Brain Connectivity Networks in Theory and Practice

Yunqi Bu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Johannes Lederer, Chair

Kwun Chan

Tyler McCormick

Program Authorized to Offer Degree:
Public Health: Biostatistics

University of Washington

Abstract

Brain Connectivity Networks
in Theory and Practice

Yunqi Bu

Chair of the Supervisory Committee:
Professor Johannes Lederer
Statistics

The main purpose of this thesis is to develop statistical methods to explore the human brain connectivity and its relationship to cognitive diseases, such as Alzheimer's disease. Among a variety of different imaging methods, resting-state functional magnetic resonance imaging (rs-fMRI) data are popularly used for estimating brain connectivity networks in the absence of tasks. In neuroimaging, Gaussian graphical models are mainstream tools for modeling statistical dependencies as functional connectivity across anatomically distinct regions of the human brain. Brain connectivity here is an undirected graph estimated from high-dimensional fMRI data.

We show, however, that standard Gaussian graphical modeling methods such as neighborhood selection and graphical lasso can fail to provide accurate and reproducible graph recovery when estimating brain connectivity networks. This problem persists even under the best circumstances of optimal tuning and sufficiently large sample sizes, which is often not achieved in real applications with these methods. In Chapter 1, we attempt to solve this problem by leveraging the three-dimensional spatial positions of the nodes into a neighborhood selection framework to gain more accurate graph estimations. These positions are incorporated into the tuning parameters of each nodes' penalized regression in the form of pairwise distances between brain regions. This approach (named SI) is motivated by the

biological rationale that direct brain connections are more likely between close regions than between distant regions.

Clinically, fMRI data is often obtained longitudinally for each subject. However, discussion for estimating networks in a longitudinal clinical setting are scarce. Human brain connectivity has been shown to be reproducible across individuals. Recent advances in data acquisition and preprocessing have also largely improved the reliability of functional magnetic resonance imaging for estimating functional brain networks. In Chapter 3, driven by these developments, we exploit the presence of shared connectivity structure in order to produce more accurate brain connectivity estimates for individual patients in clinical settings. More specifically, we propose an approach that can incorporate information from baseline fMRI assessment when estimating networks in follow-up fMRI data. For this new approach (named Geofuse), we manage to jointly estimate two graphs under one neighborhood selection model. For each regression, we add an additional fused lasso penalty on the basis of SI from Chapter 1 to encourage the two groups of parameters for the two graphs to shrink together, yielding more stable networks across repeated scans of the same individuals.

Both approaches SI and Geofuse are computationally convenient and efficient. Using data from an Alzheimer's disease dataset and the Consortium for Reliability and Reproducibility, we illustrate that SI (for single graph estimation) and Geofuse (for joint graph estimation) both produce more stable brain connectivity networks than state of the art methods. These two approaches may, therefore, be of particular value to the clinical neurosciences.

In Chapter 2, we study a topic outside of neuroscience, namely personalized medicine. Sometimes, the development of a market-ready companion diagnostic test (CDx) for identifying the best treatment for individual patients may lag behind the development of the actual drug, and we use a clinical trial assay (CTA) to enroll patients into the drug pivotal clinical trial instead. Thus, when CDx becomes available, a bridging study is required to assess the drug efficacy in the CDx intended use population (CDx IU). Due to randomization

related missingness of the CDx results, one challenge we face is covariate imbalance between treatment arms for the subpopulation with both positive CDx and CTA. In Chapter 2, we address this challenge in bridging studies under a causal inference framework and evaluate the performance of two methods: 1) the propensity score method with doubly robust estimation. 2) the optimal matching method.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to Prof. Johannes Lederer and Prof. Andrew Zhou for years of mentorship and support. We thank Dantao Peng, Yanlei Mu, and Xiao Zhang for providing the AD fMRI data. We would like to express our appreciation to Min Zhang and Tobias Kaufmann for their neuroimaging expertise and for preprocessing the fMRI data. We thank Dr. Meijuan Li and Dr. Gene Pennello for their valuable support and advice. Thanks to Rosemary Adams for contributions to the visualizations, and Benjamin Phillips for contributions to the `R` code. We also thank Noah Simon, Gary Chan, Tyler McCormick, Marcia Ciol, Sam Koelle, Mengjie Pan, Yuxi Wu, Néhémy Lim, and Rui Zhuang for helpful suggestions. Thanks to Anqi Cheng for her friendship and companionship. Thank you to Jiaqi Guo, for all his love and support. Finally, I wish to thank my parents for their love and encouragement, without whom I would never have enjoyed so many opportunities.

Chapter 1

INTEGRATING ADDITIONAL KNOWLEDGE INTO THE ESTIMATION OF GRAPHICAL MODELS FOR BRAIN CONNECTIVITY NETWORKS

This chapter is an adapted version of my arXiv paper [6].

Abstract

Brain connectivity networks derived from fMRI data are considered a prime gateway to understanding cognitive diseases of the human brain. We show, however, that standard methods for Gaussian graphical modeling can fail to provide accurate graph recovery for estimating brain connectivity. This problem persists even when optimal tuning is applied and large sample sizes are available. We attempt to solve this problem by leveraging information that is readily available but typically neglected: the spatial positions of the measurements. These positions are incorporated into the tuning parameter of neighborhood selection in the form of pairwise distances between brain regions. This approach is computationally convenient and efficient, carries a lucid Bayesian interpretation in concordance with biological rationale, and improves standard methods in terms of statistical stability. Applied to data about Alzheimer's disease, our approach allows us to highlight the central role of lobes in the connectivity structure of the brain and an increased connectivity within the cerebellum for Alzheimer's patients compared to other subjects.

1.1 Introduction

Brain connectivity networks derived from fMRI data are considered a prime gateway to understanding cognitive diseases. Along with other fields, brain research has boosted the

interest in statistical methodology for uncovering dependence networks. A standard framework for dependence networks is Gaussian graphical models [40]. Gaussian graphical models have become particularly popular after the development of methods and algorithms that can handle large and high-dimensional data.

Two widely-used approaches to Gaussian graphical models are neighborhood selection and graphical lasso. Neighborhood selection aims at graph reconstruction by aggregating local estimates [49]. Graphical lasso, on the other hand, is based on a global objective function [2, 18, 70]. Both approaches are now accompanied by a bulk of literature on theory and computation; we refer to [8, 31] and references therein. However, we show that even with optimal tuning and large sample sizes, standard methods for Gaussian graphical modeling can fail to provide accurate graph recovery when estimating brain connectivity networks.

Accordingly, the purpose of this paper is to estimate brain connectivity networks more accurately. In detail, we are interested in estimating brain connectivity networks by analyzing resting-state functional magnetic resonance imaging (fMRI) data that describe the levels of co-activation between brain regions [63] as measured by changes in blood flow [36]. Brain regions have spatial coordinates, so in addition to the fMRIs, there is information in terms of pairwise distances between brain regions. Our key idea is to leverage this additional knowledge for more effective graph estimation. Our main strategy for this is to strengthen the role of tuning parameters. Commonly, tuning parameters are considered an inconvenience, because they need to be calibrated for each data set specifically. We instead think of this adaptability as an asset that can make tuning parameters a potent instrument for funneling external information into the estimation process. More specifically, for our goal of brain network estimation, we use tuning parameters to make neighborhood selection receptive to additional knowledge. We first show numerically that without such an integration, the sample size (or the number of fMRI scans per person) needed for accurate graph recovery of the brain network is exceedingly large. We then show that adopting our notion can lead to four general improvements:

Accuracy: We show by simulations that our approach can render graph estimation more

accurate.

Stability: We adopt notions of stability and show the reliability of our approach with the fMRI data.

Accessibility: We point out that our approach is amenable to straightforward and efficient implementations based on existing software. In particular, our approach preserves the general forms of the standard penalties and does not introduce any additional penalty terms.

Interpretability: We demonstrate that our approach has a lucid, biological Bayesian interpretation.

On actual brain fMRI data collected from patients with and without Alzheimer's disease, the estimated brain connectivity networks are stable and in agreement with biological reasoning. Also compared to competing methods, our approach manifests the central role of the lobes in the connectivity structure more clearly. We find that connections are significantly more likely within the lobes than between the lobes, despite that we provide our method only with the pairwise distances and not with the lobe affiliations of the regions.

The remainder of the paper is structured as follows. In Section 1.2, we show how standard methods can fail to provide accurate graph recovery of brain connectivity networks and explain the biological motivations of incorporating additional spatial knowledge into the estimation process. Then in Section 3.2, we introduce our general concept for this incorporation. Next, in Section 1.4, we confirm that our approach can harvest additional knowledge to improve graph estimation for brain connectivity networks with simulations and a sensitivity analysis under three different scenarios of model misspecifications. And in Section 1.5, we study the brain networks estimated on real fMRI data and their stability for our method as well as for competing methods. In Section 1.6, we conclude with a discussion.

Data, details on the preprocessing, and implementations of our methods can be found on `https://github.com/LedererLab/fMRI`.

