



Functional Graphical Models

Xinghao Qiao, Shaojun Guo & Gareth M. James

To cite this article: Xinghao Qiao, Shaojun Guo & Gareth M. James (2019) Functional Graphical Models, Journal of the American Statistical Association, 114:525, 211-222, DOI: [10.1080/01621459.2017.1390466](https://doi.org/10.1080/01621459.2017.1390466)

To link to this article: <https://doi.org/10.1080/01621459.2017.1390466>



View supplementary material [↗](#)



Published online: 15 Aug 2018.



Submit your article to this journal [↗](#)



Article views: 3920



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 20 View citing articles [↗](#)



Functional Graphical Models

Xinghao Qiao^a, Shaojun Guo^b, and Gareth M. James^c

^aDepartment of Statistics, London School of Economics, London, UK; ^bInstitute of Statistics and Big Data, Renmin University of China, Beijing, P.R. China;

^cDepartment of Data Sciences and Operations, University of Southern California, Los Angeles, CA

ABSTRACT

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

ARTICLE HISTORY

Received February 2017
Revised October 2017

KEYWORDS

Block coordinate descent algorithm; Block sparse precision matrix; Functional data; Functional principal component analysis; Graphical models

1. Introduction

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


For Gaussian data, where \mathbf{X} follows a multivariate Gaussian distribution, one can show that estimating the edge set is equivalent to identifying the locations of the nonzero elements in the *precision matrix*, that is, the inverse covariance matrix of \mathbf{X} . Hence, the literature has mainly focused on two approaches for estimating high-dimensional *Gaussian graphical models*. One type of method, proposed by Meinshausen and Bühlmann (2006), considers neighborhood selection. It adopts a lasso (Tibshirani 1996) or Dantzig selector (Candes and Tao 2007) type of penalized regression approach whereby each variable is regressed on the other variables, thus identifying the locations of nonzero entries in the precision matrix column by column, see also Peng et al. (2009), Cai, Liu, and Luo (2011), and Sun and Zhang (2013). Another method, proposed by Yuan and Lin (2007), optimizes the glasso criterion, essentially a Gaussian

log-likelihood with the addition of a lasso type penalty on the entries of the precision matrix. The glasso has arguably proved the most popular of these two methods, in part because a number of efficient algorithms have been developed to minimize the convex glasso criterion (Friedman, Hastie and Tibshirani 2007; Boyd et al. 2010; Witten, Friedman and Simon 2011; Mazumder and Hastie 2012a, 2012b). 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, Lafferty and Wasserman (2010), Kolar and Xing (2011), Danaher, Wang and Witten (2014), and Zhu, Shen and Pan (2014), and the references therein.

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

CONTACT Xinghao Qiao  X.Qiao@lse.ac.uk  Department of Statistics, London School of Economics, U.K.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

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

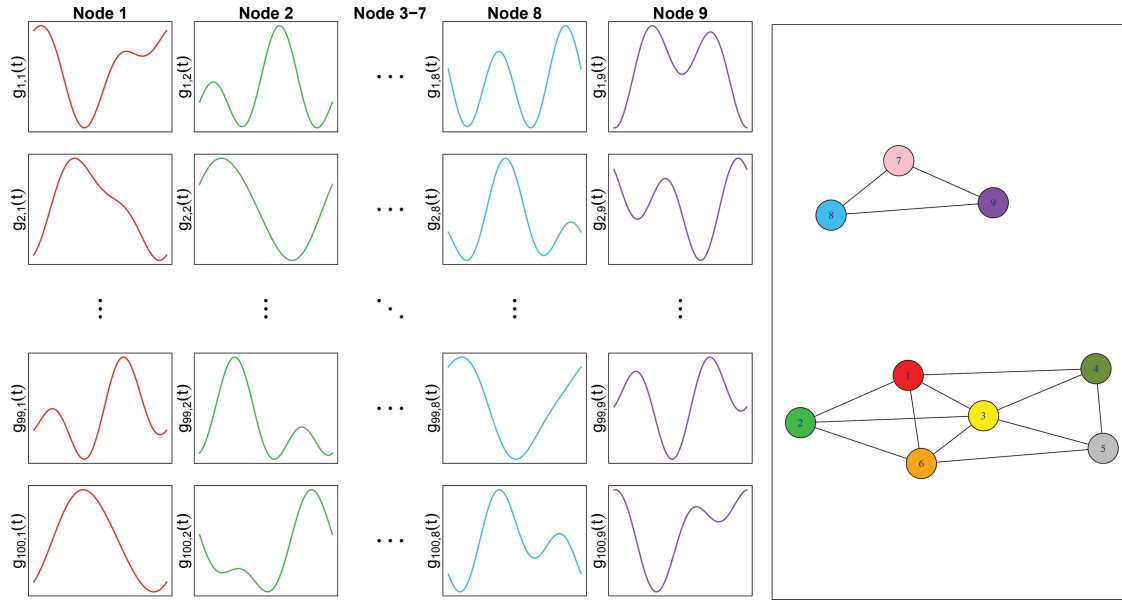


Figure 1. The illustrative example. Left: The data, $n = 100$ observations of $g_{ij}(t)$ for $j = 1, \dots, 9$ nodes. Right: The true underlying network structure.

and hence conditionally independent of, the first six 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) dataset taken from an alcoholism study (Zhang et al. 1995; Ingber 1997). The study consists of $n = 122$ subjects split between an alcoholic group and a control group. Each subject was exposed to either a single stimulus or two stimuli. The resulting EEG activity was recorded at 256 time points over a one second interval using electrodes placed at 64 standard locations on the subject's scalp. Hence, each observation, or subject, involves $p = 64$ different functions observed at 256 time points. It is of scientific interest to identify differences in brain EEG activity filtered at α frequency bands (Hayden et al. 2006) between the two groups, so we construct FGM's for each group and explore the differences. Functional data of this sort can arise in a number of other contexts. For example, rather than recording only a static set of p gene expression levels at a single point in time, it is now becoming common to observe multiple expression levels over time (Storey et al. 2005), so $g_{ij}(t)$ would represent the expression level of gene j for subject i at time t . Alternatively, in a marketing context, it is now possible to observe online purchase patterns among a basket of p different products for each of n individual customers over a period of time, so $g_{ij}(t)$ might represent the cumulative purchase history of product j by customer i at time t .

