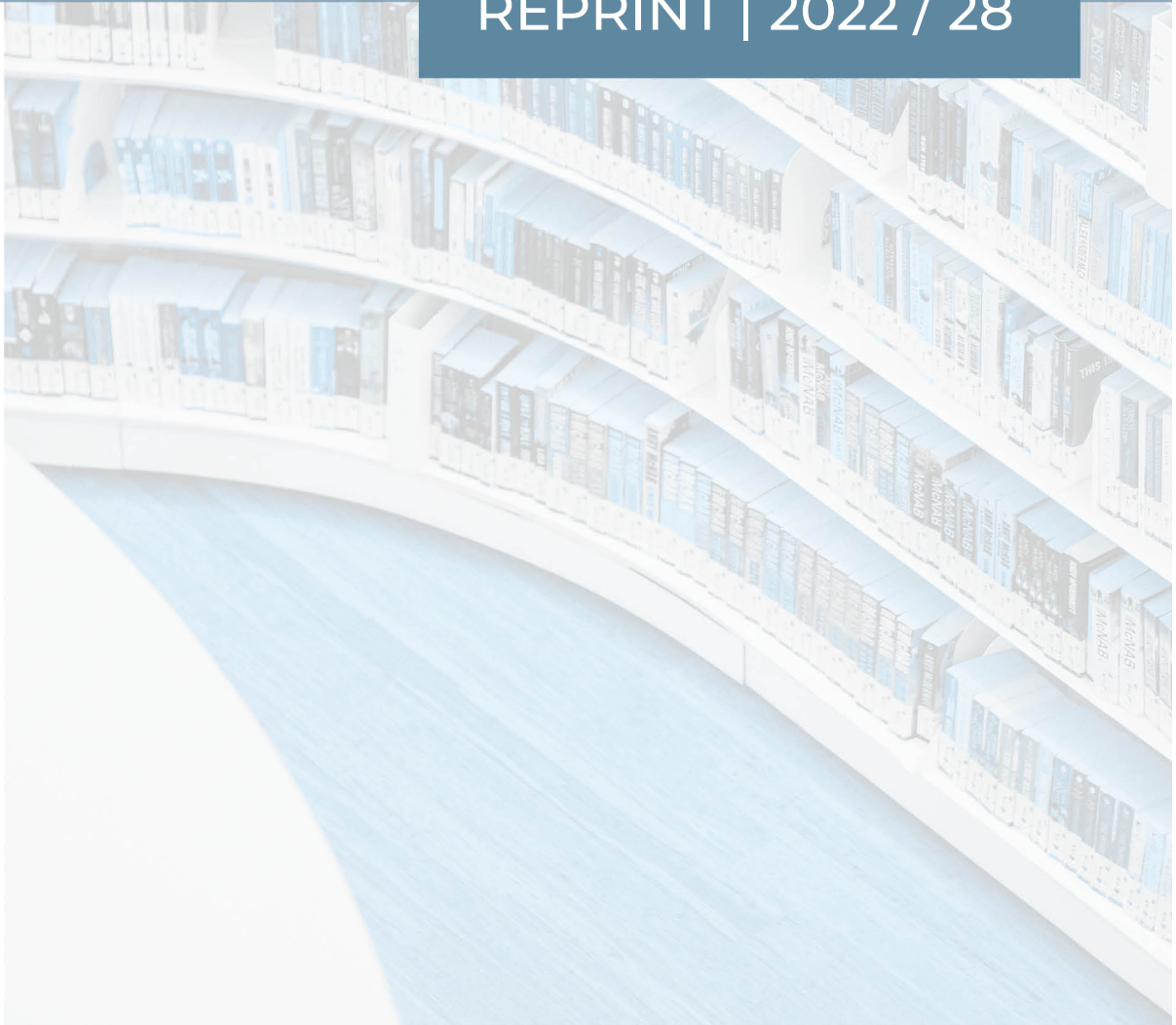


LINEAR MANIFOLD MODELING AND GRAPH ESTIMATION BASED ON MULTIVARIATE FUNCTIONAL DATA WITH DIFFERENT COARSENESS SCALES

Eugen Pircalabelu, Gerda Claeskens

REPRINT | 2022 / 28



ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

<https://uclouvain.be/en/research-institutes/lidam/isba/publication.html>

Linear manifold modeling and graph estimation based on multivariate functional data with different coarseness scales

Eugen Pircalabelu *

Institute of Statistics, Biostatistics and Actuarial Sciences, LIDAM, UCLouvain

Gerda Claeskens

ORSTAT and Leuven Statistics Research Center, KU Leuven

September 6, 2022

Abstract

We develop a high-dimensional graphical modeling approach for functional data where the number of functions exceeds the available sample size. This is accomplished by proposing a sparse estimator for a concentration matrix when identifying linear manifolds. As such, the procedure extends the ideas of the manifold representation for functional data to high-dimensional settings where the number of functions is larger than the sample size. By working in a penalized setting it enriches the functional data framework by estimating sparse undirected graphs that show how functional nodes connect to other functional nodes. The procedure allows multiple coarseness scales to be present in the data and proposes a simultaneous estimation of several related graphs. Its performance is illustrated using a real-life fMRI dataset and with simulated data.

Keywords: Multiscale functional data, Graphical lasso, Joint estimation, Group penalty

1 Introduction

Multivariate functional data are nowadays increasingly encountered in neurological research, due to the development of non-invasive techniques such as resonance imaging which facilitate the collection of large amounts of data (Lazar, 2008; Huettel et al., 2009).

*The authors gratefully acknowledge the computational resources provided by the supercomputing facilities of the UCLouvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region. We acknowledge as well the support of the KU Leuven Research Fund C1-project C16/20/002.

This work is motivated by a functional magnetic resonance imaging (fMRI) study where brain activity measurements (as represented by the BOLD signals) at voxel level were recorded for six healthy subjects, in an exploratory study where the subjects viewed pictures of faces, cats and five different objects: houses, chairs, scissors, shoes, and bottles. For each stimulus a total of 108 fMRI measurements are recorded. The data are described in Haxby et al. (2001). Using the multiscale functional brain parcellation proposed in Bellec et al. (2010), the fMRI images were then first segmented into $M^1 = 64$ regions of interest (ROIs, i.e., sets of voxels that form non-overlapping parts of the brain) and then into $M^2 = 122$ ROIs. We call the 64 regions as coarse scale regions and the 122 regions as finer scale regions, as each of the finer partitions at scale 2 has precisely one parent in the set of M^1 ROIs. It is of interest to study the relation between these 64 and 122 ROIs or put differently, between the ROIs at each of the coarseness scales.

For the fMRI example, it is of interest to determine if ROIs are ‘functionally connected’, by which in brain mapping terminology it is understood that distinct ROIs in the brain contribute together to perform a functional task such as hearing, seeing, memory, etc. For vector valued data, this is generally done using a form of correlation analysis, but in this work we address that data for each ROI are themselves functional objects and thus opt to model the neurological signal using functional approaches in order to describe the smooth functional behavior of the generating process. Advantages of using functional data analysis (FDA), have been illustrated in several works such as Ullah and Finch (2013) for general biomedical data, Viviani et al. (2005), Tian (2010), Mejia et al. (2017), Lila et al. (2016) and Jiang et al. (2016) for fMRI data, among many other works. See also Wang et al. (2016) for a comprehensive review on FDA.

For multivariate functional data, Chiou and Müller (2014) proposed a linear manifold model to estimate a dynamic dependence structure between the functional components of the multivariate process. The model is characterized by a set of time-varying linear combinations of the component trajectories (subject to certain constraints) and requires information about the covariance structure of a set of random coefficients obtained from the Karhunen-Loève (KL) decomposition. An estimator for the covariance structure is not directly obtainable in that context when $n \ll M^c$ for $c = 1, \dots, C$. In this paper we develop a novel statistical framework that aims at discovering the dynamic dependence structure between the functional components of the multivariate process by (1) estimating time-varying linear combinations of the component trajectories as in Chiou and Müller (2014) when (2) using a high-dimensional graphical modeling approach for functional data where the number of functions might greatly exceed the available sample size. As will be detailed later, the linear manifold builds data-driven linear combinations that quantify which ROIs show more similar/dissimilar patterns of activation across time. It puts into balance the evolution of all ROIs with the purpose of signalling out the key ones and to do so it uses covariance information related to the joint behavior of functions. As such, we propose a sparse estimator for a concentration matrix (i.e. the inverse of the covariance matrix) and we introduce for the linear manifold model, techniques from the ℓ_1 -penalized Gaussian graphical modeling literature treated in the works of Yuan and Lin (2007), Banerjee et al. (2008), Friedman et al. (2008), Bickel and Levina (2008), Cai et al. (2010) and Leng and Tang (2012) amongst many others.

Our purpose is to *simultaneously* estimate linear manifolds based on time-varying linear combinations of the multivariate trajectories and at the same time model the dependence among the functional components with the help of *sparse* graphs that share similarities across the scales. Thus, the linear combinations of functions in our proposed procedure are constructed based on a quantification of the covariance information that describes the dependence between multiple functional objects. To this end, the random coefficients obtained from the KL decomposition are mapped to nodes in a graph (one node per coefficient) for which, based on data we estimate edges connecting two nodes at a time. If two nodes are connected by an edge it denotes that they are, conditionally on all remaining nodes, dependent. Since the nodes in the graph relate to ROIs it will also imply that ROIs are dependent. However, for interpretation purposes the added value of the graphical model procedures to multivariate functional data analysis is their increased ability to represent a large number of conditional *independence* relations with a graph. If two nodes a and b (a node is attributed to each coefficient) are *not* connected by an edge in the graph, then there exist other nodes which when conditioned upon, make nodes a and b conditionally independent. The graph represents thus, a compact and easy way to visualize a list of conditional independence relations between the components of multivariate functional data. The sparser the graph is, the more conditional independencies exist between the functional components and the easier the graph is to interpret since only strongly dependent nodes are connected by edges. Since sparsity is in itself useful for interpretability, the models we propose here shrink sufficiently small coefficients (that correspond to edges in the graph) directly to zero by using sufficiently large penalties, thus enforcing sparse graphs. For the estimation of an undirected graphical model for functional data we take a similar approach as in Qiao et al. (2015), Qiao et al. (2019) and Zapata et al. (2022) where a multivariate normality assumption is used for modeling the pooled vector of random coefficients coming from the Karhunen-Loève expansions of each function.

By using penalized procedures to model the dependence between the functional nodes of graphs, we also extend the linear manifold procedure to high-dimensional settings where M , the number of nodes, is larger than the available sample size n . Note that the above fMRI example, used throughout the manuscript, fits perfectly in the high-dimensional framework as one has $M^1 = 64$ (coarse scale) and $M^2 = 122$ (fine scale) ROIs, while $n = 6$. We extend to the case where the simultaneous joint estimation of several related graphs is desired, due to multiple coarseness scales in the data. By ‘related graphs’ we understand graphs that are describing the same underlying connectivity pattern, with data available at multiple coarseness scales as in Piricalabelu et al. (2016). More details are offered in Section 2.

By using graph-informed linear manifolds one can better understand the dynamics related to the dependence structure (through the sparse graph estimation step) between ROIs of the brain and since it is generally low-dimensional (as only the first K most relevant linear combinations are considered for interpretation) it also leads to simpler representations of the multivariate component trajectories as linear combinations. These linear combinations, obtained under integrated variance and orthonormality constraints, define data-driven time-varying relationships between the different ROIs. For each such linear combination, the relationships between ROIs are interpreted relative to one another. Such interpretability benefits for fMRI settings are illustrated in Section 5.

The paper is structured as follows: in Section 2 we set up the model and tackle the estimation aspects. In Section 3 we complement the estimation strategy by analytical derivations and an algorithmic implementation. In Section 4 the theoretical properties of the procedure are investigated and in Section 5 its application to the fMRI data is performed. A controlled, simulation-based study is proposed in Section 6, while Section 7 extends the model to explicitly account for dependencies across scales. We close in Section 8 with a discussion on the method.

2 Graph-informed linear manifold

We use in this section the framework of Chiou and Müller (2014) where we keep only the essential information and for ease of readability the notation is kept similar where possible.