1.1.1 A brief review of Gaussian graphical models

We give a brief review of Gaussian graphical models, our main tool for estimating brain connectivity graphs. Gaussian graphical models assume samples from a centered p-dimensional normal distribution $\mathcal{N}_p(\mathbf{0}, \Sigma)$ with a symmetric, positive definite covariance matrix Σ. The distribution is then complemented with an associated undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that has node set $\mathcal{V} = \{1, \ldots, p\}$ and edge set $\mathcal{E} = \{(i, j) \in \mathcal{V} \times \mathcal{V} \,|\, i \neq j \text{ and } (\Sigma^{-1})_{ij} \neq 0\}$. The corresponding adjacency matrix is defined by $A \in \mathbb{R}^{p \times p}$, $A_{ij} = 1$ if $(i, j) \in \mathcal{E}$ and $A_{ij} = 0$ otherwise.

The crux of Gaussian graphical models is that the conditional and marginal dependence structures of the samples are concisely captured by the edge set $\mathcal{E}$. Indeed, the Hammersley-Clifford theorem [5, 25] states that the ith and jth coordinate of a sample from $\mathcal{N}_p(\mathbf{0}, \Sigma)$ are conditionally independent given all other coordinates if and only if $(\Sigma^{-1})_{ij} = 0$, that is, $(i, j) \notin \mathcal{E}$. Moreover, the ith and jth coordinate are independent if and only if one cannot construct a chain of the form $(i, k_1), (k_1, k_2), \ldots, (k_l, j)$ by using only elements in $\mathcal{E}$. Our goal is consequently to uncover the edge set $\mathcal{E}$ from data. For this aim, we develop estimators $\hat{\mathcal{E}} \equiv \hat{\mathcal{E}}(X)$ of $\mathcal{E}$ based on independent identically distributed Gaussian samples $X_1, \ldots, X_n \in \mathbb{R}^p$ summarized in the data matrix $X = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times p}$.

1.2 Motivation

Functional Magnetic Resonance Imaging (fMRI) is a promising gateway to understanding the human brain [63]. Our specific goal is to use brain activity records from resting-state fMRI to infer co-activation networks among brain regions. For this, we rely on data collected from outpatients at the Department of Neurology at Beijing Hospital. The imaging and preprocessing details can be found in Appendix C.

The preprocessed data set comprises 37 subjects in total: 22 patients with Alzheimer's disease (AD), 5 patients with mild cognitive impairment (MCI), and 10 patients with normal cognition (NC). Each subject's data contains $n = 210$ consecutive scans. Each of the $p =$

116 variables is an average intensity over all voxels in an anatomical volume of interest defined by Automated Anatomical Labeling [61]. Autocorrelation was accounted for with an autoregressive integrated moving average model [24, 32].

The brain regions have spatial coordinates, so aside from the intensity fMRI data, we have additional knowledge in terms of pairwise distances between the 116 regions. How and why should we incorporate this additional knowledge into estimating the brain connectivity network? Our biological motivation is that *direct* connections are more likely between close brain regions than between distant brain regions. Distant brain regions are more likely to be connected *indirectly*, that is, through other regions. Our concept in Section 3.2 based on graphical models is ideal for incorporating this understanding.

1.2.1 The need for additional knowledge

Here, we show that even with optimal tuning and large sample sizes, standard methods for Gaussian graphical modeling can fail to provide accurate graph recovery. For this, we conduct a simulation study comparing four of the most popular methods: thresholding the partial correlation matrix (THR); neighborhood selection with the "or-rule" (MB(or)); neighborhood selection with the "and-rule" (MB(and)); and graphical lasso (GLASSO).

We simulate data with the aim of mimicking settings encountered in the brain fMRI data. For this, the number of nodes is set to $p = 116$, which equals the number of brain regions. Then, a standard preferential attachment algorithm [3] is used to construct 115 edges between these nodes. The resulting edge set $\mathcal{E}$ determines which off-diagonal entries of the inverse covariance matrix Σ^{-1} are non-zero. The values of these entries are then independently sampled uniformly at random from $[-1, -0.2] \cup [0.2, 1]$. The diagonal entries of Σ^{-1} are set to a common value such that the condition number equals 100. With Σ^{-1} constructed this way, vectors are independently sampled from the Gaussian distribution $\mathcal{N}_p(\mathbf{0}, \Sigma)$ and summarized in data sets with sample sizes $n \in \{50, 100, 200, 400, 600, 800, 1000\}$.

The accuracy of graph recovery is assessed via Hamming distance. Hamming distance between the estimated edge set $\hat{\mathcal{E}}$ and the true edge set $\mathcal{E}$ is defined by $d_{\mathrm{H}}(\hat{\mathcal{E}}, \mathcal{E}) := \big|\{(i, j) \mid$

$(i,j) \in \hat{\mathcal{E}}, (i,j) \notin \mathcal{E}\} \cup \{(i,j) \mid (i,j) \notin \hat{\mathcal{E}}, (i,j) \in \mathcal{E}\}\big|$. In terms of the corresponding adjaceny matrices $\hat{A}$ and A, this equals $d_{\mathrm{H}}(\hat{\mathcal{E}}, \mathcal{E}) = |\!|\!|\hat{A} - A|\!|\!|_1$, where $|\!|\!| \cdot |\!|\!|_1$ is the entrywise ℓ_1-norm; larger Hamming distances indicate less accurate estimation. The tuning parameters of all methods are calibrated to minimize Hamming distance, noting that the true graphs are known in simulations. This "oracle" tuning allows us to study the maximal potential of the standard methods' accuracies.

Hamming Distance

Method	**Sample Size** n						
	50	**100**	**200**	**400**	**600**	**800**	**1000**
THR	115	115	112	64	46	36	29
MB(or)	108	107	99	98	96	94	90
MB(and)	104	81	76	48	40	31	27
GLASSO	103	84	77	58	44	34	32

Table 1.1: Graph estimation with optimally-calibrated standard methods becomes more accurate as the sample size n increases. However, even with $n = 1000$, which is much larger than the number of nodes $p = 116$, the minimum Hamming distance is still 27. Given that there are only 115 true edges in total, 27 wrongly assigned edges mark a poor performance. Thus, graph estimation with standard methods can still be highly inaccurate even when the sample sizes are very large.

Table 1.1 contains the Hamming distances averaged over 50 repetitions. For $n = 1000$, MB(and) is the most accurate approach among all four methods, having an average Hamming distance of 27. Observe, however, that 27 wrongly assigned edges still mark a poor performance given that there are only 115 true edges in total. For smaller sample sizes, the accuracies of the methods decline even further. In view of real data being commonly high-dimensional, where p is of the same order as n – or even larger, these observations thus provide substantial motivation for the inclusion of additional knowledge.

1.3 Method

Now, we are ready to introduce and discuss how to amend the estimation process with additional knowledge. We first explain our statistical scheme to incorporate additional knowledge into graph recovery. The resulting method can be computed efficiently with standard software packages. Then, we establish a theory for exact graph recovery for our method. Last, we provide a direct Bayesian interpretation of the method to bridge our biological motivation with the additional spatial knowledge of the brain.

1.3.1 Neighborhood selection with additional knowledge

A basis particularly suited for our approach is neighborhood selection. The main idea of neighborhood selection is that graph recovery can be established through a sequence of regressions. The corresponding regression parameter $\beta^j \in \mathbb{R}^{p-1}$ for a given node j determines the edges between node j and the other nodes: $(\beta^j_+)_i \neq 0 \;\Leftrightarrow (i,j) \in \mathcal{E}$, where $\beta^j_+ := (\beta^j_1, \cdots, \beta^j_{j-1}, 0, \beta^j_j, \cdots, \beta^j_{p-1}) \in \mathbb{R}^p$ is β^j with a zero added as the jth entry. With this in mind, the standard regression estimators are of the form

$$\hat{\beta}^{j,r^j} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^{p-1}} \left\{ \frac{1}{2} \|X^j - X^{-j}\beta\|_2^2 + r^j \|\beta\| \right\}, \quad (\, j \in \{1, \ldots, p\}\,). \tag{1.1}$$

Here, the vector $X^j \in \mathbb{R}^n$ denotes the jth column of the data matrix X, the matrix $X^{-j} \in \mathbb{R}^{n \times (p-1)}$ denotes X without the jth column, the positive numbers $r^1, \ldots, r^p \in (0, \infty)$ are tuning parameters, and $\|\cdot\|$ is a norm (or more generally, a convex, positive function, such as the Ridge penalty). A standard example is neighborhood selection with the lasso, where $\|\cdot\|$ is set to the ℓ_1-norm. Adopting the "and-rule," the edge set $\mathcal{E}$ is finally estimated by $\hat{\mathcal{E}} = \{(i,j) \mid i \neq j,\ (\hat{\beta}^{i,r^i}_+)_j \neq 0 \text{ and } (\hat{\beta}^{j,r^j}_+)_i \neq 0\}$.

The choice of MB(and) as the basis for our approach is motivated by the simulation results in Section 1.2.1 and the fact that there is a myriad of existing software packages for solving problems of the form (1.1). In principle, however, one could apply the following recipe also to MB(or) and GLASSO.