One possible approach to handle this sort of functional data would be to sample the functions over a grid of time points, t_1, \dots, t_T , estimate T separate networks, and then either report all T networks or construct a single graphical model by somehow merging the T networks. However, while conceptually simple, this strategy has several drawbacks. First, the approach is only possible if $g_1(t), \dots, g_p(t)$ are observed over a common domain $t \in \mathcal{T}$, but in many instances the domains of $g_j(t)$ and $g_l(t)$ will differ. Second, functions are often observed over a relatively sparse set of time points, which makes it impossible to create a single set of grid points, t_1, \dots, t_T , at which the functions can be sampled. Third, one of the main advantages

of a graphical network is its ability to provide a relatively simple representation of the conditional dependence structure. However, simultaneously interpreting T different networks would significantly increase the complexity of the dependence structure. Finally, each of the T networks would only correspond to dependencies among the functions at a common time point. However, it seems likely that some dependencies may only exist at different time points, that is, it may be the case that $g_j(t)$ and $g_l(t)$ are conditionally uncorrelated for every individual value of t but $g_j(s)$ and $g_l(t)$ are correlated for some $s \neq t$. A simple example that illustrates this issue is provided in Appendix A.1. In such a scenario, each of the T networks would fail to identify the correlation structure.

In this 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, for example, *functional principal component analysis* (FPCA), to approximate each function by a finite representation, which results in estimating a block precision matrix. Standard glasso algorithms for scalar data involve estimating the nonzero elements in the precision matrix. By comparison, we have developed an efficient algorithm to estimate the nonzero blocks of a higher-dimensional precision matrix. In our theoretical results, we develop the entry-wise concentration inequalities for the sample covariance matrix of the estimated principal component scores. To the best of our knowledge, this is the first result on concentration inequalities for modelling high-dimensional functional data under a FPCA framework, which provides a powerful tool to derive the nonasymptotic upper bounds. This result expands the theoretical analysis of graph selection consistency from the standard setting to the more complicated functional domain.

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

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

2. Methodology

2.1. Gaussian Graphical Models

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

$$c_{jl} = \text{cov}(X_j, X_l | \{X_k, k \neq j, l\}) \quad (1)$$

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 $\mathbf{X} = (X_1, \dots, X_p)^T$ is multivariate Gaussian with covariance matrix Σ^* , 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, $\Theta^* = \Sigma^{*-1}$. Let $G = (V, E)$ denote an undirected graph with vertex set $V = \{1, \dots, p\}$ 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, $\mathbf{x}_1, \dots, \mathbf{x}_n$, 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 nonzero elements in the precision matrix.

The glasso (Yuan and Lin 2007) considers a regularized estimator for Θ^* by adding an ℓ_1 penalty on the off-diagonal entries of the precision matrix to the Gaussian log-likelihood (up to constants)

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

where $\Theta \in \mathbb{R}^{p \times p}$ is symmetric positive definite, \mathbf{S} is the sample covariance matrix of $\mathbf{x}_1, \dots, \mathbf{x}_n$, and γ_n is a nonnegative tuning parameter. In a similar fashion to the standard lasso, the ℓ_1 penalty in (2) both regularizes the estimate and ensures that $\hat{\Theta}$ is sparse, that is, has many zero elements.

2.2. Functional Graphical Models

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

$$C_{jl}(s, t) = \text{cov}(g_j(s), g_l(t) | \{g_k(\cdot), k \neq j, l\}), \quad (3)$$

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

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

Suppose we observe $\mathbf{g}_i = (g_{i1}, \dots, g_{ip})^T$, $i = 1, \dots, n$, and for each i , $g_{ij}(t)$ $t \in \mathcal{T}$ is a realization from a mean zero Gaussian process.³ The Karhunen–Loève expansion (Theorem 1.5 in Bosq 2000) allows us to represent each function in the form

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

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

² Here, we can use the projection theorem for Hilbert spaces (Hsing and Eubank 2015, chap. 2.5) to rigorously define the relevant conditional expectation terms, for example, $E(g_j(s) | \{g_k(\cdot), k \neq j, l\})$. See also the definition of the conditional joint probability measure within Hilbert spaces in Zhu, Strawn, and Dunson (2016).

³ Our methodological and theoretical results can be extended to the case of Gaussian processes with nonzero means but for clarity of the exposition, we do not investigate that case here.

component functions and form an infinite dimensional basis representation for $g_{ij}(t)$.

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

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

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

$$C_{jl}^M(s, t) = \text{cov}(g_j^M(s), g_l^M(t) | \{g_k^M(\cdot), k \neq j, l\}). \quad (6)$$

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

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

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

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

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

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

2.3. Functional Graphical Lasso

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

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

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

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

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

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

Let $\hat{\mathbf{a}}_i^M = ((\hat{\mathbf{a}}_{i1}^M)^T, \dots, (\hat{\mathbf{a}}_{ip}^M)^T)^T \in \mathbb{R}^{Mp}$, where $\hat{\mathbf{a}}_i^M = (\hat{a}_{ij1}, \dots, \hat{a}_{ijM})^T$ and \mathbf{S}^M be the sample covariance matrix of $\hat{\mathbf{a}}_i^M$. Motivated by **Lemma 1**, we propose the *fglasso* to estimate the network structure. The *fglasso* modifies the glasso by incorporating a group lasso penalty (Yuan and Lin 2006) to produce a block sparsity structure. Specifically, the *fglasso* is defined as the solution to

$$\hat{\Theta}^M = \underset{\Theta^M}{\text{argmax}} \left\{ \log \det \Theta^M - \text{trace}(\mathbf{S}^M \Theta^M) - \gamma_n \sum_{j \neq l} \|\Theta_{jl}^M\|_F \right\}, \quad (9)$$

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

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

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

2.4. Joint Functional Graphical Lasso

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

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

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

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

to performing Q uncoupled fglasso problems (9). Here, we consider using a group lasso penalty for P_2 , that is,

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

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

3. Computation

3.1. fglasso Algorithm

A number of efficient algorithms (Friedman, Hastie and Tibshirani 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 that mirrors recent techniques for optimizing the glasso criterion (Zhu, Shen and Pan 2014).

