WB-GRAPHS: A WITHIN VERSUS BETWEEN GROUP SIMILARITY INTERPLAY

Eugen Pircalabelu

LIDAM Discussion Paper ISBA
2022 / 07

UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

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

WB-graphs: a within versus between group similarity interplay

Eugen Pircalabelu

UCLouvain, Institute of Statistics, Biostatistics and Actuarial Sciences
Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium
eugen.pircalabelu@uclouvain.be

Abstract

A method for balancing the within/between group similarity is proposed in the framework of estimating graphs when data from multiple groups or classes are available. The method leverages connections between the Karush-Kuhn-Tucker (KKT) conditions for estimating ℓ_1 penalized graphs in order to define an optimization problem that can be solved with already existing, known algorithms from the literature. The method is illustrated on an fMRI dataset and with a simulated, controlled experiment. Statistical guarantees are as well provided.

Keywords: Within/between group similarity, Penalized graphical models, Differential networks

1 Introduction

A probabilistic graphical model (PGM) is a pair $(G,\ P)$ where G is a graph based on a set of vertices V and a set of edges E, while P is the probability distribution of a random vector $\boldsymbol{Y} = (Y_1, ..., Y_p)^{\mathsf{T}} \sim P$ where each component $Y_j, j = 1, \ldots, p$ corresponds to a vertex from the set $V = \{1, \ldots, p\}$. The graph summarizes the information coming from multivariate data in a graphical format where nodes, corresponding to features, are linked by edges that indicate dependence relations between the nodes. The goal is to estimate the structure of the graph (which nodes connect to which other nodes) from data.

A very convenient and computationally tractable assumption is to let $\boldsymbol{Y} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1})$. Under the Gaussian assumption, it is well known that if $\boldsymbol{\Theta}_{ab} \neq 0$ then this corresponds to an undirected edge linking nodes a and b in the graph G where E is a set of undirected edges of the form $a - b$ between a pair of nodes (a, b) with $a \neq b$. By $\boldsymbol{\Theta}_{ab}$ we denote the element on line a and column b of the matrix $\boldsymbol{\Theta}$.

Given a sample of n iid vectors, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ from $\boldsymbol{Y}$, if p is close to n or if $p > n$, an estimator for $\boldsymbol{\Theta}$ is obtained by minimizing a penalized log-likelihood function of the form

$$\hat{\boldsymbol{\Theta}} = \arg\min_{\Theta \succ 0} \Big\{ \text{trace}(\boldsymbol{S\Theta}) - \log\det \boldsymbol{\Theta} + P_\lambda(\boldsymbol{\Theta}) \Big\} \tag{1}$$

under the constraint that $\boldsymbol{\Theta} \succ 0$ (positive definite) and where $\boldsymbol{S} = (1/n)\sum_{i=1}^{n} \boldsymbol{Y}_i \boldsymbol{Y}_i^{\top}$ is the empirical covariance matrix, assuming without loss of generality that the data are centred, while $P_\lambda(\boldsymbol{\Theta})$ is a suitable penalty function applied to the entries of $\boldsymbol{\Theta}$, which depends on a regularization parameter λ. Probably the most well known and used penalty function for graphs is the ℓ_1 sparsity enforcing penalty $P_\lambda(\boldsymbol{\Theta}) := \lambda \sum_{a \neq b} |\boldsymbol{\Theta}_{ab}|$. For the estimation of penalized sparse undirected graphs we refer to Friedman et al. (2008), Ravikumar et al. (2008), Bickel and Levina (008a,b), Boyd et al. (2011), Witten et al. (2011), Mazumder and Hastie (2012) and Pircalabelu and Claeskens (2020) among many others.

The optimization problem in (1) assumes that all the observed information comes from the same multivariate normal distribution. Nevertheless, in many applications information could come from multiple different groups of subjects. It is expected that the estimated graphs for similar groups share a large number of edges, but at the same time, dissimilar groups share a

low number of edges. For instance, a disease could be studied through the information gathered from healthy subjects and patients, followed by the estimation of graphs. It seems logical to expect that if the disease is present, the graphs estimated on patients will be similar to those of other patients. Similar graphs would also be expected for the group of healthy subjects, however if the disease is active and influences the functioning and response of the organism, then it might be that the graphs under these two conditions (healthy/non-healthy) can have very different topologies. Multiple works have proposed procedures that estimate multiple graphs while enforcing similarity as in the works of Guo et al. (2011) and Danaher et al. (2014) or estimate differential edges as in the work of Zhao et al. (2014). For a recent survey on differential networks we refer the reader to Shojaie (2020).

The proposed procedure in this article is motivated by an fMRI study. Recently, Richardson et al. (2018) investigated the relationship between neural development and changes in reasoning about others for 122 children between the ages of 3 to 12. A reference group of 33 adults was also present in the study. All 155 participants were asked to watch a short movie focusing on mental states and physical sensations of the characters, while they were subjected to fMRI. One of the conclusions of the study was that correlations between two brain networks (one focusing on how children develop understanding of others' minds and one brain network focusing on perceiving the physical pain of others) vastly differ between groups of participants. Leveraging this conclusion, one expects that brain pathways estimated with undirected graphs, one on the group of adults and one on the group of older children between the ages of 8 to 12 might share similarity. At the same time, one expects the brain pathways of 3 year old subjects to differ more relative to the other two groups (adults and 8-12 yo).

2 Proposed procedure

Consider $\boldsymbol{Y}^{(1)} = (Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_p^{(1)})^\mathsf{T}$ as a vector of features of interest for the first group, that will be mapped to nodes in a graph. We assume further that $\boldsymbol{Y}^{(1)} \sim N(\mathbf{0}, \boldsymbol{\Sigma}^{(1)} = [\boldsymbol{\Theta}^{(1)}]^{-1})$.

Similarly, we consider as well a vector of features of interest for the second group $\boldsymbol{Y}^{(2)} \sim N(\mathbf{0}, [\boldsymbol{\Theta}^{(2)}]^{-1})$ and assume the existence of centred sample data $\boldsymbol{Y}_1^{(1)}, \ldots, \boldsymbol{Y}_{n_1}^{(1)}$ and $\boldsymbol{Y}_1^{(2)}, \ldots, \boldsymbol{Y}_{n_2}^{(2)}$, where n_1 and n_2 denote the two sample sizes of the two groups.

Using the first order KKT optimality conditions for minimizers $\hat{\boldsymbol{\Theta}}^{(1)}$ and $\hat{\boldsymbol{\Theta}}^{(2)}$ of (1) for each sample, we have that:

$$\begin{cases} \boldsymbol{S}^{(1)} - [\hat{\boldsymbol{\Theta}}^{(1)}]^{-1} + \lambda \boldsymbol{Z}^{(1)} = \mathbf{0} & (2) \\ \boldsymbol{S}^{(2)} - [\hat{\boldsymbol{\Theta}}^{(2)}]^{-1} + \lambda \boldsymbol{Z}^{(2)} = \mathbf{0} & (3) \end{cases}$$

where $\boldsymbol{S}^{(1)} = (1/n_1)\sum_{i=1}^{n_1} \boldsymbol{Y}_i^{(1)}[\boldsymbol{Y}_i^{(1)}]^\top$ and $\boldsymbol{S}^{(2)} = (1/n_2)\sum_{i=1}^{n_2} \boldsymbol{Y}_i^{(2)}[\boldsymbol{Y}_i^{(2)}]^\top$ are the sample covariance matrices, while $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ belong to the subdifferential of the norm evaluated at $\hat{\boldsymbol{\Theta}}^{(1)}$ and $\hat{\boldsymbol{\Theta}}^{(2)}$ respectively. It is clear that (2) and (3) are completely disconnected from each other and estimators could be obtained solving twice the same optimization problem, but with different sample data, namely $\boldsymbol{S}^{(1)}$ and $\boldsymbol{S}^{(2)}$. We assume for simplicity that the regularization parameter λ is the same for both optimization problems.

