Submitted to the Annals of Statistics

FUNCTIONAL GRAPHICAL MODELS*

By Xinghao Qiao[†], Gareth James[†] and Jinchi Lv[†]

University of Southern California[†]

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 p 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 and our theoretical results demonstrate that, with high probability, the fglasso will correctly 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.

1. Introduction. Recently there has been a great deal of interest in constructing graphical models, or networks, from high-dimensional data. A Gaussian graphical model is used to depict the conditional dependence structure among p multivariate Gaussian 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 depict 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 correlated, conditional on the other $p-2$ variables.,

One can show that, for Gaussian data, 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

*Lv's research was supported by NSF CAREER Award DMS-0955316.

MSC 2010 subject classifications: Primary 62F12, 62F30; secondary 62M99

Keywords and phrases: Functional Data, Graphical Model, Functional Principal Component Analysis, Block Sparse Precision Matrix.

concentrated on two main approaches for estimating high dimensional Gaussian graphical models. One method involves a penalized regression approach whereby each variable is regressed on all the remaining variables, thus identifying the locations of non-zero entries in the precision matrix column by column (Meinshausen and Buhlmann, 2006; Cai et al., 2011). 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 in 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. (2014b) and the references therein.

In this paper we are interested in estimating a graphical network in a somewhat more complicated setting. For $t \in \mathcal{T}$, where $\mathcal{T}$ is a closed subset of the real line, let $g_1(t), \ldots, g_p(t)$ represent p Gaussian random functions. Our goal is to construct a *functional graphical model* (FGM) depicting the conditional dependence structure among these p random functions. 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. 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 locations on the subject's scalp. Hence, each observation, or subject, involves $p = 64$ different functions

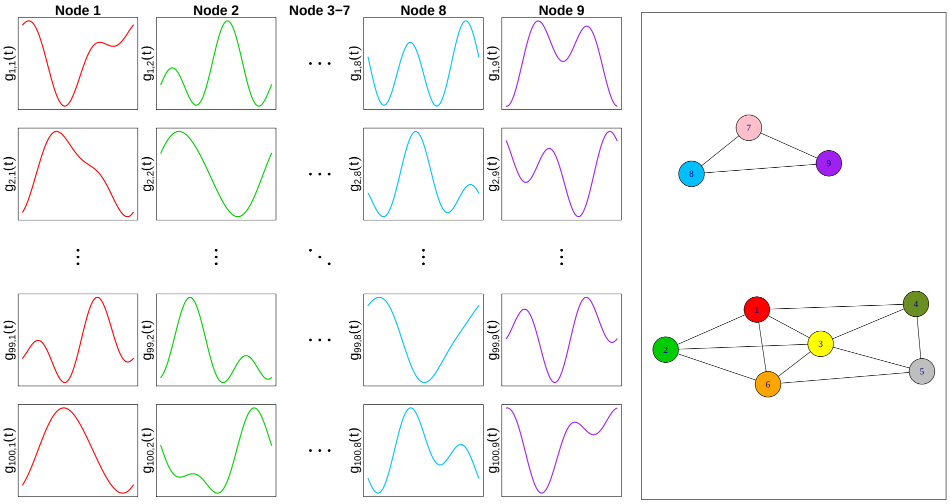

Fig 1. *The illustrative example. Left: The data,* $n = 100$ *observations of* $g_{ij}(t)$ *for* $j = 1, \ldots, 9$ *nodes. Right: the true underlying network structure.*

observed at 256 time points. It is of scientific interest to identify differences in brain activity between the two groups, so we construct FGM's 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, 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, in many instances the functions are not observed at a uniform set of time points and are often not measured at the same set of points. The grid approach would fail in this setting. Second, 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 $s \neq t$. In such a scenario each of the T networks would miss the correlation structure. Some recent research in graphical models considered estimating a time varying graphical model through a nonparametric approach which constructs similar graphs from adjacent time points (Zhou et al., 2010; Kolar and Xing, 2011). However, while this approach does consider dependencies across time it will miss correlations among points that are far apart and also requires a common grid of points.

In this article, 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 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 investigate the structure of the underlying functional network to obtain the rate of convergence for blocks in the precision matrix and expand previous model selection consistency results from the standard setting to the more complicated functional domain.

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 criterion provides a sparse estimate, $\widehat{E}$, for the edge set E. An efficient block coordinate descent algorithm for minimizing the fglasso criterion is presented in Section 3. The fglasso algorithm is extended to handle even larger values of p by applying the partition approach of Witten et al. (2011). Section 4 provides our theoretical results. Here we show that the estimated edge set $\widehat{E}$ is the same as the true edge set E with probability converging to one, even for $p > n$. 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.

We conclude our paper with some possible future directions for this work in Section 6.

2. Methodology.

2.1. *Gaussian Graphical Models.* As discussed in the previous section, the edges depict the conditional dependence structure of the p variables. Specifically, let

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

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 $\boldsymbol{\Sigma}$, 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, $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. Let $G = (V, E)$ denote an undirected graph with vertex set $V = \{1, \cdots, p\}$ 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.

One natural approach for estimating $\boldsymbol{\Theta}$ is to use the maximum likelihood estimator (MLE). Standard calculations show that the Gaussian log likelihood (up to constants) is given by

$$\log\det\boldsymbol{\Theta} - \operatorname{trace}(\mathbf{S}\boldsymbol{\Theta}),$$

which, for $p < n$, is maximized by setting $\widehat{\boldsymbol{\Theta}} = \mathbf{S}^{-1}$, when $\mathbf{S}$, the sample covariance matrix, is nonsingular. However, this estimator has two deficiencies. First, for $n \approx p$ the MLE will provide a poor estimate for $\boldsymbol{\Theta}$ and for $p > n$ the MLE is not even unique. Second, even in lower dimensional settings, where the MLE may be more accurate, none of the elements of $\widehat{\boldsymbol{\Theta}}$ will be exactly zero—a serious limitation given that our ultimate goal is to identify the pairs (j,l) such that $\Theta_{jl} = 0$.

The graphical lasso (Yuan and Lin, 2007) addresses both of these limitations by adding a penalty term to the log likelihood function:

$$\widehat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}}\{\log\det\boldsymbol{\Theta} - \operatorname{trace}(\mathbf{S}\boldsymbol{\Theta}) - \lambda\|\boldsymbol{\Theta}\|_1\}, \tag{2}$$

where $\mathbf{\Theta} \in \mathcal{R}^{p\times p}$ is symmetric positive definite, $\lambda \geq 0$ is a tuning parameter and $\|\mathbf{\Theta}\|_1 = \sum_{jl} |\Theta_{jl}|$. 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, $g_1, \cdots, g_p$, following from a *multivariate Gaussian process* (MGP), belong to an undirected graph $G = (V, E)$ with vertex set $V = \{1, \cdots, p\}$ 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 $(s, t) \in \mathcal{T}^2$, the cross-sectional conditional covariance function is defined by

$$C_{jl}(s,t) = \text{Cov}\left(g_j(s), g_l(t) | g_k(u), k \neq j, l, \ \forall u \in \mathcal{T}\right), \tag{3}$$

which represents the covariance between $g_j(s)$ and $g_l(t)$ conditional on the remaining $p-2$ functions. We consider g_j and g_l to be conditionally independent if $C_{jl}(s,t) = 0$ for all $(s,t) \in \mathcal{T}^2$. Hence our ultimate goal is to estimate the edge set

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

2.3. *Functional Graphical Lasso.* Suppose we observe $\mathbf{g}_i = (g_{i1}, \ldots, g_{ip})^T$, $i = 1, \ldots, n$, and each $g_{ij}(t)$, $t \in \mathcal{T}$ is a realization from a mean zero Gaussian process. The Karhunen-Loève expansion allows us to represent each function in the form

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

where $a_{ijk} \sim N(0, \lambda_{jk})$, 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 $\phi_{j1}(t), \phi_{j2}(t), \ldots$ represent *functional principal component* (FPC) functions and form an infinite dimensional basis representation for $g_{ij}(t)$. If we denote the covariance function by $K_{jj}(s,t) = Cov(g_j(s), g_j(t))$, then the corresponding eigen pairs satisfy

$$\int_{\mathcal{T}} K_{jj}(s,t)\phi_{jk}(t)dt = \lambda_{jk}\phi_{jk}(s), \tag{5}$$

where $\int_{\mathcal{T}} \phi_{jk}(t)^2 dt = 1$ and $\int_{\mathcal{T}} \phi_{jk}(t)\phi_{jm}(t)dt = 0$ for $m < k$.

Let

$$\widetilde{g}_{ij}^M(t) = \sum_{k=1}^{M} a_{ijk}\phi_{jk}(t) = (\mathbf{a}_{ij}^M)^T \boldsymbol{\phi}_j^M(t) \tag{6}$$

represent the M-truncated version of $g_{ij}(t)$, where $\mathbf{a}_{ij}^M = (a_{ij1}, \cdots, a_{ijM})^T$ and $\boldsymbol{\phi}_j^M = (\phi_{j1}, \cdots, \phi_{jM})^T$. Then $\widetilde{g}_{ij}^M(t)$ provides the best M-dimensional approximation to $g_{ij}(t)$ in terms of mean squared error. Analogously to the conditional covariance function (3), we define the M-truncated conditional covariance function by

$$\widetilde{C}_{jl}^M(s,t) = \mathrm{Cov}\left(\widetilde{g}_j^M(s), \widetilde{g}_l^M(t) | \widetilde{g}_k^M(u), k \neq j, l, \ \forall u \in \mathcal{T}\right), \tag{7}$$

and the corresponding truncated edge set as

$$\widetilde{E}^M = \left\{\{j,l\} : \widetilde{C}_{jl}^M(s,t) \neq 0 \,\text{for some}\, s \text{ and } t, (j,l) \in V^2, j \neq l\right\}. \tag{8}$$

The basic intuition to our approach is to develop a method for estimating $\widetilde{E}^M$ which, for large enough M, will provide an accurate estimate for E. Our theoretical results in Section 4 formalize this idea.

To form our estimate for $\widetilde{E}^M$ we first let $\mathbf{a}_i^M = \left((\mathbf{a}_{i1}^M)^T, \ldots, (\mathbf{a}_{ip}^M)^T\right)^T \in \mathcal{R}^{Mp}$ represent the first M FPC scores for the ith set of functions, $i = 1, \ldots, n$. 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 $\widetilde{\boldsymbol{\Sigma}}^M = \left(\widetilde{\boldsymbol{\Theta}}^M\right)^{-1}$.

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

$$\widetilde{E}^M = \left\{\{j,l\} : \|\widetilde{\boldsymbol{\Theta}}_{jl}^M\|_F \neq 0, (j,l) \in V^2, j \neq l\right\}, \tag{9}$$

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

Lemma 1 suggests that the problem of estimating $\widetilde{E}^M$, and hence E, can be reduced to one of accurately estimating the block sparsity structure in $\widetilde{\boldsymbol{\Theta}}^M$.

Thus, we estimate the network structure using the *functional graphical lasso (fglasso)*, which modifies the graphical lasso by incorporating a group lasso penalty to produce a block sparsity structure. The fglasso is defined as the solution to

$$\widehat{\mathbf{\Theta}}^M = \arg\max_{\widetilde{\mathbf{\Theta}}^M} \left\{ \log\det \widetilde{\mathbf{\Theta}}^M - \operatorname{trace}(\widetilde{\mathbf{S}}^M \widetilde{\mathbf{\Theta}}^M) - \lambda \sum_{j \neq l} \|\widetilde{\mathbf{\Theta}}^M_{jl}\|_F \right\}, \tag{10}$$

where $\widetilde{\mathbf{\Theta}}^M \in \mathcal{R}^{Mp \times Mp}$ is symmetric positive definite, $\lambda \geq 0$ is a tuning parameter and $\widetilde{\mathbf{S}}^M$ is the sample covariance matrix of $\mathbf{a}^M$. The penalty in (10) is equivalent to a group lasso penalty (Yuan and Lin, 2006; Friedman et al., 2010) and forces the elements of $\mathbf{\Theta}^M_{jl}$ to either all be zero (a sparse solution) or not all zero (a connected edge between g_j and g_l). Hence, as λ increases $\widehat{\mathbf{\Theta}}^M$ grows sparser in a blockwise fashion. Our final estimated edge set is then defined as

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

We calculate $\widetilde{\mathbf{S}}^M$ directly from $\mathbf{a}^M$. Although $\mathbf{a}^M$ is not directly observed there are a number of standard methods for computing the functional principal component scores from a set of functions. See the Supplementary Material for further details. Note $\widehat{\mathbf{\Theta}}$, $\widetilde{\mathbf{\Theta}}$, $\widetilde{\mathbf{S}}$, $\mathbf{a}_j$ and $\boldsymbol{\phi}_j$, $j = 1 \cdots, p$, all depend on M, but for simplicity of notation we will omit the corresponding superscripts where the context is clear.

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., 2014b).

