random variables. As a result of its computational efficiency the *graphical lasso* (glasso) has become one of the most popular approaches for fitting high dimensional graphical models. In this article we extend the graphical models concept to model the conditional dependence structure among p random functions. In this setting, not only is p large, but each function is itself a high dimensional object, posing an additional level of statistical and computational complexity. We develop an extension of the glasso criterion (fglasso), which estimates the *functional graphical model* by imposing a block sparsity constraint on the precision matrix, via a group lasso penalty. The fglasso criterion can be optimized using an efficient block coordinate descent algorithm. We establish the concentration inequalities of the estimates, which guarantee the desirable graph support recovery property, i.e. with probability tending to one, the fglasso will correctly

*Shaojun Guo was partially supported by the National Natural Science Foundation of China (No. 11771447) , the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China.

identify the true conditional dependence structure. Finally we show that the fglasso significantly outperforms possible competing methods through both simulations and an analysis of a real world EEG data set comparing alcoholic and non-alcoholic patients.

Some key words: Functional data; Graphical models; Functional principal component analysis; Block sparse precision matrix, Block coordinate descent algorithm.

1 Introduction

The graphical model is used to depict the conditional dependence structure among p random variables, $\mathbf{X} = (X_1, \ldots, X_p)^T$. Such a network consists of p *nodes*, one for each variable, and a number of *edges* connecting a subset of the nodes. The edges describe the conditional dependence structure of the p variables, i.e. nodes j and l are connected by an edge if and only if X_j and X_l are dependent, conditional on the other $p-2$ variables. Recently there has been a lot of interest in fitting high dimensional graphical models, where p is very large.

For Gaussian data, where $\mathbf{X}$ follows a multivariate Gaussian distribution, one can show that estimating the edge set is equivalent to identifying the locations of the non-zero elements in the *precision matrix*, i.e. the inverse covariance matrix of $\mathbf{X}$. Hence, the literature has mainly focused on two approaches for estimating high dimensional *Gaussian graphical models*. One type of method, proposed by Meinshausen and Buhlmann (2006), considers neighbourhood selection. It adopts a lasso (Tibshirani, 1996) or Dantzig selector (Candes and Tao, 2007) type of penalized regression approach whereby each variable is regressed on the other variables, thus identifying the locations of non-zero entries in the precision matrix column by column, see also Peng et al. (2009); Cai et al. (2011); Sun and Zhang (2013). Another method, proposed by Yuan and Lin (2007), optimizes the *graphical lasso* (glasso) criterion, essentially a Gaussian log likelihood with the addition of a lasso type penalty on the entries of the precision matrix. The glasso has arguably proved the most popular of these two methods, in part because a number of efficient algorithms have been developed to

minimize the convex glasso criterion (Friedman et al., 2007; Boyd et al., 2010; Witten et al., 2011; Mazumder and Hastie, 2012a,b). Its theoretical properties have also been well studied (Lam and Fan, 2009; Ravikumar et al., 2011), and several variants and extensions of the glasso have been proposed, see Zhou et al. (2010); Kolar and Xing (2011); Danaher et al. (2014); Zhu et al. (2014) and the references therein.

In this paper we are interested in estimating a graphical network in a somewhat more complicated setting. Let $g_1(t), \ldots, g_p(t)$ jointly follow from a p-dimensional *multivariate Gaussian process* (MGP) where $t \in \mathcal{T}$ and $\mathcal{T}$ is a closed subset of the real line[1]. Our goal is to construct a *functional graphical model* (FGM) depicting the conditional dependence structure among these p random functions. The left panel of Figure 1 provides an illustrative example with $p = 9$ functions, or nodes. We have 100 observations of each function, corresponding to 100 individuals. In other words our data consists of functions, $g_{ij}(t)$ where $i = 1, \ldots, 100$ and $j = 1, \ldots, 9$. The right panel of Figure 1 illustrates the conditional dependence structure of these functions i.e. the FGM. For example, we observe that the last 3 functions are disconnected from, and hence conditionally independent of, the first 6 functions. We wish to take the observed functions in the left panel and estimate the FGM in the right panel.

Our motivating example is an electroencephalography (EEG) data set taken from an alcoholism study (Zhang et al., 1995; Ingber, 1997). The study consists of $n = 122$ subjects split between an alcoholic group and a control group. Each subject was exposed to either a single stimulus or two stimuli. The resulting EEG activity was recorded at 256 time points over a one second interval using electrodes placed at 64 standard locations on the subject's scalp. Hence, each observation, or subject, involves $p = 64$ different functions observed at 256 time points. It is of scientific interest to identify differences in brain EEG activity filtered at α frequency bands (Hayden et al., 2006) between the two groups, so we construct FGM's

[1]Here we assume the same time domain, $\mathcal{T}$, for all random functions to simplify the notation, but our methodological and theoretical results extend naturally to the more general case where each function corresponds to a different time domain.

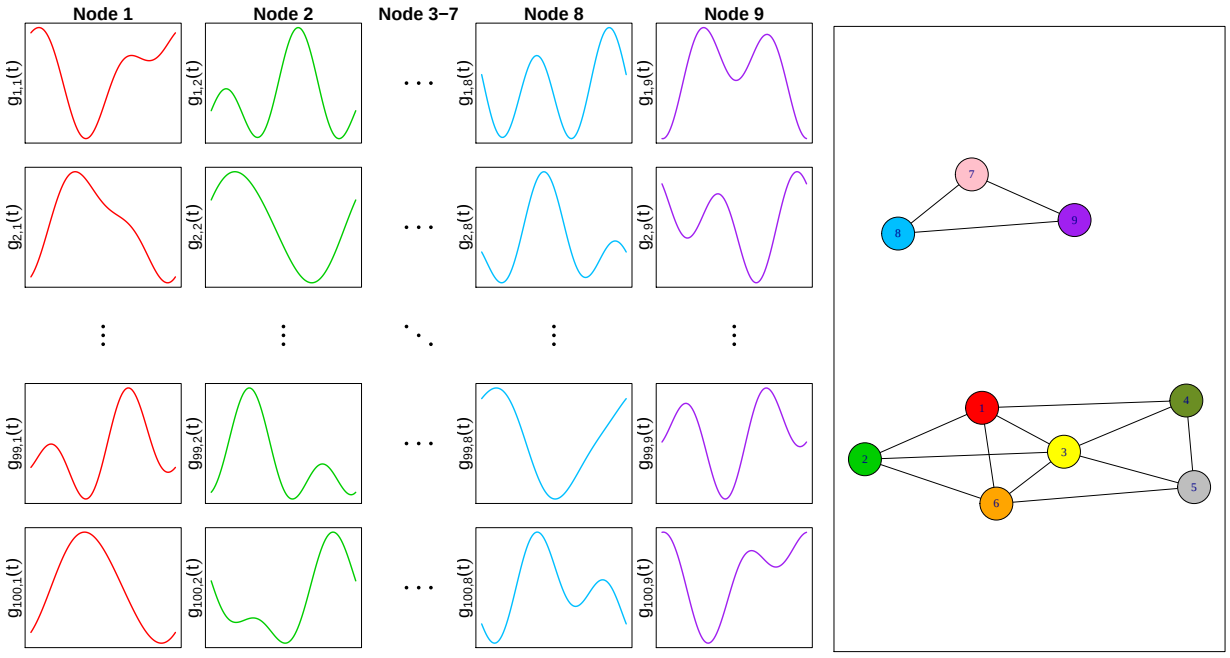

Figure 1: *The illustrative example. Left: The data,* $n = 100$ *observations of* $g_{ij}(t)$ *for* $j = 1, \ldots, 9$ *nodes. Right: the true underlying network structure.*

for each group and explore the differences. Functional data of this sort can arise in a number of other contexts. For example, rather than recording only a static set of p gene expression levels at a single point in time, it is now becoming common to observe multiple expression levels over time (Storey et al., 2005), so $g_{ij}(t)$ would represent the expression level of gene j for subject i at time t. Alternatively, in a marketing context it is now possible to observe online purchase patterns among a basket of p different products for each of n individual customers over a period of time, so $g_{ij}(t)$ might represent the cumulative purchase history of product j by customer i at time t.

One possible approach to handle this sort of functional data would be to sample the functions over a grid of time points, $t_1, \ldots, t_T$, estimate T separate networks, and then either report all T networks or construct a single graphical model by somehow merging the T networks. However, while conceptually simple, this strategy has several drawbacks. First, the approach is only possible if $g_1(t), \ldots, g_p(t)$ are observed over a common domain $t \in \mathcal{T}$,

but in many instances the domains of $g_j(t)$ and $g_l(t)$ will differ. Second, functions are often observed over a relatively sparse set of time points which makes it impossible to create a single set of grid points, $t_1, \ldots, t_T$, at which the functions can be sampled. Third, one of the main advantages of a graphical network is its ability to provide a relatively simple representation of the conditional dependence structure. However, simultaneously interpreting T different networks would significantly increase the complexity of the dependence structure. Finally, each of the T networks would only correspond to dependencies among the functions at a common time point. However, it seems likely that some dependencies may only exist at different time points i.e. it may be the case that $g_j(t)$ and $g_l(t)$ are conditionally uncorrelated for every individual value of t but $g_j(s)$ and $g_l(t)$ are correlated for some $s \neq t$. A simple example that illustrates this issue is provided in Appendix A.1. In such a scenario each of the T networks would fail to identify the correlation structure.

In this paper, we propose a FGM which is able to estimate a single network and overcome these disadvantages. The functional network still contains p nodes, one for each function, but in order to construct the edges we extend the conditional covariance definition to the functional domain, and then use this extended covariance definition to estimate the edge set, E. There exist several challenges involved in estimating the FGM. Since each function is an infinite dimensional object, we need to adopt a dimension reduction approach, e.g. *functional principal component analysis* (FPCA), to approximate each function by a finite representation, which results in estimating a block precision matrix. Standard glasso algorithms for scalar data involve estimating the non-zero elements in the precision matrix. By comparison we have developed an efficient algorithm to estimate the non-zero blocks of a higher dimensional precision matrix. In our theoretical results we develop the entry-wise concentration inequalities for the sample covariance matrix of the estimated principal component scores. To the best of our knowledge, this is the first result on concentration inequalities for modelling high dimensional functional data under a FPCA framework, which provides a powerful tool to derive the non-asymptotic upper bounds. This result expands

the theoretical analysis of graph selection consistency from the standard setting to the more complicated functional domain.

Some recent research in graphical models for time-dependent data considered estimating a time varying graphical model through a nonparametric approach which constructs graphs that are similar across adjacent time points (Zhou et al., 2010; Kolar and Xing, 2011; Qiu et al., 2016). This approach has similarities to the grid approach discussed previously in that it estimates a separate graph at each time point. However, in addition, it also assumes correlation across time in the graph structures. By contrast our FGM estimates a single graph by considering the global correlation structures among the functions over all time points. Both approaches are useful but aim to answer different questions. One other relevant work of Zhu et al. (2016) proposed decomposable graphical models for multivariate functional data from a Bayesian perspective without investigating the graph selection consistency. Their framework is based on extending Markov distributions and hyper Markov laws from Gaussian random variables to Gaussian random functions, which significantly differs from our approach.

The paper is set out as follows. In Section 2 we propose a convex penalized criterion which has connections to both the graphical lasso (Yuan and Lin, 2007) and the group lasso (Yuan and Lin, 2006). Minimizing our functional graphical lasso (fglasso) criterion provides an estimate, $\widehat{E}$, for the edge set, E. We also propose a joint fglasso approach for the case of estimating multiple graphs simultaneously. An efficient block coordinate descent algorithm for minimizing the fglasso criterion is presented in Section 3. In addition, we demonstrate a method to extend the fglasso algorithm to handle even larger values of p by applying the partition approach of Witten et al. (2011). Section 4 provides our theoretical results. Specifically, we show that the estimated edge set $\widehat{E}$ is the same as the true edge set E with probability converging to one. The finite sample performance of the fglasso is examined in Section 5 through a series of simulation studies. Section 5 also provides a demonstration of the fglasso on the EEG data set. Further discussion of our approach, as well as some extensions and limitations, are presented in Section 6. We relegate all the technical proofs

to the Supplementary Material.

2 Methodology

2.1 Gaussian Graphical Models

As discussed in the previous section, the edges in a graphical model depict the conditional dependence structure of the p variables. Specifically, let

$$c_{jl} = \text{Cov}\left(X_j, X_l \middle| \{X_k, k \neq j, l\}\right) \tag{1}$$

represent the covariance of X_j and X_l conditional on the remaining variables. Then nodes j and l are connected by an edge if and only if $c_{jl} \neq 0$.

Under the assumption that $\mathbf{X} = (X_1, \ldots, X_p)^T$ is multivariate Gaussian with covariance matrix $\mathbf{\Sigma}^*$, one can show that $c_{jl} = 0$ if and only if $\Theta^*_{jl} = 0$, where Θ^*_{jl} is the (j,l)-th component of the precision matrix, $\mathbf{\Theta}^* = \mathbf{\Sigma}^{*-1}$. Let $G = (V, E)$ denote an undirected graph with vertex set $V = \{1, \ldots, p\}$ and edge set $E = \{(j,l) : c_{jl} \neq 0, (j,l) \in V^2, j \neq l\} = \{(j,l) : \Theta^*_{jl} \neq 0, (j,l) \in V^2, j \neq l\}$. In practice Θ^*_{jl}, and hence the network structure, must be estimated based on a set of n observed p-dimensional realizations, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, of the random vector $\mathbf{X}$. Hence, much of the research in this area involves various approaches for estimating E, which for Gaussian data is equivalent to identifying the locations of the non-zero elements in the precision matrix.

The graphical lasso (Yuan and Lin, 2007) considers a regularized estimator for $\mathbf{\Theta}^*$ by adding an ℓ_1 penalty on the off-diagonal entries of the precision matrix to the Gaussian log-likelihood (up to constants):

$$\widehat{\mathbf{\Theta}} = \underset{\mathbf{\Theta}}{\text{argmax}} \left\{ \log\det \mathbf{\Theta} - \text{trace}(\mathbf{S}\mathbf{\Theta}) - \gamma_n \sum_{j \neq l} |\Theta_{jl}| \right\}, \tag{2}$$

where $\mathbf{\Theta} \in \mathbb{R}^{p \times p}$ is symmetric positive definite, $\mathbf{S}$ is the sample covariance matrix of $\mathbf{x}_1, ..., \mathbf{x}_n$ and γ_n is a non-negative tuning parameter. In a similar fashion to the standard lasso, the

ℓ_1 penalty in (2) both regularizes the estimate and ensures that $\widehat{\mathbf{\Theta}}$ is sparse i.e. has many zero elements.

2.2 Functional Graphical Models

The functional setting considered in this paper is more complicated than that for the standard graphical model. Suppose the functional variables, $g_1(t), \ldots, g_p(t)$, jointly following from a p-dimensional MGP, belong to an undirected graph $G = (V, E)$ with vertex set $V = \{1, \ldots, p\}$ and edge set E. Then we must first provide a definition for the conditional covariance between two functions. For each pair $(j, l) \in V^2, j \neq l$ and any $(s, t) \in \mathcal{T}^2$, we define the conditional cross covariance function by

$$C_{jl}(s,t) = \text{Cov}\left(g_j(s), g_l(t) \middle| \{g_k(\cdot), k \neq j, l\}\right), \tag{3}$$

which represents the covariance between $g_j(s)$ and $g_l(t)$ conditional on the remaining $p-2$ functions. Here we can use the projection theorem for Hilbert spaces [Chapter 2.5 in Hsing and Eubank (2015)] to rigorously define the relevant conditional expectation terms, e.g. $E\big(g_j(s)|\{g_k(\cdot), k \neq j, l\}\big)$. See also the definition of the conditional joint probability measure within Hilbert spaces in Zhu et al. (2016). Note that g_j and g_l are conditionally independent if and only if $C_{jl}(s,t) = 0$ for all $(s,t) \in \mathcal{T}^2$. Hence our ultimate goal is to recover the edge set

$$E = \left\{(j,l) : C_{jl}(s,t) \neq 0 \text{ for some } s \text{ and } t, (j,l) \in V^2, j \neq l\right\}. \tag{4}$$

Suppose we observe $\mathbf{g}_i = (g_{i1}, \ldots, g_{ip})^T$, $i = 1, \ldots, n$, and for each i, $g_{ij}(t)$ $t \in \mathcal{T}$ is a realization from a mean zero Gaussian process[2]. The Karhunen-Loève expansion (Theorem 1.5 in Bosq (2000)) allows us to represent each function in the form

$$g_{ij}(t) = \sum_{k=1}^{\infty} a_{ijk}\phi_{jk}(t),$$

[2]Our methodological and theoretical results can be extended to the case of Gaussian processes with non-zero means but for clarity of the exposition, we do not investigate that case here.

where $a_{ijk} = \int_{\mathcal{T}} g_{ij}(t)\phi_{jk}(t)dt \sim N(0, \lambda_{jk})$, a_{ijk} is independent from $a_{i'jk'}$ for $i \neq i'$ or $k \neq k'$, and $\lambda_{j1} \geq \lambda_{j2} \geq \cdots \geq 0$ for $j = 1, \ldots, p$. In this formulation $\{\phi_{jk}(t)\}_{k=1}^{\infty}$ represent principal component functions and form an infinite dimensional basis representation for $g_{ij}(t)$.

Since each functional object is either infinite or high dimensional, some form of dimension reduction is needed. Let

$$g_{ij}^{M}(t) = \sum_{k=1}^{M} a_{ijk}\phi_{jk}(t) \tag{5}$$

represent the M-truncated version of $g_{ij}(t)$. Then $g_{ij}^{M}(t)$ provides the best M-dimensional approximation to $g_{ij}(t)$ in terms of integrated mean squared error. Let $\mathbf{a}_i^M = \left((\mathbf{a}_{i1}^M)^T, \ldots, (\mathbf{a}_{ip}^M)^T\right)^T \in \mathbb{R}^{Mp}$ represent the first M principal component scores for the ith set of functions for $i = 1, \ldots, n$, where $\mathbf{a}_{ij}^M = (a_{ij1}, \ldots, a_{ijM})^T$. Then, provided $g_{ij}(t)$ is a realization from a Gaussian process, $\mathbf{a}_i^M$ will have a multivariate Gaussian distribution with covariance matrix $\mathbf{\Sigma}^{*M} = \left(\mathbf{\Theta}^{*M}\right)^{-1}$. Analogously to (3), we can define the M-truncated conditional cross covariance function by

$$C_{jl}^{M}(s,t) = \mathrm{Cov}\left(g_j^M(s), g_l^M(t) | \{g_k^M(\cdot), k \neq j, l\}\right). \tag{6}$$

Our goal is to recover the edge set E based on $\{C_{jl}^{M}(s,t), (j,l) \in V^2\}$. Since the principal component scores, $\mathbf{a}_i^M$, and hence $\mathbf{\Theta}^{*M}$, share the same information as $\{g_{ij}^{M}(t), j = 1, \ldots, p\}$, one might expect to see a connection between E and $\mathbf{\Theta}^{*M}$.

To gain some intuition, we first consider the special case where $g_{ij}(t)$ is exactly M-dimensional i.e. $g_{ij}(t) = g_{ij}^{M}(t)$. In this simplified setting the following lemma provides a precise statement of the connection between E and $\mathbf{\Theta}^{*M}$.

Lemma 1 *For $(j,l) \in V^2$, let $\mathbf{\Theta}_{jl}^{*M}$ be the $M \times M$ matrix corresponding to the (j,l)-th submatrix of $\mathbf{\Theta}^{*M}$. Then*

$$E = \left\{(j,l) : \|\mathbf{\Theta}_{jl}^{*M}\|_F \neq 0, (j,l) \in V^2, j \neq l\right\}, \tag{7}$$

where $||\cdot||_F$ denotes the Frobenius norm.

Lemma 1 suggests that the problem of recovering E can be reduced to one of accurately estimating the block sparsity structure in $\mathbf{\Theta}^{*M}$. Although Lemma 1 is not directly applicable when $M = \infty$ and it only applies to the setting where $g_{ij}(t)$ is exactly M-dimensional, even in the more general setting where the dimension of $g_{ij}(t)$ approaches infinity, our theoretical results in Section 4 can still formalize this connection. In the following section we develop an approach to estimate the block sparsity structure in $\mathbf{\Theta}^{*M}$ which, for a large enough M, provides an accurate estimate for E.

2.3 Functional Graphical Lasso

In this section, we first introduce the estimation procedure for the relevant terms in Section 2.2 and then propose our approach to estimate the true edge set, E.

If we denote the covariance function by $K_{jj}(s,t) = \text{Cov}(g_j(s), g_j(t))$, then the corresponding eigen-pairs satisfy

$$\int_{\mathcal{T}} K_{jj}(s,t)\phi_{jk}(t)dt = \lambda_{jk}\phi_{jk}(s), \tag{8}$$

where $\int_{\mathcal{T}} \phi_{jk}(t)^2 dt = 1$ and $\int_{\mathcal{T}} \phi_{jk}(t)\phi_{jm}(t)dt = 0$ for $m < k$. An empirical estimator for $K_{jj}(s,t)$ is given by

$$\widehat{K}_{jj}(s,t) = \frac{1}{n}\sum_{i=1}^{n} \left(g_{ij}(s) - \bar{g}_j(s)\right)\left(g_{ij}(t) - \bar{g}_j(t)\right)$$

where $\bar{g}_j = n^{-1}\sum_i g_{ij}$. Performing the eigen-decomposition on $\widehat{K}_{jj}(s,t)$, we obtain the estimators $(\widehat{\lambda}_{jk}, \widehat{\phi}_{jk})$ for $(\lambda_{jk}, \phi_{jk})$ as defined in (8) and the estimated principal component scores $\widehat{a}_{ijk} = \int_{\mathcal{T}} g_{ij}(t)\widehat{\phi}_{jk}(t)dt$.

Let $\widehat{\mathbf{a}}_i^M = \left((\widehat{\mathbf{a}}_{i1}^M)^T, \ldots, (\widehat{\mathbf{a}}_{ip}^M)^T\right)^T \in \mathbb{R}^{Mp}$, where $\widehat{\mathbf{a}}_{ij}^M = (\widehat{a}_{ij1}, \ldots, \widehat{a}_{ijM})^T$ and $\mathbf{S}^M$ be the sample covariance matrix of $\widehat{\mathbf{a}}_i^M$. Motivated by Lemma 1, we propose the *functional graphical lasso (fglasso)* to estimate the network structure. The fglasso modifies the graphical lasso by incorporating a group lasso penalty (Yuan and Lin, 2006) to produce a block sparsity

structure. Specifically, the fglasso is defined as the solution to

$$\widehat{\mathbf{\Theta}}^{M} = \underset{\mathbf{\Theta}^{M}}{\operatorname{argmax}} \left\{ \log\det \mathbf{\Theta}^{M} - \operatorname{trace}(\mathbf{S}^{M}\mathbf{\Theta}^{M}) - \gamma_n \sum_{j\neq l} \|\mathbf{\Theta}^{M}_{jl}\|_F \right\}, \tag{9}$$

where $\mathbf{\Theta}^{M} \in \mathbb{R}^{Mp\times Mp}$ is symmetric positive definite and γ_n is a non-negative tuning parameter. The group lasso penalty in (9) forces the elements of $\mathbf{\Theta}^{M}_{jl}$ to either all be zero (a sparse solution) or all non-zero (a connected edge between g_j and g_l). Hence, as γ_n increases $\widehat{\mathbf{\Theta}}^{M}$ grows sparser in a blockwise fashion. Our final estimated edge set is then defined as

$$\widehat{E}^{M} = \left\{ (j,l) : \|\widehat{\mathbf{\Theta}}^{M}_{jl}\|_F \neq 0, (j,l) \in V^2, j \neq l \right\}. \tag{10}$$

Note $\mathbf{\Theta}$, $\widehat{\mathbf{\Theta}}$, $\mathbf{S}$, $\widehat{E}$, $\mathbf{a}_{ij}$, $\widehat{\mathbf{a}}_{ij}$ and $\boldsymbol{\phi}_j$ all depend on M, but for simplicity of notation we will omit the corresponding superscripts where the context is clear.

2.4 Joint Functional Graphical Lasso

For scalar data, Danaher et al. (2014) proposed a *joint graphical lasso* to jointly estimate separate graphical models for each of Q different groups in situations where the groups can be assumed to share similar network structures. The joint graphical lasso attempts to borrow strength across the groups to estimate connections that are common to all Q networks while still allowing for differences among the groups. In the functional domain, given Q data sets, one would observe $\mathbf{a}^{(q)} \in \mathbb{R}^{Mp}, q = 1, \ldots, Q$, with each $\mathbf{a}^{(q)}$ following a multivariate Gaussian distribution with covariance matrix, $\mathbf{\Sigma}^{*(q)} = (\mathbf{\Theta}^{*(q)})^{-1}$. Then the *joint functional graphical lasso* would correspond to finding $\{\widehat{\mathbf{\Theta}}\} = \widehat{\mathbf{\Theta}}^{(1)}, \ldots, \widehat{\mathbf{\Theta}}^{(Q)}$ to maximize

$$\sum_{q=1}^{Q} n_q \left\{ \log\det \mathbf{\Theta}^{(q)} - \operatorname{trace}(\mathbf{S}^{(q)}\mathbf{\Theta}^{(q)}) \right\} - P_1(\{\mathbf{\Theta}\}) - P_2(\{\mathbf{\Theta}\}), \tag{11}$$

where n_q is the number of observations, and $\mathbf{S}^{(q)}$ is the sample covariance matrix of $\widehat{\mathbf{a}}_i^{(q)}$, for the qth class.