Assuming next that the graphs underlying $\boldsymbol{\Theta}^{(1)}$ and $\boldsymbol{\Theta}^{(1)}$ have a similar structure, by which it is understood that many of the edges are placed between the same pairs of nodes (a, b) in both graphs, it is reasonable to estimate jointly related graphs and propose estimators that satisfy:

$$\boldsymbol{S}^{(p)} - \hat{\boldsymbol{\Sigma}}^{(p)} + \lambda(\boldsymbol{Z}^{(1)} + \boldsymbol{Z}^{(2)}) = \mathbf{0}, \quad (4)$$

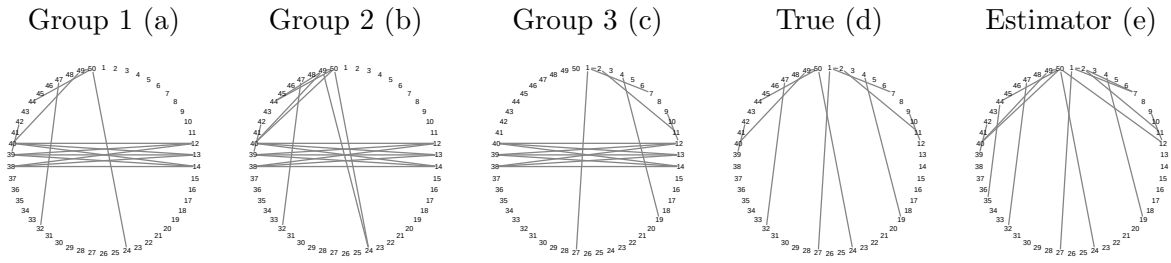

Figure 1: Toy example illustrating graphs of size 50 for 3 groups. Groups 1 and 2 are similar to each other and share most of the edges, while the graph for group 3 is different. Panel (d) illustrates edges that differentiate between the three graphs, while panel (e) illustrates the estimated WB-graph. None of the graphs in panels (a)-(d) are known in practice and the goal is to accurately estimate the graph in panel (d).

where $\boldsymbol{S}^{(p)} = \boldsymbol{S}^{(1)} + \boldsymbol{S}^{(2)}$, $\hat{\boldsymbol{\Sigma}}^{(p)} = \hat{\boldsymbol{\Sigma}}^{(1)} + \hat{\boldsymbol{\Sigma}}^{(2)}$ with $\hat{\boldsymbol{\Sigma}}^{(1)} = [\hat{\boldsymbol{\Theta}}^{(1)}]^{-1}$ and $\hat{\boldsymbol{\Sigma}}^{(2)} = [\hat{\boldsymbol{\Theta}}^{(2)}]^{-1}$, respectively. However, in practice estimators $\hat{\boldsymbol{\Theta}}^{(1)}$ and $\hat{\boldsymbol{\Theta}}^{(2)}$ cannot be obtained from (4) as it contains many unknown quantities. The similarity between structures could arise from different factors. For example, in the brain development study of Richardson et al. (2018) the similarity pattern is driven by age, with children of similar ages sharing more similarity with respect to cognitive development.

In the sequel we introduce a third group (eg. adults in the brain development example), for which we consider $\boldsymbol{Y}^{(3)} \sim N(\boldsymbol{0}, [\boldsymbol{\Theta}^{(3)}]^{-1})$ and where we assume that the structure of the underlying graph is largely different from that underlying $\boldsymbol{\Theta}^{(1)}$ and $\boldsymbol{\Theta}^{(2)}$. Building on the above we have that

$$\begin{cases} \boldsymbol{S}^{(p)} - \hat{\boldsymbol{\Sigma}}^{(p)} + \lambda(\boldsymbol{Z}^{(1)} + \boldsymbol{Z}^{(2)}) & = \boldsymbol{0} \\ \boldsymbol{S}^{(3)} - [\hat{\boldsymbol{\Theta}}^{(3)}]^{-1} + \lambda \boldsymbol{Z}^{(3)} & = \boldsymbol{0} \end{cases}$$

leading to

$$(\boldsymbol{S}^{(p)} - \boldsymbol{S}^{(3)}) - \hat{\boldsymbol{\Sigma}}^{(p)} + \hat{\boldsymbol{\Sigma}}^{(3)} + \lambda(\boldsymbol{Z}^{(1)} + \boldsymbol{Z}^{(2)} - \boldsymbol{Z}^{(3)}) = \boldsymbol{0}, \tag{5}$$

which again cannot directly be used to obtain estimators $\hat{\boldsymbol{\Theta}}^{(1)}$, $\hat{\boldsymbol{\Theta}}^{(2)}$ and $\hat{\boldsymbol{\Theta}}^{(3)}$, but offers a different intuition.

Equation (4) reflects the 'within group' similarity aspect of the graphs, while equation (5) brings in the 'between group' dissimilarity. As such, estimators in (5) reflect a balancing act between estimating edges that are similar between the graphs corresponding to $\boldsymbol{\Theta}^{(1)}$ and $\boldsymbol{\Theta}^{(2)}$ and edges that are different relative to the graph underlying $\boldsymbol{\Theta}^{(3)}$.

Figure 1 illustrates with a toy example what is meant by the interplay within/between group graphs. Panels (a) and (b) show the underlying graphs for Groups 1 and 2 when $p = 50$. The graphs are highly similar in the sense that many edges find themselves on the same positions. The graph for Group 3 in panel (c) shares edges with the graphs of Groups 1 and 2 (for all three graphs nodes labeled 12, 13, 14 connect with nodes 38, 39, 40) but contains as well edges that are specific to this condition and which find themselves on different positions relative to graphs 1 and 2. Since the goal is to discriminate between the three graphs, the edges between nodes 12, 13, 14 and nodes 38, 39, 40 are not informative since they are common to all groups. As such, the graph in panel (d) illustrates edges that are common between groups 1 and 2 and different relative to group 3. Note that there is a 'symmetry' aspect as well, in the sense that the graph in panel (d) also illustrates the specific edges pertaining to Group 3 that are different from edges

contained in graphs 1 and 2. It is crucial to mention at this stage that *none* of the graphs in panels (a)-(d) are available in practice, since they are all population quantities. As such, in panel (e) we illustrate the estimated WB-graph obtained by using the proposed procedure.

Inspired by (5), we focus further on the population quantity $\boldsymbol{\Delta} \coloneqq (\boldsymbol{\Sigma}^{(1)} + \boldsymbol{\Sigma}^{(2)})^{-1} - \boldsymbol{\Theta}^{(3)}$. The reason for it is that

$$\begin{array}{lll} (\boldsymbol{\Sigma}^{(1)} + \boldsymbol{\Sigma}^{(2)})\boldsymbol{\Delta}\boldsymbol{\Sigma}^{(3)} & = (\boldsymbol{\Sigma}^{(1)} + \boldsymbol{\Sigma}^{(2)})[(\boldsymbol{\Sigma}^{(1)} + \boldsymbol{\Sigma}^{(2)})^{-1} - \boldsymbol{\Theta}^{(3)}]\boldsymbol{\Sigma}^{(3)} & = \boldsymbol{\Sigma}^{(3)} - (\boldsymbol{\Sigma}^{(1)} + \boldsymbol{\Sigma}^{(2)}) \\ (\boldsymbol{S}^{(1)} + \boldsymbol{S}^{(2)})\boldsymbol{\Delta}\boldsymbol{S}^{(3)} & = (\boldsymbol{S}^{(1)} + \boldsymbol{S}^{(2)})[(\boldsymbol{\Sigma}^{(1)} + \boldsymbol{\Sigma}^{(2)})^{-1} - \boldsymbol{\Theta}^{(3)}]\boldsymbol{S}^{(3)} & \approx \boldsymbol{S}^{(3)} - (\boldsymbol{S}^{(1)} + \boldsymbol{S}^{(2)}) \end{array} \tag{6}$$

where the approximation is accurate enough if the sample covariances are close to their population counterparts.

The same reasoning could be extended to multiple, say L groups and in light of (6) and since (5) cannot directly be used to obtain useful estimators, we propose the optimization problem:

$$\min ||\boldsymbol{\Delta}||_1 \quad \text{s.t.}$$

$$\left|\left|\left(\sum_{k=1}^{K} \frac{n_k \boldsymbol{S}^{(k)}}{\sum_{k=1}^{K} n_k}\right)\boldsymbol{\Delta}\left(\sum_{l=K+1}^{L} \frac{n_l \boldsymbol{S}^{(l)}}{\sum_{l=K+1}^{L} n_l}\right) + \sum_{l=K+1}^{L} \frac{n_l \boldsymbol{S}^{(l)}}{\sum_{l=K+1}^{L} n_l} - \sum_{k=1}^{K} \frac{n_k \boldsymbol{S}^{(k)}}{\sum_{k=1}^{K} n_k}\right|\right|_\infty \leq \lambda_n. \tag{7}$$

In (7) by $||M||_1$ we denote the ℓ_1 norm of M defined as $||M||_1 = \sum_{ab} |M_{ab}|$ and $||M||_\infty = \max_{ab} |M_{ab}|$ for any general matrix M. It also reflects that we desire $\boldsymbol{\Delta}$ to be sparse and where we assume a total of L groups exist of which wlog the first K are similar amongst themselves, but different from the remaining $L-K$. Since the sample sizes $n_1, \ldots, n_L$ might be of different sizes, each empirical covariance matrix is weighted accordingly. An optimizer for (7) can be obtained using the ADMM algorithm of Zhao et al. (2014) and for selecting an optimal regularization parameter we propose to use similar AIC/BIC-like criteria defined as:

$$\begin{aligned} \text{AIC} &= (\sum_k n_k + \sum_l n_l)\text{Loss}(\lambda_n) + 2\nu \\ \text{BIC} &= (\sum_k n_k + \sum_l n_l)\text{Loss}(\lambda_n) + \log(\sum_k n_k + \sum_l n_l)\nu \end{aligned}$$

where $\nu = \#\{\hat{\boldsymbol{\Delta}}_{ab} : \hat{\boldsymbol{\Delta}}_{ab} \neq 0\}$ (ie. the number of non-zero elements in the estimated matrix) and where $\text{Loss}(\lambda_n) = \left|\left|\left(\sum_k \frac{n_k \boldsymbol{S}^{(k)}}{\sum_k n_k}\right)\hat{\boldsymbol{\Delta}}\left(\sum_l \frac{n_l \boldsymbol{S}^{(l)}}{\sum_l n_l}\right) + \sum_l \frac{n_l \boldsymbol{S}^{(l)}}{\sum_l n_l} - \sum_k \frac{n_k \boldsymbol{S}^{(k)}}{\sum_k n_k}\right|\right|_\infty$.