Let $\widetilde{\mathbf{\Theta}}_{-j}, \widetilde{\mathbf{\Sigma}}_{-j}$ and $\widetilde{\mathbf{S}}_{-j}$ respectively be $M(p-1) \times M(p-1)$ sub matrices excluding the jth row and column block of $\widetilde{\mathbf{\Theta}}$, $\widetilde{\mathbf{\Sigma}}$ and $\widetilde{\mathbf{S}}$, 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 $\mathbf{\Theta}_{jj}, \mathbf{\Sigma}_{jj}$ and $\mathbf{S}_{jj}$ be the (j, j)th $M \times M$ blocks in $\widetilde{\mathbf{\Theta}}, \widetilde{\mathbf{\Sigma}}$ and $\widetilde{\mathbf{S}}$ respectively. So, for instance for $j = 1$, $\widetilde{\mathbf{\Theta}} = \begin{pmatrix} \mathbf{\Theta}_{11} & \mathbf{w}_1^T \\ \mathbf{w}_1 & \widetilde{\mathbf{\Theta}}_{-1} \end{pmatrix}$. Then, for a fixed value of $\widetilde{\mathbf{\Theta}}_{-j}$, standard calculations show that (10) is minimized by setting

$$\widehat{\mathbf{\Theta}}_{jj} = \mathbf{S}_{jj}^{-1} + \widehat{\mathbf{w}}_j^T \widetilde{\mathbf{\Theta}}_{-j}^{-1} \widehat{\mathbf{w}}_j, \tag{11}$$

where

(12)
$$\widehat{\mathbf{w}}_j = \arg\min_{\mathbf{w}_j} \left\{ \mathrm{trace}(\mathbf{S}_{jj}\mathbf{w}_j^T\widetilde{\boldsymbol{\Theta}}_{-j}^{-1}\mathbf{w}_j) + 2\mathrm{trace}(\mathbf{s}_j^T\mathbf{w}_j) + 2\lambda\sum_{l=1}^{p-1}\|\mathbf{w}_{jl}\|_F \right\},$$

and $\mathbf{w}_{jl}$ represents the lth $M \times M$ block of $\mathbf{w}_j$. Computing (12) can be achieved using some matrix calculus and standard non-linear equation software. Details are provided in Appendix A.1.

This suggests a block coordinate descent algorithm where one iterates through j repeatedly computing (12) until convergence. In fact by checking the conditions of Theorem 4.1 in Tseng (2001) it is easy to verify that iteratively minimizing (12) over $\mathbf{w}_j$ and updating $\boldsymbol{\Theta}_{jj}$ by (11) for $j = 1, \cdots, p$ provides a convergent solution for globally maximizing the fglasso criterion. The main potential difficulty with this approach is that $\widetilde{\boldsymbol{\Theta}}_{-j}^{-1}$ 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 the Appendix A.2.

Algorithm 1 Functional Graphical Lasso Algorithm

```
1. Initialize $\widehat{\boldsymbol{\Theta}} = \mathbf{I}$ and $\widehat{\boldsymbol{\Sigma}} = \mathbf{I}$.
2. Repeat until convergence for $j = 1, \cdots, p$.
    (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 (12) using Algorithm 3 in Appendix A.1.
    (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., 2014b). Since the resulting glasso problems involve many fewer nodes each network 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 (10) takes a block diagonal form, i.e.* $\widetilde{\mathbf{\Theta}} = diag\left(\widetilde{\mathbf{\Theta}}^{(1)}, \cdots, \widetilde{\mathbf{\Theta}}^{(K)}\right)$*, then (10) can be solved by separately maximizing K smaller fglasso problems*

$$\widehat{\mathbf{\Theta}}^{(k)} = \arg\max_{\widetilde{\mathbf{\Theta}}^{(k)}} \left\{ \log\det \widetilde{\mathbf{\Theta}}^{(k)} - trace(\widetilde{\mathbf{S}}^{(k)} \widetilde{\mathbf{\Theta}}^{(k)}) - \lambda \sum_{j \neq l} \|\widetilde{\mathbf{\Theta}}_{jl}^{(k)}\|_F \right\}, \tag{13}$$

for $k = 1, \cdots, K$, where $\widetilde{\mathbf{S}}^{(k)}$ is the submatrix of $\widetilde{\mathbf{S}}$ corresponding to $\widetilde{\mathbf{\Theta}}^{(k)}$.

Proposition 2. *Without loss of generality, let $G_1, \cdots, G_K$ 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 $G_1, \cdots, G_K$ is that $\|\widetilde{\mathbf{S}}_{ii'}\|_F \le \lambda$ for all $i \in G_k$, $i' \in G_{k'}$, $k \neq k'$.*

The proofs of these two propositions are provided in the supplementary material. Propositions 1 and 2 suggest first performing a screening procedure on $\widetilde{\mathbf{S}}$ to identify K distinct graphs and then solving the resulting K fglasso problems. These steps are summarized in Algorithm 2.

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 $A_{ii'} = 1_{\|\widetilde{\mathbf{S}}_{ii'}\|_F > \lambda}$.
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 $k$th connected component, $k = 1, \cdots, K$.
3. For $k = 1, \cdots, K$, 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, diag$\left(\widehat{\mathbf{\Theta}}^{(1)}, \cdots, \widehat{\mathbf{\Theta}}^{(K)}\right)$.
```

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 $k = 1, \cdots, K$, hence the total computational cost for Algorithm 2 is $O\left(p^2 + \sum_{k=1}^K |G_k|^3\right)$. Algorithm 2 significantly reduces the computational cost, if $|G_1|, \cdots, |G_K|$ are much smaller than p, which is the case when the tuning parameter, λ, 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.

4. Theory.

We now investigate the sampling properties of the fglasso. To this end, several regularity and smoothness conditions are needed to facilitate our technical analysis.

CONDITION 1. *For some $M = c_1 n^{\alpha}$ with constants $c_1 > 0$ and $\alpha \in [0, 1)$, $\sum_{k=M+1}^{\infty} \lambda_{jk} = O(n^{-\beta})$ for some constant $\beta > 0$ and $\sup_{s \in \mathcal{T}} \max_{k \geq M+1} |\phi_{jk}(s)| = O(1)$ holds uniformly in $j \in V$.*

Condition 1 is a basic assumption in the functional setting. The parameter α controls the number of selected FPC functions and the parameter β determines the decay rate of any decreasing sequence $\lambda_{j1} \geq \lambda_{j2} \geq \cdots \geq 0$ for $j = 1, \cdots, p$. Larger values of α and β yield a larger M and a faster decay rate, respectively.

Before stating the next few conditions, we introduce some necessary notation. Let $\mathbf{g} = (g_1, \cdots, g_p)^T$ be a collection of random processes whose jth component g_j is a continuous function in $L^2(\mathcal{T}_j)$, where $\mathcal{T}_j$ is a closed subset of the real line. Then the domain for $\mathbf{g}$ is $\mathcal{T} = \bigcup_{j=1}^{p} \mathcal{T}_j$. Denote by $\mathcal{K} = \{K_{jl} : \mathcal{T}_j \times \mathcal{T}_l \to \mathcal{R}, (j, l) \in V^2\}$ a collection of covariance kernel functions with $K_{jl}(s, t) = \text{Cov}\,(g_j(s), g_l(t))$, $(s, t) \in \mathcal{T}_j \times \mathcal{T}_l$, and $\mathbf{g} \sim \text{MGP}(\mathbf{0}, \mathcal{K})$ a MGP, indexed by the vertex set V, with mean $\mathbf{0}$ and covariance $\mathcal{K}$. In general, let $\mathcal{A} = \{A_{jl} : \mathcal{T}_j \times \mathcal{T}_l \to \mathcal{R}, (j, l) \in V^2\}$ be a collection of functions $A_{jl}(s, t)$, $(s, t) \in \mathcal{T}_j \times \mathcal{T}_l$, and $\mathcal{X} = \{x_j : \mathcal{T}_j \to \mathcal{R}, j \in V\}$ a collection of functions $x_j(s), s \in \mathcal{T}_j$. The sets $\mathcal{A}$ and $\mathcal{X}$ can be regarded as the functional counterparts of matrices and vectors, respectively; see Appendix B.2 for more details about the corresponding operations and notation.

For any subsets W, W' of V, denote by $\mathbf{g}_W$ an MGP indexed by W, and $\mathcal{K}_{WW'}$ a subset of covariance with columns and rows in W and W', respectively. The inverse of covariance $\mathcal{K}_{WW}$ is denoted as $\mathcal{K}_{WW}^{-1}$. We define the corresponding M-truncated versions of $\mathbf{g}$, $\mathcal{K}$, $\mathcal{K}_{WW'}$, and $\mathcal{K}_{WW}^{-1}$, by $\widetilde{\mathbf{g}}$, $\widetilde{\mathcal{K}}$, $\widetilde{\mathcal{K}}_{WW'}$, and $\widetilde{\mathcal{K}}_{WW}^{-1}$, respectively. Here, for notational simplicity we drop the corresponding superscripts in the M-truncated terms.

CONDITION 2. *For each $(j, l) \in V^2$ with $j \neq l$, there exist unique inverse covariances $\mathcal{K}_{UU}^{-1}$ and $\widetilde{\mathcal{K}}_{UU}^{-1}$ for covariances $\mathcal{K}_{UU}$ and $\widetilde{\mathcal{K}}_{UU}$ of $\mathbf{g}_U$ and $\widetilde{\mathbf{g}}_U$ with $U = V \backslash \{j, l\}$ such that all kernel functions in $\mathcal{K}$, $\widetilde{\mathcal{K}}$, $\mathcal{K}_{UU}^{-1}$, and $\widetilde{\mathcal{K}}_{UU}^{-1}$ are square integrable, $\lambda_{\min}(\mathcal{K}_{UU})$ and $\lambda_{\min}(\widetilde{\mathcal{K}}_{UU}) \geq c_2$ for some constant*

$c_2 > 0$, *and* $\sup_{s\in\mathcal{T}_j}||\mathcal{K}_{jU}(s)||_{\max} = O(1)$ *and* $\sup_{t\in\mathcal{T}_l}||\mathcal{K}_{Ul}(t)||_{\max} = O(1)$ *uniformly in* (j,l), *where* $\lambda(\cdot)$ *and* $||\cdot||_{\max}$ *are defined in Appendix B.2.*

CONDITION 3. *For each* $(j,l) \in V^2$ *with* $j \neq l$, *suppose* $\|\mathbf{\Phi}_{jl}\|_F = O(1)$ *and* $\lambda_{\min}(\mathbf{D}_{jj})$, $\lambda_{\min}(\mathbf{D}_{ll}) \geq c_3$ *for some positive constant* c_3, *hold uniformly in* (j,l), *where* $\mathbf{\Phi}_{jl} = \int_{\mathcal{T}_j}\int_{\mathcal{T}_l}\boldsymbol{\phi}_j(s)\boldsymbol{\phi}_l(t)^T dsdt$ *and* $\mathbf{D}_{jj} = Var(\mathbf{a}_j) - Cov(\mathbf{a}_j, \mathbf{a}_U)\ Var(\mathbf{a}_U)^{-1}Cov(\mathbf{a}_U, \mathbf{a}_j)$.

CONDITION 4. *It holds that* $\lambda_{\max}(\mathbf{D}_{jj})$ *and* $\lambda_{\max}(\mathbf{D}_{ll}) \leq c_4$, *and*

$$\min_{(j,l)\in E}\inf_{(s,t)\in\mathcal{T}_j\times\mathcal{T}_l}|C_{jl}(s,t)| \geq \frac{c_5p^2}{n^{\beta/2}} + \frac{c_6p^2}{n^{(\beta-2\alpha)/2}}, \tag{14}$$

uniformly in $(j,l) \in E$ *for some constants* c_4, c_5, *and* $c_6 > 0$.

Condition 2 is on the eigen-structure and smoothness of the functional network, while Condition 3 is about the FPC functions and eigen-structure of FPC scores. Condition 4 is an assumption on the minimum signal strength for successful graph recovery. These conditions on the underlying functional network are crucial for obtaining the rate of convergence of $||\widetilde{\mathbf{\Theta}}_{jl}||_F$ for $(j,l) \in S^c$, the complement of $S = E \cup \{(1,1),\cdots,(p,p)\}$ in V^2, and the equivalence between the truncated and true edge sets. See Lemmas 3 and 4 in Appendix B.2 and B.3.

More specifically, Condition 2 restricts the eigen-structure and puts some smoothness requirement on the functional network such that the truncated conditional covariance function (7) converges uniformly to the true one (3) at a certain rate as M goes to infinity. In Condition 3, the bound condition on the cross-integrated FPC functions ensures the conditional cross uncertainties of $\widetilde{\mathbf{g}}$ can be captured by the FPC scores $\mathbf{a}$. The restriction on the eigen-structure of FPC scores provides a sufficient condition to obtain the uniform convergence rate of $||\widetilde{\mathbf{\Theta}}_{jl}||_F$ from that of $||\text{Cov}(\mathbf{a}_j, \mathbf{a}_l|\mathbf{a}_U)||_F$ in $(j,l) \in S^c$. Moreover, Condition 4 requires that the minimum magnitude of the conditional covariance function on the true edge set be bounded from below to ensure that the truncated edge set correctly retains all true edges.

We next introduce an additional irrepresentable-type assumption for deriving the model selection consistency of fglasso, that is, the exact truncated functional graph recovery with asymptotic probability one. Denote by $\widetilde{\mathbf{\Gamma}} = \widetilde{\mathbf{\Theta}}^{-1} \otimes \widetilde{\mathbf{\Theta}}^{-1} \in \mathcal{R}^{(Mp)^2\times(Mp)^2}$ with $\otimes$ the Kronecker product, and $\widetilde{\mathbf{\Gamma}}_{WW'}$ the $M^2|W| \times M^2|W'|$ submatrix of $\widetilde{\mathbf{\Gamma}}$ with row and column blocks in W and W', respectively, for any subsets W, W' of V^2. For any block matrix $\mathbf{B} = (\mathbf{B}_{ij})$

with $\mathbf{B}_{ij} \in \mathcal{R}^{M^2 \times M^2}, 1 \leq i, j \leq p$, define $||\mathbf{B}||_\infty^{(M^2)} = \max_i \sum_{j=1}^p ||\mathbf{B}_{ij}||_F$ as the M^2-block version of matrix ∞-norm.