The first penalty term in (11), i.e $P_1(\{\mathbf{\Theta}\}) = \gamma_{1n} \sum_q \sum_{j\neq l} \|\mathbf{\Theta}^{(q)}_{jl}\|_F$, $\gamma_{1n} \geq 0$, produces a block sparsity structure for each $\mathbf{\Theta}^{(q)}$, while P_2 encourages a common structure among the

$\boldsymbol{\Theta}^{(q)}$'s. When $P_2(\{\boldsymbol{\Theta}\}) = 0$, (11) reduces to performing Q uncoupled fglasso problems (9). Here we consider using a group lasso penalty for P_2, i.e.

$$P_2(\{\boldsymbol{\Theta}\}) = \gamma_{2n} \sum_{j \neq l} \left(\sum_q ||\boldsymbol{\Theta}_{jl}^{(q)}||_F^2 \right)^{1/2}, \tag{12}$$

where γ_{2n} is a non-negative tuning parameter and (12) encourages a similar pattern of zero blocks across all the precision matrices. In particular as γ_{2n} grows larger, then the structure of the Q networks will become more similar. In the scalar data setting Danaher et al. (2014) also considered the fused lasso (Tibshirani et al., 2005) as a candidate penalty for P_2, which encourages a stronger form of similarity across the $\boldsymbol{\Theta}^{(q)}$'s by allowing not only similar network structure but also similar edge values. This idea can be naturally extended to our functional setting, but we do not explore that here due to space considerations.

3 Computation

3.1 Fglasso Algorithm

A number of efficient algorithms (Friedman et al., 2007; Boyd et al., 2010) have been developed to solve the glasso problem, but to date none of these approaches have considered the functional domain. Here we propose an algorithm which mirrors recent techniques for optimizing the glasso crierion (Zhu et al., 2014).

Let $\boldsymbol{\Theta}_{-j}, \boldsymbol{\Sigma}_{-j}$ and $\mathbf{S}_{-j}$ respectively be $M(p-1) \times M(p-1)$ sub matrices excluding the jth row and column block of $\boldsymbol{\Theta}, \boldsymbol{\Sigma}$ and $\mathbf{S}$, and let $\mathbf{w}_j, \boldsymbol{\sigma}_j$ and $\mathbf{s}_j$ be $M(p-1) \times M$ matrices representing the jth column block after excluding the jth row block. Finally, let $\boldsymbol{\Theta}_{jj}, \boldsymbol{\Sigma}_{jj}$ and $\mathbf{S}_{jj}$ be the (j,j)th $M \times M$ blocks in $\boldsymbol{\Theta}, \boldsymbol{\Sigma}$ and $\mathbf{S}$ respectively. So, for instance for $j = 1$, $\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\Theta}_{11} & \mathbf{w}_1^T \\ \mathbf{w}_1 & \boldsymbol{\Theta}_{-1} \end{pmatrix}$. Then, for a fixed value of $\boldsymbol{\Theta}_{-j}$, standard calculations show that (9) is solved by setting

$$\widehat{\boldsymbol{\Theta}}_{jj} = \mathbf{S}_{jj}^{-1} + \widehat{\mathbf{w}}_j^T \boldsymbol{\Theta}_{-j}^{-1} \widehat{\mathbf{w}}_j, \tag{13}$$

where

$$\widehat{\mathbf{w}}_j = \arg\min_{\mathbf{w}_j} \left\{ \text{trace}(\mathbf{S}_{jj}\mathbf{w}_j^T\boldsymbol{\Theta}_{-j}^{-1}\mathbf{w}_j) + 2\text{trace}(\mathbf{s}_j^T\mathbf{w}_j) + 2\gamma_n \sum_{l=1}^{p-1} \|\mathbf{w}_{jl}\|_F \right\}, \tag{14}$$

and $\mathbf{w}_{jl}$ represents the lth $M \times M$ block of $\mathbf{w}_j$. Computing (14) can be achieved using some matrix calculus with details provided in Section B.1 of the Supplementary Material.

This suggests a block coordinate descent algorithm where one iterates through j repeatedly computing (14) until convergence. In fact by checking the conditions of Theorem 4.1 in Tseng (2001) it is easy to verify that iteratively minimizing (14) over $\mathbf{w}_j$ and updating $\boldsymbol{\Theta}_{jj}$ by (13) for $j = 1, \ldots, p$ provides a convergent solution for globally maximizing the fglasso criterion. The main potential difficulty with this approach is that $\boldsymbol{\Theta}_{-j}^{-1}$ must be updated at each step which would be computationally expensive if we performed the matrix inversion directly. However, Algorithm 1 demonstrates that the calculation can be performed efficiently. Steps 2(a) and 2(c) are derived using standard matrix results, the details of which are provided in Section B.2 of the Supplementary Material. We also develop in Section B.3 an analogous algorithm for solving (11) when jointly estimating multiple networks.

Algorithm 1 Functional Graphical Lasso Algorithm

1. Initialize $\widehat{\boldsymbol{\Theta}} = \mathbf{I}_{Mp}$ and $\widehat{\boldsymbol{\Sigma}} = \mathbf{I}_{Mp}$.

2. Repeat until convergence for $j = 1, \ldots, p$.

   (a) Compute $\widehat{\boldsymbol{\Theta}}_{-j}^{-1} \leftarrow \widehat{\boldsymbol{\Sigma}}_{-j} - \widehat{\boldsymbol{\sigma}}_j \widehat{\boldsymbol{\Sigma}}_{jj}^{-1} \widehat{\boldsymbol{\sigma}}_j^T$.

   (b) Solve for $\widehat{\mathbf{w}}_j$ in (14) using Algorithm 3 in the Supplementary Material.

   (c) Reconstruct $\widehat{\boldsymbol{\Sigma}}$ using $\widehat{\boldsymbol{\Sigma}}_{jj} = \mathbf{S}_{jj}$, $\widehat{\boldsymbol{\sigma}}_j = -\mathbf{U}_j\mathbf{S}_{jj}$ and $\widehat{\boldsymbol{\Sigma}}_{-j} = \widehat{\boldsymbol{\Theta}}_{-j}^{-1} + \mathbf{U}_j\mathbf{S}_{jj}\mathbf{U}_j^T$, where $\mathbf{U}_j = \widehat{\boldsymbol{\Theta}}_{-j}^{-1}\widehat{\mathbf{w}}_j$.

3. Set $\widehat{E} = \left\{ (j,l) : \|\widehat{\boldsymbol{\Theta}}_{jl}\|_F \neq 0, (j,l) \in V^2, j \neq l \right\}$.

3.2 Block Partitioning to Accelerate the Algorithm

A common approach to significantly speed up the glasso algorithm involves first performing a screening step on the sample covariance matrix to partition the nodes into K distinct sets and then solving K separate glasso problems (Witten et al., 2011; Mazumder and Hastie, 2012b; Danaher et al., 2014; Zhu et al., 2014). Since each resulting network involves many fewer nodes the glasso problem can be computed at a much lower computational cost.

Here we show that a similar approach can be used to significantly accelerate our proposed fglasso algorithm.

Proposition 1 *If the solution to (9) takes a block diagonal form, i.e.* $\mathbf{\Theta} = diag\,(\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_K)$, *then (9) can be computed by separately solving K smaller fglasso problems*

$$\widehat{\mathbf{\Theta}}_k = \arg\max_{\mathbf{\Theta}_k} \left\{ \log\det \mathbf{\Theta}_k - trace(\mathbf{S}_k \mathbf{\Theta}_k) - \gamma_n \sum_{j \neq l} \|\mathbf{\Theta}_{k,jl}\|_F \right\}, \tag{15}$$

for $k = 1, \ldots, K$, where $\mathbf{S}_k$ is the submatrix of $\mathbf{S}$ corresponding to $\mathbf{\Theta}_k$ and $\mathbf{\Theta}_{k,jl}$ is the (j,l)-th submatrix of $\mathbf{\Theta}_k$.

Proposition 2 *Without loss of generality, let $G_1, \ldots, G_K$ be a partition of p ordered features, hence if $i \in G_k$, $i' \in G_{k'}$, $k < k'$, then $i < i'$. Then a necessary and sufficient condition for the solution to the fglasso problem to be block diagonal with blocks indexed by $G_1, \ldots, G_K$ is that $\|\mathbf{S}_{ii'}\|_F \leq \gamma_n$ for all $i \in G_k$, $i' \in G_{k'}$, $k \neq k'$.*

Propositions 1 and 2 suggest first performing a screening procedure on $\mathbf{S}$ to identify K distinct graphs and then solving the resulting K fglasso problems. These steps are summarized in Algorithm 2.

For a fixed M, implementing Algorithm 1 requires $O(p^3)$ operations. Steps 1 and 2 in Algorithm 2 need $O(p^2)$ operations and the kth fglasso problem requires $O(|G_k|^3)$ operations for $k = 1, \ldots, K$, hence the total computational cost for Algorithm 2 is $O\left(p^2 + \sum_{k=1}^{K} |G_k|^3\right)$.

Algorithm 2 Fglasso Algorithm with Partitioning Rule

1. Let $\mathbf{A}$ be a p by p adjacency matrix, whose diagonal elements are one and off-diagonal elements take the form $A_{ii'} = 1_{\|\mathbf{S}_{ii'}\|_F > \gamma_n}$.

2. Identify K connected components of the graph based on the adjacency matrix $\mathbf{A}$. Let G_k be the index set of the features in the kth connected component, $k = 1, \ldots, K$.

3. For $k = 1, \ldots, K$, solve $\widehat{\mathbf{\Theta}}_k$ via Algorithm 1 using the nodes in G_k. The final solution to the fglasso problem $\widehat{\mathbf{\Theta}}$ is obtained by rearranging the rows/columns of the permuted version, $\text{diag}\big(\widehat{\mathbf{\Theta}}_1, \ldots, \widehat{\mathbf{\Theta}}_K\big)$.

Algorithm 2 significantly reduces the computational cost, if $|G_1|, \ldots, |G_K|$ are much smaller than p, which is the situation when the tuning parameter, γ_n, is large. This is the case we are generally interested in for real data problems since, for the sake of network interpretation and visualization, most practical applications estimate sparse networks.

3.3 Selection of Tuning Parameters

Estimating the FGM requires choosing M (number of selected principal components) and γ_n (regularization parameter to tune the block sparsity level of $\mathbf{\Theta}$). First, to select M, one can either adopt leave-one-curve-out cross validation (Rice and Silverman, 1991) or an AIC-type criteria (Yao et al., 2005) . To reach a compromise, we develop a J-fold cross-validated (CV) approach. Specifically, let h_{ijk} represent a noisy observation of $g_{ij}(t_k)$. We randomly divide the set of observed time points into J equal-size folds. We then treat one group for each $g_{ij}(t)$ as a validation data set, apply FPCA on the remaining $J-1$ groups, where each function is approximated by a L-dimensional B-spline basis [Chapter 8 of Ramsay and Silverman (2005)], calculate the squared error between h_{ijk} and the fitted values $\widehat{g}_{ij}(t_k)$ (via (5)) on the validation set, and repeat this procedure J-times to compute the CV squared error. We calculate the CV errors over a grid of $M \leq L$ values and choose the pair with the lowest

error. In general, we can select a different number of principal components for each random function that results in estimating matrices with non-square blocks. However, to simplify the computation in Algorithm 1, we use an identical number across $j \in V$ under the assumption that the corresponding covariance operators, $K_{jj}(s,t)$'s, share similar complexity structure.

Second, to choose the tuning parameter γ_n, there exist a number of possible approaches. Approaches such as AIC/BIC, cross-validation and stability selection (Meinshausen and Buhlmann, 2010) are popular and have been well studied in the graphical model literature. However, given the complicated functional structure of FGM, it is unclear how to compute the effective degrees of freedom for AIC/BIC. Alternatively, with some prior information about the targeted network density, one can select the value of γ_n that results in the network with a desired sparsity level. In our simulations, we fit our approach over a sequence of γ_n values, and generate corresponding ROC curves to explore the graph selection consistency.

4 Theoretical Properties

We now investigate the theoretical properties of the fglasso proposed in Section 2.3. The model selection consistency of the fglasso, *i.e.* the exact functional graph recovery with overwhelming probability, are established under some regularity conditions.

We begin by introducing Condition 1 as a basic assumption in our functional setting.

Condition 1 *(i) The truncated dimension of the functional data, M, satisfies $M \asymp n^{\alpha}$ with some constant $\alpha \geq 0$; (ii) The eigenvalues satisfy $\lambda_{j1} > \lambda_{j2} > \cdots > \lambda_{jM} > \lambda_{j(M+1)} \geq \cdots$ with $\max_{j \in V} \sum_{k=1}^{\infty} \lambda_{jk} < \infty$ and there exists some constant $\beta > 1$ with $\alpha\beta \leq 1/4$ such that for each $k = 1, \ldots, M$, $\lambda_{jk} \asymp k^{-\beta}$ and $d_{jk}\lambda_{jk} = O(k)$ uniformly in $j \in V$, where $d_{jk} = 2\sqrt{2}\max\{(\lambda_{j(k-1)} - \lambda_{jk})^{-1}, (\lambda_{jk} - \lambda_{j(k+1)})^{-1}\}$; (iii) The principal component functions $\phi_{jk}(s)$'s are continuous on the compact set $\mathcal{U}$ and satisfy $\max_{j \in V} \sup_{s \in \mathcal{T}} \sup_{k \geq 1} |\phi_{jk}(s)| = O(1)$.*

Here $a_n \asymp b_n$ denotes $B \leq \inf_n |a_n/b_n| \leq \sup_n |a_n/b_n| \leq A$ for some positive constants A and B. The parameter β determines the decay rate of any decreasing sequence $\lambda_{j1} >$

$\lambda_{j2} > \cdots > \lambda_{jM}$ for $j \in V$ and $d_{jk}\lambda_{jk} = O(k)$ restricts the decay rate of eigen-gaps, $d_{jk}^{-1} \geq d_0 k^{-(\beta+1)}$ with some positive constant d_0 for $j \in V$, see also Bosq (2000) for more details. The parameter α controls the number of selected principal components that provide a reasonable approximation to the infinite-dimensional process. It is easy to see that larger values of β yield a faster decay rate, while increasing α results in a value for larger M.

To show the model selection consistency of the fglasso, we first need to establish concentration bounds for all entries of $\mathbf{S} - \mathbf{\Sigma}^*$. Denote the (j,l)-th $M \times M$ submatrix of $\mathbf{S}$ by $\mathbf{S}_{jl}$ and the (k,m)-th entry of $\mathbf{S}_{jl}$ by $\widehat{\sigma}_{jlkm}$ for $j,l = 1,\ldots,p$ and $k,m = 1,\ldots,M$. Similarly, let $\mathbf{\Sigma}^* = (\mathbf{\Sigma}^*_{jl})_{1\leq j,l\leq p}$, where $\mathbf{\Sigma}^*_{jl} = (\sigma^*_{jlkm})_{1\leq k,m\leq M}$.

Theorem 1 *Suppose that Condition 1 holds. Then there exist two positive constants C_1 and C_2 such that*

(i) for $0 < \delta \leq C_1$ and each $j = 1,\ldots,p$ and $k = 1,\ldots,M$,

$$P\Big(\big|\widehat{\sigma}_{jjkk} - \sigma^*_{jjkk}\big| \geq \delta\Big) \leq C_2\Big\{\exp\big(-C_1 n k^{-2}\delta^2\big) + \exp\big(-C_1 n k^{-(2+2\beta)}\delta\big)\Big\}; \tag{16}$$

(ii) for $0 < \delta \leq C_1$ and each $(j,l) \in V^2, j \neq l$ and $k,m = 1,\ldots,M$,

$$P\Big(\big|\widehat{\sigma}_{jlkm} - \sigma^*_{jlkm}\big| \geq \delta\Big) \leq C_2 \exp\Big\{-C_1 n\big(k+m\big)^{-(2+2\beta)}\delta^2\Big\}. \tag{17}$$

In particular, there exist two positive constants C_1 and C_2 such that for any δ with $0 < \delta \leq C_1$ and for all $j,l = 1,\ldots,p$ and $k,m = 1,\ldots,M$,

$$P\Big(\big|\widehat{\sigma}_{jlkm} - \sigma^*_{jlkm}\big| \geq \delta\Big) \leq C_2 \exp\Big\{-C_1 n^{1-2\alpha(1+\beta)}\delta^2\Big\}. \tag{18}$$

Theorem 1 provides a general result for the tail probability of $\widehat{\sigma}_{jlkm} - \sigma^*_{jlkm}$ and indicates that the magnitudes of the λ_{jk}'s play an important role in their tail behavior. In particular, if each component in the MGP $\{g_{ij}, j = 1,\ldots,p\}$ is a fixed dimensional object ($\alpha = 0$), then $\widehat{\sigma}_{jlkm} - \sigma^*_{jlkm}$ behaves in a sub-Guassian fashion i.e.

$$P\Big(\big|\widehat{\sigma}_{jlkm} - \sigma^*_{jlkm}\big| \geq \delta\Big) \leq C_2 \exp\big(-C_1 n\delta^2\big),$$

for $0 < \delta \leq C_1$, where C_1 and C_2 are two positive constants.

To state our main result in Theorem 2, we present several regularity conditions.

Condition 2 *(i) The truncated dimension of the functional data, M, satisfies $M \asymp n^{\alpha}$ with some constant $\alpha \geq 0$; (ii) There exists some integer $M_n \geq M$ and constant $\beta > 1$ with $\alpha\beta \leq 1/4$ such that $\lambda_{j1} > \lambda_{j2} > \cdots > \lambda_{jM} > \lambda_{j(M+1)} \geq \cdots \lambda_{jM_n} > 0$ and $\lambda_{jk} = 0$ if $k \geq M_n + 1$, where for each $k = 1, \ldots, M$, $\lambda_{jk} \asymp k^{-\beta}$, $d_{jk}\lambda_{jk} = O(k)$, and $\sum_{k=M+1}^{M_n} \lambda_{jk} \leq O\big(n^{\alpha(1-\beta)}\big)$ uniformly in $j \in V$; (iii) The principal component functions $\phi_{jk}(s)$'s are continuous on the compact set $\mathcal{U}$ and satisfy $\sup_{s\in\mathcal{T}} \max_{1\leq k\leq M_n} |\phi_{jk}(s)| = O(1)$ uniformly in $j \in V$.*

Condition 2 is nearly the same as Condition 1 except for the incorporation of the intrinsic dimension of the functional data, M_n. Our assumption that M_n is finite simplifies the statement of Conditions 3–4 below. However, it should be noted that M_n can be made arbitrarily large relative to n, e.g. $M_n = 1000$ and $n = 200$. Hence, this assumption does not place a practical constraint on our method.

Denote $\widetilde{\mathbf{a}}_{1j} = (a_{1j1}, \ldots, a_{1jM_n})^T$ for $j = 1, \ldots, p$. Let $\mathbf{\Sigma}$ be the population covariance matrix of $(\widetilde{\mathbf{a}}_{11}^T, \ldots, \widetilde{\mathbf{a}}_{1p}^T)^T$, and $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = (\mathbf{\Omega}_{jl})_{1\leq j,l\leq p}$, where $\mathbf{\Omega}_{jl}$ is the (j,l)-th $M_n \times M_n$ submatrix. Let $\mathbf{\Omega}_{jl} = \begin{pmatrix} \mathbf{\Omega}_{jl,1}^{(k)} & \mathbf{\Omega}_{jl,2}^{(k)} \\ \mathbf{\Omega}_{jl,3}^{(k)} & \mathbf{\Omega}_{jl,4}^{(k)} \end{pmatrix}$ for $(j,l) \in V^2$, where $\mathbf{\Omega}_{jl,1}^{(k)}$ is a $k \times k$ submatrix and $\mathbf{\Omega}_{jl,4}^{(k)}$ is a $(M_n - k) \times (M_n - k)$ submatrix.

Condition 3 *There exists some positive constant $\nu > 0$ such that*

$$\max_{(j,l)\in E} \left\| \mathbf{\Omega}_{jl,2}^{(M)} \right\|_F \leq O(n^{-\alpha\nu}). \tag{19}$$

Condition 4 *With α, β and ν defined in Conditions 2 and 3, $\mathbf{\Omega}_{jl,1}^{(M)}$ satisfies*

$$\min_{(j,l)\in E} \left\| \mathbf{\Omega}_{jl,1}^{(M)} \right\|_F \gg |E|^2 n^{\alpha(1-2\nu-\beta)}. \tag{20}$$

If we let $\mathbf{C}_{jl} = \mathrm{Cov}(\widetilde{\mathbf{a}}_{1j}, \widetilde{\mathbf{a}}_{1l} | \widetilde{\mathbf{a}}_{1k}, k \neq j, l)$, then $\int_{\mathcal{T}}\int_{\mathcal{T}} \{C_{jl}(s,t)\}^2 dsdt = \|\mathbf{C}_{jl}\|_F^2$ and $\mathbf{\Omega}_{jl} = -\mathbf{C}_{jj}^{-1}\mathbf{C}_{jl}\mathbf{C}_{ll}^{-1}$ for each $(j,l) \in V^2$ with $j \neq l$. In this sense, Condition 3 controls the effect of

biases between the truncated and true processes and Condition 4 requires the minimum signal strength for successful graph recovery to be much larger than $|E|^2 n^{\alpha(1-2\nu-\beta)}$. Conditions 3 and 4 are crucial for obtaining the rate of convergence of $||\mathbf{\Theta}^*_{jl}||_F$ for $(j,l) \in V^2$ and the equivalence between the truncated and true edge sets, respectively. See Lemmas 2 and 3 in the Supplementary Material for details. In particular, when $E \asymp pd$, we need ν to be large enough so as to satisfy Condition 4. In this case, $\max_{(j,l)\in E} \left\| \mathbf{\Omega}^{(k)}_{jl,2} \right\|_F$ in Condition 3 needs to be small. We provide an example satisfying Conditions 3 and 4 in Appendix A.2.

We next introduce an irrepresentable-type condition for deriving the exact functional graph recovery with overwhelming probability. Before stating the condition, we begin with some notation. Denote by $\mathbf{\Gamma}^* = \mathbf{\Theta}^{*-1} \otimes \mathbf{\Theta}^{*-1} \in \mathbb{R}^{(Mp)^2 \times (Mp)^2}$ with $\otimes$ the Kronecker product, and $\mathbf{\Gamma}^*_{JJ'}$ the $M^2|J| \times M^2|J'|$ submatrix of $\mathbf{\Gamma}^*$ with row and column blocks in J and J', respectively, for any subsets J, J' of V^2. For any block matrix $\mathbf{A} = (\mathbf{A}_{ij})$ with $\mathbf{A}_{ij} \in \mathbb{R}^{M\times M}, 1 \le i \le p, 1 \le j \le q$, define $||\mathbf{A}||^{(M)}_\infty = \max_{1\le i\le p} \sum_{j=1}^q ||\mathbf{A}_{ij}||_F$, $||\mathbf{A}||^{(M)}_{\max} = \max_{1\le i\le p} \max_{1\le j\le q} ||\mathbf{A}_{ij}||_F$ as the M-block versions of the matrix ℓ_∞-norm and elementwise ℓ_∞ norm, respectively. Let the augmented edge set be $S = E \cup \{(1,1), \ldots, (p,p)\}$.

Condition 5 *There exists some constant $\eta \in (0,1]$ such that*

$$||\mathbf{\Gamma}^*_{S^cS}(\mathbf{\Gamma}^*_{SS})^{-1}||^{(M^2)}_\infty \le 1 - \eta. \tag{21}$$

Our remark on Condition 5 is provided in Appendix A.2. We are now ready to present the main theorem on the model selection consistency of the fglasso for estimating FGM. Denote by $\kappa_{\mathbf{\Gamma}^*} = ||(\mathbf{\Gamma}^*_{SS})^{-1}||^{(M^2)}_\infty$, $\kappa_{\mathbf{\Sigma}^*} = ||\mathbf{\Sigma}^*||^{(M)}_\infty$, $\kappa_{\mathbf{B}^*} = ||\mathbf{\Theta}^{*-1}\mathbf{B}^*||^{(M)}_\infty \kappa^{-1}_{\mathbf{\Sigma}^*}$, where $\mathbf{B}^* = (\mathbf{B}^*_{jl})$ with $\mathbf{B}^*_{jl} = \mathbf{\Theta}^*_{jl}$ for $(j,l) \in S^c$, and $\mathbf{B}^*_{jl} = \mathbf{0}$ for $(j,l) \in S$. Here $\mathbf{B}^*$ represents the bias matrix caused by the truncated approximation using (5). Let $d = \max_{j\in V} |\{l \in V : (j,l) \in E\}|$, the maximum degree of the graph in the underlying FGM.