Let Θ_{-j} , Σ_{-j} , and S_{-j} , respectively, be $M(p-1) \times M(p-1)$ sub matrices excluding the j th row and column block of Θ , Σ , and S , and let w_j , σ_j , and s_j be $M(p-1) \times M$ matrices representing the j th column block after excluding the j th row block. Finally, let Θ_{jj} , Σ_{jj} , and S_{jj} be the (j, j) th $M \times M$ blocks in Θ , Σ , and S , respectively. So, for instance for $j = 1$, $\Theta = \begin{pmatrix} \Theta_{11} & w_1^T \\ w_1 & \Theta_{-1} \end{pmatrix}$. Then, for a fixed value of Θ_{-j} , standard calculations show that (9) is solved by setting

$$\hat{\Theta}_{jj} = S_{jj}^{-1} + \hat{w}_j^T \Theta_{-j}^{-1} \hat{w}_j, \quad (13)$$

where

$$\hat{w}_j = \arg \min_{w_j} \left\{ \text{trace}(S_{jj} w_j^T \Theta_{-j}^{-1} w_j) + 2\text{trace}(s_j^T w_j) + 2\gamma_n \sum_{l=1}^{p-1} \|w_{jl}\|_F \right\}, \quad (14)$$

and w_{jl} represents the l th $M \times M$ block of w_j . Computing (14) can be achieved using some matrix calculus with details provided in Section B.1 of the supplementary material.

This suggests a block coordinate descent algorithm, where one iterates through j repeatedly computing (14) until convergence. In fact, by checking the conditions of Theorem 4.1 in Tseng (2001), it is easy to verify that iteratively minimizing (14) over w_j and updating Θ_{jj} by (13) for $j = 1, \dots, p$ provides a convergent solution for globally maximizing the fglasso criterion. The main potential difficulty with this approach is that Θ_{-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 (c) are derived using standard matrix results, the details of which are provided in Section B.2 of the supplementary material. We also develop in Section B.3, an analogous algorithm for solving (11) when jointly estimating multiple networks.

Algorithm 1 Functional Graphical Lasso Algorithm

1. Initialize $\hat{\Theta} = I_{Mp}$ and $\hat{\Sigma} = I_{Mp}$.
 2. Repeat until convergence for $j = 1, \dots, p$.
 - (a) Compute $\hat{\Theta}_{-j}^{-1} \leftarrow \hat{\Sigma}_{-j} - \hat{\sigma}_j \hat{\Sigma}_{jj}^{-1} \hat{\sigma}_j^T$.
 - (b) Solve for \hat{w}_j in (14) using Algorithm 3 in the supplementary material.
 - (c) Reconstruct $\hat{\Sigma}$ using $\hat{\Sigma}_{jj} = S_{jj}$, $\hat{\sigma}_j = -U_j S_{jj}$, and $\hat{\Sigma}_{-j} = \hat{\Theta}_{-j}^{-1} + U_j S_{jj} U_j^T$, where $U_j = \hat{\Theta}_{-j}^{-1} \hat{w}_j$.
 3. Set $\hat{E} = \{(j, l) : \|\hat{\Theta}_{jl}\|_F \neq 0, (j, l) \in V^2, j \neq l\}$.
-

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, Friedman, and Simon 2011; Mazumder and Hastie 2012b; Danaher, Wang, and Witten 2014; Zhu, Shen and Pan 2014). Since each resulting network involves many fewer nodes the glasso problem can be computed at a much lower computational cost.

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

Proposition 1. If the solution to (9) takes a block diagonal form, that is, $\Theta = \text{diag}(\Theta_1, \dots, \Theta_K)$, then (9) can be computed by separately solving K smaller fglasso problems

$$\hat{\Theta}_k = \arg \max_{\Theta_k} \left\{ \log \det \Theta_k - \text{trace}(S_k \Theta_k) - \gamma_n \sum_{j \neq l} \|\Theta_{k,jl}\|_F \right\}, \quad (15)$$

for $k = 1, \dots, K$, where S_k is the submatrix of S corresponding to Θ_k and $\Theta_{k,jl}$ is the (j, l) th submatrix of Θ_k .

Proposition 2. Without loss of generality, let G_1, \dots, 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, \dots, G_K is that $\|S_{i' i}\|_F \leq \gamma_n$ for all $i \in G_k, i' \in G_{k'}, k \neq k'$.

Propositions 1 and 2 suggest first performing a screening procedure on 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 A be a p by p adjacency matrix, whose diagonal elements are one and off-diagonal elements take the form $A_{i' i} = 1_{\|S_{i' i}\|_F > \gamma_n}$.

2. Identify K connected components of the graph based on the adjacency matrix \mathbf{A} . Let G_k be the index set of the features in the k th connected component, $k = 1, \dots, K$.
3. For $k = 1, \dots, K$, solve $\hat{\Theta}_k$ via Algorithm 1 using the nodes in G_k . The final solution to the fglasso problem $\hat{\Theta}$ is obtained by rearranging the rows/columns of the permuted version, $\text{diag}(\hat{\Theta}_1, \dots, \hat{\Theta}_K)$.

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 k th fglasso problem requires $O(|G_k|^3)$ operations for $k = 1, \dots, K$, hence, the total computational cost for Algorithm 2 is $O(p^2 + \sum_{k=1}^K |G_k|^3)$. Algorithm 2 significantly reduces the computational cost, if $|G_1|, \dots, |G_K|$ are much smaller than p , which is the situation when the tuning parameter, γ_n , is large. This is the case, we are generally interested in for real-data problems since, for the sake of network interpretation and visualization, most practical applications estimate sparse networks.