CONDITION 5. *For $j \in V$ and $m \in \{1, \cdots, M\}$, there exist some constants $c_8 > 0$, $\gamma \in (0, 1 - c_8 p^2 n^{(\alpha-\beta)/2}]$, $\eta > 0$, and $\tau > 2$ such that*

$$||\widetilde{\mathbf{\Gamma}}_{S^cS}(\widetilde{\mathbf{\Gamma}}_{SS})^{-1}||_\infty^{(M^2)} \leq 1 - \gamma - c_8 p^2 n^{(\alpha-\beta)/2}, \tag{15}$$

$$||\widetilde{\mathbf{\Gamma}}_{S^cS}(\widetilde{\mathbf{\Gamma}}_{SS})^{-1}\widetilde{\mathbf{\Gamma}}_{SS^c} - \widetilde{\mathbf{\Gamma}}_{S^cS^c}||_\infty^{(M^2)} \leq \left[\frac{\gamma n^{(\beta-\alpha)/2}}{2c_8 p^2} + 1\right]\lambda, \tag{16}$$

$$||\widetilde{\mathbf{\Gamma}}_{SS^c}||_\infty^{(M^2)} \leq 40\sqrt{2}\eta c_1 n^\alpha \max_{j,m} \lambda_{jm} \sqrt{\frac{\tau \log(c_1 n^\alpha p) + \log 4}{n}}. \tag{17}$$

It is worth noting that $\widetilde{\mathbf{\Gamma}}$ is the Hessian of $-\log\det(\mathbf{\Theta})$ evaluated at $\mathbf{\Theta} = \widetilde{\mathbf{\Theta}}$. Hence the entry $\widetilde{\mathbf{\Gamma}}_{(j,j')(l,l')}^{(k,k')(m,m')}$ is equal to the partial derivative $\frac{\partial(-\log\det(\mathbf{\Theta}))}{\partial\mathbf{\Theta}_{jj'}^{kk'}\partial\mathbf{\Theta}_{ll'}^{mm'}}$ evaluated at $\mathbf{\Theta} = \widetilde{\mathbf{\Theta}}$, where $\mathbf{\Theta}_{jj'}^{kk'}$ is the (k,k')th entry of the $M \times M$ submatrix $\mathbf{\Theta}_{jj'}$, $1 \leq j, j', l, l' \leq p$, $1 \leq k, k', m, m' \leq M$. Since $\mathbf{a}$ is multivariate Gaussian, some standard calculations show that $\widetilde{\mathbf{\Gamma}}_{(j,j')(l,l')}^{(k,k')(m,m')} = \mathrm{Cov}\left(a_{jk}a_{j'k'}, a_{lm}a_{l'm'}\right)$. The Hessian of the negative log-determinant in the scalar case of $M = 1$ was studied in Ravikumar et al. (2011). We extend their work by viewing $\widetilde{\mathbf{\Gamma}}$, the Fisher information of the model, as an edge-based M^2-block covariance matrix instead of the node-based covariance matrix $\widetilde{\mathbf{\Theta}}$.

For each $(j, j') \in V^2$ with $j \neq j'$, denote by $\mathbf{b}_{jj'} = \mathbf{a}_j \otimes \mathbf{a}_{j'} \in \mathcal{R}^{M^2}$ the edge-based vector, where $\mathbf{a}_j$, $\mathbf{a}_l$ are the node-based vectors. Then we have $\widetilde{\mathbf{\Gamma}}_{(j,j')(l,l')} = E(\mathbf{b}_{jj'}\mathbf{b}_{ll'}^T)$, 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 inequality (15) is equivalent to $||E(\mathbf{b}_{S^c}\mathbf{b}_S^T)E(\mathbf{b}_S\mathbf{b}_S^T)^{-1}||_\infty^{(M^2)} \leq 1 - \gamma - c_8 p^2 n^{(\alpha-\beta)/2}$, 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. Moreover, since $\mathbf{b}_S$ and $\mathbf{b}_{S^c}$ are multivariate Gaussian with mean zero, then inequality (16) is equivalent to $||E(\mathbf{b}_{S^c}\mathbf{b}_S^T)E(\mathbf{b}_S\mathbf{b}_S^T)^{-1}E(\mathbf{b}_S\mathbf{b}_{S^c}^T) - E(\mathbf{b}_{S^c}\mathbf{b}_{S^c}^T)||_\infty^{(M^2)} = ||\mathrm{Var}(\mathbf{b}_{S^c}|\mathbf{b}_S)||_\infty^{(M^2)} \leq [\frac{\gamma n^{(\beta-\alpha)/2}}{2c_8 p^2} + 1]\lambda$, which bounds the variance of $\mathbf{b}_{jj'}$'s outside S conditional on those within S. Finally, (17) imposes an upper bound on $||\widetilde{\mathbf{\Gamma}}_{SS^c}||_\infty^{(M^2)}$, which constrains the covariance between $\mathbf{b}_S$ and $\mathbf{b}_{S^c}$

to ensure a bounded error for the estimate $\widehat{\boldsymbol{\Theta}}$. In particular, for the scalar case of $M = 1, \alpha = 0, \beta = \infty$, (16), (17) are not required anymore and (15) reduces to the irrepresentable condition in Ravikumar et al. (2011), since η can be relaxed to be ∞ in such a case.

We are now ready to present the main theorem on the model selection consistency of our proposed approach fglasso for estimating a FGM with different time domains of functions. Denote by $\kappa_{\widetilde{\Gamma}} = ||(\widetilde{\boldsymbol{\Gamma}}_{SS})^{-1}||_{\infty}^{(M^2)}$, $\widetilde{\zeta} = ||\widetilde{\boldsymbol{\Gamma}}_{SS^c}||_{\infty}^{(M^2)}$, $\kappa_{\widetilde{\Sigma}} = ||\widetilde{\boldsymbol{\Sigma}}||_{\infty}^{(M)}$, and

$$d = \max_{j \in V} \left| \{ l \in V : C_{jl}(s,t) \neq 0 \ \text{ for some } s \text{ and } t \} \right|, \tag{18}$$

the maximum degree of the graph in the underlying FGM.

Theorem 1. *Assume that Conditions 1–5 hold. For $j \in V$ and $m \in \{1, \cdots, M\}$, let $\widehat{\boldsymbol{\Theta}}$ be the unique solution to the fglasso problem (10) with regularization parameter $\lambda = (320\sqrt{2}c_1 n^{\alpha}/\gamma)\max\limits_{j,m}\lambda_{jm}\sqrt{\frac{\tau \log(c_1 n^{\alpha} p) + \log 4}{n}}$. If the sample size n satisfies the lower bound $n > c_7 p^{4/\beta}$ and*

(19)

$$n^{1-2\alpha} > \max\{C_1 d^2, C_2 \Theta_{\min}^{-2}\} \left[1 + \frac{8}{\gamma} + \frac{\eta c_8 p^2}{n^{(\beta-\alpha)/2}}\right]^2 [\tau\alpha \log n + \tau \log p + \log(4c_1^{\tau})],$$

with some constant $c_7 > 0$, $\Theta_{\min} = \min\limits_{(j,l)\in E} ||\widetilde{\boldsymbol{\Theta}}_{jl}||_F$, $C_2 = \{80\sqrt{2}c_1 \max\limits_{j,m}\lambda_{jm}\kappa_{\widetilde{\Gamma}}\}^2$, and

$$\begin{aligned} C_1 \quad = \quad & \Bigg\{ 240\sqrt{2}c_1 \max_{j,m}\lambda_{jm} \max\Bigg\{ \frac{\kappa_{\widetilde{\Sigma}}\kappa_{\widetilde{\Gamma}}}{1 - 3(p-d)c_8 p^2 n^{(\alpha-\beta)/2}\kappa_{\widetilde{\Sigma}}}, \\ & \frac{\kappa_{\widetilde{\Sigma}}^3 \kappa_{\widetilde{\Gamma}}^2 [1 + 8\gamma^{-1} + \eta c_8 p^2 n^{(\alpha-\beta)/2}]}{1 - 3(p-d)c_8 p^2 n^{(\alpha-\beta)/2}\kappa_{\widetilde{\Sigma}}^3 \kappa_{\widetilde{\Gamma}}[1 + 8\gamma^{-1} + \eta c_8 p^2 n^{(\alpha-\beta)/2}]} \Bigg\} \Bigg\}^2, \end{aligned}$$

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.

For any $(j,l) \in S^c$, Lemma 3 implies $||\widetilde{\boldsymbol{\Theta}}_{jl}||_F = c_8 p^2/n^{(\beta-\alpha)/2}$. For a larger truncation error $c_9 = c_8 p^2/n^{(\beta-\alpha)/2}$, (15) and (16) become more stringent. Since in view of (15) c_9 is bounded from above by $1-\gamma$, we see that a sample size $n \gtrsim p^{\frac{4}{\beta-\alpha}}$ is sufficient to satisfy all the technical assumptions, where $\gtrsim$ means asymptotic lower bound. In such a case, the sample size condition (19) can be simplified as $n^{1-2\alpha} > \max\{C_1 d^2, C_2 \Theta_{min}^{-2}\}(1 + 8\gamma^{-1} +$

$\eta c_9)^2[\tau\alpha\log n + \tau\log p + \log(4c_1^\tau)]$ with
$C_1 = \left\{240\sqrt{2}c_1\max_{j,m}\lambda_{jm}\max\{\frac{\kappa_{\widetilde{\Sigma}}\kappa_{\widetilde{\Gamma}}}{1-3(p-d)c_9\kappa_{\widetilde{\Sigma}}}, \frac{\kappa_{\widetilde{\Sigma}}^3\kappa_{\widetilde{\Gamma}}^2(1+8\gamma^{-1}+\eta c_8)}{1-3(p-d)c_9\kappa_{\widetilde{\Sigma}}^3\kappa_{\widetilde{\Gamma}}(1+8\gamma^{-1}+\eta c_9)}\}\right\}^2$. Let us further assume that $\kappa_{\widetilde{\Sigma}}$, $\kappa_{\widetilde{\Gamma}}$, $\widetilde{\zeta}$, γ remain constant with respect to n, p, d, and $n^\alpha = o(p)$. Then a sample size

$$n \gtrsim \left[(d^2 + \Theta_{\text{min}}^{-2})\tau\log p\right]^{\frac{1}{1-2\alpha}} \tag{20}$$

guarantees the model selection consistency of 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 scalar case of $M = 1, \alpha = 0, \beta = \infty$, the sample size condition (19) reduces to $n \gtrsim (d^2 + \Theta_{\text{min}}^{-2})\tau\log p$, which is consistent with the results in Ravikumar et al. (2011).

It is easy to see that a sample size $n \gtrsim \left(d^2\tau\log p\right)^{1/(1-2\alpha)}$ is sufficient for ensuring model selection consistency as long as the minimum Frobenius norm within the true edge set $\Theta_{\text{min}} \gtrsim \sqrt{\frac{\log p}{n^{1-2\alpha}}}$. In functional data analysis, one usually considers using the first several FPC scores in the case when λ_{jk}'s decay fast to zero so that a small α and a large β can be appropriate. When $\beta - \alpha > 4$ and the maximum node degree $d = o\left(\sqrt{\frac{p^{1-2\alpha}}{\log p}}\right)$, the 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 $g_{ij}(t) = \mathbf{s}(t)^T\boldsymbol{\eta}_{ij}$, where $\mathbf{s}(t)$ was a 5-dimensional orthogonal Fourier basis function, and $\boldsymbol{\eta}_{ij} \in R^5$ was a random Gaussian vector with mean zero. Hence, $\boldsymbol{\eta}_i = (\boldsymbol{\eta}_{i1}^T, \cdots, \boldsymbol{\eta}_{ip}^T)^T \in R^{5p}$ followed from a multivariate Gaussian distribution with covariance $\mathbf{\Sigma} = \mathbf{\Theta}^{-1}$. Different block sparse patterns in the precision matrix, $\mathbf{\Theta}$, correspond to different conditional dependence structures. We considered two general structures.

- Model 1: An AR(2) model with $\mathbf{\Theta}_{jj} = \mathbf{I}$, $\mathbf{\Theta}_{j,j-1} = \mathbf{\Theta}_{j-1,j} = 0.4\mathbf{I}$, $\mathbf{\Theta}_{j,j-2} = \mathbf{\Theta}_{j-2,j} = 0.2\mathbf{I}$ for $j = 1, \cdots, p$, and $\mathbf{\Theta}$ equal to zero 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 $\mathbf{\Theta}$ came from Model 1 with $p = 10$, indicating every alternating block of 10 nodes are connected by an AR(2) model. For $j = 11, \ldots, 20, 31, \ldots, 40, \ldots$, we only let $\mathbf{\Theta}_{jj} = \mathbf{I}$, so the remaining nodes are fully isolated.

In both settings, we generated $n = 100$ samples of $\boldsymbol{\eta}_i$ 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 100 equally spaced time points, $0 = t_1, \cdots, t_{100} = 1$.