Theorem 2 *Suppose that Conditions 2–5 hold, there exists some positive constant c_1 such that $M = c_1 n^\alpha$ and the bias term satisfies $||\mathbf{B}^*||^{(M)}_{\max} \le \gamma_n \eta \kappa^{-2}_{\mathbf{\Sigma}^*}/16$. Let $\widehat{\mathbf{\Theta}}$ be the unique solu-*

tion to the fglasso problem (9) with $\gamma_n = 16\eta^{-1}\left(c_1 C_1^{-1/2} n^{2\alpha(2+\beta)-1}(\tau \log c_1 n^{\alpha} p + \log C_2)^{1/2}\right)$. *for some* $\tau > 2$. *Suppose the sample size* n *satisfies the lower bound*

$$n^{1-2\alpha(2+\beta)} > \max\{C_3 d^2, C_4 \Theta_{\min}^{*-2}\}[\tau\alpha \log n + \tau \log p + \log(C_2 c_1^{\tau})], \tag{22}$$

with $c_\eta = 2+16\eta^{-1}$, $\Theta_{\min}^* = \min_{(j,l)\in E} ||\mathbf{\Theta}_{jl}^*||_F$, $C_3 = \left\{6c_1 c_\eta C_1^{-1/2} \max\left\{\frac{\kappa_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*}}{1-3\kappa_{\mathbf{B}^*}\kappa_{\mathbf{\Sigma}^*}}, \frac{\kappa_{\mathbf{\Sigma}^*}^3 \kappa_{\mathbf{\Gamma}^*}^2 c_\eta}{1-3\kappa_{\mathbf{B}^*}\kappa_{\mathbf{\Sigma}^*}^3\kappa_{\mathbf{\Gamma}^*}c_\eta}\right\}\right\}^2$ *and* $C_4 = \{2c_1 c_\eta C_1^{-1/2}\kappa_{\mathbf{\Gamma}^*}\}^2$, *then with probability at least* $1-(c_1 n^{\alpha} p)^{2-\tau}$, *the estimated edge set* $\widehat{E}$ *is the same as the true edge set* E.

Let us further assume that $\kappa_{\mathbf{\Sigma}^*}$, $\kappa_{\mathbf{\Gamma}^*}$, $\kappa_{\mathbf{B}^*}$, η remain constant with respect to n, p, d. Let $\gtrsim$ denote the asymptotic lower bound. Then a sample size

$$n \gtrsim \left[(d^2 + \Theta_{\min}^{*-2})\tau \log p\right]^{\frac{1}{1-2\alpha(2+\beta)}} \tag{23}$$

guarantees the model selection consistency of the fglasso with probability at least $1-(c_1 n^{\alpha} p)^{2-\tau}$. Note that a larger value of parameter τ enables a higher functional graph recovery probability, at the expense of a larger sample size. In particular, for the case where M is bounded ($\alpha = 0$), the sample size condition (23) reduces to $n \gtrsim (d^2 + \Theta_{\min}^{*-2})\tau \log p$, which is consistent with the results for scalar data in Ravikumar et al. (2011). It is easy to see that a sample size $n \gtrsim (d^2 \tau \log p)^{1/(1-2\alpha(2+\beta))}$ is sufficient for ensuring model selection consistency as long as the minimum Frobenius norm within the true edge set $\Theta_{\min} \gtrsim \sqrt{\frac{\log p}{n^{1-2\alpha(2+\beta)}}}$. When the maximum node degree $d = o\left(\sqrt{\frac{p^{1-2\alpha(2+\beta)}}{\log p}}\right)$, model selection consistency can hold even in the $p \gg n$ regime.

5 Empirical Analysis

5.1 Simulations

We performed a number of simulation studies to compare the fglasso to potential competing methods. In each setting we generated $n \times p$ functional variables via $g_{ij}(t) = \mathbf{s}(t)^T \boldsymbol{\delta}_{ij}$, where

$\mathbf{s}(t)$ was a 5-dimensional Fourier basis function, and $\boldsymbol{\delta}_{ij} \in \mathbb{R}^5$ was a mean zero Gaussian random vector. Hence, $\boldsymbol{\delta}_i = (\boldsymbol{\delta}_{i1}^T, \ldots, \boldsymbol{\delta}_{ip}^T)^T \in \mathbb{R}^{5p}$ followed from a multivariate Gaussian distribution with covariance $\boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1}$. Different block sparsity patterns in the precision matrix, $\boldsymbol{\Theta}$, correspond to different conditional dependence structures. We considered three general structures.

- Model 1: We generate a block banded matrix $\boldsymbol{\Theta}$ with $\boldsymbol{\Theta}_{jj} = \mathbf{I}_5$, $\boldsymbol{\Theta}_{j,j-1} = \boldsymbol{\Theta}_{j-1,j} = 0.4\mathbf{I}_5$, $\boldsymbol{\Theta}_{j,j-2} = \boldsymbol{\Theta}_{j-2,j} = 0.2\mathbf{I}_5$ for $j = 1, \ldots, p$, and $\mathbf{0}$ at all other locations. Hence, only the adjacent two nodes are connected.

- Model 2: For $j = 1, \ldots, 10,\ 21, \ldots, 30, \ldots$, the corresponding submatrices in $\boldsymbol{\Theta}$ came from Model 1 with $p = 10$, indicating every alternating block of 10 nodes are connected by Model 1. For $j = 11, \ldots, 20,\ 31, \ldots, 40, \ldots$, we set $\boldsymbol{\Theta}_{jj} = \mathbf{I}_5$, so the remaining nodes were fully isolated.

- Model 3: We generate block sparse matrices without any special patterns. Specifically, we let each off-block-diagonal component in $\boldsymbol{\Theta}$ be generated independently and equals $0.5\mathbf{I}_5$ with probability 0.1 or $\mathbf{0}$ with probability 0.9. We also set each block-diagonal element to be $\delta'\mathbf{I}_5$, where δ' is chosen to guarantee the positive definiteness of $\boldsymbol{\Theta}$.

In all settings, we generated $n = 100$ observations of $\boldsymbol{\delta}_i$ from the associated multivariate Gaussian distribution, and the observed values, h_{ijk}, were sampled using

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

where each function was observed at $T = 100$ equally spaced time points, $0 = t_1, \ldots, t_{100} = 1$.

To implement the fglasso we must compute $\widehat{\mathbf{a}}_{ij}$, the first M estimated principal component scores of g_{ij}. As mentioned previously, this is a standard problem and there are a number of possible approaches one could use for the calculation. In order to mimic a real data setting we chose to fit each function using a L-dimensional B-spline basis (rather than using the

Fourier basis which was used to generate the data) and then compute $\widehat{\mathbf{a}}_{ij}$ from the basis coefficients. We used 5-fold cross-validation to choose both L and M, the details of which are discussed in Section 3.3. Typically, $L = 5$ to 10 basis functions and $M = 4, 5$ or 6 principal components were selected in our simulations.

We compared fglasso to four competing methods. In the first three methods we fit the standard glasso T times, once on each time point, producing T different network structures. We then used one of three possible rules to combine the T networks into a single FGM. ALL involved identifying an edge if it was selected in all T networks, NEVER identified an edge unless it appeared in none of the T networks, and HALF identifying an edge if it was selected in more than half of the T networks. The final approach, PCA, transformed the functional data into a standard format by computing the first estimated principal component score on each $g_{ij}(t)$ and then running the standard glasso on this data. The dimension of the B-spline basis function, L, was still selected by 5-fold cross validation after setting $M = 1$. We also implemented the competing approaches using CLIME (Cai et al., 2011), rather than glasso. This gave only a very slight improvement over glasso, so we do not report the results here.

For each method and tuning parameter, γ, we calculated the true positive rate (TPR_γ) and false positive rate (FPR_γ), in terms of network edges correctly identified. These quantities are defined by $\text{TPR}_\gamma = \text{TP}_\gamma/(\text{TP}_\gamma + \text{FN}_\gamma)$ and $\text{FPR}_\gamma = \text{FP}_\gamma/(\text{FP}_\gamma + \text{TN}_\gamma)$, where TP_γ and TN_γ respectively stand for true positives/negatives, and respectively FP_γ and FN_γ represent false positives/negatives. Plotting TPR_γ versus FPR_γ over a fine grid of values of γ produces a ROC curve, with curves close to the top left corner indicating a method that is performing well.

We considered different settings with $p = 50$, 100, 150, and ran each simulation 100 times. Figure 2 plots the median best ROC curves for each of the five comparison methods, respectively for Models 1–3. The fglasso (black curve) clearly had the best overall performance in recovering support of the functional network. Table 1 provides the area under the ROC curves (average over the 100 simulation runs) along with standard errors. Larger numbers

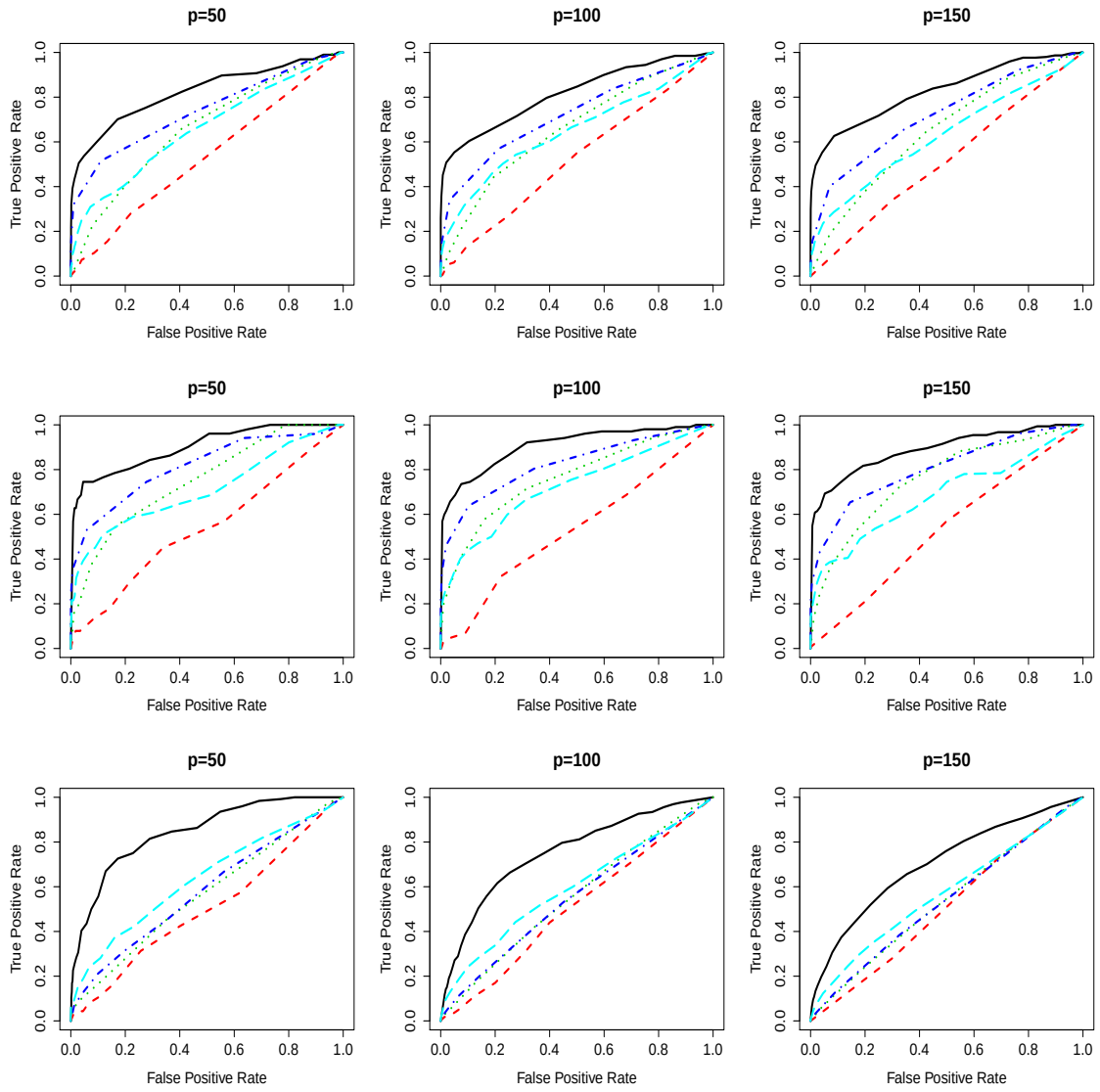

Figure 2: *Model* 1 *(top row), Model* 2 *(middle row) and Model* 3 *(bottom row) for* $p = 50, 100$ *and* 150*: Comparison of median estimated ROC curves over* 100 *simulation runs for fglasso (black solid), ALL (red dashed), NEVER (green dotted), HALF (blue dash dotted), and PCA (cyan long dashed).*

indicate superior estimates of the true network structure. Again we see that the fglasso provided highly significant improvements in accuracy for graph recovery over the competing methods in all the settings we considered. Among the four competing approaches, HALF performed the best for Models 1 and 2, and PCA slightly outperformed others for Model 3.

Table 1: The mean area under the ROC curves. Standard errors are shown in parentheses.

	fglasso	ALL	NEVER	HALF	PCA
p			Model 1		
50	0.82(0.02)	0.53(0.02)	0.66(0.02)	0.73(0.03)	0.66(0.03)
100	0.82(0.02)	0.52(0.02)	0.66(0.02)	0.73(0.03)	0.64(0.03)
150	0.83(0.02)	0.52(0.02)	0.65(0.02)	0.73(0.03)	0.64(0.03)
p			Model 2		
50	0.90(0.02)	0.54(0.02)	0.75(0.03)	0.82(0.03)	0.71(0.04)
100	0.90(0.02)	0.53(0.02)	0.75(0.03)	0.82(0.03)	0.72(0.04)
150	0.89(0.02)	0.53(0.03)	0.76(0.02)	0.81(0.03)	0.70(0.04)
p			Model 3		
50	0.85(0.02)	0.51(0.02)	0.56(0.03)	0.58(0.03)	0.63(0.04)
100	0.76(0.02)	0.51(0.02)	0.56(0.03)	0.55(0.03)	0.60(0.04)
150	0.71(0.03)	0.51(0.02)	0.54(0.03)	0.53(0.03)	0.57(0.04)

5.2 EEG Data

We test the performance of the fglasso on the EEG data set from the alcoholism study discussed in Section 1. The study consists of 122 subjects, 77 in the alcoholic group and 45 in the control group. For each subject, voltage values were measured from 64 electrodes placed on the scalp which were sampled at 256 Hz (3.9-ms epoch) for one second. Each subject completed 120 trials under either a single stimulus or two stimuli. The electrodes were located at standard positions (Standard Electrode Position Nomenclature, American Electroencephalographic Association (1990)). Zhang et al. (1995) discuss the data collection process in detail. Li et al. (2010); Zhou and Li (2014) analyze the data treating each covariate as a 256×64 matrix. We focus on the EEG signals filtered at α frequency bands between 8 and 12.5Hz, the case considered in Knyazev (2007), Hayden et al. (2006) and Zhu et al.

(2016). Using 4 representative electrodes from the frontal and parietal region of the scalp Hayden et al. (2006) found evidence of regional asymmetric patterns between the two groups. Zhu et al. (2016) discussed connectivity and asymmetry of electrodes selected from 5 different regions. Many authors used multiple samples per subject in order to obtain a sufficiently large sample, violating the independence assumption inherent in most methods. Following the analysis in Li et al. (2010) we only consider the average of all trials for each subject under the single stimulus condition. Thus we have at most $n = 77$ observations and aim to estimate a network involving $p = 64$ nodes/electrodes.

We first performed a preprocessing step using the *eegfilt* function (part of the *eeglab* toolbox) to perform α band filtering on the signals. The fglasso was then fitted to the filtered data. The dimension of the B-spline basis function, L, was selected using the same cross-validation approach as for the simulation study. We set $M = 6$ for this data since 6 principal components already explained more than 90% of the variation in the signal trajectories. Note that since our goal was to provide interpretable visualizations and investigate differences in brain connectivity between the alcoholic and control groups we computed sparse networks with approximately 5% connected edges. To assess the variability in the fglasso fit we performed a bootstrap procedure by randomly selecting n observations with replacement from the functional data, finding a tuning parameter γ_n to yield 5% sparsity level, applying the fglasso approach to the bootstrapped data, and repeating the above process 50 times. The "bootstrapped fglasso" was then constructed from the edges that occurred in at least 50% of the bootstrap replications.

Figure 3 plots the estimated network using the fglasso and the bootstrapped fglasso for both the alcoholic and the control groups. The bootstrapped fglasso estimated a sparser network with sparsity level 4.1% for the alcoholic group and 2.5% for the control group. We observe a few apparent patterns. First, electrodes from the frontal region are densely connected in both groups but the control group has increased connectivity relative to the alcoholic group. Second, the left central and parietal regions of the alcoholic group includes

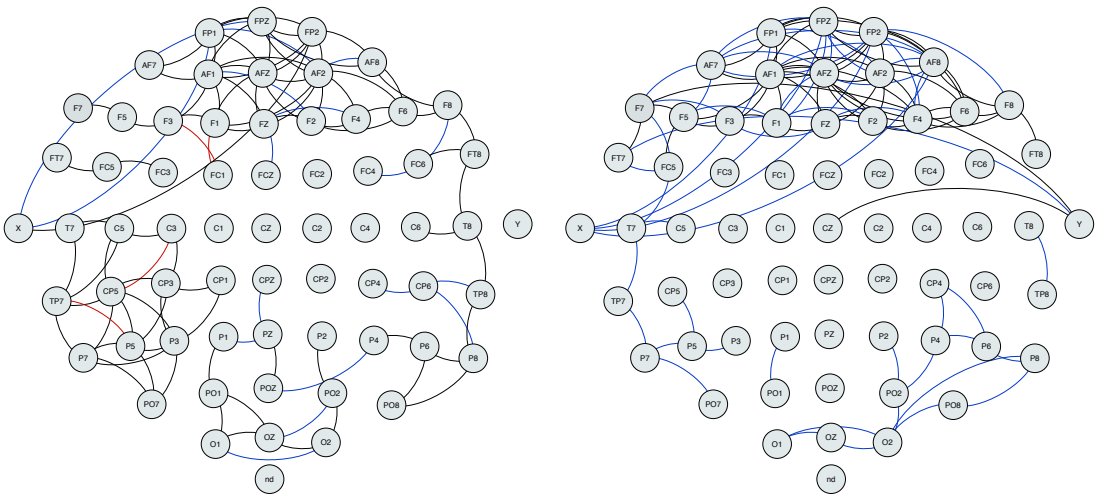

Figure 3: *Left graph plots the estimated network for the alcoholic group and right graph plots the estimated network for the control group. Black lines denote edges identified by both fglasso and bootstrapped fglasso, blue lines denote edges identified by fglasso but not selected by the bootstrapped fglasso and red lines denote edges identified by the bootstrapped fglasso but missed by the fglasso.*

more connected edges. Third, electrodes from other regions of the scalp tend to be only sparsely connected. Finally, the fraction of black to red and blue edges provides a proxy for the level of confidence in any given estimated network. For the alcoholic group this is fairly high, suggesting an accurately estimated network. However, the ratio is somewhat lower for the control group, suggesting a less accurate estimate. This is not surprising given the challenging data set with $p = 64$ nodes, corresponding to estimating graphs for $64 \times 6 = 384$ variables based on only 45 observations.

To identify edges that were clearly different between the two groups we selected edges that occurred at least 50% more often in the bootstrap replications for one group relative to the other group. Figure 4 plots the edges only identified by either the alcoholic group or the control group. We observe that some edges in the left central and parietal regions were identified by the alcoholic group but missed by the control group, while one edge in the frontal region was identified by the control group but missed by the alcoholic group. Both

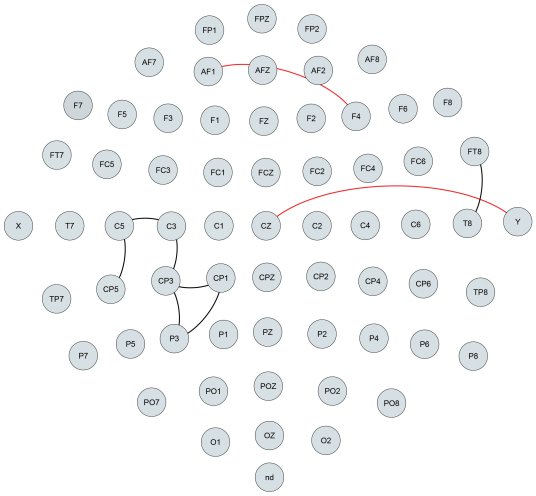

Figure 4: *Black lines denote edges identified only by the alcoholic group and red lines denote edges identified only by the control group.*

findings provide confirmation for our informal observations from Figure 3.

6 Discussion

We conclude the paper by discussing three extensions. Here we have assumed that the trajectories of the functional variables are fully observed, although our results could be extended to the setting of densely observed curves under extra regularity and smoothness conditions. Hence, the first possible extension involves constructing a graphical model for sparse, irregular and noisy functional data, a common situation in functional data analysis (FDA). This extension could be achieved by performing FPCA on sparsely sampled functional data using either a mixed effects model (James et al., 2000) or a local smoother method (Yao et al., 2005), and then implementing the fglasso on the conditional expectations of the principal component scores.

Second, one referee was concerned that, since the inverse of the covariance operator of $(g_1, \ldots, g_p)$ is unbounded (Bosq, 2000), then the minimum eigenvalue of the covariance matrix $\mathbf{\Sigma}^{*M}$ goes to zero as $M \to \infty$. He/she suggested that the true edge set E could

instead be recovered based on the bounded inverse correlation operator, i.e. using the block sparsity pattern in $\mathbf{Q}^{*M}$, the inverse correlation matrix of $\mathbf{a}_i^M$. This could be implemented using an alternative criterion by replacing the sample covariance matrix, $\mathbf{S}^M$, in (9) with the sample correlation matrix, $\mathbf{R}^M$. Specifically, we propose to solve the following optimization problem

$$\widehat{\mathbf{Q}}^M = \underset{\mathbf{Q}^M}{\operatorname{argmax}} \left\{ \log\det(\mathbf{Q}^M) - \operatorname{trace}(\mathbf{R}^M\mathbf{Q}^M) - \gamma_n \sum_{j\neq l} \|\mathbf{Q}_{jl}^M\|_F \right\}, \tag{24}$$

where the optimization is restricted to be over symmetric positive definite matrices in $\mathbb{R}^{Mp\times Mp}$ such that the diagonal elements of $(\mathbf{Q}^M)^{-1}$ are one and γ_n is a non-negative tuning parameter. We could then use the identified block sparsity structure in $\widehat{\mathbf{Q}}^M$ to estimate E. We discuss the connection between our fglasso approach and (24) in Section D.2 of the Supplementary Material and develop an algorithm to solve (24) in Section D.3. Theoretically, however, derivations on the entrywise concentration inequalities for $\mathbf{R}^M$ are needed, posing additional challenges. It is worth noting that in FDA one usually selects only the first few principal components, so this issue does not pose any practical concern for the fglasso approach.

Third, the main theoretical limitation in this paper is to treat the dimension of the functional variables as approaching infinity rather than truly infinite dimensional functional objects ($M_n \to \infty$ rather than $M_n = \infty$). It is challenging, under the current framework, to relax the assumption $M_n < \infty$ to the fully functional situation with $M_n = \infty$, since we would need to write Conditions 3-4 and the relevant proofs in terms of abstract functional analysis language in Hilbert space rather than the current compact matrix forms. We next present another way to understand the conditional dependence structure when $M_n < \infty$. From Peng et al. (2009), a_{ijk} can be expressed as

$$a_{ijk} = \sum_{l\neq j}^{p} \sum_{m=1}^{M_n} \beta_{jlkm} a_{ilm} + \epsilon_{ijk}, i = 1, \ldots, j \in V, k = 1, \ldots, M_n, \tag{25}$$

such that $\{\epsilon_{ijk}, k = 1, \ldots, M_n\}$ is uncorrelated with $\{a_{ilm}, l \in V, l \neq j, m = 1, \ldots, M_n\}$ if

and only if

$$\boldsymbol{\beta}_{jl} = -(\boldsymbol{\Theta}_{jj}^{*M_n})^{-1}\boldsymbol{\Theta}_{jl}^{*M_n}, (j,l) \in V^2, l \neq j, \tag{26}$$

with its (k,m)-th entry given by β_{jlkm}. In other words, both $\{\boldsymbol{\beta}_{jl}, (j,l) \in V^2, j \neq l\}$ and $\{\boldsymbol{\Theta}_{jl}^{*M_n}, (j,l) \in V^2, j \neq l\}$ can be used to identify the true edge set. When $M_n = \infty$, although (26) cannot be written in the compact matrix form, the analogy to (25) still holds and $\{\beta_{jlkm}, k, m = 1, \ldots, \infty\}$ reflects the network structure between nodes j and l. The expression (25) with $M_n = \infty$ provides an alternative approach for estimating the FGM, but would require new algorithms and theoretical guarantees.

Another potential approach to tackle the finite dimensional limitation is to find a large enough value of $M_n' < \infty$ such that

$$\max_{(j,l)\in V^2, j\neq l} \|C_{jl}^{M_n'}(s,t) - C_{jl}(s,t)\|_* \leq O(n^{-\omega}), \tag{27}$$

where $\|\cdot\|_*$ denotes some functional norm and ω is some positive value. Intuitively, if $\max_{(j,l)\in V^2, j\neq l} \|C_{jl}^{M_n'}(s,t) - C_{jl}(s,t)\|_*$ is small enough, $C_{jl}^{M_n'}(s,t)$ provides a good approximation to $C_{jl}(s,t)$, hence $C_{jl}^{M_n'}(s,t)$ can still be used to identify the graph structure. This formulation then reduces to the model considered in our paper, which assumes large but finite dimensional functional data and our theoretical results become applicable in the more general setting. However, it appears challenging to prove (27) with suitable choices of M_n' and ω.

These are all fruitful topics for future research but are beyond the scope of this paper.

Acknowledgements.

We are grateful to the Editor, the Associate Editor and three referees for their useful comments and suggestions.

A Appendix

Appendix A.1 contains a counterexample where the grid method described in Section 1 fails. Further remarks on some regularity conditions are provided in Appendix A.2.