A few links with already existing literature can be noted. In Lee and Liu (2015) the authors also propose a joint estimation of multiple concentration matrices that share a common structure. However, the assumption there is that given L separate groups, for each group the precision matrix can be written as an additive process between a mean matrix that is common to all groups and a unique (pertaining to that group only) residual matrix. The goal of that work is thus, to estimate jointly $L+1$ matrices (a mean matrix and L individual matrices). Quite differently, in this paper we are interested in estimating one single matrix that balances the within group similarity relative to the between group dissimilarity.

In the discussion section of their paper, as a possible future extension of their method, Zhao et al. (2014) briefly suggested also something similar to the idea of Lee and Liu (2015), in the sense that they propose to first construct a pooled matrix across all groups and then estimate L differential networks, where each network quantifies its 'distance' from the pooled matrix. In the final step they consider comparing the estimates pairwise, between two groups at a time. Quite differently, in this work we estimate only one single matrix, hence avoid the pairwise comparison and L separate estimations, that accounts for all similarities/differences at the same time. Our proposal is built up from the KKT conditions, as in the spirit of Cai et al. (2011).

3 Computational aspects

In Section 2 we have stated that an optimizer for (7) can directly be obtained using the CLIME-based differential approach of Zhao et al. (2014). The reason for our claim is developed in this section.

Assuming $\boldsymbol{\Delta}$ is symmetric, (7) can be recast as:

$$\min ||\boldsymbol{\Delta}||_1 \quad \text{s.t.}$$

$$\left|\left|\left(\sum_l \frac{n_l \boldsymbol{S}^{(l)}}{\sum_l n_l}\right) \otimes \left(\sum_k \frac{n_k \boldsymbol{S}^{(k)}}{\sum_k n_k}\right) \text{vec}(\boldsymbol{\Delta}) - \text{vec}\left(\sum_k \frac{n_k \boldsymbol{S}^{(k)}}{\sum_k n_k} - \sum_l \frac{n_l \boldsymbol{S}^{(l)}}{\sum_l n_l}\right)\right|\right|_\infty \leq \lambda_n \tag{8}$$

where $M \otimes Q$ denotes the Kronecker product between two general matrices M and Q and where $\text{vec}(M)$ denotes the vectorization procedure of the matrix M.

We further have that

$$\begin{aligned}\left(\sum_l \frac{n_l \boldsymbol{S}^{(l)}}{\sum_l n_l}\right) \otimes \left(\sum_k \frac{n_k \boldsymbol{S}^{(k)}}{\sum_k n_k}\right) \text{vec}(\boldsymbol{\Delta}) &= \left(\sum_{l,k} \frac{n_l n_k}{\sum_{l,k} n_l n_k} \boldsymbol{S}^{(l)} \otimes \boldsymbol{S}^{(k)}\right) \text{vec}(\boldsymbol{\Delta}) \\ &= \left(\sum_{l,k} \frac{n_l n_k}{\sum_{l,k} n_l n_k} \boldsymbol{S}^{(l)} \otimes \boldsymbol{S}^{(k)}\right) \mathcal{S}\boldsymbol{\beta},\end{aligned}$$

where $\mathcal{S}$ is a fully known, duplication matrix (see Chapter 16 of Harville, 1997) containing only the values $\{0,1\}$ and $\boldsymbol{\beta}$ is a $p(p+1)/2$ vector containing the unique elements of $\boldsymbol{\Delta}$. As the duplication matrix is used only for the unknown parameters, since by Lemma 2 there exists a unique solution, call it $\boldsymbol{\beta}_0$ to the problem

$$\mathcal{S}^\mathsf{T}\left[\left(\frac{1}{L-K}\sum_l \boldsymbol{\Sigma}^{(l)}\right) \otimes \left(\frac{1}{K}\sum_k \boldsymbol{\Sigma}^{(k)}\right)\right]\mathcal{S}\boldsymbol{\beta} - \mathcal{S}^\mathsf{T}\text{vec}\left(\frac{1}{K}\sum_k \boldsymbol{\Sigma}^{(k)} - \frac{1}{L-K}\sum_l^L \boldsymbol{\Sigma}^{(l)}\right) = \mathbf{0}$$

and since Lemma 1 provides an explicit control over the maximal deviation between the average empirical covariances and the target, one can safely proceed as in Zhao et al. (2014) and estimate a sparse $\boldsymbol{\beta}_0$ by solving with the ADMM algorithm:

$$\hat{\boldsymbol{\beta}} = \arg\min ||\boldsymbol{\beta}||_1 \quad \text{s.t.}$$

$$\left|\left|\mathcal{S}^\mathsf{T}\left[\left(\sum_l \frac{n_l \boldsymbol{S}^{(l)}}{\sum_l n_l}\right) \otimes \left(\sum_k \frac{n_k \boldsymbol{S}^{(k)}}{\sum_k n_k}\right)\right]\mathcal{S}\boldsymbol{\beta} - \mathcal{S}^\mathsf{T}\text{vec}\left(\sum_k \frac{n_k \boldsymbol{S}^{(k)}}{\sum_k n_k} - \sum_l \frac{n_l \boldsymbol{S}^{(l)}}{\sum_l n_l}\right)\right|\right|_\infty \leq \lambda_n. \tag{9}$$

4 Properties

We list below four needed regularity conditions, but introduce first some notation.

Let σ_{ab}^K and σ_{ab}^L denote the (a,b) entries of the two average matrices $\frac{1}{K}\sum_{k=1}^K \boldsymbol{\Sigma}^{(k)}$ and $\frac{1}{L-K}\sum_{l=K+1}^L \boldsymbol{\Sigma}^{(l)}$. Denote next by $\sigma_{\max}^K = \max_a \sigma_{aa}^K$ and $\sigma_{\max}^L = \max_a \sigma_{aa}^L$ the largest elements on the diagonal of each matrix. Similarly, let $\mu_{\max}^K = \max_{a\neq b} |\sigma_{ab}^K|$ and $\mu_{\max}^L = \max_{a\neq b} |\sigma_{ab}^L|$ the largest, absolute off-diagonal elements for each of the two matrices. Furthermore, let $\sigma_{\min}^{\mathcal{S}} = \min_{ab}(\sigma_{aa}^L\sigma_{aa}^K, \sigma_{bb}^L\sigma_{aa}^K + 2\sigma_{ba}^L\sigma_{ab}^K + \sigma_{aa}^L\sigma_{bb}^K)$.

C1 (adapted from Cai et al., 2011): There exists some $0 < \eta < 1/4$ and a constant M such that $\frac{L \log p}{\sum_{r=1}^{L} n_r} \leq \eta$ and $\mathrm{E}(\exp(t(\boldsymbol{Y}_i^{(r)}))) \leq M < \infty$ for any $|t| \leq \eta$, any $r = 1, \ldots, L$ and any $i = 1, \ldots n_r$.

C2 (Zhao et al., 2014): The true difference matrix $\boldsymbol{\Delta}_0$ has $s < p$ non-zero entries and there exists a constant R (that does not depend on p) such that $||\boldsymbol{\Delta}_0||_1 < R$.

C3 (Zhao et al., 2014): With s as defined in C2, suppose that $\mu = 4 \max(\mu_{\max}^K \sigma_{\max}^K, \mu_{\max}^L \sigma_{\max}^L) < \sigma_{\min}^{\mathcal{S}}/(2s)$.

C4 As $L \to \infty$ suppose there exist constants R'_{ab} such that $\frac{1}{L} \sum_{r=1}^{L} \boldsymbol{\Sigma}_{ab}^{(r)} \leq R'_{ab}$ for any pair (a, b).

Condition (1) is a standard condition as in the spirit of that of Cai et al. (2011) requiring exponential-tail behaviour for the components of $\boldsymbol{Y}^{(k)}$, while accounting for the fact that L groups are used. Condition (2) specifies that the ℓ_1 norm of the true matrix $\boldsymbol{\Delta}$ stays bounded, while similarly condition (4) specifies that as the total number of groups grows, for any pair (a, b) the sum $\sum_{r=1}^{L} \boldsymbol{\Sigma}_{ab}^{(r)}$ cannot grow at a faster rate than L. Condition (3) is a technical condition to obtain the desired convergence rates.

Proposition 1. *Under conditions (1)-(4) if* $\frac{L \log p}{n_\dagger} = o(1)$ *and* $\lambda_n \geq C(\frac{L \log p}{n_\dagger})^{1/2}$ *for a sufficiently large constant* C*, then with probability greater than* $1 - 4p^{-\tau}$ *we have*

$$||\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}_0||_\infty \leq \kappa_{s,\mu,\sigma_{\min}^{\mathcal{S}}} \left(\frac{L \log p}{n_\dagger} \right)^{1/2},$$

where κ *is a constant that depends on the sparsity* s *and the constants* $\mu, \sigma_{\min}^{\mathcal{S}}$ *and where* $n_\dagger = \min(n_1, n_2, \ldots, n_L)$.

Proposition 1 shows that in sup-norm the estimator that is obtained based on the interplay within/between group similarity, converges at the usual rate for such penalized problems $\left(\log p/n_\dagger\right)^{1/2}$ times a $\sqrt{L}$ factor, denoting the price to pay for having L groups involved in the estimation. The proof and supporting Lemmas are presented in Appendix I.