To implement the fglasso we must compute $\mathbf{a}_{ij}$, the first M functional 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 K-dimensional B-spline basis (rather than using the Fourier basis which was used to generate the data) and then compute $\mathbf{a}_{ij}$ from the basis coefficients. We used 5-fold cross-validation to choose both K and M. In particular, for a given candidate pair $\{K, M\}$, we removed 1/5th of the observed time points for each $g_{ij}(t)$ as a validation data set, applied FPCA on the remaining data, calculated the squared error between $h_{ij}(t_k)$ and the fitted values, $\widehat{g}_{ij}(t)$, on the validation set, and repeated 5 times to compute the cross-validated squared error. We computed the cross-validated error over a grid of $M \leq K$ values and chose the pair with the lowest error. Typically, $K = 5, 6$ or 7 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 FGL; 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 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, K, was still selected by 5-fold cross validation after setting $M = 1$.

For each method and tuning parameter, λ, we calculated the true positive rate (TPR$_\lambda$) and false positive rate (FPR$_\lambda$), in terms of network edges correctly identified. These quantities are defined by

$$TPR_\lambda = \frac{TP}{TP + FN} \quad \text{and} \quad FPR_\lambda = \frac{FP}{FP + TN},$$

where TP and TN respectively stand for true positives/negatives, and respectively FP and FN represent false positives/negatives. Plotting TPR$_\lambda$

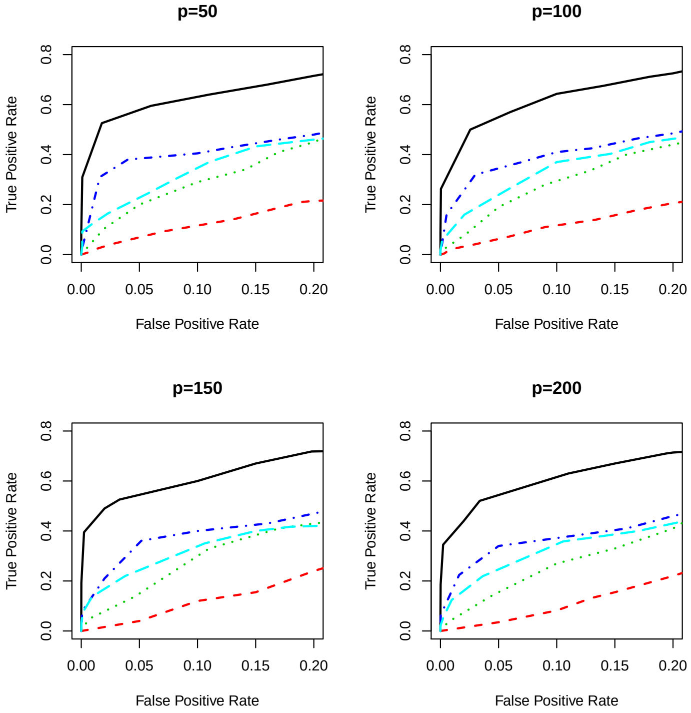

Fig 2. *Model 1 for p=50, 100, 150 and 200: Comparison of median estimated ROC curves over 100 simulation runs with false positive rates up to 20%. fglasso (black solid), ALL (red dashed), NEVER (green dotted), HALF (blue dash dotted) and PCA (cyan long dashed).*

versus FPR_λ over a fine grid of values of λ produces a ROC curve, with curves close to the top left corner indicating a method that is performing well.

We consider different settings with $p = 50$, 100, 150 and 200, and ran each simulation 100 times. Figures 2 and 3 respectively plot the median best ROC curves for each of the five comparison methods of Models 1 and 2. We restrict to consider networks with $FPR \leq 20\%$ since this already requires estimating thousands of edges. Beyond this point the networks become so dense that the computational cost becomes prohibitive. The fglasso (black curve) clearly provides the best overall performance in estimating the functional network. Table 1 provides the area under the ROC curves (average over the 100 simulation runs) along with standard errors. Larger numbers indicate superior estimates of the true network structure. Again we see that the fglasso provides highly significant improvements in accurary

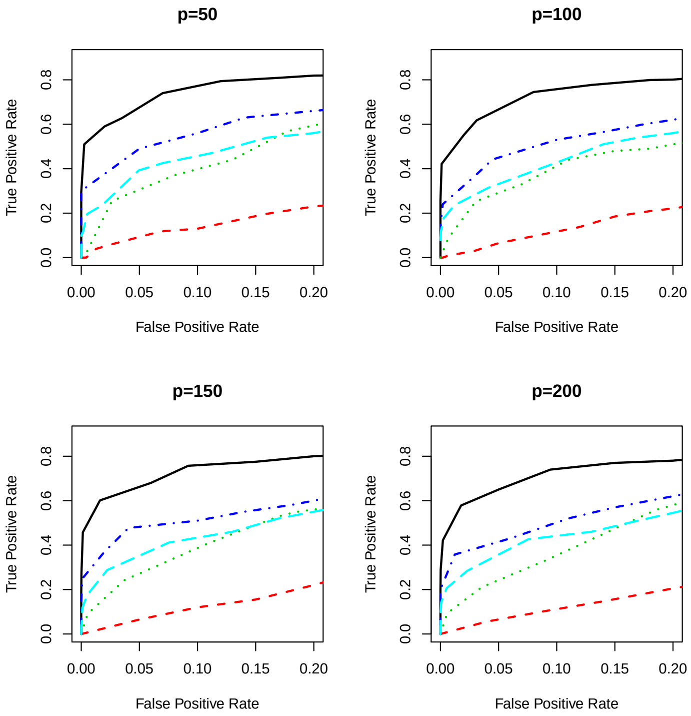

Fig 3. *Model 2 for p=50, 100, 150 and 200: Comparison of median estimated ROC curves over 100 simulation runs with false positive rates up to 20%. fglasso (black solid), ALL (red dashed), NEVER (green dotted), HALF (blue dash dotted) and PCA (cyan long dashed).*

over the competing methods and HALF performs the best among the other approaches. We also consider examining the full ROC curves in the lower dimensional setting with $p = 50$ and $p = 100$. The relative performance of various methods are unchanged, the details of which are provided in the Supplementary Material.

5.2. *EEG Data.* We test the performance of fglasso on the EEG data set from an alcoholism study. 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) discussed the data collection process in detail. Li et al. (2010), Zhou and Li (2014),

	fglasso	ALL	NEVER	HALF	PCA
p	Model 1				
50	1.224(0.010)	0.228(0.003)	0.546(0.013)	0.781(0.015)	0.651(0.006)
100	1.208(0.009)	0.221(0.003)	0.543(0.013)	0.747(0.015)	0.641(0.007)
150	1.181(0.009)	0.218(0.003)	0.540(0.014)	0.721(0.014)	0.625(0.007)
200	1.168(0.008)	0.214(0.002)	0.535(0.012)	0.703(0.014)	0.602(0.007)
p	Model 2				
50	1.485(0.011)	0.266(0.003)	0.769(0.019)	1.031(0.020)	0.858(0.007)
100	1.458(0.010)	0.234(0.003)	0.753(0.018)	0.995(0.016)	0.831(0.007)
150	1.422(0.010)	0.222(0.003)	0.736(0.019)	0.992(0.016)	0.826(0.007)
200	1.398(0.008)	0.216(0.003)	0.702(0.017)	0.970(0.018)	0.824(0.008)

TABLE 1

The mean area ($\times 10^{-1}$) under the ROC curves with false positive rates up to 20%. Standard errors are shown in parentheses.

and Hung and Wang (2013) analyzed 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. (2014a). 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. (2014a) discussed connectivity and asymmetry of 13 electrodes selected from 5 different regions. However, both sets of 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 $p = 64$ edges/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, K, 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 λ to yield 5% sparsity level,

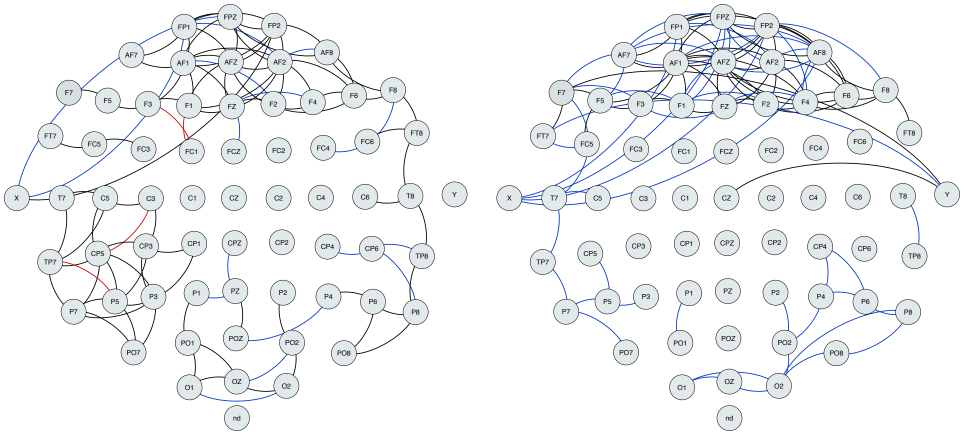

Fig 4. *Left graph plots the estimated network for the alcoholic group and right graph plots the estimated network for the control group. Black lines denote edges identified by both fglasso and bootstrapped fglasso, blue lines denote edges identified by fglasso but not selected by the bootstrapped fglasso and red lines denote edges identified by the bootstrapped fglasso but missed by the fglasso.*

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 4 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. As one would expect, estimates for the control group, with a smaller sample size, entail a larger degree of variability. 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 temporal region of the alcoholic group includes more connected edges. Third, electrodes from other regions of the scalp tend to be only sparsely connected.

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 5 plots the edges only identified by either the alcoholic group or the control group. We observe that edges in the left temporal region were identified by the alcoholic group but missed by the control group, while some edges in the frontal region were identified by the control group but missed by the alcoholic group. Both

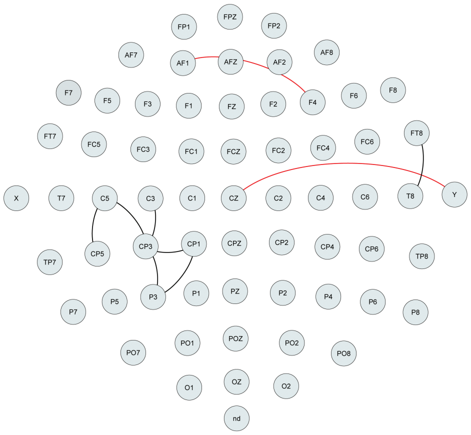

FIG 5. *Black lines denote edges identified only by the alcoholic group and red lines denote edges identified only by the control group.*

findings provide confirmation for our informal observations from Figure 4.

6. Conclusions.

The FGM extends Gaussian graphical models to describe the conditional dependence structure of functional data, while the fglasso provides an approach for estimating such networks. The fglasso enjoys several advantages. First, we are able to estimate a single network for functional data, providing a more interpretable solution. Second, we develop an efficient algorithm to maximize the fglasso criterion and a partition rule that can be applied to accelerate the algorithm even in high dimensional settings. Third, our theoretical results demonstrate the equivalence between the estimated and true edge set with a high probability, and the empirical analysis demonstrates strong finite sample performance.

There are several possible extensions for future research. The first involves constructing a graphical model for sparsely and irregularly observed functional data with measurement error. This extension could be implemented by performing FPCA on sparsely sampled data using either a mixed effects model (James et al., 2000) or a local smoother method (Yao et al., 2005), and then applying the fglasso on the estimated principal component scores. The second possible extension concerns jointly estimating multiple functional graphical models. For scalar data, Danaher et al. (2014) propose a joint graphical lasso to estimate Q different graphical models where there is an assumption that they have some structure in common. The joint graphical lasso attempts to borrow strength across all groups to estimate the connections that are common while still allowing for differences among the groups.

In our setting the joint functional graphical lasso would correspond to finding $\widetilde{\mathbf{\Theta}}^{(1)}, \cdots, \widetilde{\mathbf{\Theta}}^{(Q)}$ to maximize

(21)
$$\sum_{q=1}^{Q} n_q \left(\log\det \widetilde{\mathbf{\Theta}}^{(q)} - \text{trace}(\widetilde{\mathbf{S}}^{(q)}\widetilde{\mathbf{\Theta}}^{(q)})\right) - \lambda \left(\sum_q \sum_{j \neq l} \|\widetilde{\mathbf{\Theta}}_{jl}^{(q)}\|_F + P(\widetilde{\mathbf{\Theta}}) \right)$$

where n_q is the number of observations for the qth class and P is a penalty to encourage similar structure among the $\widetilde{\mathbf{\Theta}}^{(q)}$'s.

APPENDIX A: DERIVATIONS FOR FGLASSO ALGORITHM

A.1. Step 2(b) of Algorithm 1. To derive the estimation for $\mathbf{w}_j$, we need to use Lemma 2, whose proof is provided in the Supplementary Material.