A.1 Counterexample

We create a counterexample, in which the grid method is not able to identify the true conditional dependence structure while our approach can. Take $M = 1$, $p = 3$, $\mathcal{T} = [0, 1]$ and let $g_j(t) = a_j\phi_j(t), j = 1, 2, 3$, where the a_j's are standard normal, a_1, a_2 are correlated conditional on a_3, $\phi_1(t) = f_1(t)I_{\{0\le t\le 1/2\}}$ and $\phi_2(t) = f_2(t)I_{\{1/2<t\le 1\}}$ with $\int_0^{1/2} f_1(t)^2dt = \int_{1/2}^1 f_2(t)^2dt = 1$. Then $\mathrm{Cov}\big(g_1(s), g_2(t)|g_3(\cdot)\big) = \mathrm{Cov}\big(a_1, a_2|g_3(\cdot)\big)\phi_1(s)\phi_2(t)$, which equals zero for all $s = t$, $(s, t) \in \mathcal{T}^2$, but is nonzero for some $s \neq t$.

A.2 Further Remarks on Some Regualrity Conditions

Remark on Conditions 3–4. We provide an example satisfying Conditions 3 and 4. For convenience, denote $\mathbf{a}_{ij} = (\mathbf{x}_{ij}^T, \mathbf{y}_{ij}^T)^T, i = 1, \ldots, n, j = 1, \ldots, p$, where $\mathbf{x}_{ij} = (a_{ij1}, \ldots, a_{ijM})^T$ and $\mathbf{y}_{ij} = (a_{ij(M+1)}, \ldots, a_{ijM_n})^T$. Define $\widetilde{\mathbf{\Sigma}} \in \mathbb{R}^{pM_n\times pM_n}$ to be the covariance matrix of $(\mathbf{x}_{11}^T, \ldots, \mathbf{x}_{1p}^T, \mathbf{y}_{11}^T, \ldots, \mathbf{y}_{1p}^T)^T$. Then we can find a permutation matrix $\mathbf{P}_\pi$ satisfying $\mathbf{P}_\pi^{-1} = \mathbf{P}_\pi^T$ such that $\mathbf{P}_\pi\mathbf{\Sigma}\mathbf{P}_\pi^T = \widetilde{\mathbf{\Sigma}}$ and $\widetilde{\mathbf{\Omega}} = \mathbf{P}_\pi\mathbf{\Omega}\mathbf{P}_\pi^T$, which indicates that $\widetilde{\mathbf{\Omega}}$ is a permutation of $\mathbf{\Omega}$. Let $\widetilde{\mathbf{\Sigma}} = \begin{pmatrix} \widetilde{\mathbf{\Sigma}}_{11} & \widetilde{\mathbf{\Sigma}}_{12} \\ \widetilde{\mathbf{\Sigma}}_{21} & \widetilde{\mathbf{\Sigma}}_{22} \end{pmatrix}$ and $\widetilde{\mathbf{\Omega}} = \begin{pmatrix} \widetilde{\mathbf{\Omega}}_{11} & \widetilde{\mathbf{\Omega}}_{12} \\ \widetilde{\mathbf{\Omega}}_{21} & \widetilde{\mathbf{\Omega}}_{22} \end{pmatrix}$, where $\widetilde{\mathbf{\Sigma}}_{11}$, $\widetilde{\mathbf{\Omega}}_{11}$ are $pM \times pM$ submatrices and $\widetilde{\mathbf{\Sigma}}_{22}$, $\widetilde{\mathbf{\Omega}}_{22}$ are $p(M_n - M) \times p(M_n - M)$ submatrices. If we consider $(\mathbf{x}_{11}^T, \ldots, \mathbf{x}_{1p}^T)^T$ to be independent of $(\mathbf{y}_{11}^T, \ldots, \mathbf{y}_{1p}^T)^T$, then $\widetilde{\mathbf{\Omega}}_{12}^{(k)} = 0$ for $M \le k < M_n$, i.e. $\mathbf{\Omega}_{jl,2}^{(k)} = 0$, for $M \le k < M_n$, which satisfies Condition 3. On the other hand, $\min_{(j,l)\in E}\|\mathbf{\Omega}_{jl}\|_F$ corresponds to the minimal signal strength. In our example, $\|\mathbf{\Omega}_{jl}\|_F^2 = \|\mathbf{\Omega}_{jl,1}^{(M)}\|_F^2 + \|\mathbf{\Omega}_{jl,4}^{(M)}\|_F^2$, so Condition 4 presents a sufficient condition of minimal signal strength.

Remark on Condition 5. It is worth noting that $\mathbf{\Gamma}^*$ is the Hessian of $-\log\det(\mathbf{\Theta})$ evaluated at $\mathbf{\Theta} = \mathbf{\Theta}^*$. Hence the entry $\mathbf{\Gamma}^{*(k,k')(m,m')}_{(j,j')(l,l')}$ equals $\frac{\partial(-\log\det(\mathbf{\Theta}))}{\partial\Theta_{jj'kk'}\partial\Theta_{ll'mm'}}$ evaluated at $\mathbf{\Theta} = \mathbf{\Theta}^*$, where $\Theta_{jj'kk'}$ is the (k,k')th entry of the $M \times M$ submatrix $\mathbf{\Theta}_{jj'}$, $1 \le j, j', l, l' \le p$, $1 \le k, k', m, m' \le M$. Since $\mathbf{a}$ is multivariate Gaussian, some standard calculations show that $\mathbf{\Gamma}^{*(k,k')(m,m')}_{(j,j')(l,l')} = \mathrm{Cov}\left(a_{jk}a_{j'k'}, a_{lm}a_{l'm'}\right)$. The Hessian of the negative log-determinant for the scalar data was studied in Ravikumar et al. (2011). We extend their work by viewing $\mathbf{\Gamma}^*$, the Fisher information of the model, as an edge-based M^2-block covariance matrix instead of the node-based covariance matrix $\mathbf{\Sigma}^*$. For each $(j,j') \in V^2$, denote by $\mathbf{b}_{jj'} = \mathbf{a}_j \otimes \mathbf{a}_{j'} \in \mathbb{R}^{M^2}$ the edge-based vector, where $\mathbf{a}_j$, $\mathbf{a}_{j'}$ are the node-based vectors. Then we have $\mathbf{\Gamma}^*_{(j,j')(l,l')} = E(\mathbf{b}_{jj'}\mathbf{b}^T_{ll'})$, which indicates that Condition 5 is the population version of the irrepresentable-type condition. Define the edge-based vector within S by $\mathbf{b}_S = \{\mathbf{b}_{jj'}, (j,j') \in S\}$. Then (21) is equivalent to $||E(\mathbf{b}_{S^c}\mathbf{b}_S^T)E(\mathbf{b}_S\mathbf{b}_S^T)^{-1}||^{(M^2)}_\infty \le 1-\eta$, which bounds the effects of non-edges in S^c on the edges in S, and restricts $\mathbf{b}_{jj'}$'s outside the true edge set S to be weakly correlated with those within S.

References

Bosq, D. (2000). *Linear Processes in Function Spaces*, Springer, New York.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers, *Foundations and Trends in Machine Learning* **3(1)**: 1–122.

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

Candes, E. and Tao, T. (2007). The dantzig selection: Statistical estimation when p is much larger than n, *The Annals of Statistics* **35**: 2313–2351.

Danaher, P., Wang, P. and Witten, D. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes, *Journal of the Royal Statistical Society: Series B* **76**: 373–397.

Friedman, J., Hastie, T. and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso, *Biostatistics* **2**: 432–441.

Hayden, E., Wiegand, R., Meyer, E., Bauer, L., O's Connor, S., J.I., N., D.B., C., B., P. and H., B. (2006). Patterns of regional brain activity in alcohol-dependent subjects, *Alcoholism: Clinical and Experimental Research* **30**: 1986–1991.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, John Wiley & Sons, Ltd.

Ingber, L. (1997). Statistical mechanics of neocortical interactions: Canonical momenta indicators of electroencephalography, *Physical Review E* **55**: 4578–4593.

James, G., Hastie, T. and Sugar, C. (2000). Principal component models for sparse functional data, *Biometrika* **87**: 587–602.

Knyazev, G. (2007). Motivation, emotion, and their inhibitory control mirrored in brain osclillations, *Neuroscience Biobehavioral Reviews* **131**: 377–395.

Kolar, M. and Xing, E. (2011). On time varying undirected graphs, *Proceedings of Machine Learning Research* **15**: 407–415.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation, *The Annals of Statistics* **37**: 4254–4278.

Li, B., Kim, M. and Altman, N. (2010). On dimension folding of matrix-or array-valued statistical objects, *The Annals of Statistics* **38**: 1094–1121.

Mazumder, R. and Hastie, T. (2012a). The graphical lasso: New insights and alternatives, *Electronic Journal of Statistics* **6**: 2125–2149.

Mazumder, R. and Hastie, T. (2012b). Exact covariance thresholding into connected components for large-scale graphical lasso, *Journal of Machine Learning Research* **13**: 781–794.

Meinshausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with lasso, *The Annals of Statistics* **34**: 1436–1462.

Meinshausen, N. and Buhlmann, P. (2010). Stability selection, *Journal of the Royal Statistical Society, Series B* **72**: 417–473.

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

Qiu, H., Han, F., Liu, H. and Caffo, B. (2016). Joint estimation of multiple graphical models from high dimensional dependent data, *Journal of the Royal Statistical Society: Series B* **78**: 487–504.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2 edn, Springer.

Ravikumar, P., Wainwright, M., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing l_1-penalized log-determinant deivergence, *Electronic Journal of Statistics* **5**: 935–980.

Rice, J. and Silverman, B. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves, *Journal of the Royal Statistical Society, Series B* **53**: 233–243.

Storey, J. D., Xiao, W., Leek, T. J., Tompkins, R. G. and Davis, R. W. (2005). Significance analysis of time course microarray experiments, *Proceedings of the National Academy of Sciences* **102**: 12837–12842.

Sun, T. and Zhang, C. (2013). Sparse matrix inversion with scaled lasso, *Journal of Machine Learning Research* **14**: 3385–3418.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B* **58**: 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso, *Journal of the Royal Statistical Society, Series B* **67**: 91–108.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization, *Journal of Optimization Theory and Applications* **109**: 475–494.

Witten, D., Friedman, J. and Simon, N. (2011). New insights and faster computations for the graphical lasso, *Journal of Computational and Graphical Statistics* **20**: 892–900.

Yao, F., Muller, H. and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data, *Journal of the American Statistical Association* **100**: 577–590.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B* **68**: 49–67.

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

Zhang, X., Begleiter, B., Porjesz, B., Wang, W. and Litke, A. (1995). Event related potentials during object recognition tasks, *Brain Research Bulletin* **38**: 531–538.

Zhou, H. and Li, L. (2014). Regularized matrix regression, *Journal of the Royal Statistical Society: Series B* **76**: 463–483.

Zhou, S., Lafferty, J. and Wasserman, L. (2010). Time varying undirected graphs, *Machine Learning Journal* **80**: 295–319.

Zhu, H., Strawn, N. and Dunson, D. (2016). Bayesian graphical models for multivariate functional data, *Journal of Machine Learning Research* **17**: 1–27.

Zhu, Y., Shen, X. and Pan, W. (2014). Structural pursuit over multiple undirected graphs, *Journal of the American Statistical Association* **109**: 1683–1696.

Supplementary Material to "Functional Graphical Models"

Xinghao Qiao, Shaojun Guo, and Gareth M. James

This supplementary material contains the details of the algorithms with derivations in Appendix B, technical proofs of Propositions 1–2, Theorems 1–4, Lemmas 1-15 in Appendix C, and further discussion in Appendix D.

B Derivations for the Fglasso Algorithm

In Appendix B, we provide some further details about the fglasso algorithm and the joint fglasso algorithm.

B.1 Step 2(b) of Algorithm 1

Note (14) is equivalent to finding $\mathbf{w}_{j1}, \cdots, \mathbf{w}_{j(p-1)}$ to minimize

$$\text{trace}\left(\sum_{l=1}^{p-1}\sum_{k=1}^{p-1}\mathbf{S}_{jj}\mathbf{w}_{jl}^T(\boldsymbol{\Theta}_{-j}^{-1})_{lk}\mathbf{w}_{jk} + 2\sum_{k=1}^{p-1}\mathbf{s}_{jk}^T\mathbf{w}_{jk}\right) + 2\gamma_n\sum_{k=1}^{p-1}\|\mathbf{w}_{jk}\|_F. \tag{B.1}$$

Setting the derivative of (B.1) with respect to $\mathbf{w}_{jk}$ to be zero and applying Lemma 4 yields

$$\begin{aligned}\frac{\partial(B.1)}{\partial\mathbf{w}_{jk}} &= (\boldsymbol{\Theta}_{-j}^{-1})_{kk}\mathbf{w}_{jk}\mathbf{S}_{jj} + (\boldsymbol{\Theta}_{-j}^{-1})_{kk}^T\mathbf{w}_{jk}\mathbf{S}_{jj}^T \\ &\quad + \sum_{l\neq k}\left((\boldsymbol{\Theta}_{-j}^{-1})_{lk}^T\mathbf{w}_{jl}\mathbf{S}_{jj}^T + (\boldsymbol{\Theta}_{-j}^{-1})_{kl}\mathbf{w}_{jl}\mathbf{S}_{jj}\right) + 2\mathbf{s}_{jk} + 2\gamma_n\boldsymbol{\nu}_{jk} \\ &= 2\left((\boldsymbol{\Theta}_{-j}^{-1})_{kk}\mathbf{w}_{jk}\mathbf{S}_{jj} + \sum_{l\neq k}(\boldsymbol{\Theta}_{-j}^{-1})_{lk}^T\mathbf{w}_{jl}\mathbf{S}_{jj} + \mathbf{s}_{jk} + \gamma_n\boldsymbol{\nu}_{jk}\right) = \mathbf{0},\end{aligned}$$

where $\boldsymbol{\nu}_{jk} = \frac{\mathbf{w}_{jk}}{\|\mathbf{w}_{jk}\|_F}$ if $\mathbf{w}_{jk} \neq \mathbf{0}$, and $\boldsymbol{\nu}_{jk} \in \mathbb{R}^{M\times M}$ with $\|\boldsymbol{\nu}_{jk}\|_F \leq 1$ otherwise, $k = 1, \ldots, p-1$. We define the block "residual" by

$$\mathbf{r}_{jk} = \sum_{l\neq k}(\boldsymbol{\Theta}_{-j}^{-1})_{lk}^T\mathbf{w}_{jl}\mathbf{S}_{jj} + \mathbf{s}_{jk}. \tag{B.2}$$

If $\mathbf{w}_{jk} = \mathbf{0}$, then $\|\mathbf{r}_{jk}\|_F = \gamma_n \|\boldsymbol{\nu}_{jk}\|_F \leq \gamma_n$. Otherwise we need to solve for $\mathbf{w}_{jk}$ in the following equation

$$(\boldsymbol{\Theta}_{-j}^{-1})_{kk}\mathbf{w}_{jk}\mathbf{S}_{jj} + \mathbf{r}_{jk} + \gamma_n \frac{\mathbf{w}_{jk}}{\|\mathbf{w}_{jk}\|_F} = \mathbf{0}. \tag{B.3}$$

We replace (B.3) by (B.4), and standard packages in R/MatLab can be used to solve the following M^2 by M^2 nonlinear equation

$$((\boldsymbol{\Theta}_{-j}^{-1})_{kk} \otimes \mathbf{S}_{jj})\text{vec}(\mathbf{w}_{jk}) + \text{vec}(\mathbf{r}_{jk}) + \gamma_n \frac{\text{vec}(\mathbf{w}_{jk})}{\|\mathbf{w}_{jk}\|_F} = 0. \tag{B.4}$$

Hence, the block coordinate descent algorithm for solving $\mathbf{w}_j$ in (14) is summarized in Algorithm 3.

Algorithm 3 Block Coordinate Descent Algorithm for Solving $\mathbf{w}_j$

1. Initialize $\widehat{\mathbf{w}}_j$.
2. Repeat until convergence for $k = 1, \ldots, p-1$.
   (a) Compute $\widehat{\mathbf{r}}_{jk}$ via (B.2).
   (b) Set $\widehat{\mathbf{w}}_{jk} = \mathbf{0}$ if $\|\mathbf{r}_{jk}\|_F \leq \gamma_n$; otherwise solve for $\widehat{\mathbf{w}}_{jk}$ via (B.4).

B.2 Steps 2(a) and 2(c) of Algorithm 1

At the jth step, we need to compute $\boldsymbol{\Theta}_{-j}^{-1}$ in (14) given current $\boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1}$. Then step 2(a) follows by the blockwise inversion formula. Next we solve for $\mathbf{w}_j$ via Algorithm 3, and then update $\boldsymbol{\Theta}^{-1}$ given current $\mathbf{w}_j$, $\boldsymbol{\Theta}_{jj}$, and $\boldsymbol{\Theta}_{-j}^{-1}$, by applying the blockwise inversion formula again. Rearranging the row and column blocks such that the (j,j)-th block is the last one, we obtain the permuted version of $\boldsymbol{\Theta}^{-1}$ by $\begin{pmatrix} \boldsymbol{\Theta}_{-j}^{-1} + \mathbf{U}_j\mathbf{V}_j\mathbf{U}_j^T & -\mathbf{U}_j\mathbf{V}_j \\ -\mathbf{V}_j\mathbf{U}_j^T & \mathbf{V}_j \end{pmatrix}$, where $\mathbf{U}_j = \boldsymbol{\Theta}_{-j}^{-1}\mathbf{w}_j$ and $\mathbf{V}_j = (\boldsymbol{\Theta}_{jj} - \mathbf{w}_j^T\mathbf{U}_j)^{-1} = \mathbf{S}_{jj}$. Step 2(c) follows as a consequence.

B.3 Joint Fglasso Algorithm

We put superscript (q) on the terms used in Section 3.1 to denote the corresponding ones for the q-th class, $1 \leq q \leq Q$. Then, for a fixed value of $\boldsymbol{\Theta}_{-j}^{(q)}$, some calculations show that

(11) with the addition of the penalty (12) is minimized by setting

$$\widehat{\boldsymbol{\Theta}}_{jj}^{(q)} = (\mathbf{S}_{jj}^{(q)})^{-1} + (\widehat{\mathbf{w}}_j^{(q)})^T (\boldsymbol{\Theta}_{-j}^{(q)})^{-1} \widehat{\mathbf{w}}_j^{(q)}, \tag{B.5}$$

where $\widehat{\mathbf{w}}_j^{(1)}, \ldots, \widehat{\mathbf{w}}_j^{(Q)}$ are obtained by minimizing

$$\begin{aligned} &\sum_{q=1}^{Q} \text{trace}\left(\mathbf{S}_{jj}^{(q)} (\mathbf{w}_j^{(q)})^T (\boldsymbol{\Theta}_{-j}^{(q)})^{-1} \mathbf{w}_j^{(q)} + 2(\mathbf{s}_j^{(q)})^T \mathbf{w}_j^{(q)}\right) \\ &+2\gamma_{1n} \sum_{l=1}^{p-1} \sum_{q=1}^{Q} \|\mathbf{w}_{jl}^{(q)}\|_F + 2\gamma_{2n} \sum_{l=1}^{p-1} \sqrt{\sum_{q=1}^{Q} ||\mathbf{w}_{jl}^{(q)}||_F^2}, \end{aligned} \tag{B.6}$$

and $\mathbf{w}_{jl}^{(q)}$ represents the lth $M \times M$ block of $\mathbf{w}_j^{(q)}$. Analogously to the fglasso algorithm, we summarize the joint fglasso algorithm, which is developed to solve the optimization problem (11) in Algorithm 4.

Algorithm 4 Joint Functional Graphical Lasso Algorithm

1. Initialize $\widehat{\boldsymbol{\Theta}}^{(q)} = \mathbf{I}$ and $\widehat{\boldsymbol{\Sigma}}^{(q)} = \mathbf{I}$, $q = 1, \ldots, Q$.

2. Repeat until convergence for $j = 1, \ldots, p$, $q = 1, \ldots, Q$.

   (a) Compute $(\widehat{\boldsymbol{\Theta}}_{-j}^{(q)})^{-1} \leftarrow \widehat{\boldsymbol{\Sigma}}_{-j}^{(q)} - \widehat{\boldsymbol{\sigma}}_j^{(q)} (\widehat{\boldsymbol{\Sigma}}_{jj}^{(q)})^{-1} (\widehat{\boldsymbol{\sigma}}_j^{(q)})^T$.

   (b) Solve for $\widehat{\mathbf{w}}_j^{(q)}$ in (B.6) using Algorithm 5.

   (c) Reconstruct $\widehat{\boldsymbol{\Sigma}}^{(q)}$ using $\widehat{\boldsymbol{\Sigma}}_{jj}^{(q)} = \mathbf{S}_{jj}^{(q)}$, $\widehat{\boldsymbol{\sigma}}_j^{(q)} = -\mathbf{U}_j^{(q)} \mathbf{S}_{jj}^{(q)}$ and $\widehat{\boldsymbol{\Sigma}}_{-j}^{(q)} = (\widehat{\boldsymbol{\Theta}}_{-j}^{(q)})^{-1} + \mathbf{U}_j^{(q)} \mathbf{S}_{jj}^{(q)} (\mathbf{U}_j^{(q)})^T$, where $\mathbf{U}_j^{(q)} = (\widehat{\boldsymbol{\Theta}}_{-j}^{(q)})^{-1} \widehat{\mathbf{w}}_j^{(q)}$.

3. Set $\widehat{E}^{(q)} = \left\{(j,l) : \|\widehat{\boldsymbol{\Theta}}_{jl}^{(q)}\|_F \neq 0, (j,l) \in V^2, j \neq l\right\}$, $q = 1, \ldots, Q$.

Setting the derivative of (B.6) with respect to $\mathbf{w}_{jk}^{(q)}$ to be zero and applying Lemma 4 yield

$$\begin{aligned}\frac{\partial(B.6)}{\partial \mathbf{w}_{jk}^{(q)}} &= ((\mathbf{\Theta}_{-j}^{(q)})^{-1})_{kk}\mathbf{w}_{jk}^{(q)}\mathbf{S}_{jj}^{(q)} + ((\mathbf{\Theta}_{-j}^{-1})^{(q)})_{kk}^{T}\mathbf{w}_{jk}^{(q)}(\mathbf{S}_{jj}^{(q)})^{T} \\ &\quad + \sum_{l\neq k}\left(((\mathbf{\Theta}_{-j}^{(q)})^{-1})_{lk}^{T}\mathbf{w}_{jl}^{(q)}(\mathbf{S}_{jj}^{(q)})^{T} + ((\mathbf{\Theta}_{-j}^{(q)})^{-1})_{kl}\mathbf{w}_{jl}^{(q)}\mathbf{S}_{jj}^{(q)}\right) \\ &\quad + 2\mathbf{s}_{jk}^{(q)} + 2\lambda\boldsymbol{\nu}_{jk}^{(q)} \\ &= 2\left(((\mathbf{\Theta}_{-j}^{(q)})^{-1})_{kk}\mathbf{w}_{jk}^{(q)}\mathbf{S}_{jj}^{(q)}\right. \\ &\quad \left. + \sum_{l\neq k}((\mathbf{\Theta}_{-j}^{(q)})^{-1})_{lk}^{T}\mathbf{w}_{jl}^{(q)}\mathbf{S}_{jj}^{(q)} + \mathbf{s}_{jk}^{(q)} + \gamma_{1n}\boldsymbol{\nu}_{jk}^{(q)} + \gamma_{2n}\boldsymbol{\mu}_{jk}^{(q)}\right) = \mathbf{0},\end{aligned}$$

where

$$\begin{cases} ||\boldsymbol{\nu}_{jk}^{(q)}||_F \leq 1, \sum_{q=1}^{Q}||\boldsymbol{\mu}_{jk}^{(q)}||_F^2 \leq 1, & \text{if } \sum_{q=1}^{Q}||\mathbf{w}_{jk}^{(q)}||_F^2 = 0. \\ ||\boldsymbol{\nu}_{jk}^{(q)}||_F \leq 1, \boldsymbol{\mu}_{jk}^{(q)} = \frac{\mathbf{w}_{jk}^{(q)}}{\sqrt{\sum_{q=1}^{Q}\|\mathbf{w}_{jk}^{(q)}\|_F^2}}, & \text{if } \sum_{q=1}^{Q}||\mathbf{w}_{jk}^{(q)}||_F^2 \neq 0 \text{ and } \mathbf{w}_{jk}^{(q)} = \mathbf{0}. \\ \boldsymbol{\nu}_{jk}^{(q)} = \frac{\mathbf{w}_{jk}^{(q)}}{\|\mathbf{w}_{jk}^{(q)}\|_F}, \boldsymbol{\mu}_{jk}^{(q)} = \frac{\mathbf{w}_{jk}^{(q)}}{\sqrt{\sum_{q=1}^{Q}\|\mathbf{w}_{jk}^{(q)}\|_F^2}}, & \text{if } \sum_{q=1}^{Q}||\mathbf{w}_{jk}^{(q)}||_F^2 \neq 0 \text{ and } \mathbf{w}_{jk}^{(q)} \neq \mathbf{0}. \end{cases}$$

We define the qth block "residual" by

$$\mathbf{r}_{jk}^{(q)} = \sum_{l\neq k}((\mathbf{\Theta}_{-j}^{(q)})^{-1})_{lk}^{T}\mathbf{w}_{jl}^{(q)}\mathbf{S}_{jj}^{(q)} + \mathbf{s}_{jk}^{(q)}. \tag{B.7}$$

If $\mathbf{w}_{jk}^{(q)} = \mathbf{0}$ for all Q classes, then $\sum_{q=1}^{Q}\|\mathbf{r}_{jk}^{(q)}\|_F \leq \sum_{q=1}^{Q}(\gamma_{1n}||\boldsymbol{\nu}_{jk}^{(q)}\|_F + \gamma_{2n}\|\boldsymbol{\mu}_{jk}^{(q)}\|_F) \leq \gamma_{1n}Q + \gamma_{2n}$. Otherwise if $\mathbf{w}_{jk}^{(q)} = \mathbf{0}$, then $||\mathbf{r}_{jk}^{(q)}||_F \leq \gamma_{1n}$; if $\mathbf{w}_{jk}^{(q)} \neq \mathbf{0}$ we need to solve for $\mathbf{w}_{jk}^{(q)}$ in the following equation

$$((\mathbf{\Theta}_{-j}^{(q)})^{-1})_{kk}\mathbf{w}_{jk}^{(q)}\mathbf{S}_{jj}^{(q)} + \mathbf{r}_{jk}^{(q)} + \gamma_{1n}\frac{\mathbf{w}_{jk}^{(q)}}{\|\mathbf{w}_{jk}^{(q)}\|_F} + \gamma_{2n}\frac{\mathbf{w}_{jk}^{(q)}}{\sqrt{\sum_{q=1}^{Q}\|\mathbf{w}_{jk}^{(q)}\|_F^2}} = \mathbf{0}. \tag{B.8}$$

Hence, the block coordinate descent algorithm for solving $\mathbf{w}_{j}^{(q)}$ in (B.6) is summarized in Algorithm 5.