Now our proposal is to incorporate additional knowledge by "upgrading" the univariate tuning parameters $r^1, \dots, r^p$ to vectors. For this, we assume additional knowledge in the form of a matrix $D \in \mathbb{R}^{p \times p}$. In our case, D_{ij} is the pairwise distances between brain regions i and j. In other applications, D_{ij} could be the Euclidean distance between nodes i and j where the nodes correspond to brain regions, countries, galaxies, etc. D_{ij} could also be the (conditional) correlation between nodes i and j estimated on the present data set with an initial estimator (specializing our method to adaptive lasso-type approaches [73], for example) or more interestingly, with the same or different estimators on other data sources. Next, to incorporate D into neighborhood selection, we transform the one-dimensional tuning parameters $r^j \in (0, \infty)$ into multi-dimensional tuning parameters $\mathbf{r}^j \in (0, \infty)^p$. We then enrich these tuning parameters with additional knowledge by setting $(\mathbf{r}^j)_i = \bar{r}^j \cdot f(D_{ij})$, where $f : \mathbb{R} \to (0, \infty)$ is a positive link function. In our brain connectivity application, the link function is defined as $f(x) = x^3$ to capture the three-dimensionality of the coordinates. This yields tuning parameters of the form $(\mathbf{r}^j)_i = \bar{r}^j \cdot D_{ij}^3$. Here $\bar{r}^j$ is an overall tuning parameter for the jth node's regression. As customary in high-dimensional statistics, the free parameter $\bar{r}^j$ balances the weights of the data (X^{-j} in our case) and the structural assumptions (captured by $\| \cdot \|$ and D in our case). In practice, $\bar{r}^j$ can be calibrated via 10-fold cross-validation. The above node-wise regression (1.1) is then generalized to

$$\hat{\beta}^{j,\mathbf{r}^j} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^{p-1}} \left\{ \frac{1}{2} \|X^j - X^{-j}\beta\|_2^2 + \|\mathbf{r}^j_{-j} \circ \beta\| \right\}, \quad (j \in \{1, \dots, p\}), \tag{1.2}$$

where the circle $\circ$ indicates element-wise multiplication, $\mathbf{r}^j_{-j} \in \mathbb{R}^{p-1}$ is $\mathbf{r}^j \in \mathbb{R}^p$ omitting the jth element, and the estimator $\hat{\mathcal{E}}$ is generalized accordingly.

Let us discuss three practical and methodological aspects of (1.2). The first important observation is that while there are now p "tuning" parameters per regression, there is still only *one* free parameter (namely $\bar{r}^j$) that requires calibration. This is highly desirable: for one parameter, efficient calibration schemes are known [1, 10, 56]; in contrast, the calibration of multiple free parameters, which would appear when adding additional penalty terms instead of adopting our approach, remains a major challenge in both theory and computations.

Second, for the link function f, there are often natural choices, such as our choice to reflect the three-dimensionality of distances. Third, note that our approach retains the "flavor" of the original penalty. This means that our concept maintains the general properties of the penalty, such as the sparsity generating effect of ℓ_1-penalties [60], and it means that the practical implementation of (1.2) can be based on standard software packages, such as `glmnet` [17] in the case of ℓ_1-penalization.

In the following, we set the penalty in (1.2) to the ℓ_1-norm and refer to this specification of our method as SI. An `R` implementation of SI is provided in Appendix A.

1.3.2 Theory

Here, we show our approach is amenable to the standard proof techniques in high-dimensional statistics. We apply the primal-dual witness technique to SI to show that we can exactly recover the true graph under classical assumptions. Note first that a sufficient condition for exact graph recovery is that the estimators (1.2) provide exact variable selection at each node j, cf. [20, Section 7.3]. We denote the corresponding target regression vector by β_*^j and the noise by ε^j. We can then formulate versions of the standard irrepresentability and beta-min conditions and bounds for the tuning parameters [8, 31, 62, 71]. To this end, denote by X_*^{-j} the matrix X restricted to the columns that have index in the set $S_*^j := \Big\{k \in \{1,\dots,p\} \mid k \neq j \text{ and } (k,j) \in \mathcal{E}\Big\}$ and by X_{-*}^{-j} the matrix X restricted to the columns that have index in $\Big\{k \in \{1,\dots,p\} \mid k \neq j \text{ and } (k,j) \notin \mathcal{E}\Big\}$. This means that X_*^{-j} corresponds to X^{-j} on the support of β_*^j and X_{-*}^{-j} to the remaining parts of X^{-j}. In the same spirit, we allow ourselves to use the analogous notations $\mathbf{r}_*^j$, $\mathbf{r}_{-*}^j$, and $(\beta_*^j)_*$. For each $j \in \{1,\dots,p\}$, we then assume that $(X_*^{-j})^\top X_*^{-j}$ is invertible and that there is a constant $b_j > 0$ such that

$$\|\mathbf{r}_*^j \circ (X_*^{-j\top} X_*^{-j})^{-1} X_*^{-j\top} X^i\|_1 / (\mathbf{r}^j)_i \leq 1 - b_j$$

for all $i \in \{k \in \{1,\dots,p\} \mid k \neq j \text{ and } (k,j) \notin \mathcal{E}\}$. One can check that the more informative D is, the weaker is the condition on the design. Moreover, we assume that for each $j \in \{1,\dots,p\}$

and corresponding tuning parameter $\mathbf{r}^j$, the regression vectors satisfy

$$\min_{i\in\{1,\dots,|S_*^j|\}} \left(\frac{1}{(\mathbf{r}_*^j)_i} \left| ((\beta_*^j)_*)_i \right| \right) \;>\; \frac{1}{n} \left[1 + \max_{\substack{\nu\in\{\pm 1\}^{|S_*^j|} \\ k\in\{1,\dots,|S_*^j|\}}} \frac{\left| \left((X_*^{-j\top} X_*^{-j}/n)^{-1} \nu \right)_k \right|}{2(\mathbf{r}_*^j)_k} \right].$$

These irrepresentability and beta-min conditions are precisely the standard ones for lasso regression except for the generalization to vector-valued tuning parameters $\mathbf{r}^j$. Finally, we assume that the tuning parameters are chosen such that

$$1 \;>\; \max_{i\in\{1,\dots,p-1-|S_*^j|\}} \frac{2\left| \left(X_{-*}^{-j\top} \left(\mathrm{I}_n - X_*^{-j} (X_*^{-j\top} X_*^{-j})^{-1} X_*^{-j\top} \right) \varepsilon^j \right)_i \right|}{(\mathbf{r}_{-*}^j)_i b_j}$$

and

$$1 \;>\; \max_{i\in\{1,\dots,|S_*^j|\}} \frac{n\left| \left((X_*^{-j\top} X_*^{-j})^{-1} X_*^{-j\top} \varepsilon^j \right)_i \right|}{(\mathbf{r}_*^j)_i}\,.$$

This again is an extension of standard bounds for tuning parameters to our setting.

We can now prove the following result.

theorem 1.3.1 (Exact Graph Recovery With SI)**.** *Under the mentioned assumptions, SI provides exact graph recovery, that is,*

$$\widehat{\mathcal{E}} \;=\; \mathcal{E}\,.$$

This result implies that SI can provide exact graph recovery if conditions that are analogous to the ones formulated in the literature are met. The proof follows the primal-dual witness rationale [65], see also [31, Chapter 11.4], and is deferred to Appendix B. Importantly, the assumptions, result, and proof follow well-established standards in high-dimensional theory.

1.3.3 Bayesian interpretation

Finally, our approach has a direct Bayesian interpretation. For a given index j, we consider the hierarchical Bayesian model

$$X^j \mid \beta, \sigma \;\sim\; \mathcal{N}(X^{-j}\beta\,, \sigma^2\,\mathrm{I}_{n\times n})\,, \tag{1.3}$$

$$\mathbb{P}\big(\beta_i \mid \sigma, (\mathbf{r}^j_{-j})_i\big) \;=\; \frac{(\mathbf{r}^j_{-j})_i}{2\sigma^2} e^{-\frac{(\mathbf{r}^j_{-j})_i}{\sigma^2}|\beta_i|}\,, \qquad (\, i \in \{1,\ldots,p-1\}\,)\,. \tag{1.4}$$

This model is a generalization of the model considered in the seminal paper that introduces the Bayesian lasso [52]. The parameters $(\mathbf{r}^j)_i$ in our general model are hyperparameters that specify the shape of the prior distribution of each of the regression coefficients $(\beta^j)_i$. The negative log-posterior distribution of β is now given by

$$-\log \mathbb{P}(\beta \mid X, \sigma, \mathbf{r}^j_{-j}) \;=\; \frac{1}{\sigma^2}\left(\frac{1}{2}\|X^j - X^{-j}\beta\|_2^2 + \sum_{i=1}^{p}(\mathbf{r}^j_{-j})_i|\beta_i|\right) + c\,,$$

where c is a term independent of β. The mode of this distribution is

$$\hat{\beta}^{j,\mathbf{r}^j} \;\in\; \operatorname*{argmin}_{\beta\in\mathbb{R}^{p-1}} \left\{\frac{1}{2}\|X^j - X^{-j}\beta\|_2^2 + \sum_{i=1}^{p}(\mathbf{r}^j_{-j})_i|\beta_i|\right\}, \qquad (\, j \in \{1,\ldots,p\}\,)\,,$$

which equals the estimator yielded by our approach SI, that is, (1.2) with the penalty norm set to the ℓ_1-norm. Similarly, replacing the double-exponential distribution in (1.4) with a Gaussian distribution, the posterior mode equals the estimators yielded by our approach when the penalty in (1.2) is set to the ℓ_2-norm.