Lemma 2. *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*

$$\frac{\partial trace(\mathbf{A}\mathbf{X}^T\mathbf{B}\mathbf{X})}{\partial \mathbf{X}} = \mathbf{B}\mathbf{X}\mathbf{A} + \mathbf{B}^T\mathbf{X}\mathbf{A}^T. \tag{22}$$

Note (12) is equivalent to finding $\mathbf{w}_{j1}, \cdots, \mathbf{w}_{j(p-1)}$ to minimize

$$\text{trace}\left(\sum_{l=1}^{p-1}\sum_{k=1}^{p-1} \mathbf{S}_{jj}\mathbf{w}_{jl}^T(\widetilde{\mathbf{\Theta}}_{-j}^{-1})_{lk}\mathbf{w}_{jk} + 2\sum_{k=1}^{p-1} \mathbf{s}_{jk}^T\mathbf{w}_{jk} \right) + 2\lambda \sum_{k=1}^{p-1} \|\mathbf{w}_{jk}\|_F. \tag{23}$$

Setting the derivative of (23) with respect to $\mathbf{w}_{jk}$ to be zero and applying Lemma 2 yield

$$\begin{aligned}
\frac{\partial(23)}{\partial \mathbf{w}_{jk}} &= (\widetilde{\mathbf{\Theta}}_{-j}^{-1})_{kk}\mathbf{w}_{jk}\mathbf{S}_{jj} + (\widetilde{\mathbf{\Theta}}_{-j}^{-1})_{kk}^T\mathbf{w}_{jk}\mathbf{S}_{jj}^T \\
&\quad + \sum_{l\neq k}\left((\widetilde{\mathbf{\Theta}}_{-j}^{-1})_{lk}^T\mathbf{w}_{jl}\mathbf{S}_{jj}^T + (\widetilde{\mathbf{\Theta}}_{-j}^{-1})_{kl}\mathbf{w}_{jl}\mathbf{S}_{jj} \right) + 2\mathbf{s}_{jk} + 2\lambda\boldsymbol{\nu}_{jk} \\
&= 2\left((\widetilde{\mathbf{\Theta}}_{-j}^{-1})_{kk}\mathbf{w}_{jk}\mathbf{S}_{jj} + \sum_{l\neq k}(\widetilde{\mathbf{\Theta}}_{-j}^{-1})_{lk}^T\mathbf{w}_{jl}\mathbf{S}_{jj} + \mathbf{s}_{jk} + \lambda\boldsymbol{\nu}_{jk} \right) = \mathbf{0},
\end{aligned}$$

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

$$\mathbf{r}_{jk} = \sum_{l\neq k}(\widetilde{\mathbf{\Theta}}_{-j}^{-1})_{lk}^T\mathbf{w}_{jl}\mathbf{S}_{jj} + \mathbf{s}_{jk}. \tag{24}$$

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

$$(\widetilde{\boldsymbol{\Theta}}_{-j}^{-1})_{kk}\mathbf{w}_{jk}\mathbf{S}_{jj} + \mathbf{r}_{jk} + \lambda \frac{\mathbf{w}_{jk}}{\|\mathbf{w}_{jk}\|_F} = \mathbf{0}. \tag{25}$$

We replace (25) by (26), and standard package in R/MatLab can be used to solve the following M^2 by M^2 nonlinear equation

$$((\widetilde{\boldsymbol{\Theta}}_{-j}^{-1})_{kk} \otimes \mathbf{S}_{jj})\mathrm{vec}(\mathbf{w}_{jk}) + \mathrm{vec}(\mathbf{r}_{jk}) + \lambda \frac{\mathrm{vec}(\mathbf{w}_{jk})}{\|\mathbf{w}_{jk}\|_F} = 0. \tag{26}$$