The estimator in (9) can be further slightly modified to show that it is also sign consistent. The modification requires an additional thresholding step. For $\tau_n > 0$ let

$$\hat{\boldsymbol{\Delta}}_{ab}^{\tau_n} = \hat{\boldsymbol{\Delta}}_{ab} \mathrm{I}(|\hat{\boldsymbol{\Delta}}_{ab}| > \tau_n)$$

be the (a, b)-th element of the thresholded estimator $\hat{\boldsymbol{\Delta}}^{\tau_n}$, where $\mathrm{I}(\cdot)$ is the indicator function taking the value 1 if the condition is satisfied and 0, otherwise.

Proposition 2. *Under conditions (1)-(4) if in addition* $\min_{ab:\boldsymbol{\Delta}_{ab,0} \neq 0} |\boldsymbol{\Delta}_{ab,0}| > 2\tau_n$, $\frac{L \log p}{n_\dagger} = o(1)$, $\lambda_n \geq C(\frac{L \log p}{n_\dagger})^{1/2}$ *and* $\tau_n \geq \kappa'_{s,\mu,\sigma_{\min}^{\mathcal{S}}} (\frac{L \log p}{n_\dagger})^{1/2}$ *for sufficiently large constants* C *and* $\kappa'_{s,\mu,\sigma_{\min}^{\mathcal{S}}}$*, then with probability greater than* $1 - 4p^{-\tau}$ *we have that* $\mathrm{sgn}(\boldsymbol{\Delta}_{ab}^{\tau_n}) = \mathrm{sgn}(\boldsymbol{\Delta}_{ab,0})$ *for any pair* (a, b) *and where* $\mathrm{sgn}(\cdot)$ *denotes the sign function.*

Proposition 2 specifies that even though the estimator receives parts of data that are similar and parts that are dissimilar, it can still also recover correctly the sign of the elements of the target matrix. Under the specified conditions, the proof follows from that of Theorem 1 in Zhao et al. (2014) and is omitted.

5 Simulation study

In order to evaluate the performance of the proposed procedure, we have performed a finite sample, simulation study where we have generated data coming from three multivariate normal distributions: $\boldsymbol{Y}^{(r)} \sim N(\mathbf{0}, [\boldsymbol{\Theta}^{(r)}]^{-1})$, with $r \in \{1, 2, 3\}$.

For group 1 we generated first a sparse matrix containing either 0s or 1s corresponding to a random graph where the degrees of the vertices followed a power-law distribution $P(m) \sim m^{-\gamma}$ with exponent $\gamma = 2$. Then the matrix has been made symmetric and all diagonal elements were set to 1. Next, to ensure positive definiteness the elements on the diagonal have been replaced by the absolute value of the smallest eigenvalue to which the value .2 was added. In the last step, the matrix has been transformed to a correlation matrix and its inverse has been considered as $\boldsymbol{\Theta}^{(1)}$. For the second group we have created a 'similar' matrix to $\boldsymbol{\Theta}^{(1)}$ modifying five eigenvalues of $\boldsymbol{\Theta}^{(1)}$ and constructing $\boldsymbol{\Theta}^{(2)}$ based on these new eigenvalues. For the last matrix $\boldsymbol{\Theta}^{(3)}$, to make it different from the data generating process underlying groups 1 and 2, a similar process to the creation of $\boldsymbol{\Theta}^{(1)}$ was used, but reversing it to allow for low connecting nodes from $\boldsymbol{\Theta}^{(1)}$ to become high connecting nodes. In the final step for each of the three groups a bulk of common edges was added to each graph. Based on the generating process, the true graph is defined as $\boldsymbol{\Delta}_0 := (\boldsymbol{\Sigma}^{(1)} + \boldsymbol{\Sigma}^{(2)})^{-1} - \boldsymbol{\Theta}^{(3)}$ where $\boldsymbol{\Sigma}^{(1)} = [\boldsymbol{\Theta}^{(1)}]^{-1}$ and $\boldsymbol{\Sigma}^{(2)} = [\boldsymbol{\Theta}^{(2)}]^{-1}$. The sample size n was set to $n \in \{500, 1000, 10000\}$, while the number of nodes was set to $p \in \{50, 100, 175\}$. A total of 100 different datasets for each group was generated from the above generating process.

For ease of understanding, we present in Figure 2 the three adjacency matrices underlying $\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Theta}^{(2)}$ and $\boldsymbol{\Theta}^{(3)}$ as well as the true matrix $\boldsymbol{\Delta}$ when $p = 150$. The black dots in each figure correspond to edges between nodes.

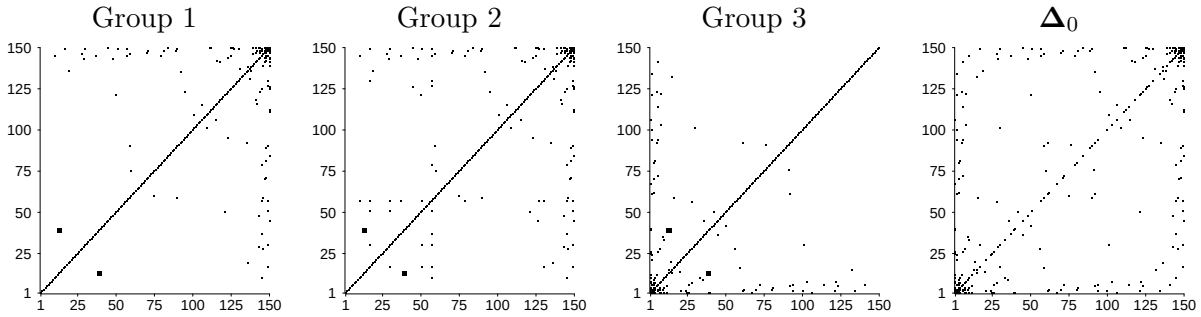

Figure 2: Simulated data. Adjacency matrices with $p = 150$ for $L = 3$ groups where the graph of Group 1 is similar to that of Group 2, while the graph of the 3rd group is different. A black dot in the graphs denotes an edge between two nodes, while all blank entries are interpreted as non-existing edges between pairs of nodes.

We benchmark the WB-graph obtained after optimizing (9) against two other competitors:

1. $\hat{\boldsymbol{\Delta}}_{\text{GL}} = (\hat{\boldsymbol{\Sigma}}^{(1)} + \hat{\boldsymbol{\Sigma}}^{(2)})^{-1} - \hat{\boldsymbol{\Theta}}^{(3)}$ where $\hat{\boldsymbol{\Sigma}}^{(1)} = [\hat{\boldsymbol{\Theta}}^{(1)}]^{-1}, \hat{\boldsymbol{\Sigma}}^{(2)} = [\hat{\boldsymbol{\Theta}}^{(2)}]^{-1}$ and $\hat{\boldsymbol{\Sigma}}^{(3)} = [\hat{\boldsymbol{\Theta}}^{(3)}]^{-1}$ are three estimates for the covariance matrices obtained based on inverting the estimated concentration matrices obtained using the Graphical lasso optimizing (1) for each group separately;

2. $\hat{\boldsymbol{\Delta}}_{\text{JGL}} = (\hat{\boldsymbol{\Sigma}}^{(1)} + \hat{\boldsymbol{\Sigma}}^{(2)})^{-1} - \hat{\boldsymbol{\Theta}}^{(1)}$ where $\hat{\boldsymbol{\Sigma}}^{(1)} = [\hat{\boldsymbol{\Theta}}^{(1)}]^{-1}, \hat{\boldsymbol{\Sigma}}^{(2)} = [\hat{\boldsymbol{\Theta}}^{(2)}]^{-1}$ are estimates for the covariance matrices obtained based on inverting the estimated concentration matrices

obtained using the Joint graphical lasso procedure of Danaher et al. (2014) and $[\hat{\boldsymbol{\Theta}}^{(3)}]^{-1}$ is obtained using the Graphical lasso for the third group alone.

It should be noted that $\hat{\boldsymbol{\Delta}}_{\text{GL}}$ does not *a priori* use any information that groups 1 and 2 share similar graph structures, hence it lets the procedure unravel it from the data, while $\hat{\boldsymbol{\Delta}}_{\text{JGL}}$ explicitly constraints the graphs of groups 1 and 2 to be similar.

All three competitors are evaluated with respect to:

1. TPR= $\frac{\#\text{estimated edges that are true edges}}{\#\text{true edges}}$ (true positive rate, larger is better);
2. FPR= $\frac{\#\text{estimated edges that are NOT true edges}}{\#\text{edges that are NOT present in the true graph}}$ (false positive rate, smaller is better);
3. FDR= $\frac{\#\text{estimated edges that are NOT true edges}}{\#\text{estimated edges}}$ (false discovery rate, smaller is better);
4. $\text{F}_1 = 2\text{PR}/(\text{P}+\text{R})$ where $\text{P} = \frac{\#\text{estimated edges that are true edges}}{\#\text{estimated edges}}$; R =TPR ($\text{F}_1$ score, larger is better);
5. Fr= $||\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}_0||_F = \sqrt{\sum_{a,b}(\hat{\boldsymbol{\Delta}}_{ab} - \boldsymbol{\Delta}_{0,ab})^2}$, where $\hat{\boldsymbol{\Delta}}_{ab}$ (resp. $\boldsymbol{\Delta}_{0,ab}$) is the $\hat{\boldsymbol{\Delta}}$ (resp. $\boldsymbol{\Delta}_0$) element on the a-th row and b-th column (Frobenius loss, smaller is better).