This analysis shows that the tuning parameters relate our frequentist and Bayesian notions about the additional knowledge. In our frequentist estimator, the larger $(\mathbf{r}^j)_i$, the more likely the edge (i, j) is excluded from the estimate. In our Bayesian view, the larger $(\mathbf{r}^j_{-j})_i$, the more the assumed distribution of $(\beta^j)_i$ is concentrated around zero. Thus, funneling the additional knowledge into the tuning parameters of the frequentist estimator can be viewed as transforming the original priors, which are the same all across the coefficient vector, into informed priors tailored to each coefficient based on our biological rationale. In general, the Bayesian view provides further support for our concept – without inflicting the computational challenges that typically come with Bayesian methods.

1.4 Simulations

We now confirm that our method SI can improve graph estimation by conducting a simulation study and a sensitivity analysis.

First, we demonstrate in a simulation study that our approach can harvest additional knowledge to improve graph estimation. To generate data, we imitate the brain connectivity application in Section 1.5: the number of nodes $p = 116$ corresponds to the number of brain regions; the number of samples $n = 210$ corresponds to the number of fMRI scans per subject; the additional knowledge D_{ij} is the Euclidian distance between brain regions i and j. The edge set is constructed based on independent Bernoulli(p_{ij}) distributions, $p_{ij} = \text{inv.logit}(10 - D_{ij}/3)$, such that an edge (i, j) is included if the corresponding Bernoulli outcome is one. The form of the distributions captures our rationale that direct connections are predominately between close regions. An anatomical map of a graph generated by the preceding scheme is displayed in the left panel of Figure 1.2. Next, the off-diagonal non-zero entries in the inverse covariance matrix Σ^{-1} as specified by the edge set are set to 0.3. The diagonal entries are set to $0.2 + 0.3 \cdot \sigma_{\min}$, where $\sigma_{\min}$ is the minimal singular value of the adjacency matrix. This construction ensures that Σ^{-1} has full rank. Finally, n i.i.d. samples are generated from $\mathcal{N}_p(\mathbf{0}, \Sigma)$.

We run our approach SI against standard methods for graph estimation. The standard methods competing are THR, MB, and GLASSO. Motivated by the simulation results in Section 1.2.1, we adopt the "and-rule" for SI and MB. A total of 40 data sets are generated, on which each method is run as a function of its free tuning parameter. The average ROC curves for graph recovery are plotted in the right panel of Figure 1.2. We find that the ROC curve of SI dominates the other curves, demonstrating that SI can improve graph recovery when additional knowledge is available.

Next, we demonstrate the performance of our approach with a sensitivity analysis under three different data generation scenarios. Again we let $n \times p = 210 \times 116$. To generate data, we start from the hierarchical Bayesian model, sample β^j from the double-exponential distribution (1.4). For the three data generation schemes, we generate data with (1) the true model $\frac{(\mathbf{r}^j_{-j})_i}{\sigma^2} \propto D^3_{ij}$, (2) model misspecification $\frac{(\mathbf{r}^j_{-j})_i}{\sigma^2} \propto D_{ij}$, (3) no additional knowledge used $\frac{(\mathbf{r}^j_{-j})_i}{\sigma^2} = c$. We set the diagonal of the precision matrix Σ^{-1} to 1, and the off-diagonal entries to $(\Sigma^{-1})_{ij} = \frac{1}{2}(\beta^i_j(\Sigma^{-1})_{ii} + \beta^j_i(\Sigma^{-1})_{jj})$. The diagonal entries of Σ^{-1} are then set to a common

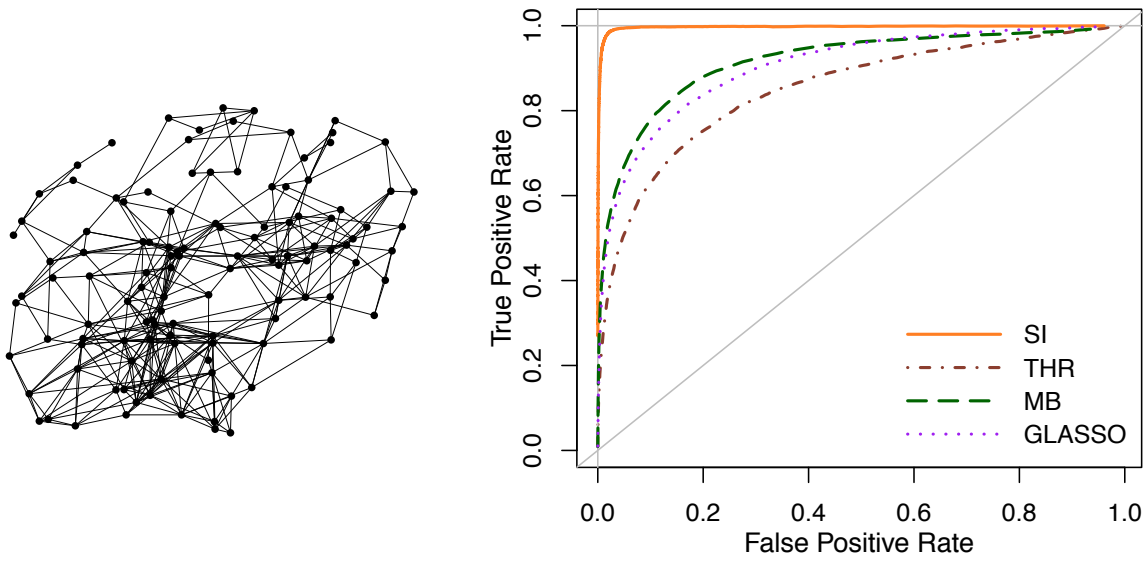
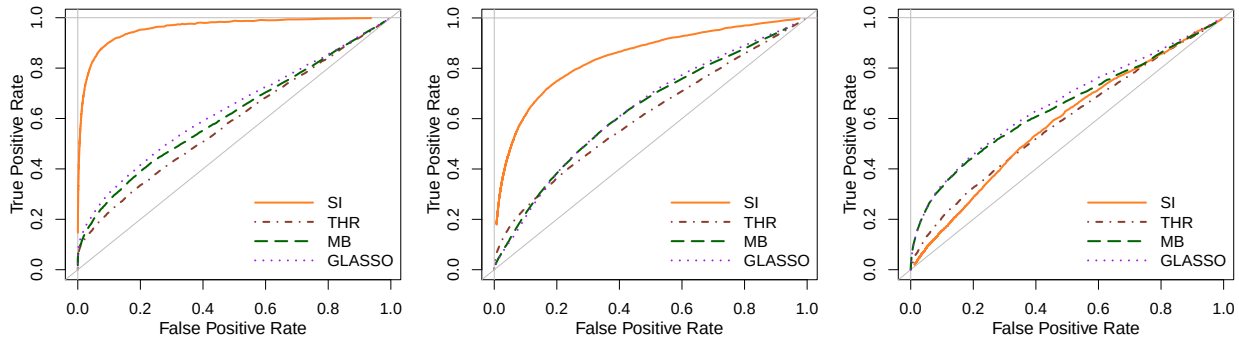


Figure 1.2: Left panel: One example of the 40 simulated graphs. Right panel: ROC curves demonstrating that SI can outperform standard methods when additional knowledge is available.

value such that the condition number equals 100. Last, i.i.d. samples are generated from the Gaussian distribution $\mathcal{N}_p(\mathbf{0}, \Sigma)$.

The average ROC curves for graph recovery over 40 repetitions are plotted in Figure 1.4. We find that the ROC curve of SI dominates the other curves when additional knowledge is used correctly. Our advantage decreases with the degree of model misspecifications, but our estimator's performance remains close to state-of-the-art methods even when the "knowledge" is completely wrong.

While the simulations provide us with some evidence, they cannot cover all cases that one could potentially encounter in practice. We thus complement the above analysis with a stability study on real fMRI data in the following section. Together, the results provide us with sufficient confidence to argue in favor of SI.



Knowledge fully correct | Knowledge partly wrong | Knowledge entirely wrong

Figure 1.4: ROC curves under three data generation schemes demonstrating that: (1) SI can largely outperform standard methods when the additional knowledge is correct; (2) SI can outperform standard methods when the additional knowledge is partly correct; (3) SI is underperforming standard methods only slightly when the additional knowledge is completely wrong.

1.5 Real data analysis

To iterate, our biological rationale for SI with the additional pairwise distances between brain regions is that direct connections are more likely between close regions than between distant regions. The SI incorporates this notion: First, graphical models, in general, distinguish between direct connections (conditional dependence) and indirect connections (marginal dependence). Second, the mentioned rationale that the distance between two nodes influences the likelihood of them being directly connected can be reflected naturally by how the link function f incorporates the additional distance information D. Importantly, we do not generally exclude direct or indirect long-range connections. This means that two distant nodes can very well be connected, but more likely, this connection will be indirect.

We now analyze the fMRI data on Alzheimer's disease with our method and compare the results to those of competing methods. We then draw insights about brain connectivity patterns from our estimates and contrast those insights with those that have been established

in Alzheimer's research.

1.5.1 Specification of the analysis

We compare our approach SI to THR, MB, and GLASSO. Again, we adopt the "and-rule" for SI and MB. The tuning parameters in MB are selected via 10-fold cross-validation. Since the tuning parameter in GLASSO is not amenable to the same cross-validation scheme, GLASSO is calibrated via BIC. In absence of established selection criteria for thresholding, THR is calibrated such that the number of connections equals the number of connections for SI.

Note that the coordinates of the observations in the present data set correspond to volume regions in the brain. The observations in another type of fMRI-induced data that received considerable attention recently correspond to the cerebral cortex [21]. We would then recommend our approach with geodesic distances among the regions as additional knowledge and a link function of form $f(x) = x^2$ to capture the two-dimensionality of the spherical coordinates.