Hence, the block coordinate descent algorithm for solving $\mathbf{w}_j$ in (12) 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 $k = 1, \cdots, p-1$.
    (a) Compute $\widehat{\mathbf{r}}_{jk}$ via (24).
    (b) Set $\widehat{\mathbf{w}}_{jk} = \mathbf{0}$ if $\|\mathbf{r}_{jk}\|_F \le \lambda$; otherwise solve for $\widehat{\mathbf{w}}_{jk}$ via (26).
```

A.2. Steps 2(a) and 2(c) of Algorithm 1. At the jth step, we need to compute $\widetilde{\boldsymbol{\Theta}}_{-j}^{-1}$ in (12) given current $\widetilde{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{\Theta}}^{-1}$. Then step 2(a) follows by the blockwise inversion formula. Next we solve for $\mathbf{w}_j$ via Algorithm 3, and then update $\widetilde{\boldsymbol{\Theta}}^{-1}$ given current $\mathbf{w}_j$, $\boldsymbol{\Theta}_{jj}$, and $\widetilde{\boldsymbol{\Theta}}_{-j}^{-1}$, by applying 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 $\widetilde{\boldsymbol{\Theta}}^{-1}$ by $\begin{pmatrix} \widetilde{\boldsymbol{\Theta}}_{-j}^{-1} + \mathbf{U}_j\mathbf{V}_j\mathbf{U}_j^T & -\mathbf{U}_j\mathbf{V}_j \\ -\mathbf{V}_j\mathbf{U}_j^T & \mathbf{V}_j \end{pmatrix}$, where $\mathbf{U}_j = \widetilde{\boldsymbol{\Theta}}_{-j}^{-1}\mathbf{w}_j$ 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.

APPENDIX B: PROOFS OF MAIN RESULTS

B.1. Proof of Lemma 1. Since both $\mathbf{a} = (\mathbf{a}_1^T, \cdots, \mathbf{a}_p^T)^T$ and $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^T, \cdots, \boldsymbol{\phi}_p^T)^T$ depend on M, we omit the corresponding superscripts to simplify the notation for readability.

For any pair $(j,l) \in V^2, j \ne l$ such that $\widetilde{C}_{jl}^M(s,t) = 0$ for all $(s,t) \in \mathcal{T}^2$, 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 (3), we have

$$\begin{aligned}
\widetilde{C}^M_{jl}(s,t) &= \mathrm{Cov}\left(\widetilde{g}^M_j(s), \widetilde{g}^M_l(t) | \widetilde{g}^M_k(u), k \neq j, l,\ \forall u \in \mathcal{T}\right) \\
&= \mathrm{Cov}\left(\mathbf{a}_j^T \boldsymbol{\phi}_j(s), \mathbf{a}_l^T \boldsymbol{\phi}_l(t) | \mathbf{a}_k^T \boldsymbol{\phi}_k(u), k \neq j, l,\ \forall u \in \mathcal{T}\right) \\
&= \mathrm{Cov}\left(\mathbf{a}_j^T \boldsymbol{\phi}_j(s), \mathbf{a}_l^T \boldsymbol{\phi}_l(t) | \mathbf{a}_k, k \neq j, l\right) \\
&= \boldsymbol{\phi}_j(s)^T \mathrm{Cov}(\mathbf{a}_j, \mathbf{a}_l | \mathbf{a}_U) \boldsymbol{\phi}_l(t) = 0. \qquad (27)
\end{aligned}$$

The third equality comes from the following argument. For any $k \in U$ and $u \in \mathcal{T}$, $\widetilde{g}^M_k(u) = \sum_{m=1}^M a_{km}\phi_{km}(u) = \mathbf{a}_k^T \boldsymbol{\phi}_k(u)$. By the orthogonality of ϕ_{km} and the fact $\sum_{m=1}^M E(a_{km}^2) = \sum_{m=1}^M \lambda_{km} < \infty$, we obtain that $a_{km} = \sum_{m'=1}^M a_{km'} \int_{\mathcal{T}} \phi_{km}(u)\phi_{km'}(u)du = \int_{\mathcal{T}} \widetilde{g}^M_k(u)\phi_{km}(u)du,\ m = 1, \cdots, M$. It then follows that there exists a one to one correspondence between $\{\mathbf{a}_k\}$ and $\{\widetilde{g}^M_k(u), \forall u \in \mathcal{T}\}$, which holds uniformly in k.

Since (27) holds for all $(s,t) \in \mathcal{T}^2$, we have $\mathrm{Cov}(\mathbf{a}_j, \mathbf{a}_l|\mathbf{a}_U) = \mathbf{0}$. Given that $\mathbf{a}_j, \mathbf{a}_l \in \mathcal{R}^M$ and $\mathbf{a}_U \in \mathcal{R}^{(p-2)M}$ are multivariate Gaussian, let the covariance matrix of their joint vector $\left(\mathbf{a}_j^T, \mathbf{a}_l^T, \mathbf{a}_U^T\right)^T$ be $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{pmatrix}$, where $\mathbf{A} = \mathrm{Var}\left((\mathbf{a}_j^T, \mathbf{a}_l^T)^T\right)$, $\mathbf{B} = \mathrm{Cov}\left((\mathbf{a}_j^T, \mathbf{a}_l^T)^T, \mathbf{a}_U\right)$, and $\mathbf{C} = \mathrm{Var}\left(\mathbf{a}_U\right)$. Then

$$\mathrm{Var}\left((\mathbf{a}_j^T, \mathbf{a}_l^T)^T | \mathbf{a}_U\right) = \mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T = \mathbf{D} = \begin{pmatrix} \mathbf{D}_{jj} & \mathbf{D}_{jl} \\ \mathbf{D}_{jl}^T & \mathbf{D}_{ll} \end{pmatrix},$$

where each term in $\mathbf{D}$ is an $M \times M$ submatrix. It follows from the blockwise inversion formula that

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{pmatrix}^{-1} = \begin{pmatrix} (\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T)^{-1} & * \\ * & * \end{pmatrix} = \begin{pmatrix} \mathbf{D}^{-1} & * \\ * & * \end{pmatrix},$$

where $\mathbf{D}^{-1} = \begin{pmatrix} (\mathbf{D}^{-1})_{jj} & (\mathbf{D}^{-1})_{jl} \\ (\mathbf{D}^{-1})_{jl}^T & (\mathbf{D}^{-1})_{ll} \end{pmatrix}$ and each term in $\mathbf{D}^{-1}$ is an $M \times M$ submatrix. Since $\mathrm{Cov}(\mathbf{a}_j, \mathbf{a}_l|\mathbf{a}_U) = \mathbf{D}_{jl} = \mathbf{0}$, $\mathbf{D}$ is block diagonal and thus $\mathbf{D}^{-1}$ is also block diagonal, which implies $(\mathbf{D}^{-1})_{jl} = \mathbf{0}$, i.e., $\widetilde{\boldsymbol{\Theta}}_{jl} = \mathbf{0}$. This completes the proof for one direction, and the other direction can be proved using similar arguments.

B.2. Convergence Rate for Submatrices in $\widetilde{\boldsymbol{\Theta}}$.

LEMMA 3. *Assume that $g_1, \cdots, g_p$ are from MGP, Conditions 1–3 hold, and $n > c_7 p^{4/\beta}$ for some positive constant c_7. Then*

$$||\widetilde{\boldsymbol{\Theta}}_{jl}||_F = O\left(\frac{p^2}{n^{(\beta-\alpha)/2}}\right) \qquad (28)$$

uniformly in $(j,l) \in S^c$.

The proof of Lemma 3 can be summarized in the following three main steps. We derive the corresponding rates of convergence for $\widetilde{K}_{j,l}(s,t)$, $\widetilde{C}_{j,l}(s,t)$, and $||\widetilde{\mathbf{\Theta}}_{jl}||_F$ under Conditions 1–3, respectively.

Step 1: For each pair $(j,l) \in V^2$ and any $(s,t) \in \mathcal{T}_j \times \mathcal{T}_l$, since

$$\begin{aligned} K_{jl}(s,t) &= \widetilde{K}_{jl}(s,t) + Cov\left(\widetilde{g}_j(s), g_l(t) - \widetilde{g}_l(t)\right) \\ &\quad + Cov\left(g_j(s) - \widetilde{g}_j(s), \widetilde{g}_l(t)\right) + Cov\left(g_j(s) - \widetilde{g}_j(s), g_l(t) - \widetilde{g}_l(t)\right), \end{aligned} \tag{29}$$

we can bound the last term by Cauchy-Schwarz inequality

$$\begin{aligned} \sup_{(s,t)\in\mathcal{T}_j\times\mathcal{T}_l} \left|Cov\left(g_j(s) - \widetilde{g}_j(s), g_l(t) - \widetilde{g}_l(t)\right)\right|^2 &\le \sup_{s\in\mathcal{T}_j} Var\left(g_j(s) - \widetilde{g}_j(s)\right) \\ &\quad \times \sup_{t\in\mathcal{T}_l} Var\left(g_l(t) - \widetilde{g}_l(t)\right). \end{aligned}$$

Moreover, under Condition 1

$$\begin{aligned} \sup_{s\in\mathcal{T}_j} Var\left(g_j(s) - \widetilde{g}_j(s)\right) &= \sup_{s\in\mathcal{T}_j} Var\left(\sum_{k=M+1}^{\infty} a_{ijk}\phi_{jk}(s)\right) \\ &= \sup_{s\in\mathcal{T}_j} \sum_{k=M+1}^{\infty} \lambda_{jk}\phi_{jk}^2(s) \\ &\le \sum_{k=M+1}^{\infty} \lambda_{jk} \sup_{s\in\mathcal{T}_j} \max_{k\ge M+1} \phi_{jk}^2(s) = O(n^{-\beta}). \end{aligned}$$

Applying this technique on the last three terms of (29), we obtain

$$\sup_{(s,t)\in\mathcal{T}_j\times\mathcal{T}_l} \left|K_{jl}(s,t) - \widetilde{K}_{jl}(s,t)\right| = O(n^{-\beta/2}). \tag{30}$$

Step 2: Recall we have defined $\mathcal{A}$ and $\mathcal{X}$, which can be respectively viewed as the functional counterparts of matrices and vectors. We need to define the corresponding operations and notation that will be used in step 2. First, define the $\mathcal{C} = \mathcal{A} \star \mathcal{B}$ such that $C_{jl}(s,t) = \sum_{k\in V} \int_{\mathcal{T}_k} A_{jk}(s,u_k)B_{kl}(u_k,t)du_k$, where $A \in \mathcal{A}$, $B \in \mathcal{B}$, and $C \in \mathcal{C}$. Note the covariance $\mathcal{K} = \{K_{jl} : \mathcal{T}_j \times \mathcal{T}_l \to \mathcal{R}, (j,l) \in V^2\}$ is a collection of covariance kernel functions such that $K_{jl}(s,t) = \text{Cov}\left(g_j(s), g_l(t)\right)$, $(s,t) \in \mathcal{T}_j \times \mathcal{T}_l$. The transpose of $\mathcal{K}$, $\mathcal{K}^T = \{K_{lj} : \mathcal{T}_l \times \mathcal{T}_j \to \mathcal{R}, (l,j) \in V^2\}$ is a collection of functions $K_{lj}(t,s) = \text{Cov}\left(g_l(t), g_j(s)\right) = K_{jl}(s,t)$. Hence $\mathcal{K} = \mathcal{K}^T$ and $\mathcal{K}$ is symmetric. The inverse covariance of $\mathcal{K}$, $\mathcal{K}^{-1} = \{(K^{-1})_{jl} : \mathcal{T}_j \times \mathcal{T}_l \to \mathcal{R}, (j,l) \in V^2\}$ is defined as a collection of functions $(K^{-1})_{jl}(s,t)$, such that $K \star K^{-1} = I$, where $I \in \mathcal{I}$ and $I_{jl}(s,t) = \delta_{jl}$ equals one if $j = l$ and zero otherwise. Moreover, define $\mathcal{Y} = \mathcal{A} \star \mathcal{X}$ and $\mathcal{Z} = \mathcal{X} \star \mathcal{A}$ such that $y_j(s) = \sum_l \int_{\mathcal{T}_l} A_{jl}(s,t)x_l(t)dt$ and $z_l(t) = \sum_j \int_{\mathcal{T}_j} x_j(s)A_{jl}(s,t)ds$, where $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $z \in \mathcal{Z}$, and $A \in \mathcal{A}$. Finally, define $\mathcal{X} \star \mathcal{Y} = \sum_j \int_{\mathcal{T}_j} x_j(s)y_j(s)ds$, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Given that $\mathbf{g} = (g_1, \cdots, g_p)^T$ are from MGP, some calculations show that the conditional covariance function (3) is given by

$$C_{jl}(s,t) = K_{jl}(s,t) - \mathcal{K}_{jU}(s) \star \mathcal{K}_{UU}^{-1} \star \mathcal{K}_{Ul}(t), \tag{31}$$

and the corresponding M-truncated conditional covariance function (7) is

$$\widetilde{C}_{jl}(s,t) = \widetilde{K}_{jl}(s,t) - \widetilde{\mathcal{K}}_{jU}(s) \star \widetilde{\mathcal{K}}_{UU}^{-1} \star \widetilde{\mathcal{K}}_{Ul}(t). \tag{32}$$

It follows from some standard algebra that

$$\begin{aligned} C_{jl}(s,t) - \widetilde{C}_{jl}(s,t) \quad = \quad & \underbrace{K_{jl}(s,t) - \widetilde{K}_{jl}(s,t)}_{S_1} + \underbrace{\mathcal{K}_{jU}(s) \star (\mathcal{K}_{UU}^{-1} - \widetilde{\mathcal{K}}_{UU}^{-1}) \star \mathcal{K}_{Ul}(t)}_{S_2} \\ & + \underbrace{(\mathcal{K}_{jU}(s) - \widetilde{\mathcal{K}}_{jU}(s)) \star \widetilde{\mathcal{K}}_{UU}^{-1} \star \mathcal{K}_{Ul}(t)}_{S_3} \\ & + \underbrace{\widetilde{\mathcal{K}}_{jU}(s) \star \widetilde{\mathcal{K}}_{UU}^{-1} \star (\mathcal{K}_{Ul}(t) - \widetilde{\mathcal{K}}_{Ul}(t))}_{S_4}. \end{aligned}$$

Note that since all kernel functions in $\mathcal{A}$ and $\mathcal{X}$ are assumed to be square integrable, i.e., $\int_{\mathcal{T}_j} \int_{\mathcal{T}_l} A_{jl}^2(s,t)dsdt < \infty$ and $\int_{\mathcal{T}_j} x_j^2(s)ds < \infty$, we can check that the properties for matrix A and vector x also hold for $\mathcal{A}$ and $\mathcal{X}$ in term of our defined operations in the functional domain. Denote $\lambda(\mathcal{A})$ to be the eigenvalues of $\mathcal{A}$ such that $\mathcal{A} \star \mathcal{X} = \lambda \mathcal{X}$ given $||\mathcal{X}||_2 = 1$, where $||\mathcal{X}||_2$ denotes the l_2 norm of $\mathcal{X}$ such that $||\mathcal{X}||_2 = \sqrt{\sum_{j \in V} \int_{\mathcal{T}_j} x_j^2(s)ds}$. Moreover, let $||\mathcal{A}||_{\max}$ be the entrywise l_∞ norm of $\mathcal{A}$ such that $||\mathcal{A}||_{\max} = \max\limits_{jl} \int_{\mathcal{T}_j} \int_{\mathcal{T}_l} |A_{jl}(s,t)|\, dsdt$, and $||\mathcal{X}||_{\max}$ the l_∞ norm of $\mathcal{X}$ such that $||\mathcal{X}||_{\max} = \max\limits_{j} \int_{\mathcal{T}_j} |x_j(s)|\, ds$. Denote $||\mathcal{A}||_2$ to be the operator norm of $\mathcal{A}$ such that $||\mathcal{A}||_2 = \sqrt{\lambda_{\max}(\mathcal{A}^T \star \mathcal{A})}$. For any symmetric positive definite $\mathcal{A}$, i.e., $\lambda_{\min}(\mathcal{A}) > 0$, it holds that $||\mathcal{A}||_2 = \lambda_{\max}(\mathcal{A})$ and $||\mathcal{A}^{-1}||_2 = 1/\lambda_{\min}(\mathcal{A})$.

It follows from $\mathcal{K}^{-1} - \tilde{\mathcal{K}}^{-1} = \mathcal{K}^{-1} \star (\tilde{\mathcal{K}} - \mathcal{K}) \star \tilde{\mathcal{K}}^{-1}$ that $S_2 = \mathcal{K}_{jU}(s) \star$