For WB-graphs the regularization parameter λ_n has been selected based on the AIC-like criterion introduced in Section 2. For the competitors (likelihood-based) the regular AIC formula was used to select the parameters and for all competitors a grid of 50 different values was considered. For JGL due to computational complexity, a bivariate grid 20×20 has been constructed for selecting the regularization parameters, where the 20 values were selected equally spaced from the original univariate grid of 50 values.

Table 1 presents the obtained results. It suggests that for the considered data generating processes:
(i) regardless of the combination (n, p) the WB-graphs have a tendency to produce smaller false positive and false discovery rates than the competitors, which in turn is translated into higher F_1 scores;
(ii) for lower sample sizes, the WB-graphs produce lower TPRs, but with a large sample its performance was on par with that of the competitors with respect to the TPR;
(iii) similarly, for low sample sizes the Frobenius norm performance is similar to that of the competitors, but with large sample sizes the WB-graphs produce a better performance relative to competitors.

All in all, for the considered DGPs there is indication that the WB-graphs tend to be more accurate than the competitors in identifying where edges should be placed in the graph (ie. between which couples of nodes a and b) and seem to benefit more from larger samples in the accuracy, as the estimated values are closer to the target as measured by the Frobenius loss.

		$n = 500$			$n = 1000$			$n = 10000$		
		$\hat{\mathbf{\Delta}}_{\text{GL}}$	$\hat{\mathbf{\Delta}}_{\text{JGL}}$	WB-graph	$\hat{\mathbf{\Delta}}_{\text{GL}}$	$\hat{\mathbf{\Delta}}_{\text{JGL}}$	WB-graph	$\hat{\mathbf{\Delta}}_{\text{GL}}$	$\hat{\mathbf{\Delta}}_{\text{JGL}}$	WB-graph
$p = 50$	TPR	1.00 (0.00)	1.00 (0.00)	0.94 (0.06)	1.00 (0.00)	1.00 (0.00)	0.99 (0.03)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
	FPR	0.96 (0.09)	0.16 (0.08)	0.01 (0.01)	0.82 (0.25)	0.09 (0.02)	0.02 (0.01)	0.05 (0.01)	0.05 (0.01)	0.02 (0.01)
	FDR	0.98 (0.01)	0.85 (0.05)	0.34 (0.14)	0.97 (0.03)	0.79 (0.04)	0.38 (0.17)	0.70 (0.05)	0.70 (0.05)	0.38 (0.18)
	F1	0.05 (0.02)	0.25 (0.07)	0.77 (0.10)	0.07 (0.05)	0.35 (0.06)	0.75 (0.12)	0.46 (0.06)	0.46 (0.06)	0.75 (0.14)
	Fr	5.99 (1.10)	5.61 (1.13)	4.37 (0.86)	5.22 (1.08)	5.10 (1.07)	2.98 (0.66)	5.25 (1.23)	5.15 (1.22)	1.05 (0.32)
$p = 100$	TPR	1.00 (0.02)	0.99 (0.02)	0.70 (0.10)	1.00 (0.02)	0.99 (0.04)	0.82 (0.07)	1.00 (0.03)	0.99 (0.03)	0.99 (0.05)
	FPR	0.84 (0.28)	0.19 (0.13)	0.01 (0.01)	0.72 (0.37)	0.09 (0.04)	0.01 (0.01)	0.07 (0.02)	0.06 (0.02)	0.02 (0.01)
	FDR	0.96 (0.05)	0.86 (0.09)	0.36 (0.13)	0.94 (0.09)	0.79 (0.07)	0.35 (0.14)	0.77 (0.06)	0.75 (0.07)	0.43 (0.12)
	F1	0.07 (0.08)	0.24 (0.13)	0.66 (0.10)	0.11 (0.13)	0.34 (0.09)	0.71 (0.11)	0.38 (0.08)	0.40 (0.08)	0.72 (0.10)
	Fr	8.99 (0.87)	8.56 (1.01)	8.33 (1.26)	8.12 (1.28)	7.97 (1.43)	5.79 (1.13)	7.74 (1.24)	7.66 (1.33)	2.14 (1.34)
$p = 175$	TPR	0.99 (0.04)	0.96 (0.08)	0.22 (0.18)	0.99 (0.05)	0.97 (0.08)	0.69 (0.15)	0.99 (0.02)	0.99 (0.04)	0.99 (0.03)
	FPR	0.86 (0.25)	0.28 (0.15)	0.00 (0.01)	0.81 (0.33)	0.20 (0.11)	0.01 (0.03)	0.13 (0.02)	0.13 (0.02)	0.01 (0.00)
	FDR	0.98 (0.02)	0.93 (0.04)	0.38 (0.23)	0.97 (0.03)	0.91 (0.04)	0.31 (0.15)	0.89 (0.02)	0.89 (0.02)	0.39 (0.09)
	F1	0.04 (0.04)	0.12 (0.07)	0.25 (0.16)	0.06 (0.06)	0.15 (0.06)	0.68 (0.14)	0.19 (0.02)	0.19 (0.03)	0.75 (0.08)
	Fr	13.56 (1.07)	12.91 (1.29)	15.97 (1.86)	12.20 (1.24)	11.80 (1.57)	11.38 (1.97)	10.60 (1.45)	10.52 (1.53)	3.14 (1.34)

Table 1: Simulated data. Average performance measures with empirical standard deviations between brackets over 100 different simulated datasets.

6 Real data analysis

We have applied the proposed procedure on the brain development dataset of Richardson et al. (2018). The dataset was retrieved from Python's `nilearn` package where the following pre-processing steps have been applied. First, voxel level time series data have been aggregated into averaged time series data for coarse regions of interest (ROIs) using the MSDL atlas (Varoquaux et al., 2011) and the well known AAL atlas (Tzourio-Mazoyer et al., 2002). The MSDL atlas contains 39 coarse ROIs, while the AAL atlas contains 116 ROIs and is hence slightly finer than the MSDL. After atlas-based aggregation, the time series have been standardized to have empirical mean 0 and variance 1, they have been detrended and confounding factors such as : motion parameters, mean framewise displacement, signal from white matter and anatomical component-based noise correction parameters have been regressed out. The total number of subjects in the dataset is 155 and for each subject a matrix of size 168×39 is summarizing the brain activation, where 168 stands for the number of time points and 39 for the number of ROIs when using the MSDL atlas, for example.

The data are then considered as final, pre-processed data used as input for the graph-based procedures. Three groups of subjects are investigated: 3yo (group 1: 17 subjects), 8-12yo (group 2: 34 subjects) and adults (group 3: 33 subjects). The expectation is that older children will have a functional connectivity topology that is closer to that corresponding to adults, hence be similar to a larger degree, and these two groups will exhibit brain pathways that are different from those of very young children.

Before running the analysis on the full dataset, we have first kept from each group 3 subjects as test data and performed training on the remaining subjects. Empirical training covariance matrices $\boldsymbol{S}_{\text{tr}}^{(r)} = \frac{1}{n_r}[\mathcal{Y}^{(r)}]^{\mathsf{T}}\mathcal{Y}^{(r)}$ with $r \in \{1,2,3\}$ have been constructed next based on data $\mathcal{Y}^{(r)}$ which represents the concatenated data across all training subjects in the $r-$th group. For example $\mathcal{Y}^{(1)}$ is a data matrix of dimension 2856×39. Similarly, test covariance matrices, denoted $\boldsymbol{S}_{\text{ts}}^{(r)}$ are created as well. We evaluate the out-of-sample performance of the WB-graphs by inspecting the loss

$$\mathcal{L}_{\infty}(\hat{\boldsymbol{\Delta}}_{\text{tr,q}}) = \left|\left|\frac{1}{2}(\boldsymbol{S}_{\text{ts}}^{(2)} + \boldsymbol{S}_{\text{ts}}^{(3)})\hat{\boldsymbol{\Delta}}_{\text{tr,q}}\boldsymbol{S}_{\text{ts}}^{(1)} + \frac{1}{2}(\boldsymbol{S}_{\text{ts}}^{(2)} + \boldsymbol{S}_{\text{ts}}^{(3)}) - \boldsymbol{S}_{\text{ts}}^{(1)}\right|\right|_{\infty}$$
$$\mathcal{L}_{\text{Fr}}(\hat{\boldsymbol{\Delta}}_{\text{tr,q}}) = \left|\left|\frac{1}{2}(\boldsymbol{S}_{\text{ts}}^{(2)} + \boldsymbol{S}_{\text{ts}}^{(3)})\hat{\boldsymbol{\Delta}}_{\text{tr,q}}\boldsymbol{S}_{\text{ts}}^{(1)} + \frac{1}{2}(\boldsymbol{S}_{\text{ts}}^{(2)} + \boldsymbol{S}_{\text{ts}}^{(3)}) - \boldsymbol{S}_{\text{ts}}^{(1)}\right|\right|_{\text{Fr}}$$

where $\hat{\boldsymbol{\Delta}}_{\text{tr,q}}$ represents the estimated matrix on the training data for competitor q with $q \in \{\text{WB-graphs}, \text{GL}, \text{JGL}\}$. To eliminate the dependence of the performance of the estimators on the splitting in training and test datasets, the splitting procedure is repeated randomly 200 times and Figure 3 presents the results when the BIC-like criterion is used for selecting λ_n. Very similar results are obtained when the AIC-like criterion is used and omitted from presentation. The figure illustrates that the WB-graph produces the best out-of-sample performance, followed by the JGL-based and GL-based estimators irrespective of which of the two atlases is being used.