1.5.2 Stability study

A reliable statistical method should provide similar outcomes across similar data sets. We thus study stability as a notion of similarity among outcomes of a method. In particular, we consider stability as a facet of reproducibility.

We employ two measures of stability. Both measures rely on data splitting with the main idea that agreement of estimates based on two parts of a data set indicates that the method is reliable, while largely disagreeing estimates raise a red flag. For the first comparison of different methods on the fMRI data, we proceed as follows. Given a patient and a method, we split the data containing the 210 scans randomly into two sets of 105 scans; then, we compute stability in terms of what we would call graph agreement

$$\text{GA} := 1 - \frac{d_{\text{H}}(\hat{\mathcal{E}}, \tilde{\mathcal{E}})}{|\hat{\mathcal{E}}| + |\tilde{\mathcal{E}}|} = 1 - \frac{\|\hat{A} - \tilde{A}\|_1}{\|\hat{A}\|_1 + \|\tilde{A}\|_1}$$

for the graphs $\hat{\mathcal{E}}$ and $\tilde{\mathcal{E}}$ (with corresponding adjacency matrices $\hat{A}$ and $\tilde{A}$) that are estimated on the two halves of the data. We finally take averages over 20 random splits each for all patients.

For the second comparison, we adopt the estimation stability concept of [68]. To this end, we split the data into v equal subsets of scans and compute the corresponding adjacency matrices $\hat{A}_1, \ldots, \hat{A}_v$. We then compute

$$\mathrm{ES}_v \;:=\; 1 - \frac{\frac{1}{v}\sum_{k=1}^{v} \|\hat{A}_k - \frac{1}{v}\sum_{l=1}^{v} \hat{A}_l\|_{\mathrm{F}}^2}{\|\frac{1}{v}\sum_{m=1}^{v} \hat{A}_m\|_{\mathrm{F}}^2}\,,$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. This definition of ES_v corresponds to [68, Page 1491], other than our definition is one minus the original one and [68]'s regression accuracy is replaced by a graph recovery accuracy here. Compare also to [44, Page 473]. We finally average ES_v over all patients.

Method	**GA**	**ES$_2$**	**ES$_5$**	**ES$_{10}$**
SI	0.62	0.75	0.39	0.02
MB	0.42	0.53	-0.27	-1.43
GLASSO	0.55	0.69	0.13	-0.59

Table 1.2: SI has higher stability on the fMRI data set than standard methods.

The values of GA, ES_2, ES_5, and ES_{10} are summarized in Table 1.2. (The sample sizes in the split data are not sufficiently large for the comparison of THR.) Note that for both GA and ES, large values mean high stability. Thus, SI is consistently the most stable method, suggesting that it can integrate additional knowledge effectively and make graph estimation more reproducible.

We have shown that SI can outperform standard methods when additional knowledge is available, can improve stability of graph estimation, is supported by theory, and has a Bayesian interpretation that matches our biological understanding while still being computationally convenient.

1.5.3 Results for brain connectivity

Having confirmed our approach's potential, we can now analyze the fMRI data on Alzheimer's disease. We compare our method with existing techniques and then work out the differences between brain connectivity networks of AD patients and normal aging subjects. For the latter task, we particularly focus on 42 regions of interest (ROI) for AD that have been identified in the literature [34, 35]. These regions are located in the frontal lobe, parietal lobe, occipital lobe, and temporal lobe. We provide evidence for three main claims: (1) connections are significantly more likely within the lobes than between the lobes among; (2) AD and NC differ mainly in their intra-lobe connections; (3) the connectivities of AD and NC differ from each other significantly within each of the four lobes. These insights could not be derived from competing methods.

To compare methods, graphs for one subject from each of the disease groups are displayed in Figure 1.5. The graphs are estimated based on all the data available for one subject. Each column in the figure corresponds to one (fixed) subject of the three groups AD, MCI, and NC; each row corresponds to one of the four methods SI, THR, MB, and GLASSO. The plots show anatomical maps, that is, the nodes are positioned according to their 3D coordinates. Edges between close regions (distances in the lower quartile) are colored orange; edges between more remote regions are colored blue. The total number of edges is stated below each map. While showing only one set of graphs per group for illustration, we find two patterns across all subjects: (i) With cross-validation as the most common calibration scheme applied, SI yields the most sparse networks. This finding shows that SI can lead to more manageable models. (ii) SI yields some edges between distant regions, but most edges are between spatially close regions (orange). In strong contrast, the other methods do not exhibit such a preference. This finding confirms our expectations about the four methods, and it shows that the estimates provided by SI are in agreement with the biological rationale.

To highlight contrasts among the study groups, the within-group averages of the esti-

mated adjacency matrices are compared:

$$\bar{A}^{\text{group}} := \frac{1}{N^{\text{group}}} \sum_{k \in \text{group}} \hat{A}^k \,,$$

where group $\in \{\text{AD}, \text{MCI}, \text{NC}\}$ is the study group, $\hat{A}^k$ is the SI estimate of the adjacency matrix for subject k, and N^{group} is the total number of subjects in the group. We first consider the entire brain to obtain an overview of the results. The contrast between the average graphs of AD and NC, that is, $\bar{A}^{\text{AD}} - \bar{A}^{\text{NC}}$, is displayed in Figure 1.6. The entries in the heat-map thus denote the frequencies of the edges in the AD group minus the corresponding frequencies in the NC group. The columns and rows are arranged such that spatially close regions tend to be close in the figure. Our main observation is that the overall graphs for the two groups seem to be similar, but there is high variability between AD and NC concentrated in some regions (see blue and yellow entries close to the diagonal). In particular, we find an increase in connectivity within the cerebellum for AD patients compared to NC patients, in concordance with the findings in [39].

For further analysis, we now focus on the mentioned 42 ROI. The corresponding subgraphs for the AD and NC groups are provided in Figure 1.8. The left panel displays $\bar{A}^{\text{AD}}$, the average graph estimates for the AD group, and the right panel $\bar{A}^{\text{NC}}$, the average graph estimates for the NC group, both restricted to the 42 ROI. The red squares highlight the lobes. The two plots indicate that connections are more likely within the lobes than between the lobes (the majority of gray and black cells are inside the red squares). To bring this visual observation on statistical grounds, we give the results for testing intra versus inter connectivities in Table 1.3. Precisely, we consider for each group $\in \{\text{AD}, \text{MCI}, \text{NC}\}$ and lobe $\in \{$frontal, parietal, occipital, temporal$\}$, the quantity

$$\Delta_{\text{intra-inter}} := \frac{1}{p_{\text{lobe}}} \sum_{i \in \left\{ \text{ ROI} \cap \text{lobe} \right\}} \left(\frac{1}{p_{\text{lobe}}} \sum_{j \in \left\{ \text{ ROI} \cap \text{lobe} \right\}} \bar{A}^{\text{group}}_{ij} - \frac{1}{42 - p_{\text{lobe}}} \sum_{j \in \left\{ \text{ ROI} \setminus \text{lobe} \right\}} \bar{A}^{\text{group}}_{ij} \right),$$

where p_{lobe} is the number of nodes in the lobe in the 42 ROI. This means that $\Delta_{\text{intra-inter}}$ is the difference of the intra (within-lobe) and inter (between-lobes) connectivities averaged over each region in the lobe. The (individual) p-values correspond to the t-tests of the null-hypothesis $\mathrm{H}_0 : \Delta_{\text{intra-inter}} = 0$. These findings provide evidence for our first claim that connections are significantly more likely within the lobes than between the lobes.

Competing methods such as GLASSO do not relate to a lobe structure, see Figure 1.11. In clear contrast, looking again at Figures 1.8 and 1.9, we see much more within-lobe connections in our plots (black cells within the red squares). Importantly, other short connections that are between-lobe connections are not picked up by our method even though the penalty for all the short connections are the same. This means that SI accentuates the biological lobe structure of the brain rather than just selecting arbitrary short distance connections, which provides us with further confidence in our method's compatibility with the data.

The 42 region subgraph of Figure 1.6 is provided in Figure 1.9. The heat map shows $\bar{A}^{\text{AD}} - \bar{A}^{\text{NC}}$, the contrast between the average graphs of AD and NC, restricted to the 42 ROI. The figure indicates that AD and NC differ mainly in their intra-lobe connections (the majority of yellow and blue cells are inside the red squares). The results of a statistical analysis of this observation are stated in Table 1.4. Precisely, we consider the quantity $\Delta_{\text{intra-inter}}$ with $\bar{A}^{\text{group}} = |\bar{A}^{\text{AD}} - \bar{A}^{\text{NC}}|$. This means that $\Delta_{\text{intra-inter}}$ is the difference of the intra and inter absolute connectivity contrast between AD and NC averaged over the regions in the lobe. The p-values correspond to the t-tests of the null-hypothesis $\mathrm{H}_0 : \Delta_{\text{intra-inter}} = 0$. These findings provide evidence for our second claim that AD and NC differ mainly in their intra-lobe connections.