Algorithm 5 Block Coordinate Descent Algorithm for Solving $\mathbf{w}_j^{(q)}$

1. Initialize $\widehat{\mathbf{w}}_j^{(1)}, \ldots, \widehat{\mathbf{w}}_j^{(Q)}$.

2. Repeat until convergence for $k = 1, \ldots, p-1$, $q = 1, \ldots, Q$.

   (a) Compute $\widehat{\mathbf{r}}_{jk}^{(q)}$ via (B.7).

   (b) Set $\widehat{\mathbf{w}}_{jk}^{(q)} = \mathbf{0}$ for all Q classes if $\sum_{q=1}^{Q} \|\mathbf{r}_{jk}^{(q)}\|_F \leq \gamma_{1n} Q + \gamma_{2n}$; otherwise go to (c)

   (c) For $q = 1, \ldots, Q$, set $\widehat{\mathbf{w}}_{jk}^{(q)} = \mathbf{0}$ if $\|\mathbf{r}_{jk}^{(q)}\|_F \leq \gamma_{1n}$; otherwise solve for $\widehat{\mathbf{w}}_{jk}^{(q)}$ via (B.8).

C Proofs of Technical Details

C.1 Proof of Proposition 1

Substituting $\boldsymbol{\Theta} = \text{diag}(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K)$ into (9) yields

$$\max_{\boldsymbol{\Theta}_1, \cdots, \boldsymbol{\Theta}_K} \left\{ \sum_{k=1}^{K} \log\det \boldsymbol{\Theta}_k - \sum_{k=1}^{K} \text{trace}(\mathbf{S}_k \boldsymbol{\Theta}_k) - \gamma_n \sum_{j \neq l} \sum_{k=1}^{K} \|\boldsymbol{\Theta}_{k,jl}\|_F \right\}, \tag{C.1}$$

which is equivalent to K separate fglasso problems in (15).

C.2 Proof of Proposition 2

If $\boldsymbol{\Theta}$ is block diagonal, and i and i' belong to separate index sets G_k and $G_{k'}$, then $\boldsymbol{\Theta}_{ii'} = \mathbf{0}$ and hence $(\boldsymbol{\Theta}^{-1})_{ii'} = \mathbf{0}$. By (C.12), we have $\|\mathbf{S}_{ii'}\|_F \leq \gamma_n \|\mathbf{Z}_{ii'}\|_F \leq \gamma_n$. This completes the proof for the sufficient condition.

Next we prove the condition is necessary. We construct $\boldsymbol{\Theta}_k$ by solving the fglasso problem (9) applied to the symmetric submatrix of $\mathbf{S}$ given by index set G_k for $k = 1, \ldots, K$, and let $\bar{\boldsymbol{\Theta}} = \text{diag}\left(\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_K\right)$. Since $\|\mathbf{S}_{ii'}\|_F \leq \gamma_n$ for all $i \in G_k$, $i' \in G_{k'}$, $k \neq k'$, and $\bar{\boldsymbol{\Theta}}_{ii'} = \mathbf{0}$ by construction, we have $(\bar{\boldsymbol{\Theta}}^{-1})_{ii'} = \mathbf{0}$ and hence the (i, i')-th equation of (C.12) is satisfied. Moreover, the (k, k)-th equation of (C.12) is satisfied by construction. Therefore, $\bar{\boldsymbol{\Theta}}$ satisfies the KKT condition (C.12) and is the solution to the fglasso problem (9).

C.3 Proof of Theorem 1

We begin with some notation. For any $H(s,t), (s,t) \in \mathcal{T}^2$ with the corresponding Karhunen-Loève decomposition $H(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t)$, define

$$\|H\|_{\mathcal{S}} = \big(\sum_{j\geq 1} \lambda_j^2\big)^{1/2}.$$

For two square-integrable functions $f(t), g(t)$, define $\langle f, g\rangle = \int_{t\in\mathcal{T}} f(t)g(t)dt$ and $\|f\|^2 = \langle f, f\rangle$. Denote also $a_{ijk} = \lambda_{jk}^{1/2}\xi_{ijk}$, where $\xi_{ijk} \sim N(0,1)$ and $\lambda_0 = \sup_{j\leq p}\sum_{k=1}^{\infty}\lambda_{jk}$.

We now prove Theorem 1. We first consider $\widehat{\sigma}_{jjkk}$ for $j = 1,\ldots,p$ and $k = 1,\ldots,M$. Note that $n\widehat{\sigma}_{jjkk} = \sum_{i=1}^n \widehat{a}_{ijk}^2 - n\bar{a}_{jk}^2$ and $\sigma^*_{jjkk} = Ea_{1jk}^2$ with $\bar{a}_{jk} = n^{-1}\sum_{i=1}^n \widehat{a}_{ijk}$, and, for each (i,j,k), $\widehat{a}_{ijk} = \langle g_{ij}, \widehat{\phi}_{jk}\rangle = a_{ijk} + \langle g_{ij}, \widehat{\phi}_{jk} - \phi_{jk}\rangle$. Then $n(\widehat{\sigma}_{jjkk} - \sigma_{jjkk})$ is rewritten as

$$\begin{aligned} n(\widehat{\sigma}_{jjkk} - \sigma^*_{jjkk}) &= \lambda_{jk}\sum_{i=1}^n \big(\xi_{ijk}^2 - 1\big) + 2\lambda_{jk}^{1/2}\sum_{i=1}^n \xi_{ijk}\langle g_{ij}, \widehat{\phi}_{jk} - \phi_{jk}\rangle \\ &\quad + \sum_{i=1}^n \langle g_{ij}, \widehat{\phi}_{jk} - \phi_{jk}\rangle^2 - n\bar{a}_{jk}^2 = I_1 + I_2 + I_3 + I_4. \end{aligned}$$

Note that for $\delta > 0$, $P\Big(\big|\widehat{\sigma}_{jjkk} - \sigma^*_{jjkk}\big| \geq 4\delta\Big) \leq \sum_{m=1}^4 P\Big(\big|I_m\big| \geq n\delta\Big)$. To derive the concentration inequality of $n(\widehat{\sigma}_{jjkk} - \sigma^*_{jjkk})$, it suffices to derive the tail behaviors of all I_m's $(m = 1,\ldots,4)$.

(a) Since ξ_{ijk}'s are independent $N(0,1)$, we have that all $0 < \delta \leq 1$,

$$P\Big\{\Big|\sum_{i=1}^n \big(\xi_{ijk}^2 - 1\big)\Big| \geq n\delta\Big\} \leq 2\exp\Big(-\frac{n\delta^2}{32 + 4\delta}\Big) \leq 2\exp\Big(-\frac{n\delta^2}{36}\Big).$$

Hence, it follows that there exists a constants C_1 such that for $0 < \delta \leq C_1$,

$$P\Big(|I_1| \geq n\delta\Big) = P\Big(\Big|\sum_{i=1}^n \Big(\xi_{ijk}^2 - 1\Big)\Big| \geq \frac{n\delta}{\lambda_0}\Big) \leq 2\exp\Big(-C_1 n\delta^2\Big).$$

(b) First, the term I_2 can be bounded by $|I_2| \leq 2\lambda_{jk}^{1/2}\Big\|\sum_{i=1}^n \xi_{ijk}g_{ij}\Big\|\|\widetilde{\phi}_{jk} - \phi_{jk}\|$. Let $Y_{n1} = \big\{\big(\big(\sum_{i=1}^n \xi_{ijk}^2\big)^2\big\}^{1/2}$, $Y_{n2} = \Big\{\sum_{m\neq k}\lambda_{jm}\big(\sum_{i=1}^n \xi_{ijk}\xi_{ijm}\big)^2\Big\}^{1/2}$. Then,

$$\Big\|\sum_{i=1}^n \xi_{ijk}g_{ij}\Big\|^2 = \lambda_{jk}\Big(\sum_{i=1}^n \xi_{ijk}^2\Big)^2 + \sum_{m\neq k}^{\infty}\lambda_{jm}\Big(\sum_{i=1}^n \xi_{ijk}\xi_{ijm}\Big)^2 = \lambda_{jk}Y_{n1}^2 + Y_{n2}^2,$$

which implies that $\|\sum_{i=1}^n \xi_{ijk} g_{ij}\| \le \lambda_{jk}^{1/2} Y_{n1} + Y_{n2}$. By the condition $\lambda_{jk} \asymp k^{-\beta}$ and $d_{jk}\lambda_{jk} = O(k)$, we have that $d_{jk}\lambda_{jk} \le d_0 k$ and $d_{jk} \le d_0 k^{1+\beta}$ for some positive constant d_0. By Lemma 8, $\|\widehat{\phi}_{jk} - \phi_{jk}\| \le d_{jk}\|\widehat{K}_{jj} - K_{jj}\|_{\mathcal{S}}$, where, w.l.o.s., $\widehat{\phi}_{jk}$ can be chosen to satisfy $sgn\langle\widehat{\phi}_{jk}, \phi_{jk}\rangle = 1$. As a result, $|I_2|$ can be further bounded by

$$|I_2| \le 2d_0 k Y_{n1}\|\widehat{K}_{jj} - K_{jj}\|_{\mathcal{S}} + 2d_0\lambda_{jk}^{-1/2} k Y_{n2}\|\widehat{K}_{jj} - K_{jj}\|_{\mathcal{S}}. \tag{C.2}$$

We first bound Y_{n1} and Y_{n2}. On one hand,

$$P\Big(Y_{n1} \ge 2n\Big) = P\Big\{\sum_{i=1}^n (\xi_{ijk}^2 - 1) \ge n\Big\} \le \exp\Big(-\frac{n}{36}\Big). \tag{C.3}$$

On the other hand, since $\xi_{ij1}, \xi_{ij2} \sim N(0,1)$ for each j, k, $\sum_{i=1}^n E|\xi_{ij1}\xi_{ij2}|^k \le nE\xi_{1j1}^{2k} \le k! n 2^k$. As a result, it follows that for all $\delta > 0$

$$P\Big(\Big|\sum_{i=1}^n \xi_{ij1}\xi_{ij2}\Big| \ge \delta\Big) \le 2\exp\Big(-\frac{\delta^2}{16n + 4\delta}\Big) \le 2\exp\Big(-\frac{\delta^2}{32n}\Big) + 2\exp\Big(-\frac{\delta}{8}\Big).$$

Consequently, using integration by parts, there exist two positive constants L_1 and L_2 not depending on n such that

$$E\Big|\sum_{i=1}^n \xi_{ij1}\xi_{ij2}\Big|^{2k} \le k!(nL_1)^k + (2k)!L_2^{2k}, k = 1, 2, 3, \ldots.$$

This further implies $E\big(Y_{n2} - EY_{n2}\big)^{2k} \le k!(2\lambda_0 L_1 n)^k + (2k)!(2\lambda_0^{1/2}L_2)^{2k}, k \ge 1$. Hence we obtain from Theorem 2.3 of Boucheron et al. (2014) that for all $\delta > 0$ and $n \ge 2L_2^2 L_1^{-1}$,

$$P\Big(Y_{n2} - EY_{n2} \ge \delta\Big) \le \exp\Big(-\frac{\delta^2}{32\lambda_0 L_1 n + 8\lambda_0^{1/2} L_2 \delta}\Big).$$

Note that $EY_{n2} \le \lambda_0 n^{1/2}$. Hence, for $\delta \ge 2\lambda_0 n^{1/2}$ and $n \ge 2L_2^2 L_1^{-1}$,

$$P\Big(Y_{n2} \ge \delta\Big) \le P\Big(Y_{n2} - EY_{n2} \ge \delta/2\Big) \le \exp\Big\{-\frac{\delta^2}{16(8\lambda_0 L_1 n + \lambda_0^{1/2} L_2 \delta)}\Big\} \tag{C.4}$$

Now consider $P\Big(\big|I_2\big| \ge n\delta\Big)$. By (C.2), we can bound this term by

$$2P\Big(\|\widehat{K}_{jj} - K_{jj}\|_{\mathcal{S}} \ge \frac{\delta}{8kd_0}\Big) + P\Big(Y_{n1} \ge 2n\Big) + P\Big(Y_{n2} \ge 2n\lambda_{jk}^{1/2}\Big).$$

Together with (C.3), (C.4) and Lemma 6, it follows that there exist two positive constants C_k $(k = 2, 3)$ free of n and p such that for $0 < \delta k^{-1} \le C_2$,

$$P\Big(\big|I_2\big| \ge n\delta\Big) \le C_3 \exp\Big(-C_2 n k^{-2}\delta^2\Big) + \exp\Big(-C_2 n k^{-\beta}\Big).$$

(c) In a similar way to I_2, we can show that there exist three positive constants C_k $(k = 4, 5, 6)$ not depending on n and p such that for $0 < \delta \le C_5$,

$$P\Big(\big|I_3\big| \ge n\delta\Big) \le 2\exp(-C_4 n) + C_6 \exp\Big(-C_5 n k^{-(2+2\beta)}\delta\Big).$$

(d) Consider the last term I_4. First, we have

$$|\bar{a}_{jk}| \le \lambda_{jk}^{1/2}|\bar{\xi}_{jk}| + \|\bar{g}_j\|\|\widehat{\phi}_{jk} - \phi_{jk}\| \le \lambda_{jk}^{1/2}|\bar{\xi}_{jk}| + d_0 k\|\bar{g}_j\|\|\widehat{K}_{jj} - K_{jj}\|_{\mathcal{S}},$$

where $\|\bar{g}_j\|^2 = \sum_{m=1}^{\infty}\lambda_{jm}\bar{\xi}_{jk}^2$. Note that the following inequalities hold for all $\delta > 0$:

$$P\Big(|\bar{\xi}_{jk}| \ge \delta\Big) \le 2\exp\Big(-C_7 n\delta^2\Big) \text{ and } P\Big(\|\bar{g}_j\| \ge \delta\Big) \le 2\exp\Big(-C_7 n\delta^2\Big).$$

for some positive constant C_7. Hence, together with Lemma 6, we obtain that $P\left(|\bar{a}_{jk}|^2 \ge \delta\right)$ can be bounded by

$$\begin{aligned}
&P\Big(|\bar{\xi}_{jk}| \ge \delta^{1/2}\lambda_{jk}^{-1/2}/2\Big) + P\left(\|\bar{g}_j\|\|\widehat{K}_{jj} - K_{jj}\|_{\mathcal{S}} \ge d_{jk}^{-1}\delta^{1/2}/2\right) \\
&\le P\Big(|\bar{\xi}_{jk}| \ge \delta^{1/2}\lambda_{jk}^{-1/2}/2\Big) + P\left\{\|\bar{g}_j\| \ge (d_{jk}^2\delta)^{1/4}/2\right\} \\
&\quad + P\left\{\|\widehat{K}_{jj} - K_{jj}\|_{\mathcal{S}} \ge (\lambda_{jk}^2\delta)^{1/4}/2\right\} \\
&\le 2\exp\left(-C_7 n k^{\beta/2}\delta\right) + C_9\exp\Big(-C_8 n k^{\beta}\delta^{1/2}\Big)
\end{aligned}$$

for all $0 < \delta \le C_8$ with some positive constants C_8 and C_9.

Combining (a), (b), (c) and (d) and choosing suitable constants, the inequality (16) follows consequently.

For general cases of (j, l, k, m) with $j \ne l$ or $m \ne k$, $\widehat{\sigma}_{jlkm} = \frac{1}{n}\sum_{i=1}^{n}\widehat{a}_{ijk}\widehat{a}_{ilm} - \bar{a}_{jk}\bar{a}_{lm}$ and $\sigma^*_{jlkm} = E(a_{ijk}a_{ilm})$. Hence $n\big(\widehat{\sigma}_{jlkm} - \sigma^*_{jlkm}\big)$ can be expressed as the sum of the following five terms:

$$\begin{aligned}
&\sum_{i=1}^{n}\left(a_{ijk}a_{ilm} - \sigma^*_{jklm}\right) + \lambda_{jk}^{1/2}\sum_{i=1}^{n}\xi_{ijk}\langle g_{il}, \widehat{\phi}_{lm} - \phi_{lm}\rangle \\
&+\lambda_{lm}^{1/2}\sum_{i=1}^{n}\xi_{ilm}\langle g_{ij}, \widehat{\phi}_{jk} - \phi_{jk}\rangle + \sum_{i=1}^{n}\langle g_{ij}, \widehat{\phi}_{jk} - \phi_{jk}\rangle\langle g_{il}, \widehat{\phi}_{lm} - \phi_{lm}\rangle - n\bar{a}_{jk}\bar{a}_{lm} \\
&= I_1 + I_2 + \ldots + I_5.
\end{aligned}$$

Observe that $|I_2| \leq O(1) \cdot k^{-\beta/2} m^{1+\beta} \left\| \sum_{i=1}^n \xi_{ijk} g_{il} \right\| \left\| \widehat{K}_{ll} - K_{ll} \right\|_{\mathcal{S}}$,
$|I_3| \leq O(1) \cdot m^{-\beta/2} k^{1+\beta} \left\| \sum_{i=1}^n \xi_{ilm} g_{ij} \right\| \left\| \widehat{K}_{jj} - K_{jj} \right\|_{\mathcal{S}}$, and
$|I_4| \leq O(1) \cdot (km)^{1+\beta} \sum_{i=1}^n \|g_{ij}\| \|g_{il}\| \left\| \widehat{K}_{ll} - K_{ll} \right\|_{\mathcal{S}} \left\| \widehat{K}_{jj} - K_{jj} \right\|_{\mathcal{S}}$. Hence the proof techniques for $n(\widehat{\sigma}_{jjkk} - \sigma^*_{jjkk})$ can be applied here and as a result, (17) follows. The proof is completed.

C.4 Proof of Theorem 2

First we obtain the general error bound for $\widehat{\mathbf{\Theta}}$ in Section C.4.1. Second in Section C.4.2 we present the general model selection consistency of fglasso in Theorem 4. Finally in Section C.4.3 we prove Theorem 2 based on the results of Lemma 3 and Theorem 4.

For convenient presentation, we adopt the definition of tail condition for the random variable given in Ravikumar et al. (2011).

Definition 1 (Tail condition) *The random vector* $\mathbf{a} \in \mathbb{R}^{Mp}$ *satisfies the tail condition if there exists a constant* $v_* \in (0, \infty]$ *and a function* $f : \mathcal{N} \times (0, \infty) \to (0, \infty)$, *such that for any* $(i, j) \in \{1, \ldots, Mp\}^2$, *let* S_{ij}, Σ^*_{ij} *be the* (i, j)*-th entry of* $\mathbf{S}, \mathbf{\Sigma}^*$ *respectively, then*

$$P\left(|S_{ij} - \Sigma^*_{ij}| \geq \delta\right) \leq 1/f(n, \delta) \ \text{ for all } \delta \in (0, 1/v_*]. \tag{C.5}$$

The tail function f is required to be monotonically increasing in δ and n. The inverse functions of n and δ are respectively defined as

$$\bar{\delta}_f(w; n) = \operatorname{argmax}\{\delta | f(n, \delta) \leq w\} \ \text{ and } \ \bar{n}_f(\delta; w) = \operatorname{argmax}\{n | f(n, \delta) \leq w\},$$

where $w \in [1, \infty)$. Then we assume that the Hessian of the negative log determinant satisfies the following general irrepresentable-type assumption.

Condition 6 *There exists some constant* $\eta \in (0, 1]$ *such that*

$$||\mathbf{\Gamma}^*_{S^c_\varepsilon S_\varepsilon} (\mathbf{\Gamma}^*_{S_\varepsilon S_\varepsilon})^{-1}||^{(M^2)}_\infty \leq 1 - \eta. \tag{C.6}$$

C.4.1 General Error Bound

In this section, we present Theorem 3 on the general error bound. We first begin with some notation. Denote by $\kappa_{\mathbf{\Gamma}^*_\varepsilon} = ||(\mathbf{\Gamma}^*_{S_\varepsilon S_\varepsilon})^{-1}||_\infty^{(M^2)}$,$\kappa_{\mathbf{B}^*_\varepsilon} = ||\mathbf{\Theta}^{*-1}\mathbf{B}^*_\varepsilon||_\infty^{(M)}\kappa_{\mathbf{\Sigma}^*}^{-1}$, where $\mathbf{B}^*_{\varepsilon,jl} = \mathbf{\Theta}^*_{jl}$ for $(j,l) \in S^c_\varepsilon$ and $\mathbf{B}^*_{\varepsilon,jl} = \mathbf{0}$ for $(j,l) \in S_\varepsilon$, and $d_\varepsilon = \max_{j\in V} \left| l \in V : ||\mathbf{\Theta}^*_{jl}||_F > \varepsilon \right|$.

Theorem 3 *Let $\widehat{\mathbf{\Theta}}$ be the unique solution to the fglasso problem (9) with regularization parameter $\gamma_n = 16\eta^{-1}M\bar{\delta}_f(n,(Mp)^\tau)$. Suppose that Conditions 2-4 and 6 hold, the bias term satisfies $||\mathbf{B}^*_\varepsilon||^{(M)}_{\max} \le \gamma_n\eta\kappa^{-2}_{\mathbf{\Sigma}^*}/16$ and the sample size n satisfies the lower bound*

$$n > \bar{n}_f\left(1/\max\left\{v_*, 6c_\eta M d_\varepsilon \max\left\{\frac{\kappa_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*_\varepsilon}}{1-3\kappa_{\mathbf{B}^*_\varepsilon}\kappa_{\mathbf{\Sigma}^*}}, \frac{\kappa^3_{\mathbf{\Sigma}^*}\kappa^2_{\mathbf{\Gamma}^*_\varepsilon}c_\eta}{1-3\kappa_{\mathbf{B}^*_\varepsilon}\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*_\varepsilon}c_\eta}\right\}\right\}, (Mp)^\tau\right) \quad \text{(C.7)}$$

with $c_\eta = 2 + 16\eta^{-1}$, then with probability at least $1-(Mp)^{2-\tau}$, we have

(i) The estimate $\widehat{\mathbf{\Theta}}$ satisfies the error bound

$$||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*||^{(M)}_{\max} \le 2c_\eta\kappa_{\mathbf{\Gamma}^*_\varepsilon}M\bar{\delta}_f(n,(Mp)^\tau); \quad \text{(C.8)}$$

(ii) The estimated edge set $\widehat{E}$ is a subset of E_ε.

C.4.2 General Model Selection Consistency

Theorem 4 *Let $\Theta^*_{\min} = \min_{(j,l)\in E_\varepsilon} ||\mathbf{\Theta}_{jl}||_F$. Under the same conditions as in Theorem 3, if the sample size n satisfies the lower bound*

$$n > \bar{n}_f\left(1/\max\left\{2\kappa_{\mathbf{\Gamma}^*_\varepsilon}c_\eta\Theta^{*-1}_{\min}M, v_*, 6c_\eta M d_\varepsilon \max\left\{\frac{\kappa_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*_\varepsilon}}{1-3\kappa_{\mathbf{B}^*_\varepsilon}\kappa_{\mathbf{\Sigma}^*}}, \frac{\kappa^3_{\mathbf{\Sigma}^*}\kappa^2_{\mathbf{\Gamma}^*_\varepsilon}c_\eta}{1-3\kappa_{\mathbf{B}^*_\varepsilon}\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*_\varepsilon}c_\eta}\right\}\right\}, (Mp)^\tau\right),$$

then $\{\widehat{E} = E_\varepsilon\}$ holds with probability at least $1-(Mp)^{2-\tau}$.