Figure 4 shows the estimated edges between ROIs when the full dataset is being used (no splitting in training/test data) for the WB-graphs, GL and the JGL-based graphs. For the selection of the regularization parameters the AIC and BIC-like criteria have been used. The figure suggests that the proposed procedure identifies graphs that are much sparser (especially for the BIC, as expected) than the other two competitors and hence more interpretable. In fact the GL and JGL-based graphs regardless of criterion are full and hence do not lead to any major improvement in terms of interpretability.

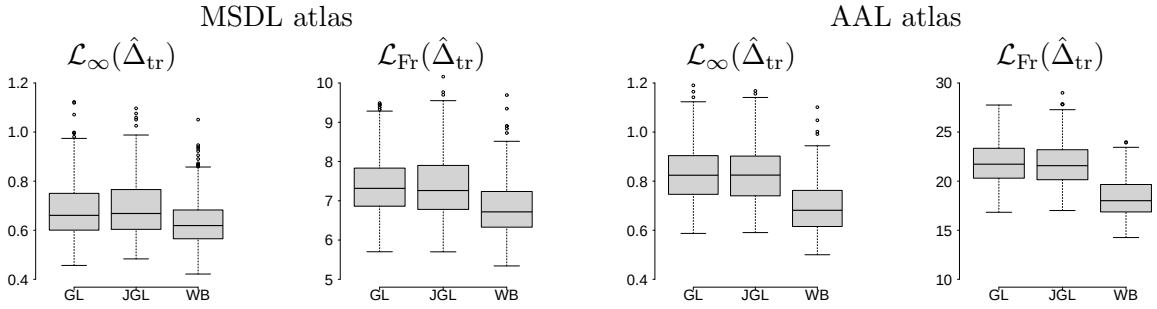

Figure 3: Boxplots of out-of-sample performance of WB-graphs relative to estimators $\hat{\Delta}_{\text{GL}}$ and $\hat{\Delta}_{\text{JGL}}$ when using the MSDL and AAL atlases and the BIC-like criterion for selecting λ_n.

To summarise how the obtained WB-graphs are informative in discovering connections that are similar between older children and adults but different from very young children, we present in Table 2 based on the MSDL atlas, the number of different edges between coarser brain parcelations (routinely also referred to as 'neuronal networks'). The table illustrates that most of the differences between these three groups of subjects involve ROIs of the Ventral attention (left and right) and Default mode network. Out of 17 unique such neuronal networks (see Table 3 in Appendix II for the full list) these three seem to act as hubs of differentiation. Quite interestingly, models selected with both criteria suggest that one key aspect in terms of differentiation between the three groups of subjects is even more localized, inside the Right ventral attention group of ROIs where edges connect these regions among themselves for some groups of subjects, but not the other.

		Aud.	Striate	DMN	Occ.	Motor	RVA	Basal	LVA	Freq.
AIC	Auditory	3		4	1	1	3	1	3	16
	Striate			1	1		1		1	4
	Default Mode Network			9	3	2	8	1	8	36
	Occipital					1	2		2	10
	Motor					1	3	1	2	11
	Right Ventral Attention						10		11	38
	Basal								2	5
	Left Ventral Attention								5	34
		Aud.	Striate	DMN	Occ.	Motor	RVA	Basal	LVA	Freq.
BIC	Auditory								1	1
	Default Mode Network			2	1	2	2		3	10
	Occipital					1	1		2	5
	Motor						1		1	5
	Right Ventral Attention						8		2	14
	Left Ventral Attention								2	11

Table 2: Real data. Number of edges estimated by WB-graphs between eight coarse neural networks (see Table 3 from Appendix II) while the remaining nine networks have been disregarded as they do not point to any difference between the three groups of subjects.

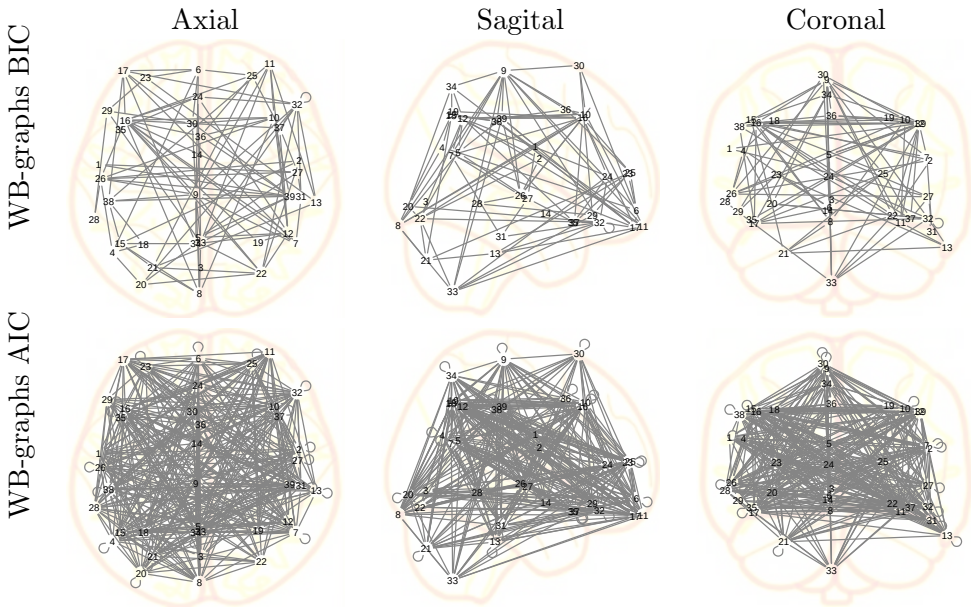

Figure 4: Estimated edges between ROIs defined based on the MSDL atlas, using WB-graphs, GL and JGL-based graphs with the AIC and BIC-like criteria used for selecting λ_n. Axial, sagital and coronal views are presented.

7 Discussion

We have introduced in this paper a method to estimate edges that differentiate between multiple groups of subjects. The method is based on the interplay between the within-group similarity and between-group dissimilarity. It is constructed so that general, off-the-shelf ADMM-based algorithms can be directly used to obtain estimates. Under fairly standard conditions we show that (i) the proposed estimator enjoys the usual $(\log p/n)^{1/2}$ rate, modulo a $\sqrt{L}$ factor accounting for having L groups in the data and that (ii) it is also sign consistent. Finite sample simulation results show a clear advantage in terms of accuracy of estimating the underlying target over standard competitors, especially so for larger sample sizes.

To illustrate the benefits of the procedure we analyse as well an fMRI dataset where we focus on three groups of subjects: adults and older children between the ages of 8 to 12 years (groups expected to share more similarity) on the one side and 3 year old subjects on the other side. In terms of brain pathways, a large number of differences between the groups are identified with our procedure at the level of the Ventral attention and DMN networks. Given that the ventral attention network is linked towards focusing attention on salient sensory events, and given that these data were collected while passively viewing a movie, this result might have been somewhat anticipated. A further study might focus deeper on determining the roots of DMN-based similarity/dissimilarity interplay as this result might not have been so easily anticipated.

Extensions of this procedure might try to quantify next the degree of uncertainty in the estimation in order to provide hypothesis testing tools, followed by a post-hoc contrasting procedure to identify from what groups the detected differences emerge.

Appendix I

Lemma 1. *Under condition (1), with probability greater than* $1-2p^{-\tau}$ *where* $\tau > 0$*, the following holds:*

$$\max_{a,b}\left|\frac{1}{N_1}\sum_{k=1}^{K} n_k \boldsymbol{S}_{ab}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\Sigma}_{ab}^{(k)}\right| \leq C_0\left(\frac{K\log p}{N_1}\right)^{1/2}$$

$$\max_{a,b}\left|\frac{1}{N_2}\sum_{l=K+1}^{L} n_l \boldsymbol{S}_{ab}^{(l)} - \frac{1}{L-K}\sum_{l=K+1}^{L}\boldsymbol{\Sigma}_{ab}^{(l)}\right| \leq C_0'\left(\frac{(L-K)\log p}{N_2}\right)^{1/2}$$

with C_0, C_0' *sufficiently large constants and* $N_1 = \sum_{k=1}^{K} n_k$, $N_2 = \sum_{l=K+1}^{L} n_l$.

Proof. We show here only the first relation as the second one is obtained analogously. The proofs are similar to those of Cai et al. (2011) and Lee and Liu (2015) and we start by considering the variables:

$$X_{m,ab}^{(k)} = Y_{m,a}^{(k)}Y_{m,b}^{(k)} - \mathrm{E}(Y_{m,a}^{(k)}Y_{m,b}^{(k)})\frac{N_1}{n_k \times K},$$

where $Y_{m,a}^{(k)}$ is the m-th component of the vector $(Y_{1,a}^{(k)}, Y_{2,a}^{(k)}, \ldots, Y_{n_k,a}^{(k)})^{\mathsf{T}}$ corresponding to feature a and of length n_k. In words, the vector collects the information from all sample points for feature a from the k-th group.