3.3. Selection of Tuning Parameters

Estimating the FGM requires choosing M (number of selected principal components) and γ_n (regularization parameter to tune the block sparsity level of Θ). First, to select M , one can either adopt leave-one-curve-out cross-validation (Rice and Silverman 1991) or an AIC-type criteria (Yao, Muller and Wang 2005). To reach a compromise, we develop a J -fold cross-validated (CV) approach. Specifically, let h_{ijk} represent a noisy observation of $g_{ij}(t_k)$. We randomly divide the set of observed time points into J equal-size folds. We then treat one group for each $g_{ij}(t)$ as a validation dataset, apply FPCA on the remaining $J - 1$ groups, where each function is approximated by a L -dimensional B-spline basis (Ramsay and Silverman 2005, chap. 8), calculate the squared error between h_{ijk} and the fitted values $\hat{g}_{ij}(t_k)$ (via (5)) on the validation set, and repeat this procedure J -times to compute the CV squared error. We calculate the CV errors over a grid of $M \leq L$ values and choose the pair with the lowest error. In general, we can select a different number of principal components for each random function that results in estimating matrices with nonsquare blocks. However, to simplify the computation in Algorithm 1, we use an identical number across $j \in V$ under the assumption that the corresponding covariance operators, $K_{jj}(s, t)$'s, share similar complexity structure.

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

4. Theoretical Properties

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

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

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

Here, $a_n \asymp b_n$ denotes $B \leq \inf_n |a_n/b_n| \leq \sup_n |a_n/b_n| \leq A$ for some positive constants A and B . The parameter β determines the decay rate of any decreasing sequence $\lambda_{j1} > \lambda_{j2} > \dots > \lambda_{jM}$ for $j \in V$ and $d_{jk}\lambda_{jk} = O(k)$ restricts the decay rate of eigen-gaps, $d_{jk}^{-1} \geq d_0 k^{-(\beta+1)}$ with some positive constant d_0 for $j \in V$, see also Bosq (2000) for more details. The parameter α controls the number of selected principal components that provide a reasonable approximation to the infinite-dimensional process. It is easy to see that larger values of β yield a faster decay rate, while increasing α results in a value for larger M .

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

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

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

$$P\left(|\hat{\sigma}_{jkk} - \sigma_{jkk}^*| \geq \delta\right) \leq C_2 \times \left\{ \exp(-C_1 n k^{-2} \delta^2) + \exp(-C_1 n k^{-(2+2\beta)} \delta) \right\}; \quad (16)$$

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

$$P\left(|\hat{\sigma}_{jlk m} - \sigma_{jlk m}^*| \geq \delta\right) \leq C_2 \times \exp\left\{-C_1 n (k+m)^{-(2+2\beta)} \delta^2\right\}. \quad (17)$$

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

$$P\left(|\hat{\sigma}_{jlk m} - \sigma_{jlk m}^*| \geq \delta\right) \leq C_2 \exp\left\{-C_1 n^{1-2\alpha(1+\beta)} \delta^2\right\}. \quad (18)$$

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

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

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

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

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

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

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

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

$$\max_{(j,l) \in E} \|\Omega_{jl,2}^{(M)}\|_F \leq O(n^{-\alpha\nu}). \quad (19)$$

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

$$\min_{(j,l) \in E} \|\Omega_{jl,1}^{(M)}\|_F \gg |E|^2 n^{\alpha(1-2\nu-\beta)}. \quad (20)$$

If we let $\mathbf{C}_{jl} = \text{cov}(\widetilde{\mathbf{a}}_{1j}, \widetilde{\mathbf{a}}_{1l} | \widetilde{\mathbf{a}}_{1k}, k \neq j, l)$, then $\int_{\mathcal{T}} \int_{\mathcal{T}} \{C_{jl}(s, t)\}^2 ds dt = \|\mathbf{C}_{jl}\|_F^2$ and $\Omega_{jl} = -\mathbf{C}_{jj}^{-1} \mathbf{C}_{jl} \mathbf{C}_{ll}^{-1}$ for each $(j, l) \in V^2$ with $j \neq l$. In this sense, **Condition 3** controls the effect of biases between the truncated and true processes and **Condition 4** requires the minimum signal strength for successful graph recovery to be much larger than $|E|^2 n^{\alpha(1-2\nu-\beta)}$. **Conditions 3** and **4** are crucial for obtaining the rate of convergence of $\|\Theta_{jl}^*\|_F$ for $(j, l) \in V^2$ and the equivalence between the truncated and true edge sets, respectively. See Lemmas 2 and 3 in the supplementary material for details. In particular, when $E \asymp pd$, we need ν to be large enough so as to satisfy **Condition 4**. In this case, $\max_{(j,l) \in E} \|\Omega_{jl,2}^{(k)}\|_F$ in **Condition 3** needs to

be small. We provide an example satisfying **Conditions 3** and **4** in **Appendix A.2**.

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

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

$$\|\Gamma_{SS}^* (\Gamma_{SS}^*)^{-1}\|_\infty^{(M^2)} \leq 1 - \eta. \quad (21)$$

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

Theorem 2. Suppose that **Conditions 2–5** hold, there exists some positive constant c_1 such that $M = c_1 n^\alpha$ and the bias term satisfies $\|\mathbf{B}^*\|_{\max}^{(M)} \leq \gamma_n \eta \kappa_{\Sigma^*}^{-2}/16$. Let $\widehat{\Theta}$ be the unique solution to the fglasso problem (9) with $\gamma_n = 16\eta^{-1} (c_1 C_1^{-1/2} n^{2\alpha(2+\beta)-1} (\tau \log c_1 n^\alpha p + \log C_2)^{1/2})$ for some $\tau > 2$.

Suppose the sample size n satisfies the lower bound

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

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

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

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

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

Frobenius norm within the true edge set $\Theta_{\min} \gtrsim \sqrt{\frac{\log p}{n^{1-2\alpha(2+\beta)}}}$.