C.4.3 Proof of Theorem 2

By (18) in Theorem 1, the sample covariance matrix satisfies the tail condition (C.5) with some constants $v_* = C_1^{-1}$ and $f(n,\delta) = C_2^{-1}\exp\{C_1 n^{1-2\alpha(1+\beta)}\delta^2\}$. Therefore, the corresponding inverse functions take the following forms

$$\bar{\delta}_f(n,(Mp)^\tau) = \sqrt{\frac{\log\{C_2(Mp)^\tau\}}{C_1 n^{1-2\alpha(1+\beta)}}} = \sqrt{\frac{\tau\log Mp + \log C_2}{C_1 n^{1-2\alpha(1+\beta)}}}, \quad \text{(C.9)}$$

$$\bar{n}_f(\delta,(Mp)^\tau)=\left\{\frac{\tau\log(Mp)+\log C_2}{C_1\delta^2}\right\}^{\{1-2\alpha(1+\beta)\}^{-1}}. \tag{C.10}$$

It follows from Lemma 3 with $\varepsilon = C|E|^2 n^{\alpha(1-2\nu-\beta)}$ that $E = E_\varepsilon$. Thus we have $S = S_\varepsilon$, $d = d_\varepsilon$, $\mathbf{B}^* = \mathbf{B}^*_\varepsilon$, $\kappa_{\mathbf{\Gamma}^*} = \kappa_{\mathbf{\Gamma}^*_\varepsilon}$ and $\kappa_{\mathbf{B}^*} = \kappa_{\mathbf{B}^*_\varepsilon}$. By substituting these terms into Theorem 4, some calculations using (C.9) and (C.10) lead to the lower bound for the sample size, i.e. $n > C_3 M^2 d^2(\tau\log(Mp)+\log C_2)/c_1^2$ and $n > C_4 M^2 \Theta_{\min}^{-2}(\tau\log(Mp)+\tau\log C_2)/c_1^2$ and the desired regularization parameter γ_n.

Under Conditions 2–4, it follows from Lemma 3 that $E = E_\varepsilon$. By satisfying Condition 6 and the lower bound condition, Theorem 4 indicates that $\{E_\varepsilon = \widehat{E}\}$ holds with probability at least $1 - 1/(c_1 n^\alpha p)^{\tau-2}$. Combining these two results completes the proof.

C.5 Proof of Theorem 3

We let the sub-differential of $\sum_{j\neq l}\|(\cdot)_{jl}\|_F$ evaluated at some $\mathbf{\Theta}$ involves all symmetric matrices $\mathbf{Z}\in\mathbb{R}^{Mp\times Mp}$ with M by M blocks defined by

$$\mathbf{Z}_{jl}=\begin{cases}\mathbf{0} & \text{if } j=l\\ \frac{\mathbf{\Theta}_{jl}}{||\mathbf{\Theta}_{jl}||_F} & \text{if } j\neq l \text{ and } \mathbf{\Theta}_{jl}\neq\mathbf{0}\\ \left\{\mathbf{Z}_{jl}\in\mathbb{R}^{M\times M}:||\mathbf{Z}_{jl}||_F\leq 1\right\} & \text{if } j\neq l \text{ and } \mathbf{\Theta}_{jl}=\mathbf{0}.\end{cases} \tag{C.11}$$

By the Karush-Kuhn-Tucker (KKT) condition, a necessary and sufficient condition for $\widehat{\mathbf{\Theta}}$ to maximize (9) is

$$\widehat{\mathbf{\Theta}}^{-1}-\mathbf{S}-\gamma_n\widehat{\mathbf{Z}}=0, \tag{C.12}$$

where $\widehat{\mathbf{Z}}$ belongs to the family of sub-differential of $\sum_{j\neq l}\|\widehat{\mathbf{\Theta}}_{jl}\|_F$ defined in (C.11).

The main idea of the proof is based on constructing the primal-dual witness solution $\widetilde{\mathbf{\Theta}}$ and $\widetilde{\mathbf{Z}}$ in the following four steps.

First, $\widetilde{\mathbf{\Theta}}$ is obtained by the following restricted fglasso problem

$$\min_{\mathbf{\Theta}_{S_\varepsilon^c}=0}\left\{\text{trace}(\mathbf{S\Theta})-\log\det\mathbf{\Theta}+\gamma_n\sum_{j\neq l}||\mathbf{\Theta}_{jl}||_F\right\}, \tag{C.13}$$

where $\boldsymbol{\Theta} \in \mathbb{R}^{Mp \times Mp}$ is symmetric positive definite. Second, for each $(j,l) \in S_\varepsilon$, we choose $\widetilde{\mathbf{Z}}_{jl}$ from the family of sub-differential of $\sum_{j \neq l} ||\boldsymbol{\Theta}_{jl}||_F$ evaluated at $\widetilde{\boldsymbol{\Theta}}_{jl}$ defined in (C.11). Third, for each $(j,l) \in S_\varepsilon^c$, where $||\boldsymbol{\Theta}^*_{jl}||_F \leq \varepsilon$, $\widetilde{\mathbf{Z}}_{jl}$ is replaced by

$$\frac{1}{\gamma_n}\left\{-\mathbf{S}_{jl} + \left(\widetilde{\boldsymbol{\Theta}}^{-1}\right)_{jl}\right\}, \tag{C.14}$$

which satisfies the KKT condition (C.12). Finally, we need to verify strict dual feasibility condition, that is, $||\widetilde{\mathbf{Z}}_{jl}||_F < 1$ uniformly in $(j,l) \in S_\varepsilon^c$.

The following terms are needed in the proof of Theorem 3. Let $\mathbf{W}$ be the noise matrix, and $\boldsymbol{\Delta}$ the difference between the primal witness matrix $\widetilde{\boldsymbol{\Theta}}$ and the truth $\boldsymbol{\Theta}^*$,

$$\mathbf{W} = \mathbf{S} - \boldsymbol{\Theta}^{*-1}\ ,\ \boldsymbol{\Delta} = \widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^* = (\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*_\varepsilon) + (\boldsymbol{\Theta}^*_\varepsilon - \boldsymbol{\Theta}^*) = \boldsymbol{\Delta}_\varepsilon + \mathbf{B}^*_\varepsilon, \tag{C.15}$$

where $\boldsymbol{\Theta}^*_{\varepsilon,jl} = \mathbf{0}$ for $(j,l) \in S_\varepsilon^c$ and $\boldsymbol{\Theta}^*_{\varepsilon,jl} = \boldsymbol{\Theta}^*_{jl}$ for $(j,l) \in S_\varepsilon$. Hence for each $(j,l) \in S_\varepsilon^c, ||\boldsymbol{\Delta}_{jl}||_F \leq \varepsilon$. Note $\mathbf{B}^*_\varepsilon$ corresponds to the bias matrix caused by M-dimensional approximation in (5) to a larger dimensional function.

The second order remainder for $\widetilde{\boldsymbol{\Theta}}^{-1}$ near $\boldsymbol{\Theta}^*$ is given by

$$\mathbf{R}(\boldsymbol{\Delta}) = \widetilde{\boldsymbol{\Theta}}^{-1} - \boldsymbol{\Theta}^{*-1} + \boldsymbol{\Theta}^{*-1}\boldsymbol{\Delta}\boldsymbol{\Theta}^{*-1}. \tag{C.16}$$

To prove Theorem 3, we need use Lemmas 9-15 as stated in Supplementary Material. We organize our proof in the following six steps.

Step 1. It follows from the tail condition (C.5) and Lemma 14 that with probability at least $1 - (Mp)^{2-\tau}$ the event $\left\{||\mathbf{W}||^{(M)}_{\max} \leq M\bar{\delta}_f(n,(Mp)^\tau)\right\}$ holds . We need to verify that the conditions in Lemma 10 hold. Choosing the regularization parameter $\gamma_n = 16\eta^{-1}M\bar{\delta}_f(n,(Mp)^\tau)$ and applying the inequalities in Lemma 15 together with the bound condition for the bias term, we have $||\mathbf{W}_\varepsilon||^{(M)}_{\max} \leq ||\mathbf{W}||^{(M)}_{\max} + ||\boldsymbol{\Theta}^{*-1}\mathbf{B}^*_\varepsilon\boldsymbol{\Theta}^{*-1}||^{(M)}_{\max} \leq ||\mathbf{W}||^{(M)}_{\max} + \kappa^2_{\boldsymbol{\Sigma}^*}||\mathbf{B}^*_\varepsilon||^{(M)}_{\max} \leq \eta\gamma_n/16 + \eta\gamma_n/16 = \eta\gamma_n/8$. It remains to prove $||\mathbf{R}(\boldsymbol{\Delta})||^{(M)}_{\max}$ is also bounded by $\eta\gamma_n/8 = 2M\bar{\delta}_f(n,(Mp)^\tau)$.

Step 2. Let $r = 2\kappa_{\boldsymbol{\Gamma}^*_\varepsilon}(||\mathbf{W}_\varepsilon||^{(M)}_{\max} + \gamma_n) \leq 2\kappa_{\boldsymbol{\Gamma}^*_\varepsilon}c_\eta M\bar{\delta}_f(n,(Mp)^\tau)$. By $\bar{\delta}_f(n,(Mp)^\tau) \leq 1/v_*$ and monotonicity of the inverse tail function, for any n satisfying the lower bound condition,

we have

$$\begin{aligned} 2\kappa_{\mathbf{\Gamma}_\varepsilon^*} c_\eta M\bar{\delta}_f(n,(Mp)^\tau) &\leq \min\left\{\frac{1-3\kappa_{\mathbf{B}_\varepsilon^*}\kappa_{\mathbf{\Sigma}^*}}{3\kappa_{\mathbf{\Sigma}^*}d_\varepsilon}, \frac{1-3\kappa_{\mathbf{B}_\varepsilon^*}\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}_\varepsilon^*}c_\eta}{3\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}_\varepsilon^*}d_\varepsilon c_\eta}\right\} \\ &\leq \min\left\{\frac{1}{3\kappa_{\mathbf{\Sigma}^*}d_\varepsilon}, \frac{1}{3\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}_\varepsilon^*}d_\varepsilon}\right\} - \frac{\kappa_{\mathbf{B}_\varepsilon^*}}{d_\varepsilon}. \end{aligned}$$

Then the conditions in Lemma 12 are satisfied, and hence the error bound satisfies $||\mathbf{\Delta}||^{(M)}_{\max} = ||\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}^*||^{(M)}_{\max} \leq r$.

Step 3. The condition $||\mathbf{\Delta}_\varepsilon||^{(M)}_{\max} \leq \frac{1}{3\kappa_{\mathbf{\Sigma}^*}d_\varepsilon} - \frac{\kappa_{\mathbf{B}_\varepsilon^*}}{d_\varepsilon}$ is satisfied by step 2. Thus by Lemma 11 and results in step 2, we have

$$\begin{aligned} ||\mathbf{R}(\mathbf{\Delta})||^{(M)}_{\max} &\leq \frac{3}{2}\kappa^3_{\mathbf{\Sigma}^*}||\mathbf{\Delta}||^{(M)}_{\max}\left(d_\varepsilon||\mathbf{\Delta}||^{(M)}_{\max} + \kappa_{\mathbf{B}_\varepsilon^*}\right) \\ &\leq \left\{3\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}_\varepsilon^*}c_\eta\left(d_\varepsilon 2\kappa_{\mathbf{\Gamma}_\varepsilon^*}c_\eta M\bar{\delta}_f(n,(Mp)^\tau) + \kappa_{\mathbf{B}_\varepsilon^*}\right)\right\}\frac{\eta\gamma_n}{8} \leq \frac{\eta\gamma_n}{8}, \end{aligned}$$

where the last inequality comes from the monotonicity of the tail function, the bound condition for the sample size n, and the fact that

$$2d_\varepsilon\kappa_{\mathbf{\Gamma}_\varepsilon^*}c_\eta M\bar{\delta}_f(n,(Mp)^\tau) \leq \frac{1-3\kappa_{\mathbf{B}_\varepsilon^*}\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}_\varepsilon^*}c_\eta}{3\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}_\varepsilon^*}c_\eta} = \frac{1}{3\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}_\varepsilon^*}c_\eta} - \kappa_{\mathbf{B}_\varepsilon^*}.$$

Step 4. Steps 1 and 3 imply the strict dual feasibility in Lemma 10, and hence $\widetilde{\mathbf{\Theta}} = \widehat{\mathbf{\Theta}}$ by Lemma 9.

Step 5. It follows from the results in steps 2 and 4 that the error bound (C.8) holds with probability at least $1-(Mp)^{2-\tau}$.

Step 6. For $(j,l) \in S_\varepsilon^c$, $||\mathbf{\Theta}^*_{jl}||_F \leq \varepsilon$. Step 4 implies $\widetilde{\mathbf{\Theta}}_{S_\varepsilon^c} = \widehat{\mathbf{\Theta}}_{S_\varepsilon^c}$. In the restricted fglasso problem (C.13), we have $\widetilde{\mathbf{\Theta}}_{S_\varepsilon^c} = \widehat{\mathbf{\Theta}}_{S_\varepsilon^c} = \mathbf{0}$. Therefore, $(E_\varepsilon)^c \subset (\widehat{E})^c$ and part (ii) follows by taking the complement.

C.6 Proof of Theorem 4

It follows from the proof and results of Theorem 3(i) that $||\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}^*||^{(M)}_{\max} \leq r \leq 2c_\eta\kappa_{\mathbf{\Gamma}_\varepsilon^*}M\bar{\delta}_f(n,(Mp)^\tau)$ and $\widehat{\mathbf{\Theta}} = \widetilde{\mathbf{\Theta}}$ hold with probability at least $1-(Mp)^{2-\tau}$. The lower bound for the sample size n in (C.9) implies $\Theta^*_{\min} > 2c_\eta\kappa_{\mathbf{\Gamma}_\varepsilon^*}M\bar{\delta}_f(n,(Mp)^\tau) \geq r$. By Lemma 13 we have $\widehat{\mathbf{\Theta}}_{jl} \neq \mathbf{0}$ for all $(j,l) \in S_\varepsilon$, which entails that $E_\varepsilon \subset \widehat{E}$. Combining this result with Theorem 3(ii) yields $E_\varepsilon = \widehat{E}$.

C.7 Proof of Lemma 1

Since both $\mathbf{a} = (\mathbf{a}_1^T, \ldots, \mathbf{a}_p^T)^T$ and $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_p^T)^T$ depend on M, we omit the corresponding superscripts to simplify the notation for readability.

Let $U = V \backslash \{j, l\}$ and $\mathbf{a}_U$, $\boldsymbol{\phi}_U$ denote $(p-2)M$-dimensional vectors excluding the jth and lth subvectors from $\mathbf{a}$ and $\boldsymbol{\phi}$, respectively. By definition (6), we have that, for any pair $(j, l) \in V^2, j \neq l$,

$$\begin{aligned} C_{jl}^M(s,t) &= \mathrm{Cov}\left(\mathbf{a}_j^T \boldsymbol{\phi}_j(s), \mathbf{a}_l^T \boldsymbol{\phi}_l(t) | \mathbf{a}_k^T \boldsymbol{\phi}_k(u), k \neq j, l, \ \forall u \in \mathcal{T}\right) \\ &= \mathrm{Cov}\left(\mathbf{a}_j^T \boldsymbol{\phi}_j(s), \mathbf{a}_l^T \boldsymbol{\phi}_l(t) | \mathbf{a}_k, k \neq j, l\right) \\ &= \boldsymbol{\phi}_j(s)^T \mathrm{Cov}(\mathbf{a}_j, \mathbf{a}_l | \mathbf{a}_U) \boldsymbol{\phi}_l(t). \end{aligned} \tag{C.17}$$

The second equality comes from the following argument. For any $k \in U$ and $u \in \mathcal{T}$, $g_k^M(u) = \sum_{m=1}^M a_{km} \phi_{km}(u) = \mathbf{a}_k^T \boldsymbol{\phi}_k(u)$. By the orthogonality of ϕ_{km}, it follows that there exists a one to one correspondence between $\{\mathbf{a}_k\}$ and $\{g_k^M(u), \forall u \in \mathcal{T}\}$, which holds uniformly in k.

Since (C.17) holds for all $(s,t) \in \mathcal{T}^2$, we have that, for fixed pair $(j,l) \in V^2, j \neq l$, $C_{jl}^M(s,t) = 0$ for all $(s,t) \in \mathcal{T}^2$ if and only if $\mathrm{Cov}(\mathbf{a}_j, \mathbf{a}_l | \mathbf{a}_U) = \mathbf{0}$. Let $\mathbf{C}_{jl} = \mathrm{Cov}(\mathbf{a}_j, \mathbf{a}_l | \mathbf{a}_U)$ for each pair (j,l). Then it follows from multivariate normal theory that, for each $(j,l) \in V^2, j \neq l$, $\mathbf{C}_{jl} = -\boldsymbol{\Theta}_{jj}^{-1} \boldsymbol{\Theta}_{jl} \boldsymbol{\Theta}_{ll}^{-1}$. Since both $\boldsymbol{\Theta}_{jj}$ and $\boldsymbol{\Theta}_{ll}$ are positive definite, we have $\mathbf{C}_{jl} = \mathbf{0}$ if and only if $\boldsymbol{\Theta}_{jl} = \mathbf{0}$ for each pair $(j,l) \in V^2, j \neq l$. This completes the proof.

C.8 Lemma 2 and its Proof

Lemma 2 *Suppose that Conditions 2–3 hold. Then, for each $(j,l) \in V^2$,*

$$\left\| \boldsymbol{\Theta}_{jl}^* - \boldsymbol{\Omega}_{jl,1}^{(M)} \right\|_F \leq O\{|E|^2 n^{\alpha(1-2\nu-\beta)}\}, \tag{C.18}$$

where $\boldsymbol{\Omega}_{jl,1}^{(M)}$ is the upper left $M \times M$ submatrix of $\boldsymbol{\Omega}_{jl}$.

Proof. First we give some notations. For any $p \times p$ matrix $\mathbf{A} = (A_{ij})_{1 \leq i,j \leq p}$, let $\mathrm{tr}(\mathbf{A}) = \sum_i A_{ii}$ and $\|\mathbf{A}\|_F = \left\{\mathrm{tr}(\mathbf{A}^T \mathbf{A})\right\}^{1/2}$. For any $(M_1 p) \times (M_2 p)$ block matrix $\mathbf{A} = (\mathbf{A}_{ij})$ with $\mathbf{A}_{ij} \in \mathbb{R}^{M_1 \times M_2}, 1 \leq i, j \leq p$, we define $\|\mathbf{A}\|_{\max}^{(M_1,M_2)} = \max_{1 \leq i,j \leq p} \|\mathbf{A}_{ij}\|_F$, and $\|\mathbf{A}\|_{\infty}^{(M_1,M_2)} =$

$\max_{1\le i\le p}\sum_{j=1}^{p}\|\mathbf{A}_{ij}\|_F$. In a special case when $M_1=M_2=M$, denote $\|\mathbf{A}\|_{\max}^{(M_1,M_1)}$ and $\|\mathbf{A}\|_{\infty}^{(M_1,M_1)}$ by $\|\mathbf{A}\|_{\max}^{(M)}$ and $\|\mathbf{A}\|_{\infty}^{(M)}$, respectively. For any block matrix $\mathbf{A}=(\mathbf{A}_{ij})$ with $\mathbf{A}_{ij}\in\mathbb{R}^{M\times M}$, $1\le i,j\le p$, we define $\|\mathbf{A}\|_{\mathrm{tr}}^{(M)}=\max_{1\le i,j\le p}\{\mathrm{tr}(\mathbf{A}_{ii})\mathrm{tr}(\mathbf{A}_{jj})\}^{1/2}$.

We now prove Lemma 2. For convenience, for $j=1,\ldots,p$, denote $\mathbf{a}_{ij}=(\mathbf{b}_{ij}^T,\mathbf{c}_{ij}^T)^T$ where $\mathbf{b}_{ij}=(a_{ij1},\ldots,a_{ijM})^T$ and $\mathbf{c}_{ij}=(a_{ij(M+1)},\ldots,a_{ijM_n})^T$. Define $\widetilde{\mathbf{\Sigma}}$ to be the covariance matrix of $(\mathbf{b}_{11}^T,\ldots,\mathbf{b}_{1p}^T,\mathbf{c}_{11}^T,\ldots,\mathbf{c}_{1p}^T)^T$. Then we can find that there exists a permutation matrix $\mathbf{P}_\pi$ such that $\mathbf{P}_\pi\mathbf{\Sigma}\mathbf{P}_\pi^T=\widetilde{\mathbf{\Sigma}}$. Since $\mathbf{P}_\pi^{-1}=\mathbf{P}_\pi^T$, $\widetilde{\mathbf{\Omega}}=\mathbf{P}_\pi\mathbf{\Omega}^{-1}\mathbf{P}_\pi^T$, which means that $\widetilde{\mathbf{\Omega}}$ is only a permutation of $\mathbf{\Theta}$. Let $\widetilde{\mathbf{\Sigma}}=\begin{pmatrix}\widetilde{\mathbf{\Sigma}}_{11}&\widetilde{\mathbf{\Sigma}}_{12}\\ \widetilde{\mathbf{\Sigma}}_{21}&\widetilde{\mathbf{\Sigma}}_{22}\end{pmatrix}$ and $\widetilde{\mathbf{\Omega}}=\begin{pmatrix}\widetilde{\mathbf{\Omega}}_{11}&\widetilde{\mathbf{\Omega}}_{12}\\ \widetilde{\mathbf{\Omega}}_{21}&\widetilde{\mathbf{\Omega}}_{22}\end{pmatrix}$, where $\widetilde{\mathbf{\Omega}}_{11}$ and $\widetilde{\mathbf{\Omega}}_{11}$ are $pM\times pM$ matrices and $\widetilde{\mathbf{\Omega}}_{11}$ and $\widetilde{\mathbf{\Omega}}_{22}$ are $pM_2\times pM_2$ matrices with $M_2=M_n-M$.

Now we apply Lemma 5 to prove this lemma. By Condition 3, we see that $\|\widetilde{\mathbf{\Omega}}_{12}\|_{\infty}^{(M_1,M_2)}\le O\big(|E|n^{-\alpha\nu}\big)$. Furthermore, since the diagonal entries of $\widetilde{\mathbf{\Sigma}}_{22}$ are eigenvalues λ_{jk}'s, we have $\|\widetilde{\mathbf{\Sigma}}_{22}\|_{tr}^{(M_2)}\le O\big\{n^{\alpha(1-\beta)}\big\}$. Hence, it follows from Lemma 5 that

$$\|\widetilde{\mathbf{\Omega}}_{11}-\mathbf{\Theta}^*\|_{\max}^{(M)}\le O\big\{|E|^2n^{\alpha(1-2\nu-\beta)}\big\}.$$

As a result, for each pair $(j,l)\in V^2$, $\big\|\mathbf{\Theta}_{jl}^*-\mathbf{\Omega}_{jl,1}^{(M)}\big\|_F\le O\big\{|E|^2n^{\alpha(1-2\nu-\beta)}\big\}$. This completes the proof for Lemma 2.

C.9 Lemma 3 and its Proof

In general, for any $\varepsilon\ge 0$, we define the corresponding truncated edge set $E_\varepsilon=\{(j,l)\in V^2: j\ne l, ||\mathbf{\Theta}_{jl}^*||_F>\varepsilon\}$. Let $S_\varepsilon=E_\varepsilon\cup\{(1,1),\cdots,(p,p)\}$. Denote S_ε^c to be the complement of S_ε in V^2 with $||\mathbf{\Theta}_{jl}^*||_F\le\varepsilon$ for $(j,l)\in S_\varepsilon^c$. Lemma 3 below ensures the equivalence between the true and truncated edge sets.

Lemma 3 *Under Conditions 2–4, let $\varepsilon=C|E|^2n^{\alpha(1-2\nu-\beta)}$ for some large constant $C>0$, we have $E=E_\varepsilon$.*

Proof. First, Lemma 2 implies that for each $(j,l)\in V^2$, $\big\|\mathbf{\Theta}_{jl}^*-\mathbf{\Omega}_{jl,1}^{(M)}\big\|_F\le O(|E|^2n^{\alpha(1-2\nu-\beta)})$. Hence, for each pair $(j,l)\in E$, $\big\|\mathbf{\Theta}_{jl}^*\big\|_F\ge\big\|\mathbf{\Omega}_{jl,1}^{(M)}\big\|_F-\big\|\mathbf{\Theta}_{jl}^*-\mathbf{\Omega}_{jl,1}^{(M)}\big\|_F\gg|E|^2n^{\alpha(1-2\nu-\beta)}$, and for $(j,l)\in S^c$, $\big\|\mathbf{\Theta}_{jl}^*\big\|_F=\big\|\mathbf{\Theta}_{jl}^*-\mathbf{\Omega}_{jl,1}^{(M)}\big\|_F\le O\big(|E|^2n^{\alpha(1-2\nu-\beta)}\big)$, since $\min_{(j,l)\in E}\big\|\mathbf{\Omega}_{jl,1}^{(M)}\big\|_F\gg$

$|E|^2 n^{\alpha(1-2\nu-\beta)}$ by Condition 4 and $\left\|\mathbf{\Omega}_{jl,1}^{(M)}\right\|_F = 0$ if $(j,l) \in S^c$. This means that for $\varepsilon = C|E|^2 n^{\alpha(1-2\nu-\beta)}$ with a large constant C, we obtain $\|\mathbf{\Theta}_{jl}^*\|_F \gg \varepsilon$ if $(j,l) \in E$ but $\|\mathbf{\Theta}_{jl}^*\|_F \leq \varepsilon$ if $(j,l) \in S^c$. Therefore, $E = E_\varepsilon$ as claimed.