Note first that if the groups are balanced then $N_1 = n_k \times K$ and we retrieve the same quantity as considered in Cai et al. (2011), slightly adapted to our context. However, generally the groups need not be balanced, hence the term $\frac{N_1}{n_k \times K} = \frac{\bar{n}}{n_k}$ acts as a weight that reflects the deviation between the sample size of the k-th group and an 'average' sample size across all K groups.

Supposing that the data are centered in each group, our interest lies in the term:

$$\sum_{k=1}^{K}\frac{n_k}{N_1}\left(\frac{1}{n_k}\sum_{m=1}^{n_k} X_{m,ab}^{(k)}\right) = \frac{1}{N_1}\sum_{k=1}^{K}\sum_{m=1}^{n_k} X_{m,ab}^{(k)} = \frac{1}{N_1}\sum_{k=1}^{K} n_k \boldsymbol{S}_{ab}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\Sigma}_{ab}^{(k)}.$$

Considering $t = \eta(\log p)^{1/2}(N_1 K)^{-1/2}$ and $C_1 = 2+\tau+M^2(\eta K)^{-1}$ as well as certain algebraic inequalities from Cai et al. (2011), we have similarly to Lee and Liu (2015) that:

$$\begin{aligned} P\left(\frac{1}{N_1}\sum_{k=1}^{K}\sum_{m=1}^{n_k} X_{m,ab}^{(k)} > \eta^{-1}C_1\big((K\log p)/n\big)^{1/2}\right) &\leq \exp(-C_1\log p)\prod_{k=1}^{K}\prod_{m=1}^{n_k}\mathrm{E}(\exp(tX_{m,ab}^{(k)})) \\ &\leq \exp\left(\log p(-C_1 + M^2(\eta K)^{-1})\right) \\ &= p^{-(\tau+2)}. \end{aligned}$$

Furthermore, it also follows that:

$$P\left(\max_{a,b}\left|\frac{1}{N_1}\sum_{k=1}^{K}\sum_{m=1}^{n_k} X_{m,ab}^{(k)}\right| > \eta^{-1}C_1\big((K\log p)/n\big)^{1/2}\right) = 2p^{-\tau},$$

as stated in the Lemma. □

Corollary 1. *Under condition (1), with probability greater than* $1-4p^{-\tau}$ *where* $\tau > 0$*, the following holds:*

$$\max_{a,b}\left|\frac{1}{N_1}\sum_{k=1}^{K} n_k \boldsymbol{S}_{ab}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\Sigma}_{ab}^{(k)}\right| \leq C_0''\left(\frac{L\log p}{n_\dagger}\right)^{1/2} \quad \textit{and}$$

$$\max_{a,b}\left|\frac{1}{N_2}\sum_{l=K+1}^{L} n_l \boldsymbol{S}_{ab}^{(l)} - \frac{1}{L-K}\sum_{l=K+1}^{L}\boldsymbol{\Sigma}_{ab}^{(l)}\right| \leq C_0''\left(\frac{L\log p}{n_\dagger}\right)^{1/2}$$

with C_0'' *a sufficiently large constant and* $n_\dagger = \min(n_1, n_2, \ldots, n_L)$.

Proof. From Lemma 1 we have that

$$\max_{a,b}\left|\frac{1}{N_1}\sum_{k=1}^{K} n_k \boldsymbol{S}_{ab}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\Sigma}_{ab}^{(k)}\right| \leq C_0\left(\frac{K\log p}{N_1}\right)^{1/2} \leq C_0''\left(\frac{L\log p}{n_\dagger}\right)^{1/2}$$

for any $C_0'' \geq C_0$, since $L > K$ and $N_1 \geq n_\dagger$. Similarly,

$$\max_{a,b}\left|\frac{1}{N_2}\sum_{l=K+1}^{L} n_l \boldsymbol{S}_{ab}^{(l)} - \frac{1}{L-K}\sum_{l=K+1}^{L}\boldsymbol{\Sigma}_{ab}^{(l)}\right| \leq C_0''\left(\frac{L\log p}{n_\dagger}\right)^{1/2}.$$

Hence

$$P\Bigg(\max_{a,b}\left|\frac{1}{N_1}\sum_{k=1}^{K} n_k \boldsymbol{S}_{ab}^{(k)} - \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\Sigma}_{ab}^{(k)}\right| \leq C_0''\left(\frac{L\log p}{n_\dagger}\right)^{1/2} \cap$$
$$\max_{a,b}\left|\frac{1}{N_2}\sum_{l=K+1}^{L} n_l \boldsymbol{S}_{ab}^{(l)} - \frac{1}{L-K}\sum_{l=K+1}^{L}\boldsymbol{\Sigma}_{ab}^{(l)}\right| \leq C_0''\left(\frac{L\log p}{n_\dagger}\right)^{1/2}\Bigg) \leq 4p^{-\tau},$$

and the proof is completed. □

Lemma 2. *The matrix* $\mathcal{S}^\mathsf{T}\big[\frac{1}{L-K}\sum_{l=K+1}^{L}\boldsymbol{\Sigma}^{(l)} \otimes \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\Sigma}^{(k)}\big]\mathcal{S}$ *is invertible where* $\boldsymbol{\Sigma}^{(r)}$ *with* $r = 1, \ldots, L$ *are the* $p \times p$ *population covariance matrices for each group and* $\mathcal{S}$ *is the duplication matrix used in Section 3.*

Proof. Using the distributed property of the Kronecker product we have that:

$$\mathcal{S}^\mathsf{T}\left[\frac{1}{L-K}\sum_{l=K+1}^{L}\boldsymbol{\Sigma}^{(l)} \otimes \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\Sigma}^{(k)}\right]\mathcal{S} = \mathcal{S}^\mathsf{T}\left[\frac{1}{K(L-K)}\sum_{l,k}\boldsymbol{\Sigma}^{(l)} \otimes \boldsymbol{\Sigma}^{(k)}\right]\mathcal{S}.$$

As $\boldsymbol{\Sigma}^{(k)}$ and $\boldsymbol{\Sigma}^{(l)}$ for any two groups k, l are positive definite, so to is their Kronecker product and the matrix $\sum_{l,k}\boldsymbol{\Sigma}^{(l)} \otimes \boldsymbol{\Sigma}^{(k)}$. Hence there exists a full rank matrix $\boldsymbol{\Gamma}^{1/2}$ with $\operatorname{rank}(\boldsymbol{\Gamma}^{1/2}) = p(p+1)/2$ such that

$$[\boldsymbol{\Gamma}^{1/2}]^\mathsf{T}\boldsymbol{\Gamma}^{1/2} = \sum_{l,k}\boldsymbol{\Sigma}^{(l)} \otimes \boldsymbol{\Sigma}^{(k)}.$$

Since by construction $\mathcal{S}$ is also full rank with $\operatorname{rank}(\mathcal{S}) = p(p+1)/2$ (see Zhao et al., 2014), we also have that

$$\begin{aligned}
\operatorname{rank}\left(\mathcal{S}^\mathsf{T}\left[\sum_{l,k}\boldsymbol{\Sigma}^{(l)} \otimes \boldsymbol{\Sigma}^{(k)}\right]\mathcal{S}\right) &= \operatorname{rank}\big(\mathcal{S}^\mathsf{T}\big[\sum_{l,k}\boldsymbol{\Sigma}^{(l)} \otimes \boldsymbol{\Sigma}^{(k)}\big]\mathcal{S}\big) \\
&= \operatorname{rank}(\mathcal{S}^\mathsf{T}\big[(\boldsymbol{\Gamma}^{1/2})^\mathsf{T}\boldsymbol{\Gamma}^{1/2}\big]\mathcal{S}) \\
&= \operatorname{rank}(\boldsymbol{\Gamma}^{1/2}\mathcal{S}) \\
&= \operatorname{rank}(\mathcal{S}),
\end{aligned}$$

hence invertible. □

Proof. Proof of Proposition 1

The proof follows directly the same approach as in Zhao et al. (2014), adapted to our context. For ease of notation, let

$$\boldsymbol{\Sigma} = \left(\frac{1}{L-K}\sum_{l}\boldsymbol{\Sigma}^{(l)}\right)\otimes\left(\frac{1}{K}\sum_{k}\boldsymbol{\Sigma}^{(k)}\right) = \frac{1}{K(L-K)}\sum_{l,k}\boldsymbol{\Sigma}^{(l)}\otimes\boldsymbol{\Sigma}^{(k)}$$

$$\hat{\boldsymbol{\Sigma}} = \left(\frac{1}{N_2}\sum_{l}n_l\boldsymbol{S}^{(l)}\right)\otimes\left(\frac{1}{N_1}\sum_{k}n_k\boldsymbol{S}^{(k)}\right) = \frac{1}{N_1N_2}\sum_{l,k}n_ln_k(\boldsymbol{S}^{(l)}\otimes\boldsymbol{S}^{(k)})$$

$$\boldsymbol{b} = \text{vec}\left(\frac{1}{K}\sum_{k}\boldsymbol{\Sigma}^{(k)} - \frac{1}{L-K}\sum_{l}\boldsymbol{\Sigma}^{(l)}\right)$$

$$\hat{\boldsymbol{b}} = \text{vec}\left(\frac{1}{N_1}\sum_{k}n_k\boldsymbol{S}^{(k)} - \frac{1}{N_2}\sum_{l}n_l\boldsymbol{S}^{(l)}\right).$$