When the maximum node degree $d = o(\sqrt{\frac{p^{1-2\alpha(2+\beta)}}{\log p}})$, 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 $g_{ij}(t) = \mathbf{s}(t)^T \delta_{ij}$, where $\mathbf{s}(t)$ was a five-dimensional Fourier basis function, and $\delta_{ij} \in \mathbb{R}^5$ was a mean zero Gaussian random vector. Hence, $\delta_i = (\delta_{i1}, \dots, \delta_{ip})^T \in \mathbb{R}^{5p}$ followed from a multivariate Gaussian distribution with covariance $\Sigma = \Theta^{-1}$. Different block sparsity patterns in the precision matrix, Θ , correspond to different conditional dependence structures. We considered three general structures.

1. Model 1: We generate a block banded matrix Θ with $\Theta_{jj} = \mathbf{I}_5$, $\Theta_{j,j-1} = \Theta_{j-1,j} = 0.4\mathbf{I}_5$, $\Theta_{j,j-2} = \Theta_{j-2,j} = 0.2\mathbf{I}_5$ for $j = 1, \dots, p$, and $\mathbf{0}$ at all other locations. Hence, only the adjacent two nodes are connected.
2. Model 2: For $j = 1, \dots, 10, 21, \dots, 30, \dots$, the corresponding submatrices in Θ came from Model 1 with $p = 10$, indicating every alternating block of 10 nodes are connected by Model 1. For $j = 11, \dots, 20, 31, \dots, 40, \dots$, we set $\Theta_{jj} = \mathbf{I}_5$, so the remaining nodes were fully isolated.
3. Model 3: We generate block sparse matrices without any special patterns. Specifically, we let each off-block-diagonal component in Θ be generated independently and equals $0.5\mathbf{I}_5$ with probability 0.1 or $\mathbf{0}$ with probability 0.9. We also set each block-diagonal element to be $\delta'\mathbf{I}_5$, where δ' is chosen to guarantee the positive definiteness of Θ .

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

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

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

To implement the fglasso, we must compute $\hat{\mathbf{a}}_{ij}$, the first M estimated principal component scores of g_{ij} . As mentioned previously, this is a standard problem and there are a number of possible approaches one could use for the calculation. In order to mimic a real-data setting, we chose to fit each function using a L -dimensional B-spline basis (rather than using the Fourier basis which was used to generate the data) and then compute $\hat{\mathbf{a}}_{ij}$ from the basis coefficients. We used 5-fold cross-validation to choose both L and M , the details of which are discussed in Section 3.3. Typically, $L = 5$ to 10 basis functions and $M = 4, 5$, or 6 principal components were selected in our simulations.

We compared fglasso to four competing methods. In the first-three methods, we fit the standard glasso T times, once on each time point, producing T different network structures. We then used one of three possible rules to combine the T networks into a single FGM. ALL involved identifying an edge if it was selected

Table 1. The mean area under the ROC curves.

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

NOTE: Standard errors are shown in parentheses.

in all T networks, NEVER identified an edge unless it appeared in none of the T networks, and HALF identifying an edge if it was selected in more than half of the T networks. The final approach, PCA, transformed the functional data into a standard format by computing the first estimated principal component score on each $g_{ij}(t)$ and then running the standard glasso on this data. The dimension of the B-spline basis function, L , was still selected by 5-fold cross-validation after setting $M = 1$. We also implemented the competing approaches using CLIME (Cai, Liu, and Luo 2011), rather than glasso. This gave only a very slight improvement over glasso, so we do not report the results here.

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

We considered different settings with $p = 50, 100, 150$, and ran each simulation 100 times. Figure 2 plots the median best ROC curves for each of the five comparison methods, respectively, for Models 1–3. The fglasso (black curve) clearly had the best overall performance in recovering support of the functional network. Table 1 provides the area under the ROC curves (average over the 100 simulation runs) along with standard errors. Larger numbers indicate superior estimates of the true network structure. Again, we see that the fglasso provided highly significant improvements in accuracy for graph recovery over the competing methods in all the settings we considered. Among the four competing approaches, HALF performed the best for Models 1 and 2, and PCA slightly outperformed others for Model 3.

5.2. EEG Data

We test the performance of the fglasso on the EEG dataset from the alcoholism study discussed in Section 1. The study consists of 122 subjects, 77 in the alcoholic group, and 45 in the control group. For each subject, voltage values were measured from 64 electrodes placed on the scalp, which were sampled at 256