C.10 Lemma 4 and its Proof

Lemma 4 *For any* $\mathbf{A} \in R^{p\times q}$, $\mathbf{B} \in R^{r\times r}$, *and* $\mathbf{X} \in R^{q\times r}$, *we have*

$$\frac{\partial trace(\mathbf{A}\mathbf{X}^T\mathbf{B}\mathbf{X})}{\partial \mathbf{X}} = \mathbf{B}\mathbf{X}\mathbf{A} + \mathbf{B}^T\mathbf{X}\mathbf{A}^T. \tag{C.19}$$

Proof. Since $d(\text{trace}(\mathbf{A}\mathbf{X}^T\mathbf{B}\mathbf{X})) = \text{trace}(d(\mathbf{A}\mathbf{X}^T)\mathbf{B}\mathbf{X}) + \text{trace}(\mathbf{A}\mathbf{X}^T d(\mathbf{B}\mathbf{X}))$, we have

$$\begin{aligned} d(\text{trace}(\mathbf{A}\mathbf{X}^T\mathbf{B}\mathbf{X})) &= \text{trace}((d\mathbf{X})^T\mathbf{B}\mathbf{X}\mathbf{A}) + \text{trace}(\mathbf{A}\mathbf{X}^T\mathbf{B}d\mathbf{X}) \\ &= \text{trace}(\mathbf{A}^T\mathbf{X}^T\mathbf{B}^T + \mathbf{A}\mathbf{X}^T\mathbf{B})d\mathbf{X}. \end{aligned}$$

Hence $\frac{\partial \text{trace}(\mathbf{A}\mathbf{X}^T\mathbf{B}\mathbf{X})}{\partial \mathbf{X}} = (\mathbf{A}^T\mathbf{X}^T\mathbf{B}^T + \mathbf{A}\mathbf{X}^T\mathbf{B})^T$, which completes the proof.

C.11 Lemma 5 and its Proof

Lemma 5 *Suppose that for a positive definite matrix* $\mathbf{H} = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22} \end{pmatrix}$, *its inverse* $\mathbf{H}^{-1}$ *is* $\begin{pmatrix} \mathbf{H}^{11} & \mathbf{H}^{12} \\ \mathbf{H}^{21} & \mathbf{H}^{22} \end{pmatrix}$, *where* $\mathbf{H}_{11}$ *and* $\mathbf{H}^{11}$ *are* $pM_1 \times pM_1$ *matrices and* $\mathbf{H}_{22}$ *and* $\mathbf{H}^{22}$ *are* $pM_2 \times pM_2$ *matrices. If* $\|\mathbf{H}_{22}\|_{tr}^{(M_2)} \leq \lambda$ *and* $\|\mathbf{H}^{12}\|_{\infty}^{(M_1,M_2)} \leq \delta$, *then*

$$\|\mathbf{H}^{11} - \mathbf{H}_{11}^{-1}\|_{\max}^{(M_1)} \leq \delta^2\lambda. \tag{C.20}$$

Proof. For a positive definite matrix $\mathbf{H} = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{12}^T & \mathbf{H}_{22} \end{pmatrix}$, its inverse $\mathbf{H}^{-1}$ is expressed as

$$\begin{pmatrix} \mathbf{H}^{11} & \mathbf{H}^{12} \\ \mathbf{H}^{21} & \mathbf{H}^{22} \end{pmatrix} = \begin{pmatrix} \mathbf{H}_{11}^{-1} + \mathbf{H}_{11}^{-1}\mathbf{H}_{12}\mathbf{D}^{-1}\mathbf{H}_{12}^T\mathbf{H}_{11}^{-1} & -\mathbf{H}_{11}^{-1}\mathbf{H}_{12}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{H}_{12}^T\mathbf{H}_{11}^{-1} & \mathbf{D}^{-1} \end{pmatrix}$$

with $\mathbf{D} = \mathbf{H}_{22} - \mathbf{H}_{12}^T\mathbf{H}_{11}^{-1}\mathbf{H}_{12}$. Since $\mathbf{D}$ is positive definite, $\|\mathbf{D}\|_{\max}^{(M_2)} \leq \|\mathbf{D}\|_{tr}^{(M_2)} \leq \|\mathbf{H}_{22}\|_{tr}^{(M_2)} \leq \lambda$. Since $\mathbf{H}^{12} = -\mathbf{H}_{11}^{-1}\mathbf{H}_{12}\mathbf{D}^{-1}$, we have

$$\|\mathbf{H}_{11}^{-1}\mathbf{H}_{12}\|_{\max}^{(M_1,M_2)} = \|\mathbf{H}^{12}\mathbf{D}\|_{\max}^{(M_1,M_2)} \leq \|\mathbf{H}^{12}\|_{\infty}^{(M_1,M_2)}\|\mathbf{D}\|_{\max}^{(M_2)} \leq \delta\lambda.$$

Hence,

$$\|\mathbf{H}^{11}-\mathbf{H}_{11}^{-1}\|_{\max}^{(M_1)} \le \|\mathbf{H}^{12}\|_{\infty}^{(M_1,M_2)}\|\mathbf{H}_{11}^{-1}\mathbf{H}_{12}\|_{\max}^{(M_1,M_2)} \le \delta^2\lambda.$$

The lemma is proved.

C.12 Lemma 6 and its Proof

Lemma 6 *Suppose that Condition 1 holds. Then there exist two positive constants $C_k (k = 1, 2)$ not depending on n and p such that, for $0 < \delta \le C_1$ and each $j = 1, \ldots, p$,*

$$P\left(\left\|\widehat{K}_{jj} - K_{jj}\right\|_{\mathcal{S}} \ge \delta\right) \le C_2 \exp\left(-C_1 n\delta^2\right),$$

where $\widehat{K}_{jj}$ and K_{jj} are defined in Section 2.3.

Proof. Without ambiguity, we drop the index j in the following. For a function $K(s,t)$, define a functional $\ell_K(\phi)(t) = \int_0^1 K(s,t)\phi(s)ds$ and its norm $\|\ell_K\|_{\mathcal{S}} = (\sum_{k\ge 1}\|\ell_K(\phi_k)\|^2)^{1/2}$. Then

$$\|\widehat{K} - K\|_{\mathcal{S}} = \|\ell_{\widehat{K}} - \ell_K\|_{\mathcal{S}}.$$

For $j = 1, \ldots, p$, let $X_{ij}(s,t) = g_{ij}(s)g_{ij}(t)$ and $D_j(s,t) = \bar{g}_j(s)\bar{g}_j(t)$ with $\bar{g}_j(t) = n^{-1}\sum_{i=1}^n g_{ij}(t)$. We know that $n(\ell_{\widehat{K}} - \ell_K) = \sum_{i=1}^n (\ell_{X_i} - \ell_K) + n\ell_D$ and hence

$$n\|\widetilde{K} - K\|_{\mathcal{S}} \le \left\|\sum_{i=1}^n (\ell_{X_i} - \ell_K)\right\|_{\mathcal{S}} + n\|\ell_D\|_{\mathcal{S}}.$$

To prove this lemma, we are going to derive the following tail inequalities:

(a) There exist two constants L_1 and L_2 such that for any $\delta > 0$,

$$P\Big\{\Big\|\sum_{i=1}^n (\ell_{X_i} - \ell_K)\Big\|_{\mathcal{S}} \ge n\delta\Big\} \le 2\exp\Big(-\frac{n\delta^2}{2L_1 + 2L_2\delta}\Big); \tag{C.21}$$

(b) There exist two positive constants L_3 and L_4 such that for $\delta > 2\lambda_0 n^{-1}$,

$$P\Big(\|\ell_D\|_{\mathcal{S}} \ge \delta\Big) \le \exp\Big(-\frac{n^2\delta^2}{8L_3 + 8L_4 n\delta}\Big). \tag{C.22}$$

After getting the above two inequalities (C.21) and (C.22), we have that for all $\delta > \lambda_0/(2n)$, $P\left(n\|\widehat{K} - K\|_{\mathcal{S}} \geq n\delta\right)$ can be bounded by

$$P\left\{\left\|\sum_{i=1}^{n}(\ell_{X_i} - \ell_K)\right\|_{\mathcal{S}} \geq \frac{n\delta}{2}\right\} + P\Big(n\|\ell_{\mathcal{D}}\|_{\mathcal{S}} \geq \frac{n\delta}{2}\Big)$$
$$\leq 2\exp\left(-\frac{n\delta^2}{8L_1 + 8L_2\delta}\right) + \exp\left(-\frac{n^2\delta^2}{32L_3 + 32L_4 n\delta}\right).$$

Take $C_1 = \min\{1, L_1L_2^{-1}, (16L_1)^{-1}, (64L_4)^{-1}\}$ and $C_2 = 3\exp(C_3^2)$ with $C_3 = \max\{2\lambda_0, L_3L_4^{-1}\}$. As a result, we obtain for any δ with $0 < \delta \leq C_1$,

$$P\Big\{\|\widehat{K} - K\|_{\mathcal{S}} \geq \delta\Big\} \leq C_2\exp\Big(-C_1 n\delta^2\Big).$$

This lemma follows.

Now we turn to prove (C.21). Note that $E\left(\ell_{X_i} - \ell_K\right) = 0$ for each i. By Lemma 7, it suffices to show that there exist two positive constants L_1 and L_2 such that

$$\sum_{i=1}^{n} E\left\|\ell_{X_i} - \ell_K\right\|_{\mathcal{S}}^k \leq \frac{1}{2}k!nL_1L_2^{k-2}, k = 2, \ldots. \tag{C.23}$$

Note that $\|\ell_{X_i} - \ell_K\|_{\mathcal{S}}^2 = \sum_{m,m'=1}^{\infty}\Big(a_{im}a_{im'} - \lambda_{mm'}\Big)^2$ where $\lambda_{mm'} = \lambda_m\delta_{mm'}$ and $\delta_{mm'} = I(m = m')$. By Jensen's inequality,

$$\begin{aligned}
E\left\|\ell_{X_i} - \ell_K\right\|_{\mathcal{S}}^k &= E\left\{\sum_{m,m'=1}^{\infty}\lambda_m\lambda_{m'}\Big(\xi_{im}\xi_{im'} - \delta_{mm'}\Big)^2\right\}^{k/2} \\
&\leq \left(\sum_{m,m'=1}^{\infty}\lambda_m\lambda_{m'}\right)^{k/2-1}\sum_{m,m'=1}^{\infty}\lambda_m\lambda_{m'}E\Big(\xi_{im}\xi_{im'} - \delta_{mm'}\Big)^k \\
&\leq \left(2\sum_{m=1}^{\infty}\lambda_m\right)^k\Big(E\xi_{i1}^{2k} + 1\Big),
\end{aligned}$$

where the inequality $E(\xi_{i1}^2 - 1)^k \leq 2^{k-1}(1 + E\xi_{i1}^{2k})(k \geq 2)$ is used. Since $\xi_{i1} \sim N(0,1)$,

$$E|\xi_{i1}|^{2k} = \pi^{-1/2}2^k\Gamma\left(\frac{2k+1}{2}\right) \leq 2^k k!.$$

Let $L_2 = 4\sum_{m=1}^{\infty}\lambda_m = 4\lambda_0 < \infty$ and $L_1 = 4L_2^2$. Then, for $k = 2, 3, \ldots,$

$$\sum_{i=1}^{n} E\left\|\ell_{X_i} - \ell_K\right\|_{\mathcal{S}}^k \leq (L_2/2)^k \cdot 2 \cdot 2^k k! \leq \frac{1}{2}k!nL_1L_2^{k-2}.$$

Next we consider to prove the inequality (C.22). Suppose that we have shown

$$E\|\ell_D\|_{\mathcal{S}}^k \le \frac{1}{2}n^{-k}k!L_3L_4^{k-2}, k = 2, 3, \ldots. \tag{C.24}$$

Then, the following inequality follows from Lemma 7:

$$P\Big(\|\ell_D\|_{\mathcal{S}} - E\|\ell_D\|_{\mathcal{S}} \ge \delta\Big) \le \exp\Big(-\frac{n^2\delta^2}{2L_3 + 2L_4 n\delta}\Big)$$

for all $\delta > 0$. Note that $\|\ell_D\|_{\mathcal{S}}^2 = n^{-2}\sum_{m,m'=1}^{\infty}\lambda_m\lambda_{m'}(\bar{\xi}_m\bar{\xi}_{m'})^2$, where $\bar{\xi}_m = n^{-1/2}\sum_{i=1}^n \xi_{im}$. Hence $E\|\ell_D\|_{\mathcal{S}} \le n^{-1}\lambda_0$. As a result, for $\delta > 2n^{-1}\lambda_0$, we have that

$$P\Big(\|\ell_D\|_{\mathcal{S}} \ge \delta\Big) \le P\Big(\|\ell_D\|_{\mathcal{S}} - E\|\ell_D\|_{\mathcal{S}} \ge \delta/2\Big) \le \exp\Big(-\frac{n^2\delta^2}{8L_3 + 8L_4 n\delta}\Big).$$

Hence, (C.22) follows.

Now we derive the upper bound of $E\|\ell_D\|_{\mathcal{S}}^k$ for $k \ge 2$ as in (C.24). By Jensen's inequality,

$$\begin{aligned}
E\|\ell_D\|_{\mathcal{S}}^k &\le \frac{1}{n^k}E\left\{\sum_{m,m'=1}^{\infty}\lambda_m\lambda_{m'}(\xi_{im}\xi_{im'})^2\right\}^{k/2} \\
&\le \frac{1}{n^k}\left(\sum_{m,m'=1}^{\infty}\lambda_m\lambda_{m'}\right)^{k/2-1}\sum_{m,m'=1}^{\infty}\lambda_m\lambda_{m'}E|\xi_{im}\xi_{im'}|^k \\
&\le \frac{1}{n^k}\left(\sum_{m=1}^{\infty}\lambda_m\right)^k E\xi_{i1}^{2k} \le \left(\frac{2}{n}\sum_{m=1}^{\infty}\lambda_m\right)^k k!.
\end{aligned}$$

Let $L_4 = 2\sum_{m=1}^{\infty}\lambda_m$ and $L_3 = 2L_4^2$. Then $E\|\ell_D\|_{\mathcal{S}}^k \le 2^{-1}n^{-k}k!L_3L_4^{k-2}$. Lemma 6 is proved.

C.13 Lemma 7 and its Proof

Lemma 7 *Let $\{X_1, \ldots, X_n\}$ be independent random variables in a separable Hilbert space with norm $\|\cdot\|$. If $EX_i = 0$ $(i = 1, \ldots, n)$ and*

$$\sum_{i=1}^n E\|X_i\|^k \le \frac{k!}{2}nL_1L_2^{k-2}, k = 2, 3, \ldots,$$

for two positive constants L_1 and L_2, then for all $\delta > 0$,

$$P\Big(\Big\|\sum_{i=1}^n X_i\Big\| \ge n\delta\Big) \le 2\exp\Big(-\frac{n\delta^2}{2L_1 + 2L_2\delta}\Big).$$

Proof. This lemma can be derived directly from Theorem 2.5 (2) of Bosq (2000) and hence its proof is omitted.

C.14 Lemma 8 and its Proof

Lemma 8 *Suppose that Condition 1 holds. Denote* $\widetilde{\phi}_{jk} = sgn\langle\widehat{\phi}_{jk}, \phi_{jk}\rangle\phi_{jk}$. *Then*

$$\|\widehat{\phi}_{jk} - \widetilde{\phi}_{jk}\| \leq d_{jk}\|\widehat{K}_{jj} - K_{jj}\|_{\mathcal{S}},$$

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

Proof. This lemma can be found in Lemma 4.3 of Bosq (2000) and hence the proof is omitted.

C.15 Lemma 9 and its Proof

Lemma 9 *For any* $\gamma_n \geq 0$, *the fglasso problem (9) has a unique solution that satisfies the optimal condition (C.12) with* $\widehat{\mathbf{Z}}$ *defined in (C.11).*

Proof. The fglasso problem can be written in the constrained form

$$\min_{\sum_{j\neq l}||\mathbf{\Theta}_{jl}||_F \leq C(\gamma_n)} \{\text{trace}(\mathbf{S\Theta}) - \log\,\det\mathbf{\Theta}\}\,, \tag{C.25}$$

where $\mathbf{\Theta} \in \mathbb{R}^{Mp\times Mp}$ is symmetric positive definite. The objective function is strictly convex in view of its Hessian and the constraint on the parameter space, if the minimum is attained then the solution is uniquely determined. We need to show that the minimum is achieved. Note the off block diagonal elements are bounded by satisfying $\sum_{j\neq l}||\mathbf{\Theta}_{jl}||_F \leq C(\lambda) < \infty$. By the fact that $\max_{i,j} A_{ij} \leq \max_i A_{ii}$ for a positive definite matrix $\mathbf{A}$, we only need to consider the possibly unbounded diagonal elements. By Hadamard's inequality for positive definite matrices, we have

$$\text{trace}(\mathbf{S\Theta}) - \log\,\det\mathbf{\Theta} \geq \sum_{i=1}^{Mp}(\mathbf{S}_{ii}\mathbf{\Theta}_{ii} - \log\,\det\mathbf{\Theta}_{ii})\,.$$

The diagonal elements of $\mathbf{S}$ are positive. The objective function goes to infinity as any sequence $(\mathbf{\Theta}_{11}^{(k)}, \ldots, \mathbf{\Theta}_{Mp,Mp}^{(k)}), k \geq 1$, goes to infinity. Thus the minimum is uniquely achieved.

C.16 Lemma 10 and its Proof

Lemma 10 *Suppose that* $\max\left\{||\mathbf{W}_\varepsilon||_{\max}^{(M)}, ||\mathbf{R}(\mathbf{\Delta})||_{\max}^{(M)}\right\} \leq \frac{\eta\gamma_n}{8}$, *where* $\mathbf{W}_\varepsilon = \mathbf{W}+\mathbf{\Theta}^{*-1}\mathbf{B}_\varepsilon^*\mathbf{\Theta}^{*-1}$. *Then* $\widetilde{\mathbf{Z}}_{S_\varepsilon^c}$ *constructed in (C.14) satisfies* $||\widetilde{\mathbf{Z}}_{S_\varepsilon^c}||_{\max}^{(M)} < 1$.

Proof. The optimal condition (C.12) can be replaced by

$$\mathbf{\Theta}^{*-1}\mathbf{\Delta}\mathbf{\Theta}^{*-1} + \mathbf{W} - \mathbf{R}(\mathbf{\Delta}) + \gamma_n\widetilde{\mathbf{Z}} = \mathbf{0},$$

and can be rewritten as

$$\mathbf{\Theta}^{*-1}\mathbf{\Delta}_\varepsilon\mathbf{\Theta}^{*-1} + \mathbf{W}_\varepsilon - \mathbf{R}(\mathbf{\Delta}) + \gamma_n\widetilde{\mathbf{Z}} = \mathbf{0}. \tag{C.26}$$

Note $\mathrm{vec}(\mathbf{\Theta}^{*-1}\mathbf{\Delta}_\varepsilon\mathbf{\Theta}^{*-1}) = (\mathbf{\Theta}^{*-1}\otimes\mathbf{\Theta}^{*-1})\mathrm{vec}(\mathbf{\Delta}_\varepsilon)$. Taking vectorization for (C.26), we have

$$\begin{pmatrix} \mathbf{\Gamma}^*_{S_\varepsilon S_\varepsilon} & \mathbf{\Gamma}^*_{S_\varepsilon S_\varepsilon^c} \\ \mathbf{\Gamma}^*_{S_\varepsilon^c S_\varepsilon} & \mathbf{\Gamma}^*_{S_\varepsilon^c S_\varepsilon^c} \end{pmatrix} \begin{pmatrix} \mathrm{vec}(\mathbf{\Delta}_{\varepsilon,S_\varepsilon}) \\ \mathrm{vec}(\mathbf{\Delta}_{\varepsilon,S_\varepsilon^c}) \end{pmatrix} + \begin{pmatrix} \mathrm{vec}(\mathbf{W}_{\varepsilon,S_\varepsilon}) \\ \mathrm{vec}(\mathbf{W}_{\varepsilon,S_\varepsilon^c}) \end{pmatrix} - \begin{pmatrix} \mathrm{vec}(\mathbf{R}_{S_\varepsilon}) \\ \mathrm{vec}(\mathbf{R}_{S_\varepsilon^c}) \end{pmatrix} + \gamma_n \begin{pmatrix} \mathrm{vec}(\widetilde{\mathbf{Z}}_{S_\varepsilon}) \\ \mathrm{vec}(\widetilde{\mathbf{Z}}_{S_\varepsilon^c}) \end{pmatrix} = \mathbf{0}. \tag{C.27}$$

We solve for $\mathrm{vec}(\mathbf{\Delta}_{\varepsilon,S_\varepsilon})$ from the first line and substitute it into the second line. Then $\mathrm{vec}(\widetilde{\mathbf{Z}}_{S_\varepsilon^c})$ can be represented as

$$\begin{aligned} \mathrm{vec}(\widetilde{\mathbf{Z}}_{S_\varepsilon^c}) \;=\; & \frac{1}{\gamma_n}\mathbf{\Gamma}^*_{S_\varepsilon^c S_\varepsilon}(\mathbf{\Gamma}^*_{S_\varepsilon S_\varepsilon})^{-1}\left(\mathrm{vec}(\mathbf{W}_{\varepsilon,S_\varepsilon}) - \mathrm{vec}(\mathbf{R}_{S_\varepsilon})\right) \\ & + \mathbf{\Gamma}^*_{S_\varepsilon^c S_\varepsilon}(\mathbf{\Gamma}^*_{S_\varepsilon S_\varepsilon})^{-1}\mathrm{vec}(\widetilde{\mathbf{Z}}_{S_\varepsilon}) \\ & - \frac{1}{\gamma_n}\left(\mathrm{vec}(\mathbf{W}_{\varepsilon,S_\varepsilon^c}) - \mathrm{vec}(\mathbf{R}_{S_\varepsilon^c})\right). \end{aligned}$$

For any vector $\mathbf{v} = (\mathbf{v}_j)$ with $\mathbf{v}_j \in \mathbb{R}^{M^2}$, $1 \leq j \leq p$, define $||\mathbf{v}||_{\max}^{(M^2)} = \max_j ||\mathbf{v}_j||_2$ as the M^2-group version of l_∞ norm. Taking the M^2-group l_∞ norm on both sides, it follows from (C.33) and (C.34) in Lemma 15 that

$$\begin{aligned} ||\mathrm{vec}(\widetilde{\mathbf{Z}}_{S_\varepsilon^c})||_{\max}^{(M^2)} \;\leq\; & \frac{1}{\gamma_n}||\mathbf{\Gamma}^*_{S_\varepsilon^c S_\varepsilon}(\mathbf{\Gamma}^*_{S_\varepsilon S_\varepsilon})^{-1}||_\infty^{(M^2)}\left(||\mathrm{vec}(\mathbf{W}_{\varepsilon,S_\varepsilon})||_{\max}^{(M^2)} + ||\mathrm{vec}(\mathbf{R}_{S_\varepsilon})||_{\max}^{(M^2)}\right) \\ & + ||\mathbf{\Gamma}^*_{S_\varepsilon^c S_\varepsilon}(\mathbf{\Gamma}^*_{S_\varepsilon S_\varepsilon})^{-1}||_\infty^{(M^2)}||\mathrm{vec}(\widetilde{\mathbf{Z}}_{S_\varepsilon})||_{\max}^{(M^2)} \\ & + \frac{1}{\gamma_n}\left(||\mathrm{vec}(\mathbf{W}_{\varepsilon,S_\varepsilon^c})||_{\max}^{(M^2)} + ||\mathrm{vec}(\mathbf{R}_{S_\varepsilon^c})||_{\max}^{(M^2)}\right). \end{aligned}$$

Note that $||\text{vec}(\widetilde{\mathbf{Z}}_{S_\varepsilon})||_{\max}^{(M^2)} \leq 1$ by construction. Applying (C.30) in Lemma 15, the bound condition for $||\mathbf{W}_\varepsilon||_{\max}^{(M)}$, $||\mathbf{R}||_{\max}^{(M)}$ and Condition 6 yield

$$\begin{aligned} ||\widetilde{\mathbf{Z}}_{S_\varepsilon^c}||_{\max}^{(M)} &\leq \frac{2-\eta}{\gamma_n}\left(||\mathbf{W}_\varepsilon||_{\max}^{(M)} + ||\mathbf{R}||_{\max}^{(M)}\right) + (1-\eta) \\ &\leq \frac{2-\eta}{\gamma_n}\left(\frac{\eta\gamma_n}{4}\right) + (1-\eta) \leq \frac{\eta}{2} + 1 - \eta < 1. \end{aligned}$$

C.17 Lemma 11 and its Proof

Lemma 11 *Suppose that* $||\mathbf{\Delta}_\varepsilon||_{\max}^{(M)} \leq \frac{1}{3\kappa_{\mathbf{\Sigma}^*} d_\varepsilon} - \frac{\kappa_{\mathbf{B}_\varepsilon^*}}{d_\varepsilon}$, *then* $||\mathbf{J}^T||_\infty \leq \frac{3}{2}$ *and*

$$||\mathbf{R}(\mathbf{\Delta})||_{\max}^{(M)} \leq \frac{3}{2}\kappa_{\mathbf{\Sigma}^*}^3 ||\mathbf{\Delta}||_{\max}^{(M)}\left(d_\varepsilon ||\mathbf{\Delta}_\varepsilon||_{\max}^{(M)} + \kappa_{\mathbf{B}_\varepsilon^*}\right),$$

where $\mathbf{J} = \sum_{k=0}^{\infty}(-1)^k(\mathbf{\Theta}^{*-1}\mathbf{\Delta})^k$ *and* $\mathbf{R}(\mathbf{\Delta}) = \mathbf{\Theta}^{*-1}\mathbf{\Delta}\mathbf{\Theta}^{*-1}\mathbf{\Delta}\mathbf{J}\mathbf{\Theta}^{*-1}$.