Suppose one has a sample of size n of observed trajectories $\{X_{1i}^c, \dots, X_{M^c i}^c\}$ for $i = 1, \dots, n$, from M^c -component functions $\{U_1^c, \dots, U_{M^c}^c\}$ at each coarseness scale c . For this application we assume a total of C coarseness scales is available for analysis and at each scale the number of available functions might differ, hence, we use the superscript c . Each function U_j^c with $j = 1, \dots, M^c$ and $c = 1, \dots, C$ is assumed to lie in the Hilbert space $L^2(\mathcal{T})$ of squared integrable functions for a compact domain $\mathcal{T}$. At each time point $t = 1, \dots, T$ using the observed trajectories at each scale c , we use the estimators $\hat{\mu}_j^c(t) = \frac{1}{n} \sum_{i=1}^n X_{ji}^c(t)$ and $\hat{G}_j^c(t_1, t_2) = \frac{1}{n} \sum_{i=1}^n (X_{ji}^c(t_1) - \hat{\mu}_j^c(t_1))(X_{ji}^c(t_2) - \hat{\mu}_j^c(t_2))$ for the mean and covariance function (for which we assume their existence) of each U_j^c : $\mu_j^c(t) = E(U_j^c(t))$ and $G_j^c(t, s) = \text{cov}(U_j^c(t), U_j^c(s)) = \sum_{l=1}^{\infty} \lambda_{jl}^c \phi_{jl}^c(t) \phi_{jl}^c(s)$, subject to the constraint that $\int_{\mathcal{T}} \phi_{jl}^c(t) \phi_{jm}^c(t) dt = \delta_{lm}$, with the Kronecker delta $\delta_{lm} = 1$ if $l = m$ and $\delta_{lm} = 0$ otherwise. As the notation suggests, we assume in this manuscript that the covariance function $\text{cov}(U_j^c(t), U_j^c(s))$ can be expanded at each scale and for each component j into an orthonormal system based on eigenfunctions $\phi_{jl}^c(t)$ and eigenvalues λ_{jl}^c . To get estimators $\hat{\lambda}_{jl}^c$ and $\hat{\phi}_{jl}^c(t)$, one can solve the eigenequation

$$\int_{\mathcal{T}} \hat{G}_j^c(t_1, t_2) \hat{\phi}_{jl}^c(t_1) dt_1 = \hat{\lambda}_{jl}^c \hat{\phi}_{jl}^c(t_2) \quad (1)$$

such that $\int_{\mathcal{T}} \hat{\phi}_{jl}^c(t) \hat{\phi}_{jm}^c(t) dt = \delta_{lm}$ and where $\hat{\lambda}_{jl}^c \geq 0$.

For FDA the eigenfunctions ϕ_{jl}^c are of crucial importance, as for each centered process $\tilde{U}_j^c(t) = U_j^c(t) - \mu_j^c(t)$ one has in light of the Karhunen-Loève expansion

$$\tilde{U}_j^c(t) = U_j^c(t) - \mu_j^c(t) = \sum_{l=1}^{\infty} \xi_{jl}^c \phi_{jl}^c(t) \text{ with } \xi_{jl}^c = \int_{\mathcal{T}} (U_j^c(t) - \mu_j^c(t)) \phi_{jl}^c(t) dt,$$

where ξ_{jl}^c are random coefficients, uncorelated and for which $E(\xi_{jl}^c) = 0$ and $\text{var}(\xi_{jl}^c) = \lambda_{jl}^c$. Using further the estimators $\hat{\phi}_{jl}^c(t)$ from (1) one can obtain the estimators

$$\hat{\xi}_{jli}^c = \int_{\mathcal{T}} (X_{ji}^c(t) - \hat{\mu}_j^c(t)) \hat{\phi}_{jl}^c(t) dt. \quad (2)$$

As the notation suggests, the estimators of the random coefficients in (2) at each scale c and for each function j are subject specific. Hence in this work, by pooling together

the information across subjects, we will use the estimates to further create an empirical covariance matrix at each scale, $\mathbf{S}^c$, for $c = 1, \dots, C$ based on which C graphs will be estimated using the optimization problem proposed in (5) later in this section.

Following Chiou and Müller (2014) we propose to model the dynamic dependence between the M^c functions at each scale c by using K linear combinations of the functions at each scale and at each time point. Using the centered process $\tilde{U}_j^c(t)$, we define at coarseness scale c , the k th linear combination as $W_k^c(t) = \sum_{j=1}^{M^c} \beta_{kj}^c(t) \tilde{U}_j^c(t)$, where $k = 1, \dots, K$ and where $\beta_{kj}^c(t)$ are unknown time-varying coefficients. These coefficients quantify the time-dependent contribution of each centered process to the linear combination, and have the purpose to determine and quantify what processes are more important than others across time. By construction, for each k th linear combination the integrated variance is assumed to vanish i.e. $\int_{\mathcal{T}} \text{var}(W_k^c(t)) dt = 0$ (implying also that $W_k^c(t) = 0$ almost everywhere on the domain $\mathcal{T}$) and for identifiability reasons, orthonormality of the functions is imposed i.e. $\sum_{j=1}^{M^c} \int_{\mathcal{T}} \beta_{kj}^c(t) \beta_{lj}^c(t) dt = \delta_{kl}$, with the Kronecker delta $\delta_{kl} = 1$ if $k = l$ and $\delta_{kl} = 0$ otherwise, for all $k, l = 1, \dots, K$.

In practice $\tilde{U}_j^c(t)$ is approximated as $\tilde{U}_j^c(t) \approx \sum_{l=1}^{L_{j;c}} \xi_{jl}^c \phi_{jl}^c(t)$ by assuming that the underlying process $\tilde{U}_j^c(t)$ can be well represented by a finite number of components $L_{j;c}$. A second approximation can be used by assuming that each function $\beta_{kj}^c(t)$ can be well approximated by using $P_{j;c}$ known basis functions $\psi_{jp}^c(t)$ with $p = 1, \dots, P_{j;c}$, leading to the overall approximation of the k th linear combination as

$$W_k^c(t) \approx \sum_{j=1}^{M^c} \sum_{p=1}^{P_{j;c}} \sum_{l=1}^{L_{j;c}} b_{kjp}^c \psi_{jp}^c(t) \xi_{jl}^c \phi_{jl}^c(t) \quad (3)$$

for some unknown coefficients b_{kjp}^c . We approximate next, in light of (3), the integrated variance of the time-varying linear combinations of the functions as

$$\int_{\mathcal{T}} \text{var}(W_k^c(t)) dt \approx \sum_{r=1}^{M^c} \sum_{s=1}^{M^c} \sum_{p=1}^{P_{r;c}} \sum_{q=1}^{P_{s;c}} b_{krp}^c b_{ksq}^c \left\{ \sum_{l=1}^{L_{r;c}} \sum_{m=1}^{L_{s;c}} E(\xi_{rl}^c \xi_{sm}^c) \int_{\mathcal{T}} \psi_{rp}^c(t) \psi_{sq}^c(t) \phi_{rl}^c(t) \phi_{sm}^c(t) dt \right\}. \quad (4)$$

Since $E(\xi_{rl}^c \xi_{sm}^c)$ is the covariance between two components of the vector $\boldsymbol{\xi}^c$ formed by the random coefficients at a given scale c , where $\boldsymbol{\xi}^c = (\xi_{11}^c, \dots, \xi_{M^c 1}^c, \dots, \xi_{M^c L_{M^c;c}}^c)^\top$ is a vector of length $\sum_{j=1}^{M^c} L_{j;c}$ (i.e., the total number of finite components needed in the approximation across all functions at scale c), under the assumption that $\tilde{U}_j^c(t)$ is a Gaussian process indexed by both j and t , then $\boldsymbol{\xi}^c \sim N(\mathbf{0}, \boldsymbol{\Sigma}^c)$. As such, we propose the estimator $\hat{E}(\xi_{rl}^c \xi_{sm}^c) = (\hat{\boldsymbol{\Theta}}^c)_{rl,sm}^{-1}$, where by $(\hat{\boldsymbol{\Theta}}^c)_{rl,sm}^{-1}$ with $r, s = 1, \dots, M^c$, $l = 1, \dots, L_{r;c}$ and $m = 1, \dots, L_{s;c}$, we denote the entry in the matrix $(\hat{\boldsymbol{\Theta}}^c)^{-1}$ from the row corresponding to the l th finite truncation point of the r th function and the column corresponding to the m th finite truncation point of the s th function, all from the c th coarseness scale.

Under a multivariate normality assumption for the vectors $\boldsymbol{\xi}^1, \dots, \boldsymbol{\xi}^C$, we propose the

estimators $\hat{\Theta}^1, \dots, \hat{\Theta}^C$ as the minimizers:

$$\begin{aligned} \{\hat{\Theta}^1, \dots, \hat{\Theta}^C\} = \arg \min_{\{\Theta^1, \dots, \Theta^C\}} & \left[\sum_{c=1}^C \left\{ \text{trace}(\mathbf{S}^c \Theta^c) - \log \det \Theta^c + \rho_{n1} \sum_{a \neq b \in V^c} |\theta_{ab}^c| \right\} \right. \\ & \left. + \rho_{n2} \sum_{h=1}^{M^1} \sum_{c=1}^C \left(\sum_{a \neq b \in V^c} (\theta_{ab;h}^c)^2 \right)^{1/2} \right], \end{aligned} \quad (5)$$

where based on (2) we have $\hat{\xi}_i^c = (\hat{\xi}_{11i}^c, \dots, \hat{\xi}_{M^c 1i}^c, \dots, \hat{\xi}_{M^c L_{M^c i}}^c)^\top$, $\Theta^c = (\Sigma^c)^{-1}$ and $\mathbf{S}^c = (1/n) \sum_{i=1}^n \hat{\xi}_i^c (\hat{\xi}_i^c)^\top$ is an empirical covariance matrix. The set $V^c = \{1, \dots, a, b, \dots, \#\xi^c\}$, represents the set of nodes associated with the components of ξ^c (one node for each component) and where $\#\xi^c$ represents the number of elements in the vector. Since C coarseness levels are used, we optimize in (5) an ‘across’-scales penalized negative log-likelihood where:

- ρ_{n1} & ρ_{n2} are regularization constants greater than 0, which can depend on n ;
- θ_{ab}^c refers to a generic entry in the matrix Θ^c on position (a, b) ;
- $\theta_{ab;h}^c$ refers to an entry in the matrix Θ^c on position (a, b) that corresponds to a certain coarse ROI h (i.e. the entries in the matrix Θ^c that pertain to modeling the inverse covariances of the vector $(\xi_{h1}^c, \dots, \xi_{hL_h}^c)^\top$ at scale 1 and $(\xi_{h'1}^c, \dots, \xi_{h'L_{h'}}^c, \xi_{h''1}^c, \dots, \xi_{h''L_{h''}}^c)^\top$ at scale $c > 1$ and where h' and h'' denote finer partitions of h). For simplicity of illustration just two finer partitions (h' and h'') are used here.

There is a well known link between Gaussian graphical models and concentration (or inverse covariance) matrices, as there is a direct correspondence between the conditional independence statements related to the components of a multivariate normal vector and a graphical structure $G(E_u, V)$, defined based on a set of undirected edges E_u that connect the nodes in the set V . The set E_u contains all pairs of nodes (a, b) that are linked by an edge, where $a \neq b \in V$. As such, if one associates to each component of ξ^c a node in the set V^c , Lauritzen (1996) showed that the pair $(a, b) \notin E_u$ (i.e. there is no edge between nodes a and b) if component ξ_a^c is independent of ξ_b^c conditionally on all remaining components in the vector (denoted by $\xi_a^c \perp \xi_b^c | \xi_{\{1, \dots, \#\xi^c\} \setminus \{a, b\}}^c$) which happens if and only if $\theta_{ab}^c = \theta_{ba}^c = 0$. This property implies that one can estimate the structure of the graph, which gives information about all the conditional dependencies that hold between components of ξ^c , by estimating the non-zero elements in the inverse covariance matrix.

Note also that in (5) the optimization function encompasses the information available at all scales, while at the same time it enforces sparsity and accounts for the grouping structure of the nodes in the estimated graphs. Sparsity is enforced through the term $\sum_{a \neq b \in V^c} |\theta_{ab}^c|$ which has the purpose of shrinking small entries in the corresponding concentration matrices to zero. Through the term $\sum_{h=1}^{M^1} \{(\sum_{a \neq b \in V^c} (\theta_{ab;h}^c)^2)\}^{1/2}$ the function accounts in the estimated graphs that some regions at finer scales are partitions of the coarse regions from the first scale. This term corresponds to a grouping penalty and it accounts for the existence of a grouping structure and similarity between certain functions at finer scales since they are all related to large, coarse ROIs by being splits thereof.