We show first that $\boldsymbol{\beta}_0$ is in the feasible set with a high probability. As such, with R as in Condition (2) and based on Corollary 1 we can obtain

$$||\mathcal{S}^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}\mathcal{S}\boldsymbol{\beta}_0 - \mathcal{S}^{\mathsf{T}}\hat{\boldsymbol{b}}||_\infty \leq 4R||\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}}||_\infty + C\left(\frac{L\log p}{n_\dagger}\right)^{1/2}.$$

Moreover as in Lemma A3 of Zhao et al. (2014) one can also write each entry of $\boldsymbol{\Sigma}$ as $\frac{1}{K(L-K)}\sum_{l,k}\boldsymbol{\Sigma}^{(k)}_{a'a}\boldsymbol{\Sigma}^{(l)}_{b'b}$ and each entry of $\hat{\boldsymbol{\Sigma}}$ as $\frac{1}{N_1N_2}\sum_{l,k}n_ln_k\boldsymbol{S}^{(k)}_{a'a}\boldsymbol{S}^{(l)}_{b'b}$ implying that

$$\begin{aligned}\left|\left(\frac{1}{N_1}\sum_{k}n_k\boldsymbol{S}^{(k)}_{a'a}\right)\left(\frac{1}{N_2}\sum_{l}\boldsymbol{S}^{(l)}_{b'b}\right) - \left(\frac{1}{K}\sum_{k}\boldsymbol{\Sigma}^{(k)}_{a'a}\right)\left(\frac{1}{L-K}\sum_{l}\boldsymbol{\Sigma}^{(l)}_{b'b}\right)\right| = \\ \left(\frac{1}{K}\sum_{k}\boldsymbol{\Sigma}^{(k)}_{a'a}\right)\left[\left(\frac{1}{N_2}\sum_{l}\boldsymbol{S}^{(l)}_{b'b}\right) - \left(\frac{1}{L-K}\sum_{l}\boldsymbol{\Sigma}^{(l)}_{b'b}\right)\right] + \\ \left(\frac{1}{L-K}\sum_{l}\boldsymbol{\Sigma}^{(l)}_{b'b}\right)\left[\left(\frac{1}{N_1}\sum_{k}n_k\boldsymbol{S}^{(k)}_{a'a}\right) - \left(\frac{1}{K}\sum_{k}\boldsymbol{\Sigma}^{(k)}_{a'a}\right)\right] + \\ \left(\frac{1}{N_1}\sum_{k}n_k\boldsymbol{S}^{(k)}_{a'a} - \frac{1}{K}\sum_{k}\boldsymbol{\Sigma}^{(k)}_{a'a}\right)\left(\frac{1}{N_2}\sum_{l}\boldsymbol{S}^{(l)}_{b'b} - \frac{1}{L-K}\sum_{l}\boldsymbol{\Sigma}^{(l)}_{b'b}\right).\end{aligned}$$

Now, if L grows with $n_\dagger$ such that $L = o(n_\dagger/\log p)$ then in view of Condition (4) we have that $\frac{1}{K}\sum_k\boldsymbol{\Sigma}^{(k)}_{a'a} = O(1)$, $\frac{1}{L-K}\sum_l\boldsymbol{\Sigma}^{(l)}_{b'b} = O(1)$ and hence for any a', a, b', b

$$\left|\left(\frac{1}{N_1}\sum_{k}n_k\boldsymbol{S}^{(k)}_{a'a}\right)\left(\frac{1}{N_2}\sum_{l}\boldsymbol{S}^{(l)}_{b'b}\right) - \left(\frac{1}{K}\sum_{k}\boldsymbol{\Sigma}^{(k)}_{a'a}\right)\left(\frac{1}{L-K}\sum_{l}\boldsymbol{\Sigma}^{(l)}_{b'b}\right)\right| = O_p\left(\frac{L\log p}{n_\dagger}\right)^{1/2}.$$

As such, if $\lambda_n \geq C\frac{L\log p}{n_\dagger}$ with C a sufficiently large constant that does not depend on p, then $\boldsymbol{\beta}_0$ is feasible with probability larger than $1-4p^{-\tau}$.

Using the results of Zhao et al. (2014) we have that

$$\begin{aligned}||\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta}_0||_\infty &\leq \kappa_{s,\mu,\sigma^{\mathcal{S}}_{\min}}||\mathcal{S}^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}\mathcal{S}(\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta}_0)||_\infty \\ ||\mathcal{S}^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}\mathcal{S}(\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta}_0)||_\infty &\leq ||\mathcal{S}^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}\mathcal{S}\hat{\boldsymbol{\beta}} - \mathcal{S}^{\mathsf{T}}\hat{\boldsymbol{b}}||_\infty + ||\mathcal{S}^{\mathsf{T}}(\boldsymbol{\Sigma}-\hat{\boldsymbol{\Sigma}})\mathcal{S}\hat{\boldsymbol{\beta}}||_\infty + ||\mathcal{S}^{\mathsf{T}}(\hat{\boldsymbol{b}}-\boldsymbol{b})||_\infty \\ &\leq \lambda_n + O_p\left(\frac{L\log p}{n_\dagger}\right)^{1/2} + O_p\left(\frac{L\log p}{n_\dagger}\right)^{1/2} \\ &= O_p\left(\frac{L\log p}{n_\dagger}\right)\end{aligned}$$

Putting it all together, the claimed result follows. □

Appendix II

x	y	z	ROI Label	ROI name	Brain network
-53.28	-8.88	32.36	1	L Aud	Aud
53.47	-6.49	27.52	2	R Aud	Aud
1.18	-74.54	10.79	3	Striate	Striate
-45.8	-64.78	31.84	4	L DMN	DMN
-0.2	-55.21	29.87	5	Med DMN	DMN
-0.15	51.42	7.58	6	Front DMN	DMN
51.66	-59.34	28.88	7	R DMN	DMN
0.41	-91.05	1.58	8	Occ post	Occ post
-1.48	-27.93	61.5	9	Motor	Motor
40.1	20.96	44.72	10	R DLPFC	R V Att
37.83	55.49	1.22	11	R Front	R V Att
47.53	-52.42	43.06	12	R Par	R V Att
62.53	-32.99	-9.14	13	R Post Temp	R V Att
-0.91	-2.75	6.15	14	Basal	Basal
-41.66	-59.04	44.61	15	L Par	L V Att
-39.04	19.28	43.27	16	L DLPFC	L V Att
-40.08	50.65	0.81	17	L Front pol	L V Att
-29.39	-59.43	44.2	18	L IPS	D Att
31.6	-58.09	45.69	19	R IPS	D Att
-30.54	-85.14	9.1	20	L LOC	Vis Sec
-24.29	-74.28	-11.74	21	Vis	Vis Sec
33.4	-77.96	4.31	22	R LOC	Vis Sec
-28.17	46.32	21.56	23	D ACC	Salience
-0.45	34.06	20.73	24	V ACC	Salience
28.38	47.72	22.13	25	R A Ins	Salience
-52.12	-17.92	13.28	26	L STS	Temporal
52.61	-13.65	12.11	27	R STS	Temporal
-55.52	-43.77	10.08	28	L TPJ	Language
-48.66	25.11	5.7	29	Broca	Language
-3.39	17.19	63.52	30	Sup Front S	Language
54.42	-29.5	-2.72	31	R TPJ	Language
52.38	29.39	2.93	32	R Pars Op	Language
1.05	-58.49	-23.91	33	Cereb	Cereb
-1.44	-59.12	55.25	34	Dors PCC	Dors PCC
-41.33	13.63	2.7	35	L Ins	Cing-Ins
1.05	9.2	46.43	36	Cing	Cing-Ins
43.01	14.3	2.79	37	R Ins	Cing-Ins
-47.85	-32.07	41.9	38	L Ant IPS	Ant IPS
48.36	-29.04	43.13	39	R Ant IPS	Ant IPS

Table 3: The x, y, z coordinates, ROI labels (corresponding to Figure 4), ROI names and coarser 'brain network parcelation'. From Varoquaux et al. (2011).

References

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

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(2):373–397.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. Springer.

Lee, W. and Liu, Y. (2015). Joint estimation of multiple precision matrices with common structures. *Journal of Machine Learning Research*, 16(31):1035–1062.

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

Pircalabelu, E. and Claeskens, G. (2020). Community-based group graphical lasso. *Journal of Machine Learning Research*, 21(64):1–32.

Ravikumar, P. D., Raskutti, G., Wainwright, M. J., and Yu, B. (2008). Model selection in Gaussian graphical models: High-dimensional consistency of l_1-regularized MLE. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1329–1336.

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., and Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications*, 9(1027).

Shojaie, A. (2020). Differential network analysis: A statistical perspective. *arXiv:*, 2003.04235.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15(1):273–289.

Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., and Thirion, B. (2011). Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *"Information Processing in Medical Imaging*, pages 562–573.

Witten, D. M., Friedman, J. H., and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900.

Zhao, S. D., Cai, T. T., and Li, H. (2014). Direct estimation of differential networks. *Biometrika*, 101(2):253–268.