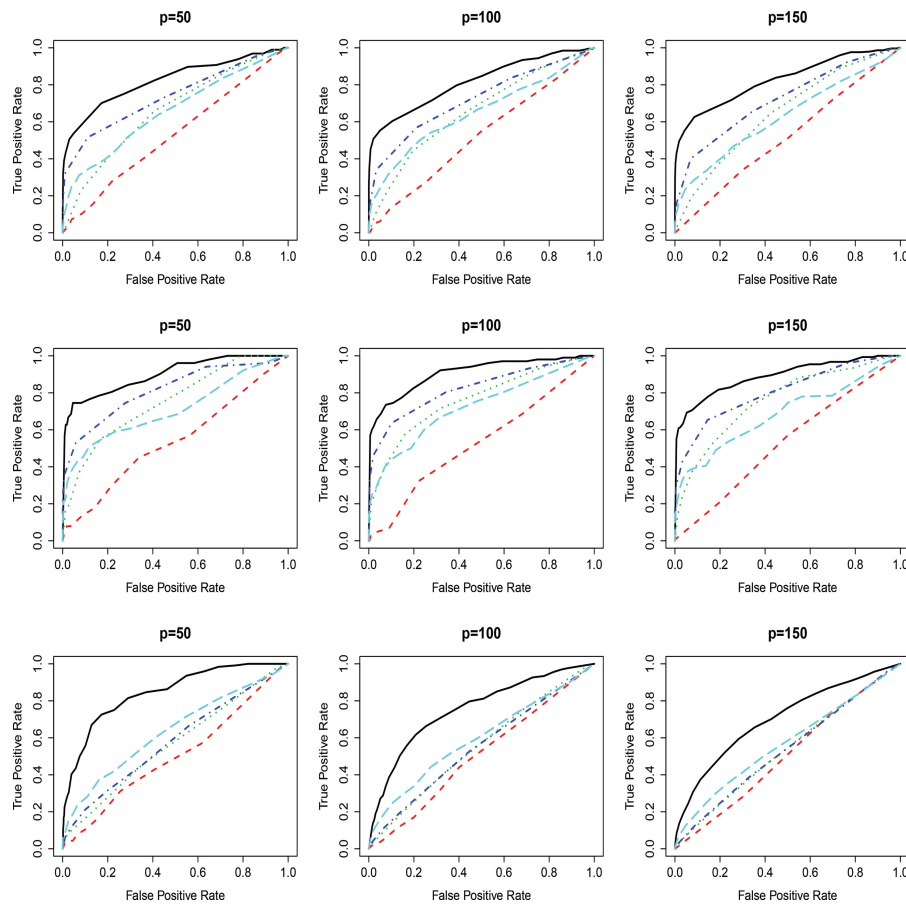


Figure 2. Model 1 (top row), Model 2 (middle row), and Model 3 (bottom row) for $p = 50, 100$, and 150 : Comparison of median estimated ROC curves over 100 simulation runs for fglasso (black solid), ALL (red dashed), NEVER (green dotted), HALF (blue dash dotted), and PCA (cyan long dashed).

Hz (3.9-ms epoch) for 1 sec. Each subject completed 120 trials under either a single stimulus or two stimuli. The electrodes were located at standard positions. Zhang et al. (1995) discuss the data collection process in detail. Li, Kim, and Altman (2010) and Zhou and Li (2014) 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 12.5 Hz, the case considered in Knyazev (2007), Hayden et al. (2006), and Zhu, Strawn, and Dunson (2016). Using four 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, Strawn, and Dunson (2016) discussed connectivity and asymmetry of electrodes selected from five different regions. Many authors used multiple samples per subject in order to obtain a sufficiently large sample, violating the independence assumption inherent in most methods. Following the analysis in Li, Kim, and Altman (2010), we only consider the average of all trials for each subject under the single stimulus condition. Thus, we have at most $n = 77$ observations and aim to estimate a network involving $p = 64$ nodes/electrodes.

We first performed a preprocessing step using the *eegfilt* function (part of the *eeglab* toolbox) to perform α band filtering on the signals. The fglasso was then fitted to the filtered data. The dimension of the B-spline basis function, L , was selected using the same cross-validation approach as for the simulation study. We set $M = 6$ for this data since six principal components already explained more than 90% of the variation in the signal

trajectories. Note that since our goal was to provide interpretable visualizations and investigate differences in brain connectivity between the alcoholic and control groups, we computed sparse networks with approximately 5% connected edges. To assess the variability in the fglasso fit, we performed a bootstrap procedure by randomly selecting n observations with replacement from the functional data, finding a tuning parameter γ_n to yield 5% sparsity level, applying the fglasso approach to the bootstrapped data, and repeating the above process 50 times. The “bootstrapped fglasso” was then constructed from the edges that occurred in at least 50% of the bootstrap replications.

Figure 3 plots the estimated network using the fglasso and the bootstrapped fglasso for both the alcoholic and the control groups. The bootstrapped fglasso estimated a sparser network with sparsity level 4.1% for the alcoholic group and 2.5% for the control group. We observe a few apparent patterns. First, electrodes from the frontal region are densely connected in both groups but the control group has increased connectivity relative to the alcoholic group. Second, the left central and parietal regions of the alcoholic group includes more connected edges. Third, electrodes from other regions of the scalp tend to be only sparsely connected. Finally, the fraction of black to red and blue edges provides a proxy for the level of confidence in any given estimated network. For the alcoholic group, this is fairly high, suggesting an accurately estimated network. However, the ratio is somewhat lower for the control group, suggesting a less accurate estimate. This is not surprising given the challenging

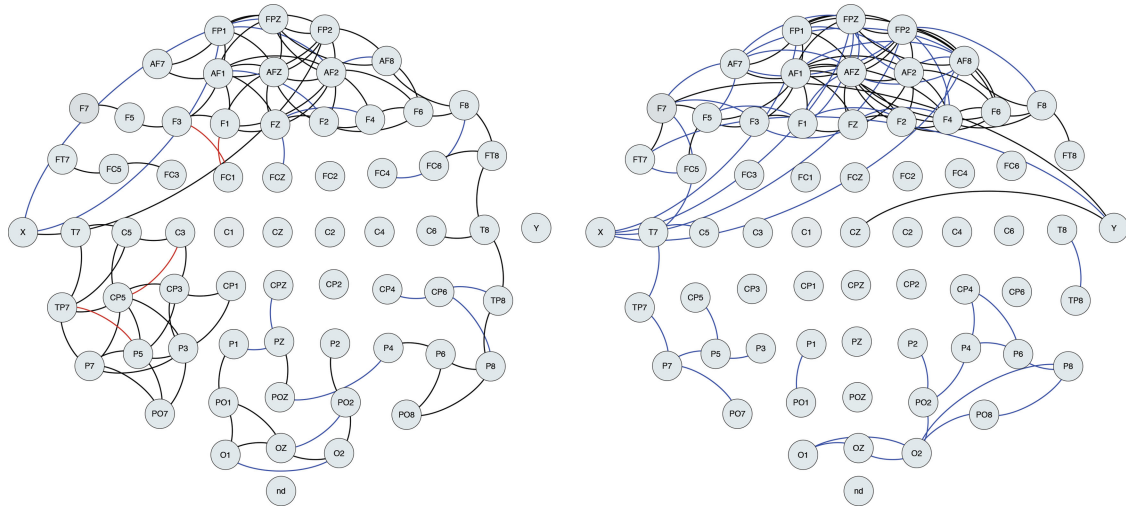


Figure 3. Left graph plots the estimated network for the alcoholic group and right graph plots the estimated network for the control group. Black lines denote edges identified by both fglasso and bootstrapped fglasso, blue lines denote edges identified by fglasso but not selected by the bootstrapped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

dataset with $p = 64$ nodes, corresponding to estimating graphs for $64 \times 6 = 384$ variables based on only 45 observations.