To illustrate the grouping structure considered in this problem, Figure 1 shows a toy example of four groups (two per each scale) informed by the functional Karhunen-Loève

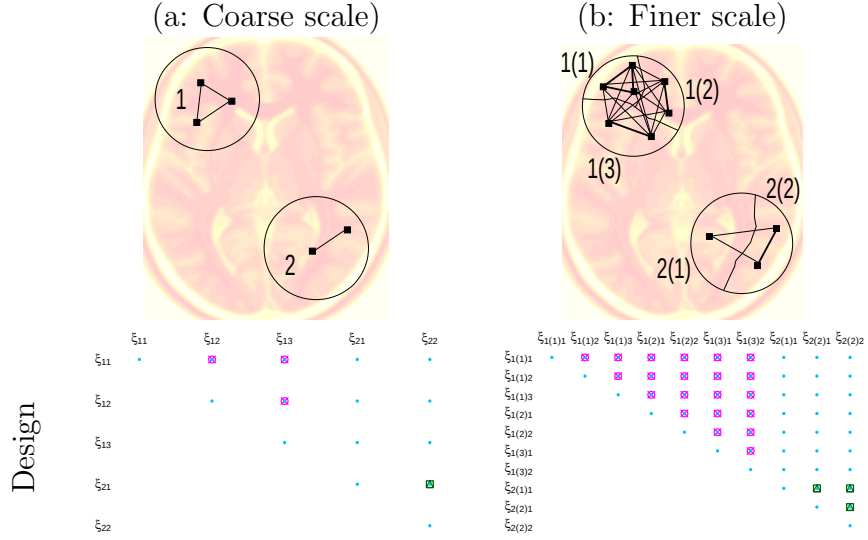


Figure 1: Toy example to illustrate functional decompositions and design-informed grouping where common symbols denote a distinct group of parameters. Panel (a) on the top row, presents three edges between ξ variables corresponding to a coarse ROI 1 and one edge between ξ variables corresponding to a coarse ROI 2. These edges form two distinct groups at scale 1. Panel (b) on the top row presents edges between ξ variables corresponding to three partitions of the coarse ROI 1 from panel (a) and to two partitions of the coarse ROI 2. Two distinct groups are formed at scale 2.

decomposition and the design of the study. In the graph in panel (a) three ξ variables (denoted as ξ_{11}, ξ_{12} and ξ_{13}) are linked by edges in a graph. They all pertain to the decomposition of the signal of a coarse ROI denoted as ‘1’ and all (three) θ coefficients corresponding to the three edges form a group. For ROI ‘2’ two ξ (denoted as ξ_{21} and ξ_{21}) variables are used in the decomposition, and the associated θ coefficient (only one in this case) corresponding to the edge between ξ_{21} and ξ_{22} forms a group at scale 1. In panel (b) of Figure 1 we represent three finer partitions of the same ROI 1 where we have three, two and two ξ variables for the partitions. This implies that the Karhunen-Loève decomposition uses three terms for the first partition, two for the second partition and again two for the last partition. For ROI 2 two finer partitions are illustrated. For this scale due to the splitting of regions, the grouping considers all θ parameters that pertain to splits of a certain ROI, as a homogeneous group. As such, at scale 2 one allows also for two groups corresponding to the coarse ROIs.

A few connections to the existing literature can be made at this point. For a setting where $C = 1$ and $\rho_{n2} = 0$, our procedure reduces to the functional graphical lasso procedure of Qiao et al. (2015). In Qiao et al. (2019) a group penalty of the form $\rho_n \sum_{a \neq b} \|\Theta_{a,b}^L\|_F$ has been used where ρ_n is a regularization parameter, L represents the number of components one retains to approximate the function in the Karhunen-Loève decomposition and a, b represent two different nodes mapping the two functions in the graph. This group penalty has as purpose to force block sparsity in the sense that all elements corresponding to an

edge between nodes a and b are forced to be all non-zero (in case an edge is estimated) or all zero (in case no edge is estimated between nodes a and b). Differently to their approach, by using the proposed two-term regularization, our procedure also allows for sparsity within a group. All relevant partitions of a coarse region h are treated as a group, but at the same time due to the presence of an ℓ_1 term, it allows for sparse estimators within a group.

In a partially separable framework, Zapata et al. (2022) explored a grouping penalty that groups together elements across the different functions, using the joint group graphical lasso procedure of Danaher et al. (2014) to perform the optimization. It should be pointed out that our estimator (see Section 3), although similar in spirit to that proposed by Danaher et al. (2014), is quite different due to the fact that the grouping in the aforementioned work is defined across scales whereas for our estimator, within each scale, multiple groups of unequal sizes are defined by the design of the problem. Our procedure departs from that of Zapata et al. (2022) as the groups are informed by both the functional decomposition and the splitting design. Moreover, the authors estimate L precision matrices, one for each truncation point in the Karhunen-Loève expansion and using an ‘OR’ rule similar to Meinshausen and Bühlmann (2006) they define an edge set linking all estimated precision matrices. Quite differently, our procedure does not make a separability assumption and estimates one full precision matrix per scale using the information contained in $\boldsymbol{\xi}^c$.

Letting

$$a_{rs,pq}^c = \sum_{l=1}^{L_{r;c}} \sum_{m=1}^{L_{s;c}} \mathbb{E}(\xi_{rl}^c \xi_{sm}^c) \int_{\mathcal{T}} \psi_{rp}^c(t) \psi_{sq}^c(t) \phi_{rl}^c(t) \phi_{sm}^c(t) dt$$

one can rewrite (4) as

$$\int_{\mathcal{T}} \text{var}(W_k^c(t)) dt \approx \sum_{r=1}^{M^c} \sum_{s=1}^{M^c} \sum_{p=1}^{P_{r;c}} \sum_{q=1}^{P_{s;c}} b_{krp}^c b_{ksq}^c a_{rs,pq}^c = (\mathbf{b}_k^c)^\top \mathbf{A}^c \mathbf{b}_k^c,$$

where $\mathbf{A}^c$ is the matrix, of dimension $(\sum_{j=1}^{M^c} P_{j;c}) \times (\sum_{j=1}^{M^c} P_{j;c})$, containing all entries $a_{rs,pq}^c$ and $\mathbf{b}_k^c = (b_{k11}^c, \dots, b_{k1P_1}^c, \dots, b_{kM^c P_{M^c}}^c)^\top$ is the vector containing all entries $b_{kjP_j}^c$.

Once estimators $\hat{\boldsymbol{\Theta}}^c$ are obtained, we proceed and estimate $\mathbf{A}^c$ elementwise as

$$\hat{a}_{rs,pq}^c = \sum_{l=1}^{L_{r;c}} \sum_{m=1}^{L_{s;c}} (\hat{\boldsymbol{\Theta}}^c)^{-1}_{rl,sm} \int_{\mathcal{T}} \psi_{rp}^c(t) \psi_{sq}^c(t) \hat{\phi}_{rl}^c(t) \hat{\phi}_{sm}^c(t) dt.$$

Using $\hat{\mathbf{A}}^c$, the eigensystem $\hat{\mathbf{A}}^c \hat{\mathbf{b}}_k^c = \hat{\gamma}_k^c \hat{\mathbf{b}}_k^c$ is solved requiring $(\hat{\mathbf{b}}_k^c)^\top \hat{\mathbf{b}}_k^c = 1$ to obtain the estimator $\hat{\mathbf{b}}_k^c$ and one finally obtains an estimator for the unknown, time-varying coefficients $\beta_{kj}^c(t)$ as

$$\hat{\beta}_{kj}^{P_{j;c}}(t) = \sum_{p=1}^{P_{j;c}} \hat{b}_{kjp}^c \psi_{jp}^c(t).$$

For simplicity we set $\psi_{jp}^c(t) = \phi_{jp}^c(t)$ for which we use the corresponding estimator $\hat{\phi}_{jp}^c(t)$.

3 Obtaining the fGGL estimator

We describe in this section an algorithm used to estimate graph-informed linear manifolds.

Step 1: Obtain $\hat{\mu}_j^c(t)$ and $\hat{G}_j^c(t_1, t_2)$ at each scale $c = 1, \dots, C$ for all functions $j = 1, \dots, M^c$.

Step 2: Using the KL decomposition obtain the vectors $\hat{\xi}_i^c$ for each $i = 1, \dots, n$.

Step 3: Using $\hat{\xi}_i^c$ calculate the empirical covariance matrix as $\mathbf{S}^c = (1/n) \sum_{i=1}^n \hat{\xi}_i^c (\hat{\xi}_i^c)^\top$. Using

$\mathbf{S}^1, \dots, \mathbf{S}^C$ optimize (5) by following steps 4 to 8 described as follows:

Step 4: Initialize $\Theta^1, \dots, \Theta^C = \mathbf{I}$, $\mathbf{U}^1, \dots, \mathbf{U}^C = \mathbf{0}$, & $\mathbf{Z}^1, \dots, \mathbf{Z}^C = \mathbf{0}$ each having the dimensions of $\mathbf{S}^1, \dots, \mathbf{S}^C$.

Step 5: At iteration (m) , for any scale c update

$$\Theta_{(m)}^c = \arg \min_{\Theta^c} \{ \text{trace}(\mathbf{S}^c \Theta^c) - \log \det \Theta^c + P(\mathbf{Z}_{(m-1)}^c) \},$$

where $P(\mathbf{Z}_{(m-1)}^c) = \rho_{n1} \sum_{a \neq b \in V^c} |\mathbf{Z}_{ab;(m-1)}^c| + \rho_{n2} \sum_{h=1}^{M^1} \{ \sum_{a \neq b \in V^c} (\mathbf{Z}_{ab;h;(m-1)}^c)^2 \}^{1/2}$. The update is obtained in closed form. First, create the augmented Lagrangian as

$$L(\Theta^c, \mathbf{Z}^c, \mathbf{U}^c) = \text{trace}(\mathbf{S}^c \Theta^c) - \log \det \Theta^c + P(\mathbf{Z}_{(m-1)}^c) + \frac{\tau}{2} \|\Theta^c - \mathbf{Z}_{(m-1)}^c + \mathbf{U}_{(m-1)}^c\|_F^2.$$

Afterward, compute the eigen decomposition $\tau(\mathbf{Z}_{(m-1)}^c - \mathbf{U}_{(m-1)}^c) - \mathbf{S}^c = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top$, where $\tau > 0$ is an arbitrary fixed scalar, $\mathbf{\Lambda} = \text{diag}(\Lambda_1, \dots, \Lambda_{M^c L_{M^c}})$ where Λ_d represents the d -th eigenvalue and then form a diagonal matrix with entries $\tilde{X}_{dd} = (\Lambda_d + \sqrt{\Lambda_d^2 + 4\tau})/2\tau$.

Update $\Theta_{(m)}^c = \mathbf{Q} \tilde{\mathbf{X}} \mathbf{Q}^\top$.

Step 6: Given $\Theta_{(m)}^c$ from Step 5, update

$$\mathbf{Z}_{(m)}^c = \arg \min_{\mathbf{Z}^c} \left\{ \frac{\tau}{2} \|\Theta_{(m)}^c - \mathbf{Z}^c + \mathbf{U}_{(m-1)}^c\|_F^2 + \rho_{n1} \sum_{a \neq b \in V^c} |\mathbf{Z}_{ab}^c| + \rho_{n2} \sum_{h=1}^{M^1} \sqrt{\sum_{a \neq b \in V^c} (\mathbf{Z}_{ab;h}^c)^2} \right\}.$$

This is done using elementwise soft thresholding for each element $\mathbf{Z}_{ab,h}^c$, corresponding to the entry on position (a, b) at scale c that pertains to the coarse region h as follows:

$$\mathbf{Z}_{ab,h,(m)}^c = \text{Soft}_{\rho_{n1}/\tau}(\Theta_{ab,h,(m)}^c + \mathbf{U}_{ab,h,(m-1)}^c) \times \left(1 - \frac{\rho_{n2}}{\tau \sqrt{\sum_{a \neq b \in V^c} \text{Soft}_{\rho_{n1}/\tau}^2(\Theta_{ab,h,(m)}^c + \mathbf{U}_{ab,h,(m-1)}^c)}} \right) +$$

where $\text{Soft}_{\rho_{n1}/\tau}(x) = \text{sgn}(x) \max\{|x| - \rho_{n1}/\tau, 0\}$ and $x_+ = \max(x, 0)$.