$\mathcal{K}^{-1}\star(\tilde{\mathcal{K}}-\mathcal{K})\star\tilde{\mathcal{K}}^{-1}\star\mathcal{K}_{Ul}(t)$. Hence, by (30) and Condition 2 we have

$$\begin{aligned}
\sup_{(s,t)\in\mathcal{T}_j\times\mathcal{T}_l}|S_2| &\leq \sup_{s\in\mathcal{T}_j}||\mathcal{K}_{jU}(s)||_2||\mathcal{K}^{-1}||_2||\tilde{\mathcal{K}}-\mathcal{K}||_2||\tilde{\mathcal{K}}^{-1}||_2\sup_{t\in\mathcal{T}_l}||\mathcal{K}_{Ul}(t)||_2\\
&\leq \sqrt{p}\sup_{s\in\mathcal{T}_j}||\mathcal{K}_{jU}(s)||_{\max}\times p||\mathcal{K}-\tilde{\mathcal{K}}||_{\max}\times\sqrt{p}\sup_{t\in\mathcal{T}_l}||\mathcal{K}_{Ul}(t)||_{\max}\\
&\qquad \times 1/\lambda_{\min}(\mathcal{K})\times 1/\lambda_{\min}(\tilde{\mathcal{K}}) = O\left(p^2/n^{\beta/2}\right). \qquad (33)
\end{aligned}$$

It follows from (30) and Condition 2 that

$$\begin{aligned}
\sup_{(s,t)\in\mathcal{T}_j\times\mathcal{T}_l}|S_3| &\leq \sup_{s\in\mathcal{T}_j}||\mathcal{K}_{jU}(s)-\widetilde{\mathcal{K}}_{jU}(s)||_2||\widetilde{\mathcal{K}}^{-1}||_2\sup_{t\in\mathcal{T}_l}||\mathcal{K}_{Ul}(t)||_2\\
&\leq \sqrt{p}\sup_{s\in\mathcal{T}_j}||\mathcal{K}_{jU}(s)-\widetilde{\mathcal{K}}_{jU}(s)||_{\max}\sqrt{p}\sup_{t\in\mathcal{T}_l}||\mathcal{K}_{Ul}(t)||_{\max}1/\lambda_{\min}(\tilde{\mathcal{K}})\\
&= O\left(p/n^{\beta/2}\right). \qquad (34)
\end{aligned}$$

By similar arguments, we can bound S_4 by

$$\sup_{(s,t)\in\mathcal{T}_j\times\mathcal{T}_l}|S_4| = \sup_{(s,t)\in\mathcal{T}_j\times\mathcal{T}_l}\left|\widetilde{\mathcal{K}}_{jU}(s)\star\widetilde{\mathcal{K}}^{-1}\star(\mathcal{K}_{Ul}(t)-\widetilde{\mathcal{K}}_{Ul}(t))\right| = O\left(p/n^{\beta/2}\right). \tag{35}$$

Combining (30) and (33)–(35), we obtain

$$\begin{aligned}
\sup_{(s,t)\in\mathcal{T}_j\times\mathcal{T}_l}\left|C_{jl}(s,t)-\widetilde{C}_{jl}(s,t)\right| &\leq \sup_{(s,t)\in\mathcal{T}_j\times\mathcal{T}_l}(|S_1|+|S_2|+|S_3|+|S_4|)\\
&= O\left(p^2/n^{\beta/2}\right). \qquad (36)
\end{aligned}$$

Step 3: It follows from the derivations in (27) and (36) that

$$\sup_{(s,t)\in\mathcal{T}_j\times\mathcal{T}_l}\left|C_{jl}(s,t)-\boldsymbol{\phi}_j(s)^T\mathrm{Cov}(\mathbf{a}_j,\mathbf{a}_l|\mathbf{a}_U)\boldsymbol{\phi}_l(t)\right| = O\left(p^2/n^{\beta/2}\right).$$

By the orthogonality property that $\int_{\mathcal{T}_j}\boldsymbol{\phi}_j(s)\boldsymbol{\phi}_j(s)^T = \int_{\mathcal{T}_l}\boldsymbol{\phi}_l(t)\boldsymbol{\phi}_l(t)^T = \mathbf{I}$, we obtain

$$\sup_{(s,t)\in\mathcal{T}_j\times\mathcal{T}_l}|C_{jl}(s,t)\boldsymbol{\Phi}_{jl}-\mathrm{Cov}(\mathbf{a}_j,\mathbf{a}_l|\mathbf{a}_U)| = O\left(p^2/n^{\beta/2}\right)\boldsymbol{\Phi}_{jl},$$

where $\boldsymbol{\Phi}_{jl}=\int_{\mathcal{T}_j}\int_{\mathcal{T}_l}\boldsymbol{\phi}_j(s)\boldsymbol{\phi}_l(t)^T dsdt$. Then under Condition 3, we obtain the rate of convergence under the Frobenius norm

$$\sup_{(s,t)\in\mathcal{T}_j\times\mathcal{T}_l}||C_{jl}(s,t)\boldsymbol{\Phi}_{jl}-\mathrm{Cov}(\mathbf{a}_j,\mathbf{a}_l|\mathbf{a}_U)||_F = O\left(p^2/n^{\beta/2}\right). \tag{37}$$

For any $(j,l) \in S^c$, $C_{jl}(s,t) = 0$ for all $(s,t) \in \mathcal{T}_j \times \mathcal{T}_l$. It follows from (37) and the derivations in Lemma 1 that $||\text{Cov}(\mathbf{a}_j, \mathbf{a}_l|\mathbf{a}_U)||_F = ||\mathbf{D}_{jl}||_F = O\left(p^2/n^{\beta/2}\right)$. Hence, we need to obtain the convergence rate for $||\widetilde{\boldsymbol{\Theta}}_{jl}||_F = ||(\mathbf{D}^{-1})_{jl}||_F$. Using the blockwise inversion formula again, we have

$$\begin{aligned}
(\mathbf{D}^{-1})_{jl} &= -\mathbf{D}_{jj}^{-1}\mathbf{D}_{jl}(\mathbf{D}_{ll} - \mathbf{D}_{jl}^T\mathbf{D}_{jj}^{-1}\mathbf{D}_{jl})^{-1} \\
&= -\mathbf{D}_{jj}^{-1}\mathbf{D}_{jl}(\mathbf{I} - \mathbf{D}_{ll}^{-1}\mathbf{D}_{jl}^T\mathbf{D}_{jj}^{-1}\mathbf{D}_{jl})^{-1}\mathbf{D}_{ll}^{-1} \\
&= -\mathbf{D}_{jj}^{-1}\mathbf{D}_{jl}\sum_{k=0}^{\infty}(\mathbf{D}_{ll}^{-1}\mathbf{D}_{jl}^T\mathbf{D}_{jj}^{-1}\mathbf{D}_{jl})^k\mathbf{D}_{ll}^{-1} \\
&= -\mathbf{D}_{jj}^{-1}\mathbf{D}_{jl}\mathbf{D}_{ll}^{-1} - \mathbf{D}_{jj}^{-1}\mathbf{D}_{jl}\sum_{k=1}^{\infty}(\mathbf{D}_{ll}^{-1}\mathbf{D}_{jl}^T\mathbf{D}_{jj}^{-1}\mathbf{D}_{jl})^k\mathbf{D}_{ll}^{-1}.
\end{aligned}$$

The convergent matrix expansion via Neumann series in the fourth equality comes from the following fact. Note $||\mathbf{D}_{jl}||_2^2 \le ||\mathbf{D}_{jl}||_F^2 = O(p^4/n^\beta) \le c_4 p^4/n^\beta$, where $c_4 > 0$ is some large enough constant. Let $c_7 > (c_4/c_3^2)^{1/\beta}$. It follows from Condition 3 and lower bound for n that $||\mathbf{D}_{ll}^{-1}\mathbf{D}_{jl}^T\mathbf{D}_{jj}^{-1}\mathbf{D}_{jl}||_2 \le ||\mathbf{D}_{jj}^{-1}||_2||\mathbf{D}_{ll}^{-1}||_2||\mathbf{D}_{jl}||_2^2 \le ||\mathbf{D}_{jl}||_F^2/(\lambda_{\min}(\mathbf{D}_{jj})\lambda_{\min}(\mathbf{D}_{ll})) \le c_4 p^4/(c_3^2 n^\beta) < c_4/(c_3^2 c_7^\beta) = c_8 < 1$.

Moreover, we have

$$||\mathbf{D}_{jj}^{-1}\mathbf{D}_{jl}\mathbf{D}_{ll}^{-1}||_2 \le ||\mathbf{D}_{jj}^{-1}||_2||\mathbf{D}_{jl}||_2||\mathbf{D}_{ll}^{-1}||_2 \le \sqrt{c_4}\frac{p^2}{n^{\beta/2}}\frac{1}{c_3^2}.$$

The higher order terms in the above expansion under matrix operator norm, $||\mathbf{D}_{jj}^{-1}\mathbf{D}_{jl}\sum_{k=1}^{\infty}(\mathbf{D}_{ll}^{-1}\mathbf{D}_{jl}^T\mathbf{D}_{jj}^{-1}\mathbf{D}_{jl})^k\mathbf{D}_{ll}^{-1}||_2$, can be bounded from above by

$$\begin{aligned}
&\le ||\mathbf{D}_{jj}^{-1}||_2||\mathbf{D}_{jl}||_2||\mathbf{D}_{ll}^{-1}||_2\sum_{k=1}^{\infty}||\mathbf{D}_{ll}^{-1}\mathbf{D}_{jl}^T\mathbf{D}_{jj}^{-1}\mathbf{D}_{jl}||_2^k \\
&< \sqrt{c_4}\frac{p^2}{n^{\beta/2}}\frac{1}{c_3^2}\sum_{k=1}^{\infty}c_8^k = \frac{\sqrt{c_4}c_8}{c_3^2(1-c_8)}\frac{p^2}{n^{\beta/2}}.
\end{aligned}$$

Hence, $||(\mathbf{D}^{-1})_{jl}||_2 = O\left(p^2/n^{\beta/2}\right)$ and $||(\mathbf{D}^{-1})_{jl}||_F \le \sqrt{M}||(\mathbf{D}^{-1})_{jl}||_2 = O\left(p^2/n^{(\beta-\alpha)/2}\right)$, which completes the proof.

B.3. Equivalence between $\boldsymbol{E}$ and $\widetilde{\boldsymbol{E}}_{\boldsymbol{\varepsilon}}$. In general, for any $\varepsilon \ge 0$, define the corresponding truncated edge set $\widetilde{E}_\varepsilon = \{(j,l) \in V^2 : j \ne l, ||\widetilde{\boldsymbol{\Theta}}_{jl}||_F > \varepsilon\}$. Let $S_\varepsilon = \widetilde{E}_\varepsilon \cup \{(1,1), \cdots, (p,p)\}$. Denote S_ε^c to be the complement of S_ε in V^2 with $||\widetilde{\boldsymbol{\Theta}}_{jl}||_F \le \varepsilon$ for $(j,l) \in S_\varepsilon^c$. Lemma 4 below, whose proof is provided in the Supplementary Material, ensures the equivalence between the true and truncated edge sets.

LEMMA 4. *Under all the conditions of Lemma 3 and Condition 4, let* $\varepsilon = c_8 p^2 / n^{(\beta-\alpha)/2}$ *for some constant* $c_8 > 0$, *we have* $E = \widetilde{E}_\varepsilon$.

B.4. Proof of Theorem 1. We first obtain the general error bound for $\widehat{\mathbf{\Theta}}$ in Section B.4.1. Then we present the general model selection consistency of fglasso, i.e., with high probability $\widehat{E} = \widetilde{E}_\varepsilon$, in Section B.4.2. Moreover in the Supplementary Material, we list Lemmas 5-10 used to prove Theorems 2 and 3, and provide their proofs therein. Finally in Section B.4.3 we prove Theorem 1 based on the results of Lemmas 3-4 and Theorem 3.

To prove the theorems, we need to use the definition of tail condition for the random variable given in Ravikumar et al. (2011).

DEFINITION 1 (Tail condition). *The random vector* $\mathbf{a} \in \mathcal{R}^{Mp}$ *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* $(i,j) \in \{1, \cdots, Mp\}^2$, *let* $\widetilde{S}_{ij}$, $\widetilde{\Sigma}_{ij}$ *be the* (i,j)*-th entry of* $\widetilde{\mathbf{S}}, \widetilde{\mathbf{\Sigma}}$ *respectively, then*

$$P\left(|\widetilde{S}_{ij} - \widetilde{\Sigma}_{ij}| \geq \delta\right) \leq 1/f(n,\delta) \ \ \textit{for all}\ \delta \in (0, 1/v_*]. \tag{38}$$

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

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

and

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

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 exist some constants* $\gamma \in (0, 1-\varepsilon]$, $\eta > 0$, *and* $\tau > 2$ *such that*

$$||\widetilde{\mathbf{\Gamma}}_{S_\varepsilon^c S_\varepsilon}(\widetilde{\mathbf{\Gamma}}_{S_\varepsilon S_\varepsilon})^{-1}||_\infty^{(M^2)} \leq 1 - \gamma - \varepsilon, \tag{39}$$

$$||\widetilde{\mathbf{\Gamma}}_{S_\varepsilon^c S_\varepsilon}(\widetilde{\mathbf{\Gamma}}_{S_\varepsilon S_\varepsilon})^{-1}\widetilde{\mathbf{\Gamma}}_{S_\varepsilon S_\varepsilon^c} - \widetilde{\mathbf{\Gamma}}_{S_\varepsilon^c S_\varepsilon^c}||_\infty^{(M^2)} \leq \left[\frac{\gamma}{2\varepsilon} + 1\right]\lambda, \tag{40}$$

$$||\widetilde{\mathbf{\Gamma}}_{S_\varepsilon S_\varepsilon^c}||_\infty^{(M^2)} \leq \eta M \bar{\delta}_f(n, (Mp)^\tau). \tag{41}$$

B.4.1. *General Error Bound.* For any block matrix $\mathbf{B} = (\mathbf{B}_{ij})$ with $\mathbf{B}_{ij} \in \mathcal{R}^{M\times M}, 1 \leq i,j \leq p$, define $||\mathbf{B}||_{\max}^{(M)} = \max\limits_{i,j} ||\mathbf{B}_{ij}||_F$ as the M-block version of entrywise l_∞-norm.

THEOREM 2. *Assume Condition 6 holds. Let $\widehat{\mathbf{\Theta}}$ be the unique solution to the fglasso (10) with regularization parameter $\lambda = (8M/\gamma)\bar{\delta}_f(n,(Mp)^\tau)$. Denote by $c_\varepsilon = (1+8\gamma^{-1}+\eta\varepsilon)$, $\kappa_{\widetilde{\Gamma}_\varepsilon} = ||(\widetilde{\mathbf{\Gamma}}_{S_\varepsilon S_\varepsilon})^{-1}||_\infty^{(M^2)}$, $\widetilde{\zeta}_\varepsilon = ||\widetilde{\mathbf{\Gamma}}_{S_\varepsilon S_\varepsilon^c}||_\infty^{(M^2)}$, and $d_\varepsilon = \max\limits_{j\in V}\Big|l \in V : ||\widetilde{\mathbf{\Theta}}_{jl}||_F > \varepsilon\Big|$. If the sample size n satisfies the lower bound*

$$
\begin{aligned}
n \quad > \quad & \bar{n}_f\left(1/\max\left\{v_*, 6c_\varepsilon M d_\varepsilon \max\{\frac{\kappa_{\widetilde{\Sigma}}\kappa_{\widetilde{\Gamma}_\varepsilon}}{1-3(p-d_\varepsilon)\varepsilon\kappa_{\widetilde{\Sigma}}},\right.\right. \\
& \left.\left.\frac{\kappa_{\widetilde{\Sigma}}^3\kappa_{\widetilde{\Gamma}_\varepsilon}^2 c_\varepsilon}{1-3(p-d_\varepsilon)\varepsilon\kappa_{\widetilde{\Sigma}}^3\kappa_{\widetilde{\Gamma}_\varepsilon}c_\varepsilon}\}\right\}, (Mp)^\tau\right),
\end{aligned}
\tag{42}
$$

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

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

$$
||\widehat{\mathbf{\Theta}} - \widetilde{\mathbf{\Theta}}||_{\max}^{(M)} \leq 2(1+8\gamma^{-1}+\eta\varepsilon)\kappa_{\widetilde{\Gamma}_\varepsilon} M\bar{\delta}_f(n,(Mp)^\tau);
\tag{43}
$$

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

By the Karush-Kuhn-Tucker (KKT) condition, a necessary and sufficient condition for $\widetilde{\mathbf{\Theta}}$ to maximize (10) is

$$
\widetilde{\mathbf{\Theta}}^{-1} - \widetilde{\mathbf{S}} - \lambda\widetilde{\mathbf{Z}} = 0,
\tag{44}
$$

where the sub-differential of $\sum_{j\neq l}||\mathbf{\Theta}_{jl}||_F$ involves all symmetric matrices $\widetilde{\mathbf{Z}}$ with $M\times M$ blocks defined by

$$
\widetilde{\mathbf{Z}}_{jl} = \begin{cases} \mathbf{0} & \text{if } j = l \\ \frac{\widetilde{\mathbf{\Theta}}_{jl}}{||\widetilde{\mathbf{\Theta}}_{jl}||_F} & \text{if } j \neq l \text{ and } \widetilde{\mathbf{\Theta}}_{jl} \neq \mathbf{0} \\ \{\widetilde{\mathbf{Z}}_{jl} \in \mathcal{R}^{M\times M} : ||\widetilde{\mathbf{Z}}_{jl}||_F \leq 1\} & \text{if } j \neq l \text{ and } \widetilde{\mathbf{\Theta}}_{jl} = \mathbf{0}. \end{cases}
\tag{45}
$$

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

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

$$