To identify edges that were clearly different between the two groups, we selected edges that occurred at least 50% more often in the bootstrap replications for one group relative to the other group. Figure 4 plots the edges only identified by either the alcoholic group or the control group. We observe that some edges in the left central and parietal regions were identified by the alcoholic group but missed by the control group, while one edge in the frontal region was identified by the control group but missed by the alcoholic group. Both findings provide confirmation for our informal observations from Figure 3.

6. Discussion

We conclude the article by discussing three extensions. Here, we have assumed that the trajectories of the functional variables are fully observed, although our results could be extended to the setting of densely observed curves under extra regularity and smoothness conditions. Hence, the first possible extension involves constructing a graphical model for sparse, irregular,

and noisy functional data, a common situation in functional data analysis (FDA). This extension could be achieved by performing FPCA on sparsely sampled functional data using either a mixed effects model (James, Hastie, and Sugar 2000) or a local smoother method (Yao, Muller, and Wang 2005), and then implementing the fglasso on the conditional expectations of the principal component scores.

Second, one referee was concerned that, since the inverse of the covariance operator of (g_1, \dots, g_p) is unbounded (Bosq 2000), then the minimum eigenvalue of the covariance matrix Σ^{*M} goes to zero as $M \rightarrow \infty$. He/she suggested that the true edge set E could instead be recovered based on the bounded inverse correlation operator, that is, using the block sparsity pattern in \mathbf{Q}^{*M} , the inverse correlation matrix of \mathbf{a}_i^M . This could be implemented using an alternative criterion by replacing the sample covariance matrix, \mathbf{S}^M , in (9) with the sample correlation matrix, \mathbf{R}^M . Specifically, we propose to solve the following optimization problem:

$$\hat{\mathbf{Q}}^M = \operatorname{argmax}_{\mathbf{Q}^M} \left\{ \log \det(\mathbf{Q}^M) - \operatorname{trace}(\mathbf{R}^M \mathbf{Q}^M) - \gamma_n \sum_{j \neq i} \|\mathbf{Q}_{ji}^M\|_F \right\}, \quad (24)$$

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

Third, the main theoretical limitation in this paper is to treat the dimension of the functional variables as approaching infinity rather than truly infinite dimensional functional objects

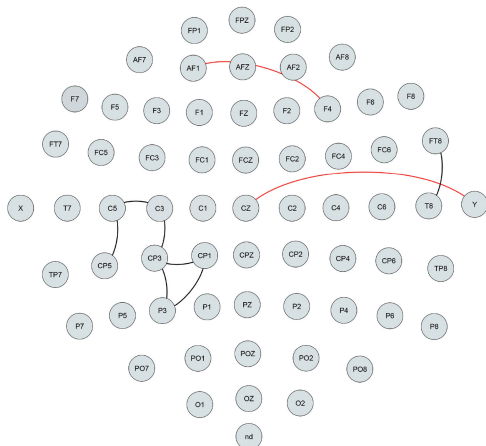


Figure 4. Black lines denote edges identified only by the alcoholic group and red lines denote edges identified only by the control group.

($M_n \rightarrow \infty$ rather than $M_n = \infty$). It is challenging, under the current framework, to relax the assumption $M_n < \infty$ to the fully functional situation with $M_n = \infty$, since we would need to write [Conditions 3](#) and [4](#) and the relevant proofs in terms of abstract functional analysis language in Hilbert space rather than the current compact matrix forms. We next present another way to understand the conditional dependence structure when $M_n < \infty$. From Peng et al. (2009), a_{ijk} can be expressed as

$$a_{ijk} = \sum_{l \neq j} \sum_{m=1}^{M_n} \beta_{jilm} a_{ilm} + \epsilon_{ijk}, \quad i = 1, \dots, j \in V, k = 1, \dots, M_n, \quad (25)$$

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

$$\beta_{jl} = -(\Theta_{jj}^{*M_n})^{-1} \Theta_{jl}^{*M_n}, \quad (j, l) \in V^2, l \neq j, \quad (26)$$

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

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

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

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

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

A. APPENDIX

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

A.1. Counterexample

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

A.2. Further Remarks on Some Regularity Conditions

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

Remark on [Condition 5](#): It is worth noting that Γ^* is the Hessian of $-\log \det(\Theta)$ evaluated at $\Theta = \Theta^*$. Hence, the entry $\Gamma_{(j,j')(l,l')}^{*(k,k')(m,m')}$ equals $\frac{\partial(-\log \det(\Theta))}{\partial \Theta_{jj'kk'} \partial \Theta_{ll'mm'}}$ evaluated at $\Theta = \Theta^*$, where $\Theta_{jj'kk'}$ is the (k, k') th entry of the $M \times M$ submatrix $\Theta_{jj'}$, $1 \leq j, j', l, l' \leq p$, $1 \leq k, k', m, m' \leq M$. Since \mathbf{a} is multivariate Gaussian, some standard calculations show that $\Gamma_{(j,j')(l,l')}^{*(k,k')(m,m')} = \text{cov}(a_{jk} a_{j'k'}, a_{lm} a_{l'm'})$. The Hessian of the negative log-determinant for the scalar data was studied in Ravikumar et al. (2011). We extend their work by viewing Γ^* , the Fisher information of the model, as an edge-based M^2 -block covariance matrix instead of the node-based covariance matrix Σ^* . For each $(j, j') \in V^2$, denote by $\mathbf{b}_{jj'} = \mathbf{a}_j \otimes \mathbf{a}_{j'} \in \mathbb{R}^{M^2}$ the edge-based vector, where $\mathbf{a}_j, \mathbf{a}_{j'}$ are the node-based vectors. Then we have $\Gamma_{(j,j')(l,l')}^{*(k,k')(m,m')} = E(\mathbf{b}_{jj'} \mathbf{b}_{ll'}^T)$, 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, (21) is equivalent to $\|E(\mathbf{b}_S \mathbf{b}_S^T) E(\mathbf{b}_S \mathbf{b}_S^T)^{-1}\|_\infty^{(M^2)} \leq 1 - \eta$, which bounds the effects of nonedges 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 .

Acknowledgment

The authors thank the Editor, the Associate Editor, and three referees for their useful comments and suggestions.