Step 7: Update $\mathbf{U}_{(m)}^c = \mathbf{U}_{(m-1)}^c + \Theta_{(m)}^c - \mathbf{Z}_{(m)}^c$.

Step 8: Iterate until convergence to obtain the estimates for $\hat{\Theta}^1, \dots, \hat{\Theta}^C$.

The algorithm stops when the difference between the values of the objective function defined in (5) when plugging-in estimates for $\hat{\Theta}^1, \dots, \hat{\Theta}^C$, at two consecutive iterations is smaller than a fixed tolerance ϵ . The value $\epsilon = 10^{-4}$ was used consistently in this manuscript. Figure 4 in the on-line Supplementary Materials shows the numerical convergence for the fMRI example discussed in Section 5. Once estimates for the concentration

matrices are obtained at each scale, we estimate further $\mathbf{A}^c$ and then obtain estimates for the time-varying coefficients $\beta_{kj}^c(t)$ as explained in Section 2.

We focus next on the minimization in Step 6 of the Algorithm. Let $\mathbf{W}_{(m)}^c = \mathbf{\Theta}_{(m)}^c + \mathbf{U}_{(m-1)}^c$ which implies that

$$\mathbf{Z}_{(m)}^c = \arg \min_{\mathbf{Z}^c} \left\{ \frac{\tau}{2} \|\mathbf{Z}^c - \mathbf{W}_{(m)}^c\|_F^2 + \rho_{n1} \sum_{a \neq b \in V^c} |\mathbf{Z}_{ab}^c| + \rho_{n2} \sum_{h=1}^{M^1} \sqrt{\sum_{a \neq b \in V^c} (\mathbf{Z}_{ab;h}^c)^2} \right\}.$$

Taking the partial derivative of the loss with respect to the element $\mathbf{Z}_{ab,h}^c \neq 0$ and setting it equal to 0, we have that

$$\mathbf{Z}_{ab,h}^c \left(1 + \frac{\rho_{n2}}{\tau} \frac{1}{\sqrt{\sum_{a \neq b \in V^c} (\mathbf{Z}_{ab;h}^c)^2}} \right) = \mathbf{W}_{ab,h}^c - \frac{\rho_{n1}}{\tau} \text{sgn}(\mathbf{Z}_{ab,h}^c) = \text{Soft}_{\rho_{n1}/\tau}(\mathbf{W}_{ab,h}^c)$$

where we can show that $\mathbf{W}_{ab,h}^c - \frac{\rho_{n1}}{\tau} \text{sgn}(\mathbf{Z}_{ab,h}^c) = \text{Soft}_{\rho_{n1}/\tau}(\mathbf{W}_{ab,h}^c)$. Now,

$$\begin{aligned} \mathbf{Z}_{ab,h}^c \left(1 + \frac{\rho_{n2}}{\tau} \frac{1}{\sqrt{\sum_{a \neq b \in V^c} (\mathbf{Z}_{ab;h}^c)^2}} \right) &= \text{Soft}_{\rho_{n1}/\tau}(\mathbf{W}_{ab,h}^c) \Leftrightarrow \\ \sqrt{\sum_{a \neq b \in V^c} (\mathbf{Z}_{ab,h}^c)^2} &= \sqrt{\sum_{a \neq b \in V^c} \text{Soft}_{\rho_{n1}/\tau}^2(\mathbf{W}_{ab,h}^c) - \frac{\rho_{n2}^2}{\tau^2}}. \end{aligned}$$

As the LHS > 0 , one needs as well that $\sqrt{\sum_{a \neq b \in V^c} \text{Soft}_{\rho_{n1}/\tau}^2(\mathbf{W}_{ab,h}^c)} > \frac{\rho_{n2}}{\tau}$. Under this condition, substituting back we have that

$$\begin{aligned} \mathbf{Z}_{ab,h}^c \left(1 + \frac{\rho_{n2}}{\tau} \frac{1}{\sqrt{\sum_{a \neq b \in V^c} (\mathbf{Z}_{ab;h}^c)^2}} \right) &= \text{Soft}_{\rho_{n1}/\tau}(\mathbf{W}_{ab,h}^c) \Leftrightarrow \\ \text{Soft}_{\rho_{n1}/\tau}(\mathbf{W}_{ab,h}^c) \left(1 - \frac{\rho_{n2}}{\tau \sqrt{\sum_{a \neq b \in V^c} \text{Soft}_{\rho_{n1}/\tau}^2(\mathbf{W}_{ab,h}^c)}} \right) &= \mathbf{Z}_{ab,h}^c. \end{aligned}$$

The $(\cdot)_+$ operator in the algorithm is used to ensure that the needed condition is satisfied.

4 Theoretical properties of the fGGL estimator

Regularity conditions.

Let $\mathbf{\Sigma}_0^c$ and $\mathbf{\Theta}_0^c$ be the true covariance and concentration matrices at scale $c = 1, \dots, C$. Let next $E_u^c = \{(a, b) | \mathbf{\Theta}_{ab,0}^c \neq 0\}$ for which $\text{card}(E_u^c) = s^c$ and denote by $\bar{E}_u^c$ the complementary set of E_u^c . In the sequel, we assume that:

- (a) there exist constants $\underline{k}$ and $\bar{k}$ such that $0 < \underline{k} \leq \text{eig}_{\min}(\mathbf{\Sigma}_0^c) \leq \text{eig}_{\max}(\mathbf{\Sigma}_0^c) \leq \bar{k} < \infty$;

(b) without loss of generality $L_{r;c} = L_{s;c} = L \geq 1$ for any scale $c = 1, \dots, C$.

Condition (a) is a standard condition that ensures the existence of invertible covariance matrices, while condition (b) specifies that without loss of generality the number of truncation points for all functions is fixed regardless of the scale.

Proposition 1 *Let $\hat{\Theta}^1, \dots, \hat{\Theta}^C$ be the minimizers of (5) for all C scales. Under conditions (a) and (b), if $\rho_{n1} \asymp O\left(\max_c \sqrt{\frac{\log(M^c L)}{n}}\right)$ and $\rho_{n2} \asymp O\left(\max_c \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{(M^1)^2 n}}\right)$ then*

$$\sum_{c=1}^C \|\hat{\Theta}^c - \Theta_0^c\|_F = O_p\left(\max_c \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}}\right).$$

Proposition 2 *Under the conditions given in Proposition 1, for minimizers $\hat{\Theta}^1, \dots, \hat{\Theta}^C$ satisfying $\|\hat{\Theta}^c - \Theta_0^c\|_F = O_p(\sqrt{n^{-1}(M^c L + s^c) \log(M^c L)})$ and $\|\hat{\Theta}^c - \Theta_0^c\|_2 = O_p(\sqrt{\eta_n})$ where $\eta_n \rightarrow 0$, we have*

$$\max_{rs,pq} |\hat{a}_{rs,pq}^c - a_{rs,pq}^c| = O_p\left(\sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}}\right).$$

Proposition 1 provides a bound on the total estimation error of the precision matrices across scales in terms of the Frobenius norm, while Proposition 2 bounds the element-wise maximal difference between the elements of the estimated $\hat{\mathbf{A}}^c$ and those of the true corresponding matrix. The proofs for both Propositions are relegated to the on-line Supplementary Materials.

Under minimal assumptions, using perturbation theory such as the ‘Bauer-Fike’ theorem (see Theorem 3.3 and Corollary 3.4 from Stewart and Sun, 1990) and the Davis-Kahan ‘sin(Θ)’ theorem (see Corollary 1 from Yu et al., 2014) we can bound both $|\hat{\gamma}_k^c - \gamma_k^c|$ for all k and the Euclidean norm $\|\hat{\mathbf{b}}_k^c - \mathbf{b}_k^c\|$ by terms of order $O_p(\|\hat{\mathbf{A}}^c - \mathbf{A}^c\|_2)$, where $\|\cdot\|_2$ denotes the spectral norm of the matrix. Using the results of Proposition 2 and the basic bound between the spectral and the max norm one can obtain bounds on these random quantities of the order $O_p(M^c L \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}})$. However, since one does not enforce sparsity on $\hat{\mathbf{A}}^c$ one pays a price in the convergence rate. Although the number of components can still grow with n , the sample size, one now needs that n must grow much faster relative to the number of components. Future extensions of the proposed procedure, could target $\mathbf{A}$ and assume it sparse, hence propose directly sparse estimators for it.

5 fMRI application

As we have argued in Section 1, accounting for the similarity between the 64 coarse and 122 finer partitions, might result in a better performance than if one were to ignore it. In Figure 2 we present the estimated graphs corresponding to $\hat{\Theta}^1$ and $\hat{\Theta}^2$ using the fMRI signal when subjects were presented the ‘faces’ and ‘house’ stimuli (for other stimuli the graphs

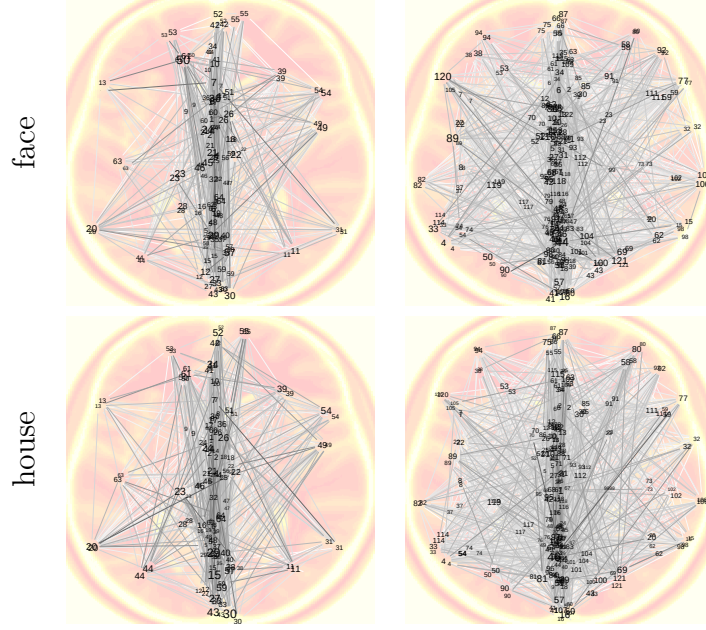


Figure 2: fMRI data. Estimated graphs based on data obtained after a two-dimensional KL decomposition at each of the 64 and 122 ROIs using the data from all $n = 6$ subjects for the ‘face’ and ‘house’ stimuli. Regularization parameters are chosen using three-fold cross-validation. For visualization, a fading effect is applied proportional to the absolute magnitude of the estimated entry, with gray/white denoting higher/lower magnitudes. Labels are proportional to the degree of the node and nodes with identical labels refer to ξ components pertaining to the same ROI.

are available upon request from the authors) and where a three-fold cross-validation scheme using the negative log-likelihood loss has been used to select the regularization parameters. A grid of 300 different couples ρ_{n1} and ρ_{n2} (20×15) has been used for selection. The grid was chosen such that the estimated graphs range from being very sparse to being very dense. The graphs estimated with the optimally selected regularization levels show that there is generally a large number of connections which is indicative of having a high degree of communication between the ROIs. The sparsity coefficients of the graphs are around 89% (scale 1) and 94% (scale 2) for the two conditions. The graphs point to the existence of a large number of common edges and a large degree of similarity between the two stimuli, but also indicate as well a differentiation with respect to what edges have more weight relative to others. Moreover, not all ROIs share the same number of connections in the graphs, as some regions are more connected than others.

To illustrate the benefits of the grouping across splits, we have next compared the fGGL procedure with a simpler functional graphical lasso (fGL) as proposed in Qiao et al. (2015) obtained when setting $\rho_{n2} = 0$. To evaluate the two procedures we investigate an out-of-sample negative log-likelihood loss (smaller is better), where we use as training set the first 81 fMRI measurements and the test set contains the remaining 27 measurements. The results are reported in Table 1.