Figure 1.9 also indicates that the connectivities of AD and NC indeed differ considerably within each lobe. Similar claims have been made in multiple papers [22, 23, 35, 59, 66]. We state the results of a corresponding statistical analysis in Table 1.5. Precisely, we consider for lobe $\in$ {frontal, parietal, occipital, temporal}, the quantity

$$\Gamma^{|\text{AD-NC}|} := \frac{1}{p_{\text{lobe}}(p_{\text{lobe}} - 1)} \sum_{i,j \in \left\{ \text{ROI} \cap \text{lobe} \right\}} \left| \bar{A}^{\text{AD}}_{ij} - \bar{A}^{\text{NC}}_{ij} \right|.$$

This means that $\Gamma^{|\text{AD-NC}|}$ is the absolute contrast between the AD and NC graphs averaged over all (potential) connections in the lobe. The p-values correspond to the t-tests of the null-hypothesis $\text{H}_0 : \Gamma^{|\text{AD-NC}|} = 0$. These findings provide evidence for our third claim that the connectivities of AD and NC differ from each other significantly within each of the four lobes.

1.6 Discussion and further research

We have shown that strengthening the role of tuning parameters based on pairwise distance is an effective approach to incorporate additional spatial knowledge in estimating brain connectivity networks more accurately. In particular, we have shown that this approach can improve stability of graph estimation, and has a clear biological Bayesian interpretation while still being computationally convenient.

The proposed method SI is particularly tailored for graph estimation of brain connectivity based on brain fMRI data and additional knowledge. More broadly, our scheme allows for the inclusion of general, application-specific information matrices D and link functions f. The information matrix D is typically predetermined in a given application. Similarly, the form of f often follows clear scientific rationales as can be seen in our brain connectivity case. One could also consider data-driven selections of f from a set of candidate functions by using score testing or related methods, cf. [26]; however, the sample sizes of typical data might be too small for a thorough non-parametric selection of the link function. Importantly, via the link function f and the information matrix D, our approach can be tailored to a wide range of applications in bioinformatics, speech recognition, computer vision, and digital communications. For example, we envision implementations in genomics, where the goal is to estimate gene regulatory networks based on gene expression levels [15]. The additional knowledge that researchers commonly want to invoke for this are previously established sub-networks or networks estimated in other studies.

Finally, we believe that our pipeline is amenable to other types of data and further inferential techniques: heavy-tailed or otherwise non-Gaussian data could be approached based

on [41]; additional inferential tools could be established by combining our ideas with [37].

From the Alzheimer's disease data considered in this paper, we find an increase in connectivity within the cerebellum for AD patients compared to NC subjects. More importantly, unlike any competing method, SI generates brain connectivity networks that accentuate the four biological lobes of the brain: connections are significantly more likely within the lobes than between the lobes, and AD patients and NC subjects differ mainly in their intra-lobe connections, with significant differences in all four lobes. This emphasizes the lobe structure's important role in brain connectivity.

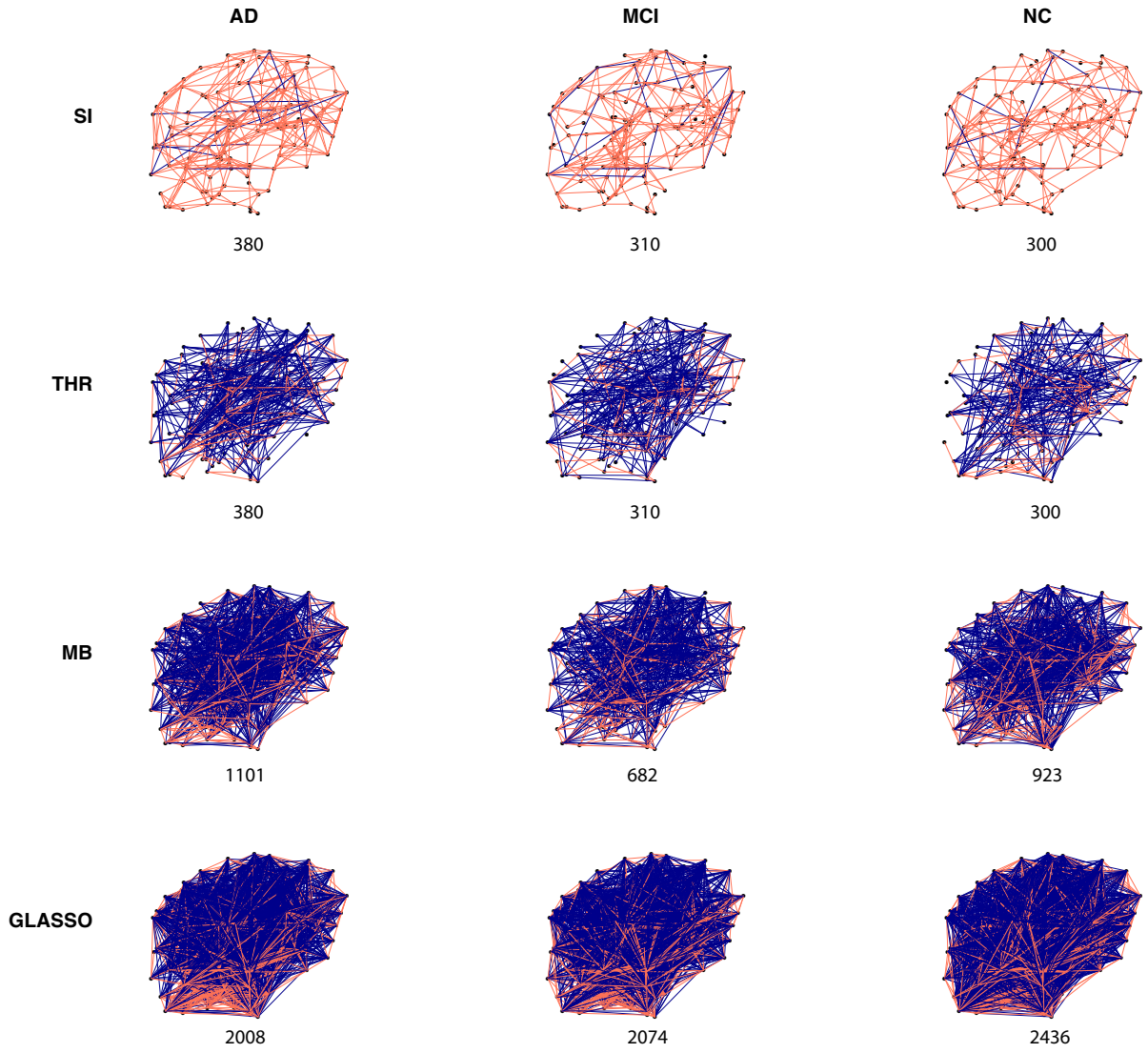


Figure 1.5: Anatomical maps of the estimated brain connectivity networks show that in contrast to standard methods, SI entails direct connections mostly between spatially close regions (orange lines).

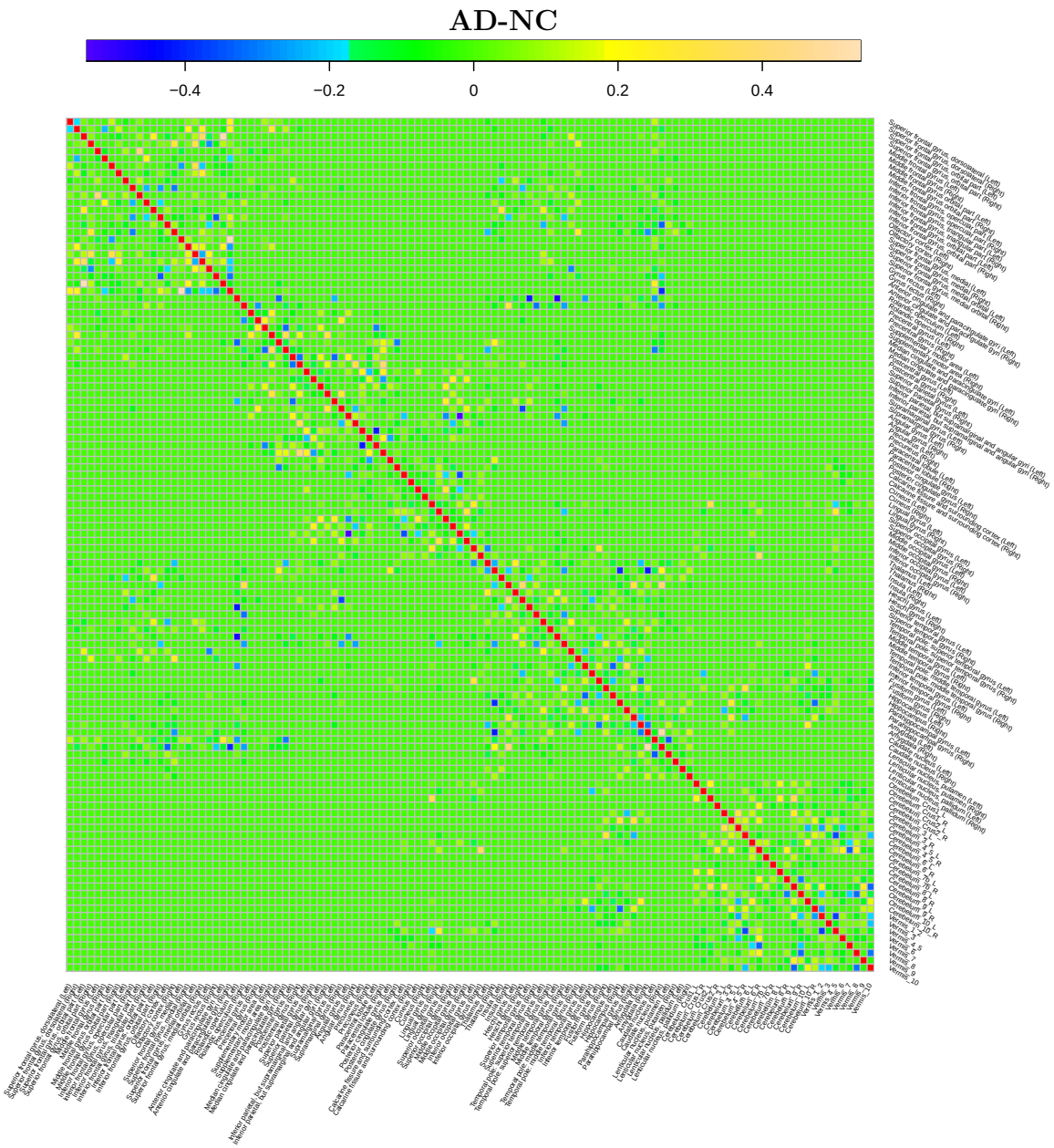


Figure 1.6: Contrast of average brain graphs between the AD and NC groups. Yellow indicates that an edge occurs more frequently in the AD group; blue indicates that an edge occurs more frequently in the NC group. The diagonal is colored in red. The graph indicates that there are considerable differences between the groups in some regions.