Funding

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

References

- Bosq, D. (2000), *Linear Processes in Function Spaces*, New York: Springer. [213,216,220]
- Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010), "Distributed Optimization and Statistical Learning Via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, 3, 1–122. [211,215]
- Cai, T., Liu, W., and Luo, X. (2011), "A Constrained l_1 Minimization Approach to Sparse Precision Matrix Estimation," *Journal of the American Statistical Association*, 106, 594–607. [211,218]
- Candes, E., and Tao, T. (2007), "The Dantzig Selection: Statistical Estimation When p is Much Larger than n ," *The Annals of Statistics*, 35, 2313–2351. [211]
- Danaher, P., Wang, P., and Witten, D. (2014), "The Joint Graphical Lasso for Inverse Covariance Estimation Across Multiple Classes," *Journal of the Royal Statistical Society, Series B*, 76, 373–397. [211,214,215]

- Friedman, J., Hastie, T., and Tibshirani, R. (2007), "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics*, 2, 432–441. [211,215]
- Hayden, E., Wiegand, R., Meyer, E., Bauer, L., O'Connor, S.J., Nurnberger, J.I., Chorlian, D.B., Porjesz, B., and Begleiter, H. (2006), "Patterns of Regional Brain Activity in Alcohol-dependent Subjects," *Alcoholism: Clinical and Experimental Research*, 30, 1986–1991. [212,219]
- Hsing, T., and Eubank, R. (2015), *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, West Sussex, UK: Wiley. [213]
- Inger, L. (1997), "Statistical Mechanics of Neocortical Interactions: Canonical Momenta Indicators of Electroencephalography," *Physical Review E*, 55, 4578–4593. [212]
- James, G., Hastie, T., and Sugar, C. (2000), "Principal Component Models for Sparse Functional Data," *Biometrika*, 87, 587–602. [220]
- Knyazev, G. (2007), "Motivation, Emotion, and Their Inhibitory Control Mirrored in Brain Oscillations," *Neuroscience Biobehavioral Reviews*, 31, 377–395. [219]
- Kolar, M., and Xing, E. (2011), "On Time Varying Undirected Graphs," *Proceedings of Machine Learning Research*, 15, 407–415. [211,213]
- Lam, C., and Fan, J. (2009), "Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation," *The Annals of Statistics*, 37, 4254–4278. [211]
- Li, B., Kim, M., and Altman, N. (2010), "On Dimension Folding of Matrix-or Array-valued Statistical Objects," *The Annals of Statistics*, 38, 1094–1121. [219]
- Mazumder, R., and Hastie, T. (2012a), "The Graphical Lasso: New Insights and Alternatives," *Electronic Journal of Statistics*, 6, 2125–2149. [211]
- (2012b), "Exact Covariance Thresholding Into Connected Components for Large-scale Graphical Lasso," *Journal of Machine Learning Research*, 13, 781–794. [211,215]
- Meinshausen, N., and Bühlmann, P. (2006), "High Dimensional Graphs and Variable Selection with Lasso," *The Annals of Statistics*, 34, 1436–1462. [211]
- (2010), "Stability Selection," *Journal of the Royal Statistical Society, Series B*, 72, 417–473. [216]
- Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial Correlation Estimation by Joint Sparse Regression Models," *Journal of the American Statistical Association*, 104, 735–746. [211,221]
- Qiu, H., Han, F., Liu, H., and Caffo, B. (2016), "Joint Estimation of Multiple Graphical Models from High Dimensional Dependent Data," *Journal of the Royal Statistical Society, Series B*, 78, 487–504. [213]
- Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis* (2nd ed.), New York: Springer. [216]
- Ravikumar, P., Wainwright, M., Raskutti, G., and Yu, B. (2011), "High-dimensional Covariance Estimation by Minimizing l_1 -penalized Log-determinant Divergence," *Electronic Journal of Statistics*, 5, 935–980. [211,217,221]
- Rice, J., and Silverman, B. (1991), "Estimating the Mean and Covariance Structure Nonparametrically When the Data are Curves," *Journal of the Royal Statistical Society, Series B*, 53, 233–243. [216]
- Storey, J. D., Xiao, W., Leek, T. J., Tompkins, R. G., and Davis, R. W. (2005), "Significance Analysis of Time Course Microarray eXperiments," *Proceedings of the National Academy of Sciences of the United States of America*, 102, 12837–12842. [212]
- Sun, T., and Zhang, C. (2013), "Sparse Matrix Inversion with Scaled Lasso," *Journal of Machine Learning Research*, 14, 3385–3418. [211]
- Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [211]
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness Via the Fused Lasso," *Journal of the Royal Statistical Society, Series B*, 67, 91–108. [215]
- Tseng, P. (2001), "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization," *Journal of Optimization Theory and Applications*, 109, 475–494. [215]
- Witten, D., Friedman, J., and Simon, N. (2011), "New Insights and Faster Computations for the Graphical Lasso," *Journal of Computational and Graphical Statistics*, 20, 892–900. [211,213,215]
- Yao, F., Muller, H., and Wang, J. L. (2005), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590. [216,220]
- Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society, Series B*, 68, 49–67. [213,214]
- (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [211,213]
- Zhang, X., Begleiter, B., Porjesz, B., Wang, W., and Litke, A. (1995), "Event Related Potentials During Object Recognition Tasks," *Brain Research Bulletin*, 38, 531–538. [212,219]
- Zhou, H., and Li, L. (2014), "Regularized matrix regression," *Journal of the Royal Statistical Society, Series B*, 76, 463–483. [219]
- Zhou, S., Lafferty, J., and Wasserman, L. (2010), "Time Varying Undirected Graphs," *Machine Learning Journal*, 80, 295–319. [211,213]
- Zhu, H., Strawn, N., and Dunson, D. (2016), "Bayesian Graphical Models for Multivariate Functional Data," *Journal of Machine Learning Research*, 17, 1–27. [213,219]
- Zhu, Y., Shen, X., and Pan, W. (2014), "Structural Pursuit Over Multiple Undirected Graphs," *Journal of the American Statistical Association*, 109, 1683–1696. [211,215]