Table 1: Out-of-sample negative log-likelihood (smaller is better) obtained when the training set contains the first 81 fMRI measurements and the test set contains the remaining 27 measurements. The analysis uses all $n = 6$ subjects and is repeated for each stimulus. Three-fold cross-validation was used to select the regularization parameters. Two scales in the data have been used and the number of truncation points is $L = 2$ and 3. The results are reported for each scale separately.

	$L = 2$				$L = 3$			
	Scale 1		Scale 2		Scale 1		Scale 2	
	fGGL	fGL	fGGL	fGL	fGGL	fGL	fGGL	fGL
bottle	345	350	1177	1186	609	638	1695	2051
cat	367	363	1241	1238	477	477	1624	2251
chair	353	352	1193	1195	462	519	1560	2022
face	391	394	1308	1327	496	491	1693	2029
house	360	358	1226	1221	685	697	1749	2166
scissors	388	369	1271	1273	738	749	1871	2371
shoe	345	352	1194	1255	596	599	1758	1977

Due to the relatively large number of functions and mainly for computational efficiency, we fixed the number of truncation points to $L_{s;c} = L_{r;c} = L = 2$ and 3 for all variables and all scales, as was also done in the work of Chiou and Müller (2014) since the proportion of explained variance was relatively high around 90% for each function when $L = 3$. The table indicates that in general accounting for the splitting design produces similar or more accurate out-of-sample results than a functional graphical lasso which is insensitive to the design grouping. The fGGL procedure is more accurate especially when $L = 3$ (hence the sizes of the graphs are larger) for both scales. This can be explained by the fact that the functional grouping of the ξ coefficients pertaining to the same region becomes informative and hence forcing similarity across scales at the level of the estimated graphs becomes beneficial. When $L = 2$ we judge the performance of the two competitors as comparable.

We investigated as well the in-sample prediction error when the original observed functions are ‘reconstructed’ using the estimated manifold. Similar conclusions are reached: both methods indicate a close reconstruction error when $L = 2$, with clear improvements for fGGL over fGL when $L = 3$.

To illustrate the added value of the linear manifold representation in the fMRI context, we plot in Figure 3 for the ‘face’ stimulus (for other stimuli, the graphs are available from the authors) the graph-informed estimates for the time-varying coefficients $\beta_{kj}^c(t)$ with $k = 1, \dots, K$, $c = 1, \dots, C$ and $j = 1, \dots, M^c$ for both scales, ie. $C = 2$ and with $K = 2$. In the figure, the dominant functions in the linear combinations are illustrated in color, whereas less important functions are illustrated in gray. Quite interestingly, the first two linear combinations at scale 1 are dominated by a contrast between the ROIs labelled 12 (cerebellum crus I), 20 (left inferior parietal lobe), 22 (inferior temporal gyrus) and 28 (lateral occipitotemporal gyrus). ROI 12 has been recently linked to ‘visual divergent thinking’ in Gao et al. (2020), while ROI 20 has been linked to the perception of emotions

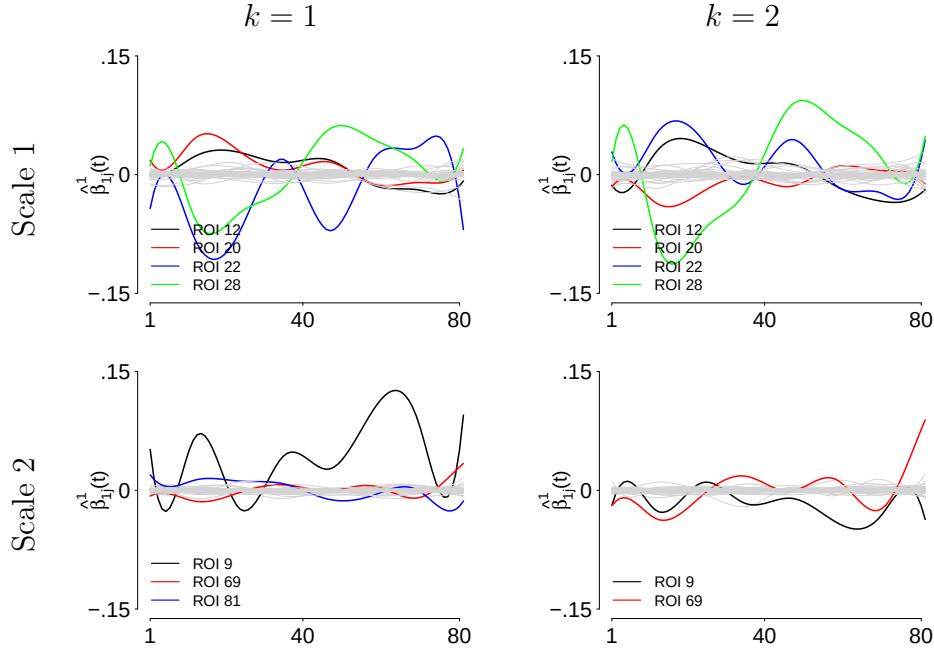


Figure 3: fMRI data. Estimated graph-based, time-varying coefficients for $K = 2$ linear constraints for both scales for the ‘face’ stimulus. For visualization purposes, a fading effect is applied where in color we present functions that dominate the linear combination by having higher absolute magnitudes for the coefficients and with gray denoting lower absolute magnitudes.

based on facial stimuli in Radua et al. (2010). Ishai et al. (1999) identified ROI 22 as being involved in visual perception, while ROI 28 (also known as the fusiform gyrus) has been repeatedly linked to vision functions as in Kanwisher and Yovel (2006). At scale 2 it is also informative to note that the linear combinations are driven by a contrast between ROIs 9 (cerebellum VI) and 69 (dorsal visual stream).

At scale 1, both constraints point to a wave-type effect for the inferior temporal gyrus and the lateral occipitotemporal gyrus but with different oscillations between these two ROIs. On the other hand, ROIs 20 and 12 have a larger absolute impact for the first part of the measurements, but their effect dissipates afterward. At scale 2, for the first constraint we see a clear dominance of cerebellum VI, while for the second constraint the dorsal visual stream has a larger absolute impact at the beginning and toward the end of the series.

All in all, we conclude that (i) the graph-based linear constraints pick up signal from relevant ROIs linked to the visual process, while discarding other ROIs as low-informative and (ii) indicate novel and interesting comparative patterns of evolution over time between these ROIs that are still to be validated by further research.

6 Simulation study

In this section we present the results obtained when evaluating the performance of the proposed method in a high and low-dimensional controlled, simulated setting.

6.1 High-dimensional, multiscale setting

To investigate the performance of the proposed procedure in finite sample settings, we have performed a simulation study that is based on the multiscale fMRI study presented in Schmittmann et al. (2015). To this end we have generated multivariate functional data for $n = 8$ subjects at $C = 2$ coarseness scales. The number of functions per scale was set to $M^1 = 68$ for scale 1 and $M^2 = 114$ for scale 2 and for each function at each scale the number of truncation points was set to $L = 2$. The number of constraints K was chosen using the criterion proposed in Chiou and Müller (2014) with $\tau = 0.05$ and the number of equally spaced time grid points was further chosen to be $T = 240$. The true matrices were set as $(\Theta^c)^{-1} = \frac{1}{n} \sum_{i=1}^n \hat{\xi}_i^c (\hat{\xi}_i^c)^\top$ where $\hat{\xi}_i^c$ represent the vectors of scores estimated using the real data of Schmittmann et al. (2015) for each case i . Using further the estimated eigenfunctions, we have created noisy trajectories as $X_{ji}^c(t) = \sum_{l=1}^L \hat{\xi}_{jli}^c \hat{\phi}_{jl}^c(t) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$ with $\sigma = .1$ or $\sigma = .5$. The noisy trajectories were then finally treated as data used in the estimation procedure. The proposed fGGL procedure was evaluated when $\rho_{n1} \in \{.25, .50\}$ and $\rho_{n2} \in \{.50, .75\}$ and compared to the performance of fGL ($\rho_{n2} = 0$). We remind the reader that when $\rho_{n2} = 0$, our procedure reduces to the functional graphical lasso procedure proposed in Qiao et al. (2015), which we consider as the competitor.

A total of 100 repetitions have been performed and the interest lies in how accurately the reconstructed trajectories based on the linear graph-informed manifold, denoted as $\hat{X}_{ji}^{K,c}(t)$ match the true, noiseless version of $X_{ji}^c(t)$ and how accurately the concentration matrices are estimated. For this purpose the following two losses (smaller is better) have been used:

$$\begin{aligned} \text{Fr.Loss} &= \sum_{c=1}^C \|\hat{\Theta}^c - (\Theta^c)^\dagger\|_F \\ \text{MSE.Loss} &= \frac{1}{C \times n \times T} \sum_{c=1}^C \sum_{j=1}^{M^c} \sum_{i=1}^n \sum_{t=1}^T (\hat{X}_{ji}^{K,c}(t) - X_{ji}^c(t))^2, \end{aligned}$$

where $(\Theta^c)^\dagger$ denotes a generalized inverse since the matrix created as $\frac{1}{n} \sum_{i=1}^n \hat{\xi}_i^c (\hat{\xi}_i^c)^\top$ might not necessarily be invertible.

For illustration, we present in Figure 1 in the Supplementary Materials the true noiseless version of $X_{ji}^c(t)$ (full, solid line) and the results of the estimated functions based on the proposed penalized graph procedure when $\rho_{n1} = .25$ and $\sigma = .5$ for a random subject and a randomly chosen ROI from scale 2. The figure suggests that generally, the procedures still accurately recover the periodic pattern when larger noise is added to the functions, but more variability occurs with respect to the estimated amplitude of the functions. Similar

Table 2: Simulated data. Average Fr.Loss and MSE.Loss (smaller is better) over 100 repetitions for fGL and fGGL (with $\rho_{n2} \in \{.50, .75\}$) when the noise level is set to $\sigma \in \{.1, .5\}$.

	fGL		fGGL($\rho_{n2} = .5$)		fGGL($\rho_{n2} = .75$)	
	Fr	MSE	Fr	MSE	Fr	MSE
$\sigma = .1$	84.6	4.8	56.7	4.5	46.9	4.4
$\sigma = .5$	84.4	60.6	56.6	58.9	45.7	57.9

conclusions hold for other functions and other subjects, and the performance averaged over all subjects and all functions, presented in Table 2 suggests an advantage for fGGL relative to fGL, both with respect to the estimation of the concentration matrices and the reconstruction of the underlying functions.

We focused further on the case where $\sigma = .5$ and to alleviate the dependence on the choice of tuning parameters, we have performed a three-fold cross-validation using the negative log-likelihood as loss and a grid of 20 different values to select ρ_{n1} for fGL. With the optimal selected value we fitted next fGL and fGGL with $\rho_{n2} \in \{.50, .75\}$. The conclusions remain roughly the same as those of Table 2: fGGL provided a more accurate graph reconstruction, while the performance on the reconstruction of the underlying functions was similar.

6.2 Low-dimensional simulation study

To empirically validate and benchmark the proposed method with the low-dimensional, non-multiscale approach (denoted C&M) of Chiou and Müller (2014), we have performed an additional low-dimensional simulation study that resembles the ‘Drosophila flies’ analysis presented in the aforementioned paper where the flying, feeding, walking and resting behavior of flies was of interest. A similar strategy as in Section 6.1 was taken, but here $n = 62$, $T = 40$ and $C = 1$. The same performance metrics were analyzed, the same σ was used for the noise level and the same regularization parameters were retained.

Figure 2 in the Supplementary Materials presents the obtained results for the graph-based procedure relative to those obtained by the C&M procedure and it suggests that generally, for low-dimensional settings the proposed methods perform similarly to the classical one. Since in this case shrinkage is not per se necessary due to low-dimensionality, as expected the Frobenius loss is slightly larger for the penalized procedures than that of the C&M method. All in all, this simulation study points to comparable results for low-dimensional settings and in line with those obtained by Chiou and Müller (2014) with the advantage that the multiscale, graph-based procedures can be applied also for cases where the number of functions is much larger than the available sample size.