\min_{\widetilde{\mathbf{\Theta}}_{S_\varepsilon^c}=0}\left\{\text{trace}(\widetilde{\mathbf{S}}\widetilde{\mathbf{\Theta}}) - \log\,\det\widetilde{\mathbf{\Theta}} + \lambda\sum_{j\neq l}||\widetilde{\mathbf{\Theta}}_{jl}||_F\right\},
\tag{46}
$$

where $\widetilde{\mathbf{\Theta}} \in \mathcal{R}^{Mp\times Mp}$ is symmetric and positive definite. Second, for each $(j,l) \in S_\varepsilon$, we choose $\breve{\mathbf{Z}}_{jl}$ from the family of sub-differential of $\sum_{j\neq l} ||\widetilde{\mathbf{\Theta}}_{jl}||_F$ evaluated at $\breve{\mathbf{\Theta}}_{jl}$ defined in (45). Third, for each $(j,l) \in S_\varepsilon^c$, i.e., $||\widetilde{\mathbf{\Theta}}_{jl}||_F \le \varepsilon$, $\breve{\mathbf{Z}}_{jl}$ is replaced by

$$\frac{1}{\lambda}\left\{-\widetilde{\mathbf{S}}_{jl} + \left(\breve{\mathbf{\Theta}}^{-1}\right)_{jl}\right\}, \tag{47}$$

which satisfies the KKT condition (44). Finally, we need to verify strict dual feasibility condition, that is, $||\breve{\mathbf{Z}}_{jl}||_F < 1$ uniformly in $(j,l) \in S_\varepsilon^c$.

The following terms are needed in the proof of Theorem 2. Let $\mathbf{W}$ be the noise matrix, and $\mathbf{\Delta}$ the difference between the primal witness matrix $\breve{\mathbf{\Theta}}$ and the truth $\widetilde{\mathbf{\Theta}}$,

$$\mathbf{W} = \widetilde{\mathbf{S}} - \widetilde{\mathbf{\Theta}}^{-1}\ , \mathbf{\Delta} = \breve{\mathbf{\Theta}} - \widetilde{\mathbf{\Theta}}. \tag{48}$$

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

$$\mathbf{R}(\mathbf{\Delta}) = \breve{\mathbf{\Theta}}^{-1} - \widetilde{\mathbf{\Theta}}^{-1} + \widetilde{\mathbf{\Theta}}^{-1}\mathbf{\Delta}\widetilde{\mathbf{\Theta}}^{-1}. \tag{49}$$

We organize the proof of Theorem 2 in the following six steps.

Step 1. It follows from the tail condition (38) and Lemma 10 that the event $\left\{||\mathbf{W}||_{\max}^{(M)} \le M\bar{\delta}_f(n,(Mp)^\tau)\right\}$ holds with probability at least $1 - 1/(Mp)^{\tau-2}$. We need to verify that the conditions in Lemma 6 hold. Choosing the regularization parameter $\lambda = \frac{8M\bar{\delta}_f(n,(Mp)^\tau)}{\gamma}$, we have $||\mathbf{W}||_{\max}^{(M)} \le \frac{\gamma\lambda}{8}$. It remains to prove $||\mathbf{R}(\mathbf{\Delta}||_{\max}^{(M)}$ is also bounded by $\frac{\gamma\lambda}{8} = M\bar{\delta}_f(n,(Mp)^\tau)$.

Step 2. Let $r = 2\kappa_{\widetilde{\Gamma}_\varepsilon}(||\mathbf{W}||_{\max}^{(M)}+\lambda+\widetilde{\zeta}_\varepsilon\varepsilon) \le 2\kappa_{\widetilde{\Gamma}_\varepsilon}(1+8/\gamma+\eta\varepsilon)M\bar{\delta}_f(n,(Mp)^\tau)$, where we have used (41) in Condition 6. By $\bar{\delta}_f(n,(Mp)^\tau) \le 1/v_*$ and monotonicity of the inverse tail function, for any n satisfying the lower bound condition, we have $2\kappa_{\widetilde{\Gamma}_\varepsilon}(1+8/\gamma+\eta\varepsilon)M\bar{\delta}_f(n,(Mp)^\tau)$

$$\begin{aligned}
&\le \min\left\{\frac{1-3(p-d_\varepsilon)\varepsilon\kappa_{\widetilde{\Sigma}}}{3\kappa_{\widetilde{\Sigma}}d_\varepsilon}, \frac{1-3(p-d_\varepsilon)\varepsilon\kappa_{\widetilde{\Sigma}}^3\kappa_{\widetilde{\Gamma}_\varepsilon}c_\varepsilon}{3\kappa_{\widetilde{\Sigma}}^3\kappa_{\widetilde{\Gamma}_\varepsilon}d_\varepsilon c_\varepsilon}\right\}\\
&\le \min\left\{\frac{1-3(p-d_\varepsilon)\varepsilon\kappa_{\widetilde{\Sigma}}}{3\kappa_{\widetilde{\Sigma}}d_\varepsilon}, \frac{1-3(p-d_\varepsilon)\varepsilon\kappa_{\widetilde{\Sigma}}^3\kappa_{\widetilde{\Gamma}_\varepsilon}}{3\kappa_{\widetilde{\Sigma}}^3\kappa_{\widetilde{\Gamma}_\varepsilon}d_\varepsilon}\right\}\\
&= \min\left\{\frac{1}{3\kappa_{\widetilde{\Sigma}}d_\varepsilon}, \frac{1}{3\kappa_{\widetilde{\Sigma}}^3\kappa_{\widetilde{\Gamma}_\varepsilon}d_\varepsilon}\right\} - \frac{p-d_\varepsilon}{d_\varepsilon}\varepsilon.
\end{aligned}$$

Then the conditions in Lemma 8 are satisfied, and hence the error bound satisfies $||\boldsymbol{\Delta}||_{\max}^{(M)} = ||\breve{\boldsymbol{\Theta}} - \widetilde{\boldsymbol{\Theta}}||_{\max}^{(M)} \leq r$.

Step 3. The condition $||\boldsymbol{\Delta}||_{\max}^{(M)} \leq \frac{1}{3\kappa_{\widetilde{\Sigma}} d_\varepsilon} - \frac{p-d_\varepsilon}{d_\varepsilon}\varepsilon$ is satisfied by step 2. Thus by Lemma 7 and results in step 2, we have $||\mathbf{R}(\boldsymbol{\Delta})||_{\max}^{(M)}$

$$\begin{aligned} &\leq \frac{3}{2}||\boldsymbol{\Delta}||_{\max}^{(M)} \left(d_\varepsilon ||\boldsymbol{\Delta}||_{\max}^{(M)} + (p-d_\varepsilon)\varepsilon\right) \kappa_{\widetilde{\Sigma}}^3 \\ &= \Big\{3\kappa_{\widetilde{\Sigma}}^3 \kappa_{\widetilde{\Gamma}_\varepsilon}(1+8/\gamma+\eta\varepsilon)\left(d_\varepsilon 2\kappa_{\widetilde{\Gamma}_\varepsilon}(1+8/\gamma+\eta\varepsilon)M\bar{\delta}_f(n,(Mp)^\tau)\right. \\ &\quad \left. +(p-d_\varepsilon)\varepsilon\right)\Big\}\,\gamma\lambda/8 \\ &\leq \gamma\lambda/8, \end{aligned}$$

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

$$\begin{aligned} 2d_\varepsilon \kappa_{\widetilde{\Gamma}_\varepsilon}(1+\frac{8}{\gamma}+\eta\varepsilon)M\bar{\delta}_f(n,(Mp)^\tau) &\leq \frac{1-3(p-d_\varepsilon)\varepsilon\kappa_{\widetilde{\Sigma}}^3\kappa_{\widetilde{\Gamma}_\varepsilon}c_\varepsilon}{3\kappa_{\widetilde{\Sigma}}^3\kappa_{\widetilde{\Gamma}_\varepsilon}c_\varepsilon} \\ &= \frac{1}{c_\varepsilon}\frac{1}{3\kappa_{\widetilde{\Sigma}}^3\kappa_{\widetilde{\Gamma}_\varepsilon}} - (p-d_\varepsilon)\varepsilon. \end{aligned}$$

Step 4. Steps 1 and 3 imply the strict dual feasibility in Lemma 6, and hence $\breve{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Theta}}$ by Lemma 5.

Step 5. It follows from the results in steps 2 and 4 that the error bound (43) holds with probability at least $1 - 1/(Mp)^{\tau-2}$.

Step 6. For $(j,l) \in S_\varepsilon^c$, $||\widetilde{\boldsymbol{\Theta}}_{jl}||_F \leq \varepsilon$. Step 4 implies $\breve{\boldsymbol{\Theta}}_{S_\varepsilon^c} = \widehat{\boldsymbol{\Theta}}_{S_\varepsilon^c}$. In the restricted fglasso problem (46), we have $\breve{\boldsymbol{\Theta}}_{S_\varepsilon^c} = \widehat{\boldsymbol{\Theta}}_{S_\varepsilon^c} = \mathbf{0}$. Therefore, $(\widetilde{E}_\varepsilon)^c \subset (\widehat{E})^c$ and part (ii) follows by taking the complement.

B.4.2. *General Model Selection Consistency.*

THEOREM 3. *Let* $\Theta_{\min} = \min\limits_{(j,l)\in\widetilde{E}_\varepsilon} ||\widetilde{\boldsymbol{\Theta}}_{jl}||_F$. *Under the same conditions as in Theorem 2, if the sample size n satisfies the lower bound*

$$\begin{aligned} n \;&>\; \bar{n}_f\left(1/\max\left\{2\kappa_{\widetilde{\Gamma}_\varepsilon}c_\varepsilon\Theta_{\min}^{-1}M, v_*, 6c_\varepsilon M d_\varepsilon \max\{\frac{\kappa_{\widetilde{\Sigma}}\kappa_{\widetilde{\Gamma}_\varepsilon}}{1-3(p-d_\varepsilon)\varepsilon\kappa_{\widetilde{\Sigma}}},\right.\right. \\ &\left.\left.\frac{\kappa_{\widetilde{\Sigma}}^3\kappa_{\widetilde{\Gamma}_\varepsilon}^2 c_\varepsilon}{1-3(p-d_\varepsilon)\varepsilon\kappa_{\widetilde{\Sigma}}^3\kappa_{\widetilde{\Gamma}_\varepsilon}c_\varepsilon}\}\right\}, (Mp)^\tau\right), \end{aligned} \tag{50}$$

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

Proof. It follows from the proof and results of Theorem 2(i) that $||\breve{\mathbf{\Theta}} - \widetilde{\mathbf{\Theta}}||_{\max}^{(M)} \le r \le 2(1+8\gamma^{-1}+\eta\varepsilon)\kappa_{\widetilde{\Gamma}_\varepsilon} M\bar{\delta}_f(n,(Mp)^\tau)$ and $\widehat{\mathbf{\Theta}} = \breve{\mathbf{\Theta}}$ hold with probability at least $1-1/(Mp)^{\tau-2}$. The lower bound for the sample size n in (50) implies $\Theta_{\min} > 2(1+8\gamma^{-1}+\eta\varepsilon)\kappa_{\widetilde{\Gamma}_\varepsilon} M\bar{\delta}_f(n,(Mp)^\tau) \ge r$. By Lemma 9, we have $\widehat{\mathbf{\Theta}}_{jl} \neq \mathbf{0}$ for all $(j,l) \in S_\varepsilon$, which entails that $\widetilde{E}_\varepsilon \subset \widehat{E}$. Combining this result with Theorem 2(ii) yields $\widetilde{E}_\varepsilon = \widehat{E}$.

B.4.3. *Proof of Theorem 1.* Recall $\mathbf{a} \in \mathcal{R}^{Mp}$. From Lemma 1 in Ravikumar et al. (2011), for any $k \in \{1,\dots,Mp\}$, the rescaled random variables $\mathbf{a}_k/\sqrt{\widetilde{\Sigma}_{kk}}$ are sub-Gaussian with parameter $\sigma = 1$, and the sample covariance matrix satisfies the tail condition (38) with $v_* = (\max_k \widetilde{\Sigma}_{kk} 40)^{-1}$ and $f(n,\delta) = (1/4)\exp(c_* n\delta^2)$, where $c_* = [128\times 5^2 \max_k(\widetilde{\Sigma}_{kk})^2]^{-1}$. Therefore the corresponding inverse functions take the following forms

$$\bar{\delta}_f(n,(Mp)^\tau) = \sqrt{\frac{\log[4(Mp)^\tau]}{c_* n}} = \sqrt{128\times 5^2 \max_k(\widetilde{\Sigma}_{kk})^2}\sqrt{\frac{\tau\log Mp + \log 4}{n}} \tag{51}$$

and

$$\bar{n}_f(\delta,(Mp)^\tau) = \frac{\log[4(Mp)^\tau]}{c_*\delta^2} = 128\times 5^2 \max_k(\widetilde{\Sigma}_{kk})^2\left(\frac{\tau\log Mp + \log 4}{\delta^2}\right) \tag{52}$$

It follows from Lemma 4 with $\varepsilon = c_8\frac{p^2}{n^{(\beta-\alpha)/2}}$ that $E = \widetilde{E}_\varepsilon$. Thus we have $S = S_\varepsilon$, $d = d_\varepsilon$, $\kappa_{\widetilde{\Gamma}} = \kappa_{\widetilde{\Gamma}_\varepsilon}$, $\widetilde{\zeta} = \widetilde{\zeta}_\varepsilon$, and $c_\varepsilon = 1 + 8\gamma^{-1} + \frac{\eta c_8 p^2}{n^{(\beta-\alpha)/2}}$. Moreover, since $\mathrm{Var}(a_{jm}) = \lambda_{jm}$, then $\sqrt{\max_k(\widetilde{\Sigma}_{kk})^2} = \sqrt{\max_k(\mathrm{Var}(\mathbf{a}_k))^2} = \max_{j,m}\lambda_{jm}$. By substituting these terms into Condition 6 and Theorem 3, some calculations using (51) and (52) lead to Condition 5, the lower bound for the sample size, i.e., $n > C_1M^2d^2(1+8\gamma^{-1}+\eta c_7p^2n^{(\alpha-\beta)/2})^2(\tau\log(Mp)+\log 4)/c_1^2$ and $n > C_2M^2\Theta_{\min}^{-2}(1+8\gamma^{-1}+\eta c_7p^2n^{(\alpha-\beta)/2})^2(\tau\log(Mp)+\tau\log 4)/c_1^2$ and the desired regularization parameter λ with $M = c_1n^\alpha$ under Condition 1.

By satisfying $n > c_7p^{4/\beta}$ and Conditions 1-4, Lemma 4 implies $E = \widetilde{E}_\varepsilon$. By satisfying (50) and Condition 6, Theorem 3 shows $\widetilde{E}_\varepsilon = \widehat{E}$ holds with probability at least $1-1/(c_1n^\alpha p)^{\tau-2}$. Combining these two results completes the proof.

SUPPLEMENTARY MATERIAL

Supplement to: "Functional Graphical Models"
(doi: xx.xxx; .pdf). We provide additional Lemmas, technical proofs and

simulation results

REFERENCES

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

T. Cai, W. Liu, and X. Luo. A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of American Statistical Association*, 106, 2011.

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

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

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and sparse group lasso. *Working Paper*, 2010.

E.P. Hayden, R.E. Wiegand, E.T. Meyer, L.O. Bauer, S.J. O's Connor, Nurbberger J.I., Chorlian D.B., Porjesz B., and Begleiter H. Patterns of regional brain activity in alcohol-dependent subjects. *Alcoholism: Clinical and Experimental Research*, 30:1986–1991, 2006.

H. Hung and C. Wang. Matrix variate logistic regression model with application to eeg data. *Biostatistics*, 14:189–202, 2013.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

H. Zhu, N. Strawn, and D. Dunson. Bayesian graphical models for multivariate functional data. *Manucript*, 2014a.

Y. Zhu, X. Shen, and W. Pan. Structural pursuit over multiple undirected graphs. *Journal of American Statistical Association*, 2014b.

Data Sciences and Operations Department
Marshall School of Business
University of Southern California
Los Angeles, CA 90089
USA
E-mail: qiaoxinghao@gmail.com
gareth@usc.edu
jinchilv@marshall.usc.edu