Proof. By the fact that $\mathbf{\Delta}_\varepsilon$ has at most d_ε $M \times M$ blocks whose Frobenius norm is at least ε for each column block, then $||\mathbf{\Delta}_\varepsilon||_\infty^{(M)} \leq d_\varepsilon ||\mathbf{\Delta}_\varepsilon||_{\max}^{(M)}$. It follows from (C.31), (C.32) in Lemma 15 and the bound condition for $||\mathbf{\Delta}_\varepsilon||_{\max}^{(M)}$ that

$$\begin{aligned} ||\mathbf{\Theta}^{*-1}\mathbf{\Delta}||_\infty^{(M)} &\leq ||\mathbf{\Theta}^{*-1}||_\infty^{(M)} ||\mathbf{\Delta}_\varepsilon||_\infty^{(M)} + ||\mathbf{\Theta}^{*-1}\mathbf{B}_\varepsilon^*||_\infty^{(M)} \\ &\leq \kappa_{\mathbf{\Sigma}^*}\left(d_\varepsilon ||\mathbf{\Delta}_\varepsilon||_{\max}^{(M)} + \kappa_{\mathbf{B}_\varepsilon^*}\right) \leq 1/3. \end{aligned}$$

Hence it follows from we have the convergent matrix expansion via Neumann series

$$(\mathbf{\Theta}^* + \mathbf{\Delta})^{-1} = \mathbf{\Theta}^{*-1} - \mathbf{\Theta}^{*-1}\mathbf{\Delta}\mathbf{\Theta}^{*-1} + \mathbf{\Theta}^{*-1}\mathbf{\Delta}\mathbf{\Theta}^{*-1}\mathbf{\Delta}\mathbf{J}\mathbf{\Theta}^{*-1}.$$

By the definitions of $\mathbf{R}(\mathbf{\Delta})$ and $\mathbf{\Delta}$, we obtain $\mathbf{R}(\mathbf{\Delta}) = \mathbf{\Theta}^{*-1}\mathbf{\Delta}\mathbf{\Theta}^{*-1}\mathbf{\Delta}\mathbf{J}\mathbf{\Theta}^{*-1}$. Let $\mathbf{e}_j \in R^{Mp\times M}$ with identity matrix in the jth block and zero matrix elsewhere, and $\mathbf{x} \in \mathbb{R}^{Mp\times M}$ with jth block $\mathbf{x}_j \in \mathbb{R}^{M\times M}$. Define $||\mathbf{x}||_{\max}^{(M)} = \max_j ||\mathbf{x}_j||_F$ and $||\mathbf{x}||_1^{(M)} = \sum_{j=1}^p ||\mathbf{x}_j||_F$. Recall that given an M-block matrix $\mathbf{A}$, we have defined M-block version of matrix ∞-norm as $||\mathbf{A}||_\infty^{(M)} = \max_i \sum_{j=1}^p ||\mathbf{A}_{ij}||_F$. Define the corresponding M-block version of matrix 1-norm

by $||\mathbf{A}||_1^{(M)} = \max_j \sum_{i=1}^p ||\mathbf{A}_{ij}||_F$. It follows from the inequalities in Lemma 15 that

$$\begin{aligned}
||\mathbf{R}(\mathbf{\Delta})||_{\max}^{(M)} &= \max_{i,j} ||\mathbf{e}_i^T \mathbf{\Theta}^{*-1} \mathbf{\Delta} \mathbf{\Theta}^{*-1} \mathbf{\Delta} \mathbf{J} \mathbf{\Theta}^{*-1} \mathbf{e}_j||_F \\
&\leq \max_i ||\mathbf{e}_i^T \mathbf{\Theta}^{*-1} \mathbf{\Delta}||_{\max}^{(M)} \max_j ||\mathbf{\Theta}^{*-1} \mathbf{\Delta} J \mathbf{\Theta}^{*-1} \mathbf{e}_j||_1^{(M)} \\
&\leq \max_i ||\mathbf{e}_i^T \mathbf{\Theta}^{*-1}||_1^{(M)} ||\mathbf{\Delta}||_{\max}^{(M)} \max_j ||\mathbf{\Theta}^{*-1} \mathbf{\Delta} \mathbf{J} \mathbf{\Theta}^{*-1} \mathbf{e}_j||_1^{(M)} \\
&= ||\mathbf{\Theta}^{*-1}||_\infty^{(M)} ||\mathbf{\Delta}||_{\max}^{(M)} ||\mathbf{\Theta}^{*-1} \mathbf{\Delta} \mathbf{J} \mathbf{\Theta}^{*-1} \mathbf{e}_j||_1^{(M)} \\
&\leq \kappa_{\mathbf{\Sigma}^*} ||\mathbf{\Delta}||_{\max}^{(M)} ||\mathbf{\Theta}^{*-1} \mathbf{J}^T \mathbf{\Delta} \mathbf{\Theta}^{*-1}||_\infty^{(M)} \\
&\leq \kappa_{\mathbf{\Sigma}^*}^2 ||\mathbf{\Delta}||_{\max}^{(M)} ||\mathbf{J}^T||_\infty^{(M)} ||\mathbf{\Theta}^{*-1} \mathbf{\Delta}||_\infty^{(M)}
\end{aligned}$$

Note that $\mathbf{J} = \sum_{k=0}^\infty (-1)^k (\mathbf{\Theta}^{*-1} \mathbf{\Delta})^k$. It follows from (C.32) in Lemma 15 that

$$||\mathbf{J}^T||_\infty^{(M)} \leq \sum_{k=0}^\infty \left(||\mathbf{\Theta}^{*-1} \mathbf{\Delta}||_\infty^{(M)} \right)^k = \frac{1}{1 - ||\mathbf{\Theta}^{*-1} \mathbf{\Delta}||_\infty^{(M)}} \leq \frac{3}{2}.$$

Hence it follows from (C.28) that we can bound the second order remainder $\mathbf{R}(\mathbf{\Delta})$ by

$$||\mathbf{R}(\mathbf{\Delta})||_{\max}^{(M)} \leq \frac{3}{2} \kappa_{\mathbf{\Sigma}^*}^3 ||\mathbf{\Delta}||_{\max}^{(M)} \left(d_\varepsilon ||\mathbf{\Delta}_\varepsilon||_{\max}^{(M)} + \kappa_{\mathbf{B}_\varepsilon^*} \right).$$

C.18 Lemma 12 and its Proof

Lemma 12 *Suppose that* $r = 2\kappa_{\mathbf{\Gamma}_\varepsilon^*} (||\mathbf{W}_\varepsilon||_{\max}^{(M)} + \gamma_n) \leq \min \left\{ \frac{1}{3\kappa_{\mathbf{\Sigma}^*} d_\varepsilon}, \frac{1}{3\kappa_{\mathbf{\Sigma}^*}^3 \kappa_{\mathbf{\Gamma}_\varepsilon^*} d_\varepsilon} \right\} - \frac{\kappa_{\mathbf{B}_\varepsilon^*}}{d_\varepsilon}$. *then* $||\mathbf{\Delta}||_{\max}^{(M)} = ||\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}^*||_{\max}^{(M)} \leq r$.

Proof. Let $G(\mathbf{\Theta}_{S_\varepsilon}) = -(\mathbf{\Theta}^{-1})_{S_\varepsilon} + \mathbf{S}_{S_\varepsilon} + \gamma_n \widetilde{\mathbf{Z}}_{S_\varepsilon}$. We define a continuous map $F : \mathbb{R}^{M^2|S_\varepsilon|} \to \mathbb{R}^{M^2|S_\varepsilon|}$ by

$$F(\text{vec}(\mathbf{\Delta}_{S_\varepsilon})) = -(\mathbf{\Gamma}_{S_\varepsilon S_\varepsilon}^*)^{-1} \text{vec}(G(\mathbf{\Theta}_{S_\varepsilon}^* + \mathbf{\Delta}_{S_\varepsilon})) + \text{vec}(\mathbf{\Delta}_{S_\varepsilon}). \tag{C.28}$$

Note that $F(\text{vec}(\mathbf{\Delta}_{S_\varepsilon})) = \text{vec}(\mathbf{\Delta}_{S_\varepsilon})$ holds if and only if $G(\mathbf{\Theta}_{S_\varepsilon}^* + \mathbf{\Delta}_{S_\varepsilon}) = G(\widetilde{\mathbf{\Theta}}_{S_\varepsilon}) = \mathbf{0}$ by construction. We need to show that the function F maps the following ball $\mathbf{B}(r)$ onto itself

$$\mathbf{B}(r) = \{\mathbf{\Theta}_{S_\varepsilon} : ||\mathbf{\Theta}_{S_\varepsilon}||_{\max}^{(M)} \leq r\}, \tag{C.29}$$

where $r = 2\kappa_{\mathbf{\Gamma}_\varepsilon^*} (||\mathbf{W}_\varepsilon||_{\max}^{(M)} + \gamma_n)$. Note F is continuous and $\mathbf{B}(r)$ is convex and compact, then by Brouwer's fixed point theorem, there exists some fixed point $\mathbf{\Delta}_{S_\varepsilon} \in \mathbf{B}(r)$, which implies

that $||\widetilde{\mathbf{\Theta}}_{S_\varepsilon} - \mathbf{\Theta}^*_{S_\varepsilon}||^{(M)}_{\max} \leq r$. It remains to prove the claim $F(\mathbf{B}(r)) \subseteq \mathbf{B}(r)$. Note that

$$\begin{aligned} G(\mathbf{\Theta}^*_{S_\varepsilon} + \mathbf{\Delta}_{S_\varepsilon}) &= -[(\mathbf{\Theta}^* + \mathbf{\Delta})^{-1}]_{S_\varepsilon} + \mathbf{S}_{S_\varepsilon} + \gamma_n \widetilde{\mathbf{Z}}_{S_\varepsilon} \\ &= -[(\mathbf{\Theta}^* + \mathbf{\Delta})^{-1} - \mathbf{\Theta}^{*-1}]_{S_\varepsilon} + [\mathbf{S} - \mathbf{\Theta}^{*-1}]_{S_\varepsilon} + \gamma_n \widetilde{\mathbf{Z}}_{S_\varepsilon} \\ &= -[\mathbf{R}(\mathbf{\Delta}) - \mathbf{\Theta}^{*-1}\mathbf{\Delta}_\varepsilon\mathbf{\Theta}^{*-1}]_{S_\varepsilon} + \mathbf{W}_{\varepsilon,S_\varepsilon} + \gamma_n \widetilde{\mathbf{Z}}_{S_\varepsilon}. \end{aligned}$$

Then (C.28) can be substituted by

$$F(\mathrm{vec}(\mathbf{\Delta}_{S_\varepsilon})) = \underbrace{(\mathbf{\Gamma}^*_{S_\varepsilon S_\varepsilon})^{-1}\mathrm{vec}(\mathbf{R}_{S_\varepsilon})}_{\mathbf{T}_1} - \underbrace{(\mathbf{\Gamma}^*_{S_\varepsilon S_\varepsilon})^{-1}\Big(\mathrm{vec}(\mathbf{W}_{\varepsilon,S_\varepsilon}) + \gamma_n \mathrm{vec}(\widetilde{\mathbf{Z}}_{S_\varepsilon})\Big)}_{\mathbf{T}_2}.$$

By the definition of $\kappa_{\mathbf{\Gamma}^*_\varepsilon}$ and (C.33) in Lemma 15, $\mathbf{T}_2$ can be bounded by

$$||\mathbf{T}_2||^{(M^2)}_{\max} \leq \kappa_{\mathbf{\Gamma}^*_\varepsilon}\left(||\mathbf{W}_{\varepsilon,S_\varepsilon}||^{(M)}_{\max} + \gamma_n\right) = r/2.$$

With the assumed bound for r, we have $||\mathbf{\Delta}_\varepsilon||^{(M)}_{\max} \leq r \leq \frac{1}{3\kappa_{\mathbf{\Sigma}^*} d_\varepsilon} - \frac{\kappa_{\mathbf{B}^*_\varepsilon}}{d_\varepsilon}$. Then an application of the bound for $\mathbf{R}(\mathbf{\Delta})$ in Lemma 11 yields

$$||\mathbf{T}_1||^{(M^2)}_{\max} \leq \frac{3}{2}\kappa_{\mathbf{\Gamma}^*_\varepsilon}\kappa^3_{\mathbf{\Sigma}^*}||\mathbf{\Delta}||^{(M)}_{\max}\left(d_\varepsilon||\mathbf{\Delta}_\varepsilon||^{(M)}_{\max} + \kappa_{\mathbf{B}^*_\varepsilon}\right) \leq \frac{||\mathbf{\Delta}||^{(M)}_{\max}}{2} \leq \frac{r}{2},$$

where we have used the assumption $||\mathbf{\Delta}_\varepsilon||^{(M)}_{\max} \leq r \leq \frac{1}{3\kappa^3_{\mathbf{\Sigma}^*}\kappa_{\mathbf{\Gamma}^*_\varepsilon} d_\varepsilon} - \frac{\kappa_{\mathbf{B}^*_\varepsilon}}{d_\varepsilon}$. Therefore, we obtain

$$||F(\mathrm{vec}(\mathbf{\Delta}_{S_\varepsilon}))||^{(M^2)}_{\max} \leq ||\mathbf{T}_1||^{(M^2)}_{\max} + ||\mathbf{T}_2||^{(M^2)}_{\max} \leq r,$$

which proves the claim.

C.19 Lemma 13 and its Proof

Lemma 13 *Suppose that all conditions in Lemma 12 hold and* $\Theta_{\min} = \min_{(j,l)\in E_\varepsilon} ||\mathbf{\Theta}^*_{jl}||_F$ *satisfies* $\Theta^*_{\min} > 2\kappa_{\mathbf{\Gamma}^*_\varepsilon}(||\mathbf{W}_\varepsilon||^{(M)}_{\max} + \gamma_n)$, *then* $\widetilde{\mathbf{\Theta}}_{jl} \neq \mathbf{0}$ *for all* $(j,l) \in S_\varepsilon$.

Proof. From Lemma 12, we have $||\widetilde{\mathbf{\Theta}}_{jl} - \mathbf{\Theta}^*_{jl}||_F \leq r$ for any $(j,l) \in S_\varepsilon$. Thus $\widetilde{\mathbf{\Theta}}_{jl} \neq \mathbf{0}$ for all $(j,l) \in S_\varepsilon$ follows immediately from the lower bound condition on $\Theta^*_{\min}$.

C.20 Lemma 14 and its Proof

Lemma 14 *For any $\tau > 2$ and sample size n such that $\bar{\delta}_f(n,(Mp)^\tau) \leq 1/v^*$, we have $P\left(||\mathbf{W}||_{\max}^{(M)} \geq M\bar{\delta}_f(n,(Mp)^\tau)\right) \leq (Mp)^{2-\tau}$.*

Proof. By the definition of the tail function in (C.5), we have $P(W_{kl} > \delta) \leq \frac{1}{f(n,\delta)}$, where $\mathbf{W} \in \mathbb{R}^{Mp\times Mp}$ and $(k,l) \in \{1,\ldots,Mp\}^2$. It then follows from union bound of probability and $\delta = \bar{\delta}_f(n,(Mp)^\tau)$ that

$$P\left(||\mathbf{W}||_{\max}^{(M)} \geq M\bar{\delta}_f(n,(Mp)^\tau)\right) = P(\max_{i,j}||\mathbf{W}_{ij}||_F > M\delta) \leq \frac{M^2p^2}{f(n,\delta)} = (Mp)^{2-\tau}.$$

C.21 Lemma 15 and its proof

Lemma 15 *Let $\mathbf{A} = (\mathbf{A}_{ij}), \mathbf{B} = (\mathbf{B}_{ij})$ with $\mathbf{A}_{ij}, \mathbf{B}_{ij} \in \mathbb{R}^{M\times M}, 1 \leq i,j \leq p$, $\mathbf{u} = (\mathbf{u}_j), \mathbf{v} = (\mathbf{v}_j)$ with $\mathbf{u}_j, \mathbf{v}_j \in \mathbb{R}^M, 1 \leq j \leq p$, and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{Mp\times M}$ with jth block $\mathbf{x}_j, \mathbf{y}_j \in \mathbb{R}^{M\times M}$, respectively. Then the following norm properties hold:*

$$||\mathbf{A}||_{\max}^{(M)} = ||vec(\mathbf{A})||_{\max}^{(M^2)}, \tag{C.30}$$

$$||\mathbf{A}+\mathbf{B}||_{\infty}^{(M)} \leq ||\mathbf{A}||_{\infty}^{(M)} + ||\mathbf{B}||_{\infty}^{(M)}, \tag{C.31}$$

$$||\mathbf{AB}||_{\infty}^{(M)} \leq ||\mathbf{A}||_{\infty}^{(M)}||\mathbf{B}||_{\infty}^{(M)}, \tag{C.32}$$

$$||\mathbf{Au}||_{\max}^{(M)} \leq ||\mathbf{A}||_{\infty}^{(M)}||\mathbf{u}||_{\max}^{(M)}, \tag{C.33}$$

$$||\mathbf{u}+\mathbf{v}||_{\max}^{(M)} \leq ||\mathbf{u}||_{\max}^{(M)} + ||\mathbf{v}||_{\max}^{(M)}, \tag{C.34}$$

$$||\mathbf{x}^T\mathbf{y}||_F^{(M)} \leq ||\mathbf{x}||_{\max}^{(M)}||\mathbf{y}||_1^{(M)}, \tag{C.35}$$

$$||\mathbf{Ax}||_{\max}^{(M)} \leq ||\mathbf{A}||_{\max}^{(M)}||\mathbf{x}||_1^{(M)}, \tag{C.36}$$

$$||\mathbf{A}||_{\infty}^{(M)} = ||\mathbf{A}^T||_1^{(M)}. \tag{C.37}$$

Proof. Here we will only prove one inequality (C.32). Other properties can be proved

using similar techniques, so we skip the details. From definition, we write

$$\begin{aligned}
||\mathbf{AB}||_{\infty}^{(M)} &= \max_i \sum_{j=1}^{p} || \sum_{k=1}^{p} \mathbf{A}_{ik}\mathbf{B}_{kj}||_F \\
&\leq \max_i \sum_{j=1}^{p} \sum_{k=1}^{p} ||\mathbf{A}_{ik}||_F ||\mathbf{B}_{kj}||_F \\
&= \max_i \sum_{k=1}^{p} |||\mathbf{A}_{ik}||_F| \sum_{j=1}^{p} |\mathbf{B}_{kj}||_F \\
&\leq \max_i \sum_{k=1}^{p} |||\mathbf{A}_{ik}||_F \max_k \sum_{j=1}^{p} ||\mathbf{B}_{kj}||_F \\
&= ||\mathbf{A}||_{\infty}^{(M)} ||\mathbf{B}||_{\infty}^{(M)},
\end{aligned}$$

which completes the proof.

D Further Discussion

D.1 Approximation for Multivaraite Functional Data

One referee was concerned that, for multivariate functional data, the truncation approach through performing FPCA separately for each individual curve does not provide the best M-dimensional approximation. We refer to Chiou et al. (2014) and Happ and Greven (2017) for some recent developments on the Karhunen-Loeve expansion for multivariate functional data with fixed p. However, this multivariate FPCA approach cannot handle high dimensional functional data when p is very large, posing additional challenges to derive the relevant concentration bounds. In contrast, our approach is easy to implement and we are able to derive the relevant concentration bounds.

Under certain regularity conditions, we can prove that our truncation approach indeed can control the bias which approaches zero as $M \to \infty$. Roughly speaking, suppose that for each $j = 1, \ldots, p$, $g_j(t) = g_j^M(t) + \xi_j(t), t \in \mathcal{T}$, with $\|\xi_j\| \to 0$ as $M \to \infty$ and $E\big(g_j^M(t)\big) = E\big(\xi_j(t)\big) = 0$. It follows from the expansion, where $\mathrm{Cov}\big(g_j(s), g_k(t)\big) - \mathrm{Cov}\big(g_j^M(s), g_k^M(t)\big) = \mathrm{Cov}\big(g_j^M(s), \xi_j(t)\big) + \mathrm{Cov}\big(\xi_j(s), g_k^M(t)\big) + \mathrm{Cov}\big(\xi_j(s), \xi_k(t)\big)$, and Cauchy-Schwarz inequality that $\int_{(s,t)\in\mathcal{T}^2} E\big\{\mathrm{Cov}\big(g_j(s), g_k(t)\big) - \mathrm{Cov}\big(g_j^M(s), g_k^M(t)\big)\big\}^2 dsdt \leq 9 \sup_j \|g_j\|^2 \sup_j \|\xi_j\|^2$. In

other words, if $\sup_j \|g_j\|^2 \leq C$ with some positive constant C, the truncated bias can be controlled at the same order as $\sup_j \|\xi_j\|^2$.

D.2 Connection between the Fglasso Approach and (24)

We discuss the connection between our proposed fglasso approach and the alternative method using the inverse correlation matrix discussed in Section 6. Let $\mathbf{S}^M = \mathbf{D}^M \mathbf{R}^M \mathbf{D}^M$, where $\mathbf{D}^M$ is the diagonal matrix of $\mathbf{S}^M$ with its j-th block given by $\mathbf{D}_j^M \in \mathbb{R}^{M\times M}, j = 1, \ldots, p$. We modify the penalty term in (9) and consider maximizing

$$\log\det(\mathbf{\Theta}^M) - \text{trace}(\mathbf{D}^M \mathbf{R}^M \mathbf{D}^M \mathbf{\Theta}^M) - \gamma_n \sum_{j\neq l} \|\mathbf{D}_j^M \mathbf{\Theta}_{jl}^M \mathbf{D}_l^M\|_F, \tag{C.38}$$

over symmetric positive definite matrices $\mathbf{\Theta}^M \in \mathbb{R}^{pM\times pM}$. Let $\mathbf{Q}^M = \mathbf{D}^M \mathbf{\Theta}^M \mathbf{D}^M$, it is clear that the solution to the optimization problem (C.38) is equivalent to (24) in Section 6.

D.3 The Algorithm to Solve (24)

Since the fglasso criterion in (9) and (24) discussed in Section 6 take a similar form, we develop Algorithm 6 to solve the optimization problem in (24) following an analogous procedure described in Section 3.1.

Let $\mathbf{Q}_{-j}, \mathbf{P}_{-j}$ and $\mathbf{R}_{-j}$ respectively be $M(p-1)\times M(p-1)$ sub matrices excluding the jth row and column block of $\mathbf{Q}, \mathbf{P} = \mathbf{Q}^{-1}$ and $\mathbf{R}$, and let $\mathbf{q}_j, \mathbf{p}_j$ and $\mathbf{r}_j$ be $M(p-1)\times M$ matrices representing the jth column block after excluding the jth row block. Finally, let $\mathbf{Q}_{jj}, \mathbf{P}_{jj}$ and $\mathbf{R}_{jj}$ be the (j,j)th $M\times M$ blocks in $\mathbf{Q}, \mathbf{P}$ and $\mathbf{R}$ respectively. Then, for a fixed value of $\mathbf{Q}_{-j}$, (24) can be solved by setting

$$\widehat{\mathbf{Q}}_{jj} = \mathbf{R}_{jj}^{-1} + \widehat{\mathbf{q}}_j^T \mathbf{Q}_{-j}^{-1} \widehat{\mathbf{q}}_j, \tag{C.39}$$

where

$$\widehat{\mathbf{q}}_j = \arg\min_{\mathbf{q}_j} \left\{ \text{trace}(\mathbf{R}_{jj}\mathbf{q}_j^T \mathbf{Q}_{-j}^{-1}\mathbf{q}_j) + 2\text{trace}(\mathbf{r}_j^T \mathbf{q}_j) + 2\gamma_n \sum_{l=1}^{p-1} \|\mathbf{q}_{jl}\|_F \right\}, \tag{C.40}$$

where $\mathbf{q}_{jl}$ represents the lth $M\times M$ block of $\mathbf{q}_j$. The algorithm to solve (24) is summarized in Algorithm 6 below.

Algorithm 6 The Algorithm to Solve (24)

1. Initialize $\widehat{\mathbf{Q}} = \mathbf{I}_{Mp}$ and $\widehat{\mathbf{P}} = \mathbf{I}_{Mp}$.

2. Repeat until convergence for $j = 1, \ldots, p$.

   (a) Compute $\widehat{\mathbf{Q}}_{-j}^{-1} \leftarrow \widehat{\mathbf{P}}_{-j} - \widehat{\mathbf{p}}_j \widehat{\mathbf{P}}_{jj}^{-1} \widehat{\mathbf{p}}_j^T$.

   (b) Solve for $\widehat{\mathbf{q}}_j$ in (C.40) using Algorithm 3 in Section B.1.

   (c) Reconstruct $\widehat{\mathbf{P}}$ using $\widehat{\mathbf{P}}_{jj} = \mathbf{R}_{jj}$, $\widehat{\mathbf{p}}_j = -\mathbf{V}_j \mathbf{R}_{jj}$ and $\widehat{\mathbf{P}}_{-j} = \widehat{\mathbf{Q}}_{-j}^{-1} + \mathbf{V}_j \mathbf{R}_{jj} \mathbf{V}_j^T$, where $\mathbf{V}_j = \widehat{\mathbf{Q}}_{-j}^{-1} \widehat{\mathbf{q}}_j$.

3. Set $\widehat{E} = \left\{ (j,l) : \|\widehat{\mathbf{Q}}_{jl}\|_F \neq 0, (j,l) \in V^2, j \neq l \right\}$.

References

Bosq, D. (2000). *Linear Processes in Function Spaces*, Springer, New York.

Boucheron, S., Lugosi, G. and Massart, P. (2014). *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press.

Chiou, J.-M., Chen, Y.-T. and Yang, Y.-F. (2014). Multivariate functional principal component analysis: a normalization approach. *Statistica Sinica.*, 24, 1571-1596.

Happ, C. and Greven, S. (2017). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, in press.

Ravikumar, P., Wainwright, M., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing l_1-penalized log-determinant deivergence. *Electronic Journal of Statistics.*, 5, 935-980.