7 Extensions

One expects that there is correlation between the coarser and finer scale measurements, but the optimization problem in (5) is insensitive to *explicitly* modeling it. However, this correlation between splits and coarser ROIs can be accounted for and estimated by pooling all functions to construct $\hat{\boldsymbol{\xi}}_i = ((\hat{\boldsymbol{\xi}}_i^1)^\top, (\hat{\boldsymbol{\xi}}_i^2)^\top, \dots, (\hat{\boldsymbol{\xi}}_i^C)^\top)^\top$ and $\mathbf{S} = n^{-1} \sum_{i=1}^n \hat{\boldsymbol{\xi}}_i (\hat{\boldsymbol{\xi}}_i)^\top$ and solving a constrained optimization problem of the form:

$$\min_{\boldsymbol{\Theta}} \{ \text{trace}(\mathbf{S}\boldsymbol{\Theta}) - \log \det \boldsymbol{\Theta} + P_{\rho_{n1}, \rho_{n2}} \} \quad (6)$$

$$\text{where } P_{\rho_{n1}, \rho_{n2}} = \begin{cases} \rho_{n1} \sum_{h=1}^{M^1} \sum_{a \neq b} |\theta_{ab;h}| + \rho_{n2} \sum_{h=1}^{M^1} (\sum_{a \neq b} (\theta_{ab;h})^2)^{1/2} & \text{for } \theta_{ab;h} \in \mathcal{B} \\ \infty, & \text{for } \theta_{ab;h} \notin \mathcal{B} \end{cases}$$

and $\mathcal{B} = \{\theta_{ab;1}, \theta_{ab;2}, \dots, \theta_{ab;M^1}\}$ is a set of coefficients corresponding to edges that are allowed to be estimated, i.e., all entries for which *both* indices a and b denote partitions of or the full ROI h , no matter the scale.

To facilitate understanding we graphically illustrate the set $\mathcal{B}$ in panel (a) of Figure 3 from the Supplementary Materials, for the fMRI two-scale example used in Section 6.1 to generate simulated data. The gray dots illustrate an allowed edge to be estimated. The top left bulk of points represents the allowed edges for the first scale, while the bottom right bulk of points represents the allowed edges for the second scale. The points on the two ‘lines’ above/below the diagonal represent edges that are allowed across the scales, between the coarser ROIs and the finer splits. In panel (b) of Figure 3 from the Supplementary Materials we present the estimated edges in the form of an adjacency matrix where black dots represent estimated edges between two components of the vector $\boldsymbol{\xi}$, and it illustrates that optimizing (6) leads to estimation of within scale edges similarly to the procedure in Section 2, but it can also explicitly model the dependencies between coarse and fine scale signals. Graph-informed linear manifolds could then subsequently be estimated using these graphs.

8 Discussion

In this paper the linear manifold model of Chiou and Müller (2014) has been extended to settings where the number of functional components is larger than the available sample size and where multiple coarseness levels in the data exist. This extension was accomplished by making use of regularized procedures that estimate sparse undirected graphical models. By doing so, one also gets insight into the conditional independence structure that governs the multivariate functional data, while at the same time it identifies linear combinations of the components. The framework we investigate is a multiscale one where multiple graphs are estimated jointly and are constrained to share similarities due to the design of the problem. Simulated and real data examples show beneficial performance pointing to the procedure being a useful tool for modeling multiscale, multivariate functional data.

SUPPLEMENTARY MATERIAL

The online Supplementary material file contains the proofs, as well as additional figures. Software in the form of R code is available from the corresponding author.

References

- Banerjee, O., L. El Ghaoui, and A. d’Aspremont (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research* 9, 485–516.
- Bellec, P., P. Rosa-Neto, O. Lyttelton, H. Benali, and A. Evans (2010). Multi-level bootstrap analysis of stable clusters in resting-state fmri. *NeuroImage* 51(3), 1126–1139.
- Bickel, P. J. and E. Levina (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics* 36(1), 199–227.
- Cai, T. T., C.-H. Zhang, and H. H. Zhou (2010). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*. 38(4), 2118–2144.
- Chiou, J.-M. and H.-G. Müller (2014). Linear manifold modelling of multivariate functional data. *Journal of the Royal Statistical Society: Series B* 76(3), 605–626.
- Danaher, P., P. Wang, and D. Witten (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society, Series B* 76(2), 373–397.
- Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3), 432–441.
- Gao, Z., D. Liu, X. and Zhang, M. Liu, and N. Hao (2020). The indispensable role of the cerebellum in visual divergent thinking. *Scientific Reports* 10(16552).
- Haxby, J. V., M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293(5539), 2425–2430.
- Huettel, S., A. Song, and G. McCarthy (2009). *Functional Magnetic Resonance Imaging* (2nd ed.). Sunderland, Massachusetts: Sinauer Associates.
- Ishai, A., L. G. Ungerleider, A. Martin, J. L. Schouten, and J. V. Haxby (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences* 96(16), 9379–9384.
- Jiang, C.-R., J. A. D. Aston, and J.-L. Wang (2016). A functional approach to deconvolve dynamic neuroimaging data. *Journal of the American Statistical Association* 111(513), 1–13.

- Kanwisher, N. and G. Yovel (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences* 361(1476), 2109–2128.
- Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.
- Lazar, N. (2008). *The Statistical Analysis of Functional MRI Data*. Springer.
- Leng, C. and C. Y. Tang (2012). Sparse matrix graphical models. *Journal of the American Statistical Association* 107(499), 1187–1200.
- Lila, E., J. A. D. Aston, and L. M. Sangalli (2016). Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *Annals of Applied Statistics* 10(4), 1854–1879.
- Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34(3), 1436–1462.
- Mejia, A. F., M. B. Nebel, A. Eloyan, B. Caffo, and M. A. Lindquist (2017). PCA leverage: outlier detection for high-dimensional functional magnetic resonance imaging data. *Biostatistics* 18(3), 521–536.
- Pircalabelu, E., G. Claeskens, and L. Waldorp (2016). Mixed scale joint graphical lasso. *Biostatistics* 17(4), 793–806.
- Qiao, X., S. Guo, and G. M. James (2019). Functional graphical models. *Journal of the American Statistical Association* 114(525), 211–222.
- Qiao, X., G. James, and J. Lv (2015). Functional graphical models. Technical report.
- Radua, J., M. L. Phillips, T. Russell, N. Lawrence, N. Marshall, S. Kalidindi, W. El-Hage, C. McDonald, V. Giampietro, M. J. Brammer, A. S. David, and S. A. Surguladze (2010). Neural response to specific components of fearful faces in healthy and schizophrenic adults. *NeuroImage* 49(1), 939–946.
- Schmittmann, V. D., D. Borsboom, S. Jahfari, A. Savi, and L. J. Waldorp (2015). Making large-scale networks from fMRI data. *PLoS One* 10(9), e0129074.
- Stewart, G. W. and J. G. Sun (1990). *Matrix Perturbation Theory*. Academic Press.
- Tian, T. S. (2010). Functional data analysis in brain imaging studies. *Frontiers in psychology* 1(35).
- Ullah, S. and C. F. Finch (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology* 13(43).
- Viviani, R., G. Grön, and M. Spitzer (2005). Functional principal component analysis of fMRI data. *Human brain mapping* 24(2), 109–129.

- Wang, J.-L., J.-M. Chiou, and H.-G. Müller (2016). Functional data analysis. *Annual Review of Statistics and Its Application* 3(1), 257–295.
- Yu, Y., T. Wang, and R. J. Samworth (2014). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* 102(2), 315–323.
- Yuan, M. and Y. Lin (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* 94(1), 19–35.
- Zapata, J., S. Oh, and A. Petersen (2022). Partial separability and functional graphical models for multivariate Gaussian processes. *Biometrika* 109(3), 665–681.

Linear manifold modeling and graph estimation based on multivariate functional data with different coarseness scales

Supplementary Materials

Eugen Pircalabelu

Institute of Statistics, Biostatistics and Actuarial Sciences, LIDAM, UCLouvain

Gerda Claeskens

ORSTAT and Leuven Statistics Research Center, KU Leuven

September 6, 2022

1 Proof of Proposition 1

Proof. Let

$$\begin{aligned}
Q(\Theta^1, \dots, \Theta^C) &= \sum_{c=1}^C \left(\text{trace}(\mathbf{S}^c \Theta^c) - \log \det \Theta^c + \rho_{n1} \sum_{a \neq b \in V^c} |\theta_{ab}^c| + \rho_{n2} \sum_{h=1}^{M^1} \sqrt{\sum_{a \neq b \in V^c} (\theta_{ab,h}^c)^2} \right. \\
&\quad \left. - \text{trace}(\mathbf{S}^c \Theta_0^c) + \log \det \Theta_0^c - \rho_{n1} \sum_{a \neq b \in V^c} |\theta_{ab,0}^c| - \rho_{n2} \sum_{h=1}^{M^1} \sqrt{\sum_{a \neq b \in V^c} (\theta_{ab,h,0}^c)^2} \right) \\
&= \sum_{c=1}^C \left(\text{trace}[\mathbf{S}^c (\Theta^c - \Theta_0^c)] - (\log \det \Theta^c - \log \det \Theta_0^c) \right. \\
&\quad \left. + \rho_{n1} \sum_{a \neq b \in V^c} (|\theta_{ab}^c| - |\theta_{ab,0}^c|) + \rho_{n2} \sum_{h=1}^{M^1} \left(\sqrt{\sum_{a \neq b \in V^c} (\theta_{ab,h}^c)^2} - \sqrt{\sum_{a \neq b \in V^c} (\theta_{ab,h,0}^c)^2} \right) \right) \\
&= \sum_{c=1}^C \left(\text{trace}[(\Theta^c - \Theta_0^c)(\mathbf{S}^c - \Sigma_0^c)] + \text{trace}[(\Theta^c - \Theta_0^c) \Sigma_0^c] - (\log \det \Theta^c - \log \det \Theta_0^c) \right. \\
&\quad \left. + \rho_{n1} \sum_{a \neq b \in V^c} (|\theta_{ab}^c| - |\theta_{ab,0}^c|) + \rho_{n2} \sum_{h=1}^{M^1} \left(\sqrt{\sum_{a \neq b \in V^c} (\theta_{ab,h}^c)^2} - \sqrt{\sum_{a \neq b \in V^c} (\theta_{ab,h,0}^c)^2} \right) \right).
\end{aligned}$$

Let estimators $\hat{\Theta}^1, \dots, \hat{\Theta}^C$ be minimizers of $Q(\Theta^1, \dots, \Theta^C)$. Consider next a collection of ‘perturbation’ matrices $\Delta^1, \dots, \Delta^C$ around the targets $\Theta^1, \dots, \Theta^C$ and let

$$G(\Delta^1, \dots, \Delta^C) \equiv Q(\Theta_0^1 + \Delta^1, \dots, \Theta_0^C + \Delta^C).$$

Consider now the estimators $\hat{\Delta}^1, \dots, \hat{\Delta}^C$ defined as $\hat{\Delta}^1 = \hat{\Theta}^1 - \Theta_0^1, \dots, \hat{\Delta}^C = \hat{\Theta}^C - \Theta_0^C$, which in turn are minimizers of $G(\Delta^1, \dots, \Delta^C)$ and for all matrices $\Delta^1, \dots, \Delta^C$ consider next the set

$$\mathcal{A}_n(\phi) = \{\Delta^1, \dots, \Delta^C \mid \Delta^1 = (\Delta^1)^\top, \dots, \Delta^C = (\Delta^C)^\top \text{ and } \|\Delta^1\|_F = \phi r_n^1, \dots, \|\Delta^C\|_F = \phi r_n^C\}$$

where $\|\Delta^c\|_F$ is the Frobenius norm of the matrix Δ^c , $\phi > 0$ is a large enough constant and for $c = 1, \dots, C$, $r_n^c = \sqrt{n^{-1}(M^c L + s^c) \log(M^c L)}$.

Since $\hat{\Delta}^1, \dots, \hat{\Delta}^C$ are minimizers of $G(\Delta^1, \dots, \Delta^C)$, it must be that

$$G(\hat{\Delta}^1, \dots, \hat{\Delta}^C) \leq G(0, \dots, 0) = Q(\Theta_0^1, \dots, \Theta_0^C) = 0.$$

For any n fixed consider now $\mathcal{G}(\phi) = \inf_{\Delta^1, \dots, \Delta^C \in \mathcal{A}_n(\phi)} G(\Delta^1, \dots, \Delta^C)$ and suppose we can show that there exists a $\phi > 0$ such that $\mathcal{G}(\phi) > 0$. The reason we consider this is because from the inequality above,

$$\mathcal{G}(\phi) > 0 \Leftrightarrow \mathcal{G}(\phi) > G(\hat{\Delta}^1, \dots, \hat{\Delta}^C). \quad (1)$$

As $G(\Delta^1, \dots, \Delta^C)$ is convex, (1) implies that using any perturbation matrices $\Delta^1, \dots, \Delta^C \in \mathcal{A}_n(\phi)$ one is not yet at the optimal values and at least one of the inequalities $\|\hat{\Delta}^1\|_F < \phi r_n^1, \dots, \|\hat{\Delta}^C\|_F < \phi r_n^C$ needs to hold for the optimizer, meaning that the Frobenius norms for the optimizer should be smaller than the rates in $\mathcal{A}_n(\phi)$.