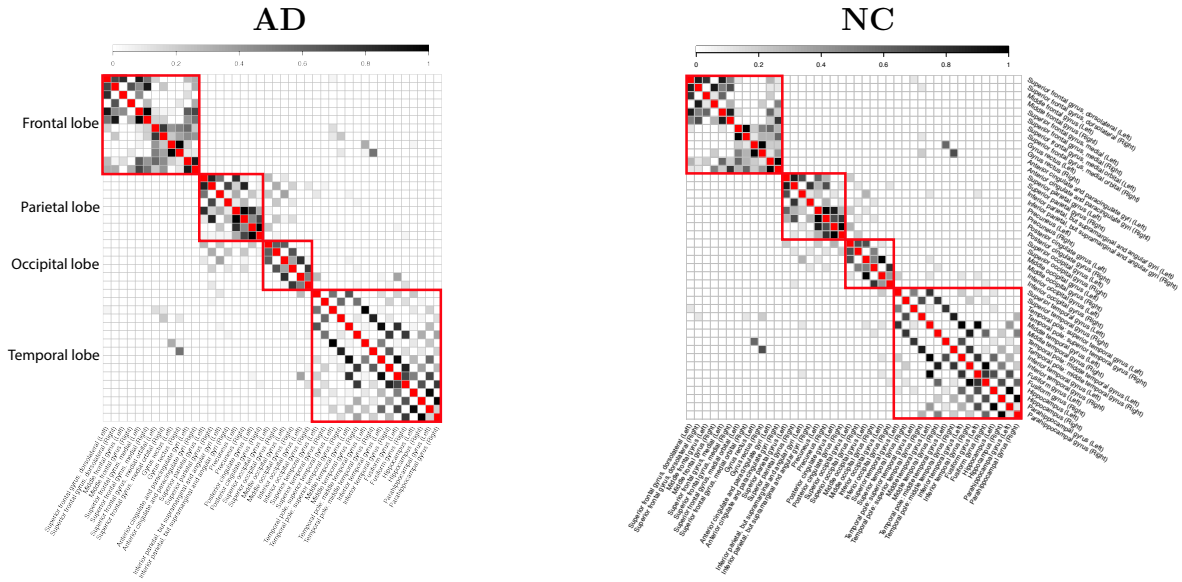


Figure 1.8: Average brain graphs within the 42 ROI for the AD and NC group. The graphs indicate that connections are more likely within-lobes than between-lobes.

Group	**Lobe**	$\Delta_{\text{intra-inter}}$	**p**
AD	Frontal lobe	0.265	< 0.0001
	Parietal lobe	0.330	0.0001
	Occipital lobe	0.280	0.0001
	Temporal lobe	0.168	< 0.0001
MCI	Frontal lobe	0.191	< 0.0001
	Parietal lobe	0.261	0.0003
	Occipital lobe	0.351	< 0.0001
	Temporal lobe	0.199	< 0.0001
NC	Frontal lobe	0.259	< 0.0001
	Parietal lobe	0.346	0.0001
	Occipital lobe	0.301	0.0002
	Temporal lobe	0.173	< 0.0001

Table 1.3: Connections are significantly more likely within-lobes than between-lobes.

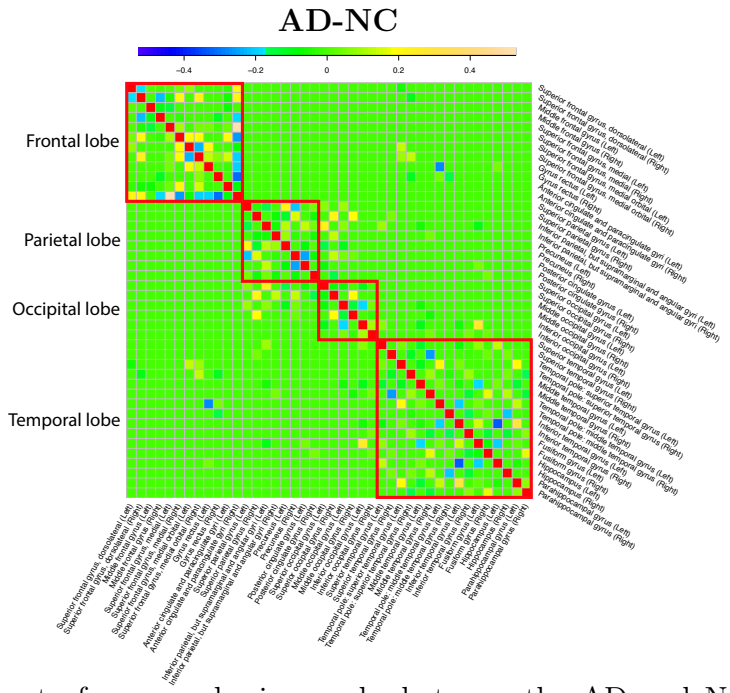


Figure 1.9: Contrast of average brain graphs between the AD and NC groups within the 42 ROI. The graph indicates that the contrasts between the groups are located mainly within the lobes and that the graphs of AD and NC indeed differ considerably from each other within the lobes.

Group	**Lobe**	$\Delta_{\text{intra-inter}}$	**p**
|AD-NC|	Frontal lobe	0.079	0.0001
	Parietal lobe	0.064	0.0001
	Occipital lobe	0.049	0.0014
	Temporal lobe	0.040	< 0.0001

Table 1.4: AD and NC differ significantly more within-lobe than between-lobes.

Lobe	$\Gamma^{\|\text{AD-NC}\|}$	**p**
Frontal lobe	0.083	< 0.0001
Parietal lobe	0.072	< 0.0001
Occipital lobe	0.060	0.0035
Temporal lobe	0.045	< 0.0001

Table 1.5: The connectivities of AD and NC differ from each other significantly within each of the four lobes.