Define

$$\begin{aligned} \text{I}^c &= \text{trace}(\Delta^c(\mathbf{S}^c - \Theta_0^c)) + g_c(\Delta^c) + \rho_{n1} \sum_{a \neq b \in V^c} (|\theta_{ab}^c + \delta_{ab,0}^c| - |\delta_{ab}^c|) \\ \text{II}^c &= \rho_{n2} \sum_{h=1}^{M^1} \left(\sqrt{\sum_{a \neq b \in V^c} (\theta_{ab,h,0}^c + \delta_{ab,h}^c)^2} - \sqrt{\sum_{a \neq b \in V^c} (\theta_{ab,h,0}^c)^2} \right). \end{aligned}$$

One can rewrite $G(\Delta^1, \dots, \Delta^C) = \sum_{c=1}^C (\text{I}^c + \text{II}^c)$. The reason for this rewriting is that at each scale c , the term I^c is identical to expression (10) of Rothman et al. (2008) for the simple graphical lasso for multivariate data, for which they show that $\text{I}^c > 0$ for a sufficiently large ϕ . In the sequel we show that for each scale c , the term $(\text{I}^c + \text{II}^c) > 0$ and as a consequence $G(\Delta^1, \dots, \Delta^C) > 0$. As in eq.(16) of Rothman et al. (2008) for some constants H_1 and H_2 one has that

$$\text{I}^c \geq \frac{1}{4} k^2 \|\Delta^c\|_F^2 - H_1 \sqrt{\frac{\log(M^c L)}{n}} |\Delta^{c-}|_1 - H_2 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}} \|\Delta^{c+}\|_F + \rho_{n1} (|\Delta_{\overline{E}_u^c}^{c-}|_1 - |\Delta_{E_u^c}^{c-}|_1)$$

where $\Delta^{c+} = \text{diag}(\Delta^c)$ is a diagonal matrix with the same diagonal as Δ^c , $\Delta^{c-} = \Delta^c - \Delta^{c+}$ is the matrix containing the off-diagonal elements of Δ^c and the value 0 on the main diagonal, while $|\Delta_{\overline{E}_u^c}^{c-}|_1$ denotes the sum of all absolute values of the elements in Δ^{c-} contained in the set $\overline{E}_u^c$. Similarly $|\Delta_{E_u^c}^{c-}|_1$ denotes the sum of all absolute values of the elements in Δ^{c-} contained in the set E_u^c .

By adding II^c on both sides one has

$$\begin{aligned} \text{I}^c + \text{II}^c &\geq \frac{1}{4} k^2 \|\Delta^c\|_F^2 - H_1 \sqrt{\frac{\log(M^c L)}{n}} |\Delta^{c-}|_1 - H_2 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}} \|\Delta^{c+}\|_F \\ &\quad + \rho_{n1} (|\Delta_{\overline{E}_u^c}^{c-}|_1 - |\Delta_{E_u^c}^{c-}|_1) - \rho_{n2} \left(\sum_{h=1}^{M^1} (\|\Theta_{0h}^{c-}\|_F - \|\Theta_{0h}^{c-} + \Delta_h^{c-}\|_F) \right). \end{aligned}$$

This is because $\{\sum_{a \neq b \in V^c} (\theta_{ab,h,0}^c)^2\}^{1/2} = \|\Theta_{0h}^{c-}\|_F$.

Focusing first on II^c we have that

$$\rho_{n2} \left(\sum_{h=1}^{M^1} (\|\Theta_{0h}^{c-}\|_F - \|\Theta_{0h}^{c-} - \Delta_h^{c-}\|_F) \right) \leq \rho_{n2} \sum_{h=1}^{M^1} (\|\Theta_{0h}^{c-} - \Theta_{0h}^{c-} + \Delta_h^{c-}\|_F) \leq \rho_{n2} M^1 \|\Delta^{c-}\|_F,$$

where the first inequality comes from the reversed triangle inequality and the second inequality holds because for all h , $\|\Delta_h^{c-}\|_F \leq \|\Delta^{c-}\|_F$. That is, the Frobenius norm of each submatrix formed by entries corresponding to a coarse ROI h has to be less than or equal to the norm of the entire matrix. As there are a total of M^1 coarse ROIs, the bound follows.

It holds then that

$$\begin{aligned} \mathbf{I}^c + \mathbf{II}^c &\geq \frac{1}{4}\underline{k}^2 \|\Delta^c\|_F^2 - H_1 \sqrt{\frac{\log(M^c L)}{n}} |\Delta^{c-}|_1 - H_2 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}} \|\Delta^{c+}\|_F \\ &\quad + \rho_{n1} (|\Delta_{E_u^c}^{c-}|_1 - |\Delta_{E_u^c}^{c-}|_1) - \rho_{n2} M^1 \|\Delta^{c-}\|_F. \end{aligned}$$

We focus further on $\mathbf{I}^c$ for which it holds that

$$\begin{aligned} \mathbf{I}^c &\geq \frac{1}{4}\underline{k}^2 \|\Delta^c\|_F^2 - H_1 \sqrt{\frac{\log(M^c L)}{n}} |\Delta^{c-}|_1 - H_2 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}} \|\Delta^{c+}\|_F + \rho_{n1} (|\Delta_{E_u^c}^{c-}|_1 - |\Delta_{E_u^c}^{c-}|_1) \\ &= \frac{1}{4}\underline{k}^2 \|\Delta^c\|_F^2 - H_1 \sqrt{\frac{\log(M^c L)}{n}} (|\Delta_{E_u^c}^{c-}|_1 + |\Delta_{E_u^c}^{c-}|_1) - H_2 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}} \|\Delta^{c+}\|_F \\ &\quad + \rho_{n1} (|\Delta_{E_u^c}^{c-}|_1 - |\Delta_{E_u^c}^{c-}|_1) \\ &= \frac{1}{4}\underline{k}^2 \|\Delta^c\|_F^2 - (H_1 \sqrt{\frac{\log(M^c L)}{n}} - \rho_{n1}) |\Delta_{E_u^c}^{c-}|_1 - (H_1 \sqrt{\frac{\log(M^c L)}{n}} + \rho_{n1}) |\Delta_{E_u^c}^{c-}|_1 \\ &\quad - H_2 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}} \|\Delta^{c+}\|_F, \end{aligned}$$

since $|\Delta^{c-}|_1 = |\Delta_{E_u^c}^{c-} + \Delta_{E_u^c}^{c-}|_1 = |\Delta_{E_u^c}^{c-}|_1 + |\Delta_{E_u^c}^{c-}|_1$.

Now, if $\rho_{n1} = \max_c \frac{H_1}{\epsilon} \sqrt{\frac{\log(M^c L)}{n}}$ with ϵ sufficiently small, we have that $(H_1 \sqrt{\frac{\log(M^c L)}{n}} - \rho_{n1}) |\Delta_{E_u^c}^{c-}|_1 \geq 0$ and can be excluded from the bound. Moreover, $|\Delta_{E_u^c}^{c-}|_1 \leq \sqrt{M^c L + s^c} \|\Delta^{c-}\|_F$ which implies that

$$\begin{aligned} \mathbf{I}^c &\geq \frac{1}{4}\underline{k}^2 \|\Delta^c\|_F^2 - (H_1 \sqrt{\frac{\log(M^c L)}{n}} + \rho_{n1}) \sqrt{M^c L + s^c} \|\Delta^{c-}\|_F - H_2 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}} \|\Delta^{c+}\|_F \\ &= \|\Delta^{c-}\|_F^2 (\frac{1}{4}\underline{k}^2 - (H_1 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}} + \rho_{n1} \sqrt{M^c L + s^c})) \|\Delta^{c-}\|_F^{-1} \\ &\quad + \|\Delta^{c+}\|_F^2 (\frac{1}{4}\underline{k}^2 - H_2 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}}) \|\Delta^{c+}\|_F^{-1} \end{aligned}$$

since $\|\Delta^c\|_F^2 = \|\Delta^{c+}\|_F^2 + \|\Delta^{c-}\|_F^2$.

Since $H_1 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}} \leq \max_c H_1 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}}$ we have that

$$\begin{aligned} \mathbf{I}^c &\geq \|\Delta^{c-}\|_F^2 (\frac{1}{4}\underline{k}^2 - \max_c (H_1 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}}) (1 + \frac{1}{\epsilon})) \|\Delta^{c-}\|_F^{-1} \\ &\quad + \|\Delta^{c+}\|_F^2 (\frac{1}{4}\underline{k}^2 - H_2 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}}) \|\Delta^{c+}\|_F^{-1}. \end{aligned}$$

Putting it all together one has

$$\begin{aligned}
\text{I}^c + \text{II}^c &\geq \|\Delta^{c-}\|_F^2 \left(\frac{1}{4} \underline{k}^2 - \max_c (H_1 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}}) (1 + \frac{1}{\epsilon}) \|\Delta^{c-}\|_F^{-1} \right) - \rho_{n2} M^1 \|\Delta^{c-}\|_F \\
&\quad + \|\Delta^{c+}\|_F^2 \left(\frac{1}{4} \underline{k}^2 - H_2 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}} \|\Delta^{c+}\|_F^{-1} \right) \\
&= \|\Delta^{c-}\|_F^2 \left(\frac{1}{4} \underline{k}^2 - \max_c (H_1 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}}) (1 + \frac{1}{\epsilon}) \|\Delta^{c-}\|_F^{-1} - \rho_{n2} M^1 \|\Delta^{c-}\|_F^{-1} \right) \\
&\quad + \|\Delta^{c+}\|_F^2 \left(\frac{1}{4} \underline{k}^2 - H_2 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}} \|\Delta^{c+}\|_F^{-1} \right).
\end{aligned}$$

Now if $\rho_{n2} = \max_c \frac{H_1}{\epsilon} \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{(M^1)^2 n}}$ then

$$\begin{aligned}
\text{I}^c + \text{II}^c &\geq \|\Delta^{c-}\|_F^2 \left(\frac{1}{4} \underline{k}^2 - \max_c (H_1 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}}) \|\Delta^{c-}\|_F^{-1} (1 + \frac{1}{\epsilon} - \frac{1}{\epsilon}) \right) \\
&\quad + \|\Delta^{c+}\|_F^2 \left(\frac{1}{4} \underline{k}^2 - H_2 \sqrt{\frac{(M^c L + s^c) \log(M^c L)}{n}} \|\Delta^{c+}\|_F^{-1} \right) \\
&= \|\Delta^{c-}\|_F^2 \left(\frac{1}{4} \underline{k}^2 - O(1/\phi) \right) + \|\Delta^{c+}\|_F^2 \left(\frac{1}{4} \underline{k}^2 - O(1/\phi) \right)
\end{aligned}$$

due to the definition of the set $\mathcal{A}_n(\phi)$. Now, for a sufficiently large ϕ we can have $(\frac{1}{4} \underline{k}^2 - O(1/\phi)) > 0$ implying that $\text{I}^c + \text{II}^c > 0$. $\square$

2 Proof of Proposition 2

Proof. We start by noting that Chiou and Müller (2014) obtained that

$$\int_{\mathcal{T}} \psi_{rp}^c(t) \psi_{sq}^c(t) \hat{\phi}_{rl}^c(t) \hat{\phi}_{sm}^c(t) dt - \int_{\mathcal{T}} \psi_{rp}^c(t) \psi_{sq}^c(t) \phi_{rl}^c(t) \phi_{sm}^c(t) dt = O_p(n^{-1/2}).$$

Consequently, if one can bound the maximal difference between $(\hat{\Theta}^c)_{rl,sm}^{-1}$ and $E(\xi_{rl}^c \xi_{sm}^c)$ the desired result follows. Note that rewritten in matrix notation, we have

$$\max_{rl,sm} |(\hat{\Theta}^c)_{rl,sm}^{-1} - E(\xi_{rl}^c \xi_{sm}^c)| = \|\hat{\Sigma}^c - \Sigma_0^c\|_{\max} \leq \|\hat{\Sigma}^c - \Sigma_0^c\|_2,$$

where $\|M\|_{\max} = \max_{ab} |M_{ab}|$ is the max-norm of a matrix and $\|M\|_2 = \sqrt{\text{eig}_{\max}(M^\top M)}$ is its spectral norm.

The following inequalities hold,

$$\begin{aligned}
\|\hat{\Sigma}^c - \Sigma_0^c\|_2 &= \|\hat{\Sigma}^c (\hat{\Theta}^c - \Theta_0^c) \Sigma_0^c\|_2 \\
&\leq \|\hat{\Sigma}^c\|_2 \times \|\hat{\Theta}^c - \Theta_0^c\|_2 \times \|\Sigma_0^c\|_2 \\
&\leq \|\hat{\Sigma}^c\|_2 \times \|\hat{\Theta}^c - \Theta_0^c\|_F \times \|\Sigma_0^c\|_2.
\end{aligned}$$

Under condition (a) $\|\Sigma_0^c\|_2 = O(1)$. Moreover, as in Lam and Fan (2009),

$$\begin{aligned}\|\hat{\Sigma}^c\|_2 &= (\text{eig}_{\min}(\hat{\Theta}^c))^{-1} = (\text{eig}_{\min}(\hat{\Theta}^c - \Theta_0^c + \Theta_0^c))^{-1} \\ &\leq (\text{eig}_{\min}(\hat{\Theta}^c - \Theta_0^c) + \text{eig}_{\min}(\Theta_0^c))^{-1} = O_p(1).\end{aligned}$$

The above inequality comes from the application of Weyl's theorem, while the final $O_p(1)$ result comes from observing that again under condition (a) $\|\Theta_0^c\|_2 = O(1)$, while the assumed rate $\|\hat{\Theta}^c - \Theta_0^c\|_2 = O_p(\sqrt{\eta_n})$ with $\eta_n \rightarrow 0$ implies that $\text{eig}_{\min}(\hat{\Theta}^c - \Theta_0^c)$ decays fast and is of the order $o_p(1)$, thus making the dominating term of order $O_p(1)$. The claim in the proposition follows by combining this result with the bound on $\|\hat{\Theta}^c - \Theta_0^c\|_F$. $\square$

3 Extra figures

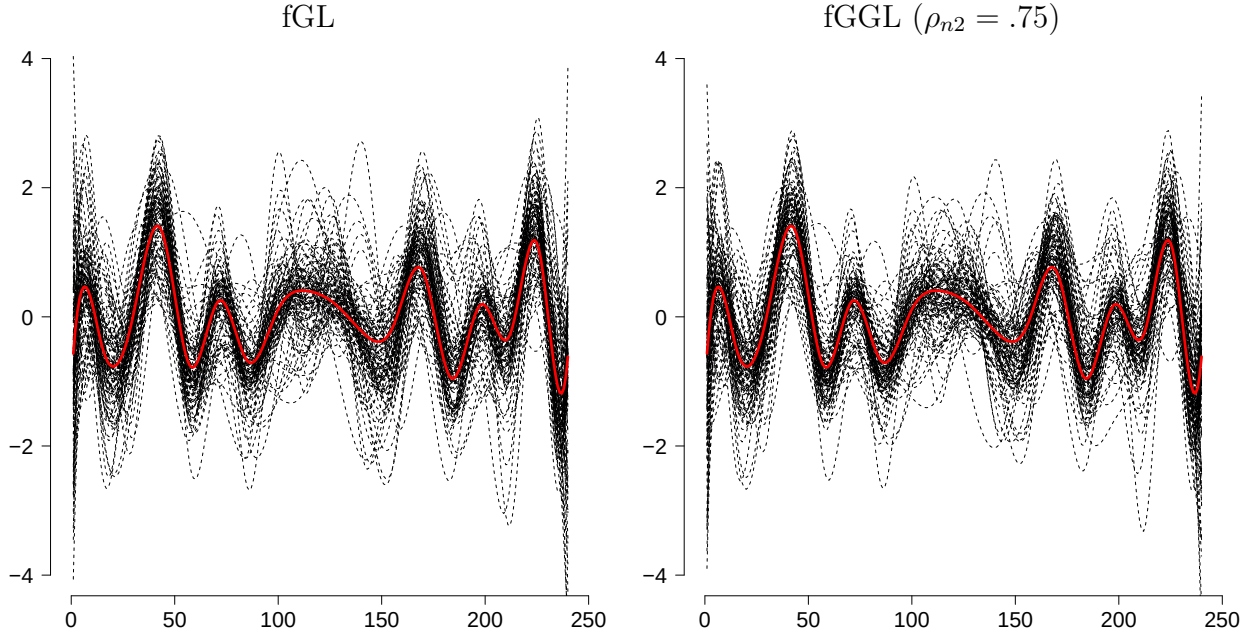


Figure 1: Simulation data with $\sigma = .5$. Estimated (dotted lines) and true (solid line) functions for fGL and fGGL($\rho_{n2} = .75$) when $\rho_{n1} = .25$ for both competitors.

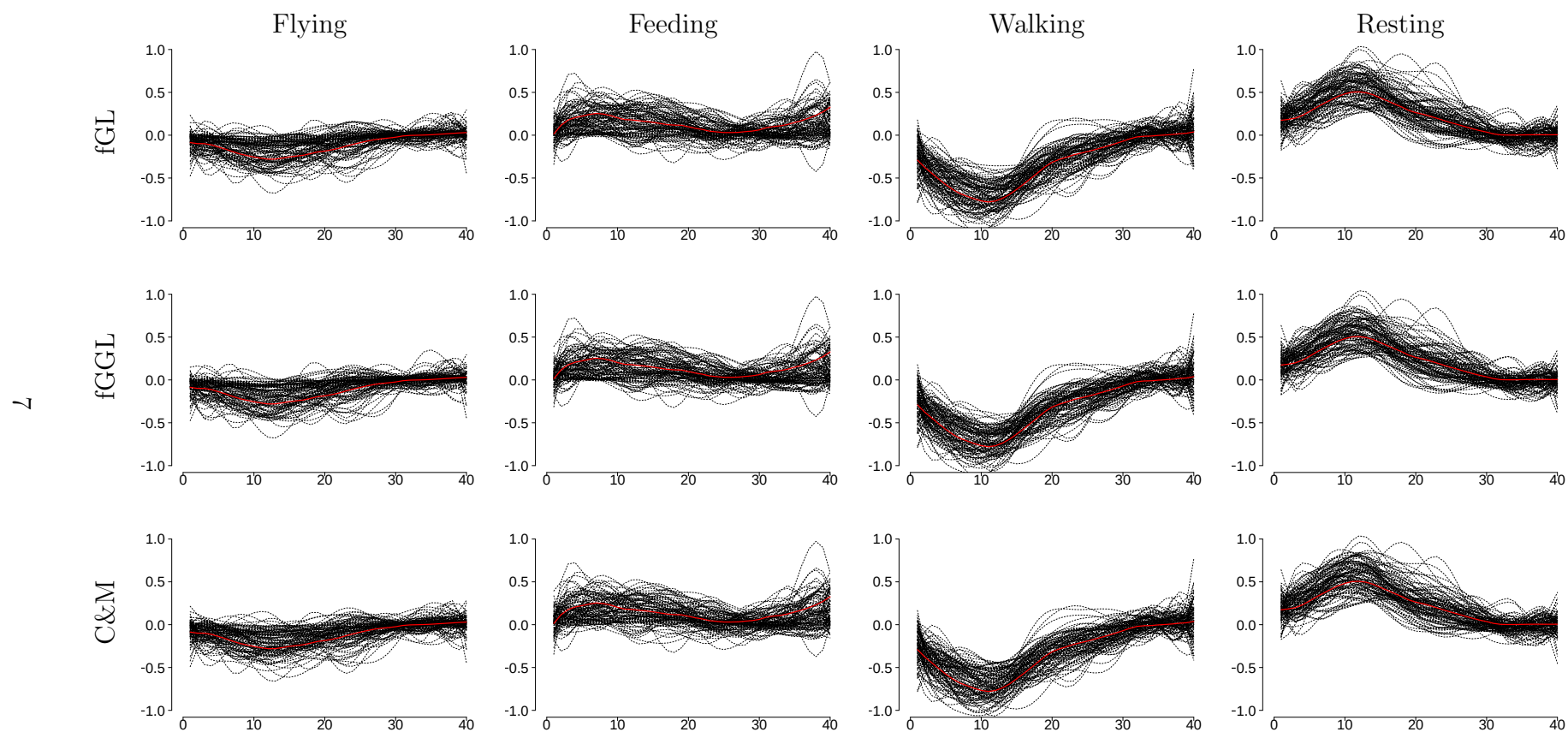


Figure 2: Simulation data with $\sigma = .5$. Estimated (dotted lines) and true (solid line) functions (Flying, Feeding, Walking and Resting) for fGL and fGGL when $\rho_{n1} = .25$ and $\rho_{n2} = .75$. The results of the C&M procedure are added for comparison.

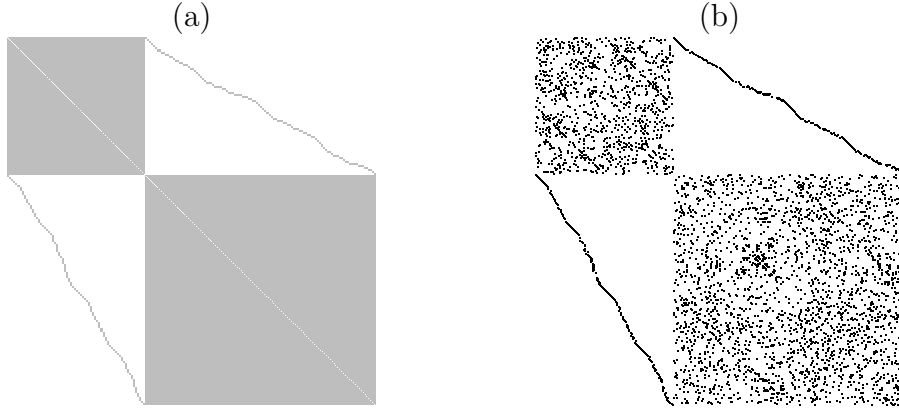


Figure 3: Schematic representation of allowed edges corresponding to the two-scale fMRI dataset used to generate simulated data in Section 6 with $L = 2$ in panel (a) and estimated graph based on eq. (6) from the main text, in panel (b) using the regularization parameters $\rho_{n1} = 2.22$ and $\rho_{n2} = .39$. Gray dots (from panel a) represent an edge allowed to be estimated between nodes, while black dots (from panel b) represent an estimated edge between nodes. In each graph the ‘smaller’ bulk of points (top left) represents edges for the first scale and the ‘larger’ bulk of points (bottom right) represents edges for the second scale. The coarser regions of scale 1 have been split in finer regions at scale 2 and the splits are denoted (in both panels) by the two ‘lines’ above/below the diagonal.

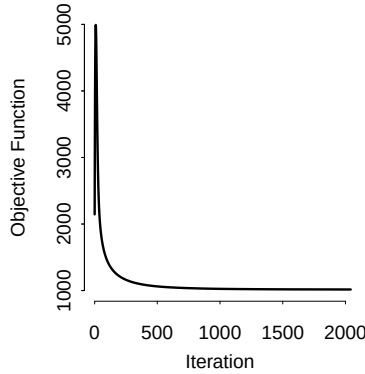


Figure 4: fMRI data. Convergence assessment of the group graph estimation procedure with regularization parameters $\rho_{n1} = 4.24$ and $\rho_{n2} = .64$. A numerical tolerance of $\epsilon = 10^{-4}$ has been used and convergence is achieved after 2042 iterations.

References

- Chiou, J.-M. and H.-G. Müller (2014). Linear manifold modelling of multivariate functional data. *Journal of the Royal Statistical Society: Series B* 76(3), 605–626.
- Lam, C. and J. Fan (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics* 37(6B), 4254–4278.

Rothman, A. J., P. J. Bickel, E. Levina, and J. Zhu (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* 2, 494–515.