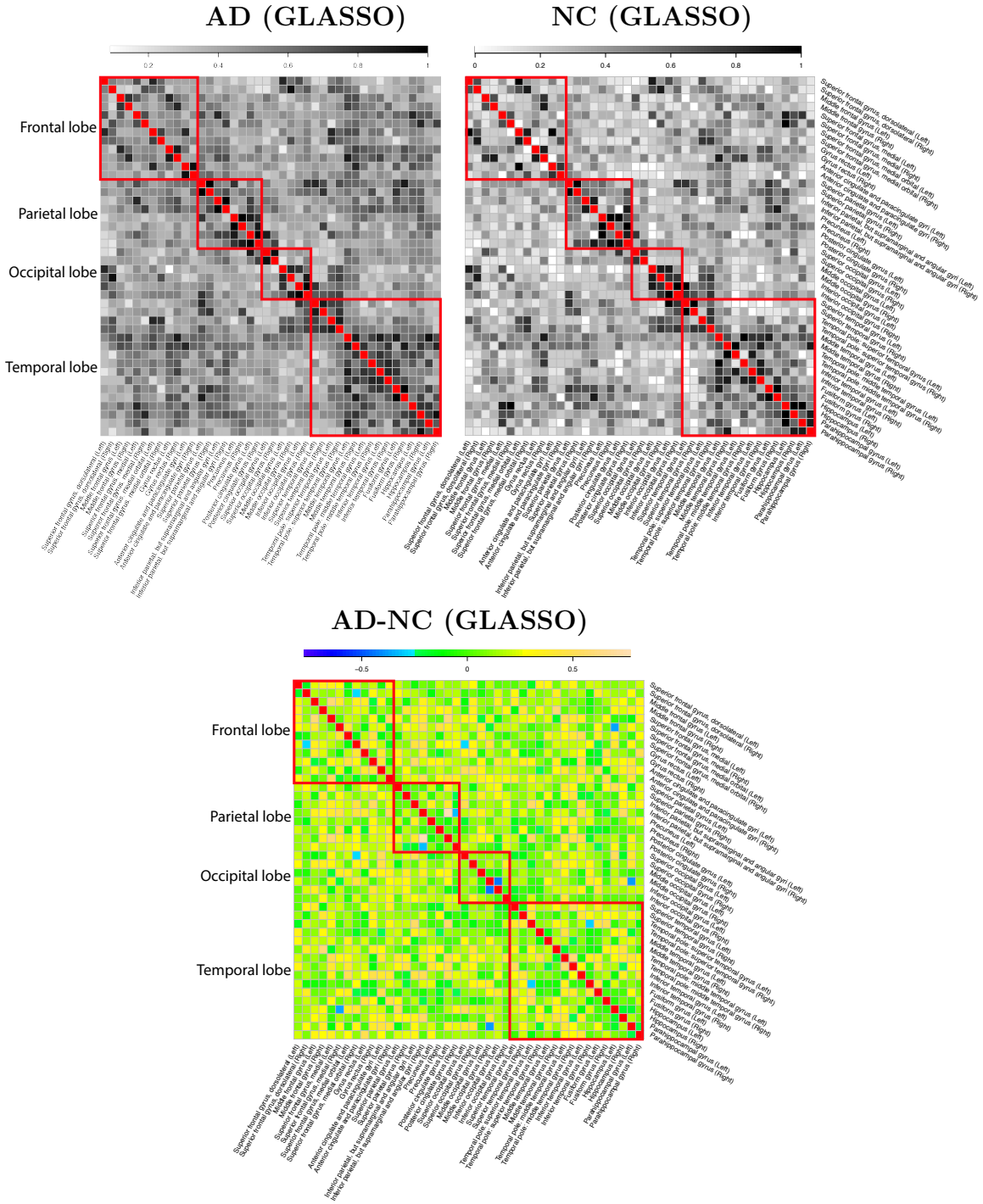


Figure 1.11: Top panel: average brain graphs from GLASSO within the 42 ROI for the AD and NC group. Bottom panel: Contrast of average brain graphs from GLASSO between the AD and NC groups within the 42 ROI. The plots do not exhibit a strong connection with the lobe structure.

Chapter 2

DRUG EFFICACY FOR BRIDGING STUDY IN COMPANION DIAGNOSTIC TEST TRIALS

This chapter is an adapted version of my published paper [7].

Abstract

Personalized medicine is an area of growing attention in medical research and practice. A market-ready companion diagnostic test (CDx) is used in personalized medicine for identifying the best treatment for an individual patient. Unfortunately, development of CDx may lag behind the development of the drug, and consequently we use a clinical trial assay (CTA) to enroll patients into the drug pivotal clinical trial instead. Thus, when CDx becomes available, a bridging study will be required to assess the drug efficacy in the CDx intended use population (CDx IU). Due to missingness of the CDx results that could be associated with randomization, one challenge we face in a bridging study is covariate imbalance between treatment arms for the subpopulation with both positive CDx and CTA. In this paper, we will evaluate the performance of two methods in bridging studies under a causal inference framework. Particularly, we aim to use the propensity score method with doubly robust estimation and optimal matching to address the challenge. We work under the framework suggested by Meijuan Li (2015) on how to estimate drug efficacy in the CDx intended use population, using data from both the bridging study and the CTA drug pivotal clinical trial. Both approaches are discussed in the context of a randomized bridging study, and a targeted design clinical trial with simulations, followed by analyzing simulated data that mimics a real ongoing clinic trial.

2.1 Introduction

Personalized medicine is an area of growing attention in medical research and practice. A market-ready companion diagnostic assay (CDx) that measures biomarkers is used in personalized medicine for choosing the best treatment for an individual patient. Such companion diagnostic devices for treatments include: Vysis ALK Break Apart FFPE FISH Test for Crizotinib, HER2 FISH pharmDx and IHC HercepTest for Herceptin, and cobas 4800 BRAF V600 Mutation Test for Vemurafenib [43]. CDx is essential for the both safety and efficacy aspects, and one main function of CDx is to identify the subpopulation who are most likely to benefit from the therapeutic product [16].

Ideally, we use CDx for patient enrollment into the device-drug pivotal clinical trial. This is to ensure appropriate clinical and analytical validation studies are planned ahead and carried out for CDx with the therapeutic product simultaneously. Unfortunately, the development of CDx may sometimes lag behind the development of the drug, consequently, CDx is unavailable during the clinical trial. Under this circumstance, instead of CDx, a clinical trial assay (CTA) is used for enrolling patients into the clinical trial. We also save the remaining specimen material, so that when CDx is ready to market, the kept specimen material can be retested by CDx for the bridging study. The purpose of retesting is to evaluate drug efficacy in the CDx intended use population (CDx IU) in an effort to "bridge" the drug efficacy results from CTA to CDx without repeating the clinical trial again with CDx.

Since the clinical trial has a targeted design, only patients with positive CTA results are enrolled into the randomized trial. Idealy, we would retest all patients tested by CTA for the clinical trial using CDx. However, retesting may not be possible for some cases, and yield only a subset of patients with valid CDx results. Reasons for missing CDx results include, for example, consent for retesting of specimen could not be obtained from patient due to bad treatment outcome, no remaining or insufficient amount of specimen left for retesting, and poor quality of remaining specimen to obtain a valid retest result. With missingness of

CDx results that could be associated with randomization, we face the challenge of covariate imbalance between treatment arms for the subgroup of patients with both positive CDx and CTA test results in the bridging study. Methods to deal with missing data, such as imputing CDx results can be done, but due to very limited information, the quality of imputation results can be unreliable and model driven. Thus, in this paper, we take another approach and evaluate the performance of two methods in bridging studies under a causal inference framework. Particularly, we aim to use the propensity score method with doubly robust estimation and optimal matching to address the challenge. We work under the framework suggested by Meijuan Li [43] directed at estimating drug efficacy in the CDx IU population, using data from the bridging study and CTA-drug pivotal clinical trial. Both approaches are discussed in the context of a random sample design bridging study and a targeted design clinical trial with simulations, followed by an illustration of analyzing simulated data that mimics a real ongoing clinic trial.

2.2 Methodology

2.2.1 Notation

Let D be the true biomarker status for a patient, where $D = 1$ if the true biomarker status is positive (denoted as $D+$) and $D = 0$ if the true biomarker status is negative (denoted as $D-$). The true biomarker status D is measured by both CDx and CTA tests, where CDx = 1 if a patient has positive CDx test results (denoted as CDx+) and CDx = 0 if a patient has negative test results (denoted as CDx−). Similar notation is used for CTA. The following parameters defined are population dependent and the population under consideration in this paper is the CDx IU population. Let $\varphi_{11} = Pr(\text{CTA}+|\text{CDx}+)$ denote the proportion of CTA+ patients that are CDx+ in the CDx IU population. Similarly we define $\varphi_{10} = Pr(\text{CTA}+|\text{CDx}-)$, $\phi_{11} = Pr(\text{CDx}+|\text{CTA}+)$ and $\phi_{10} = Pr(\text{CDx}+|\text{CTA}-)$ in the CDx IU population.

Let N be the total number of patients screened for the CTA+ drug pivotal clinical trial

with valid CTA results (*i.e.* eligible for bridging study). Of these N patients, N_1 of them are CTA+, thus, enrolled into the randomized trial based on a targeted study design. Let n be the number of patients enrolled for the bridging study. Idealy, we would retest all N patients screened for the pivotal trial using CDx and obtain valid results, so that $n = N$. However, retesting may not be possible for some cases and yield only a subset of n patients with valid CDx test result, so $n \leq N$.

	CDx		
	Positive	Negetive	Total
CTA			
Positive	n_{11}	n_{10}	$n_{1\cdot} = n_{11} + n_{10}$
Negetive	n_{01}	n_{00}	$n_{0\cdot} = n_{01} + n_{00}$
Total	$n_{\cdot 1} = n_{11} + n_{01}$	$n_{\cdot 0} = n_{10} + n_{00}$	$n = n_{1\cdot} + n_{0\cdot} = n_{\cdot 1} + n_{\cdot 0}$

Table 2.1: Cross-tabulation of the test results from CDx and CTA for the bridging study sample.

A bridging study is done with this size n sample. Test results for these patients from CDx can be compared to those previously obtained using CTA. The number of patients within each of the four combinations of the CDx and CTA test results from the bridging study is represented in Table 2.1. Let δ denote drug efficacy, in other words, the therapeutic effect difference between treatment and control on clinical outcomes. Let $\delta_{11} = \delta_{CTA+\cap CDx+}$ be the drug efficacy for CTA+ and CDx+. Similarly, let $\delta_{01} = \delta_{CTA-\cap CDx+}$ and $\delta_{\cdot 1} = \delta_{CDx+}$.

2.2.2 Estimate drug efficacy δ_{CDx+} and Var($\hat{\delta}_{CDx+}$)

In bridging study, our main goal is to estimate drug efficacy for the CDx+ patients (*i.e.* $\delta_{\cdot 1}$). Easily, $\delta_{\cdot 1} = \delta_{CDx+}$ can be written as a weighted average of $\delta_{11} = \delta_{CTA+\cap CDx+}$ and $\delta_{01} = \delta_{CTA-\cap CDx+}$, with weights $\varphi_{11} = Pr(\text{CTA+}|\text{CDx+})$ and $1 - \varphi_{11} = Pr(\text{CTA}-|\text{CDx+})$

respectively. Thus, we have

$$\delta_{\cdot 1} = \varphi_{11}\delta_{11} + (1 - \varphi_{11})\delta_{01}. \tag{2.1}$$

Assuming the estimates of φ_{11}, δ_{11} and δ_{01} are unbiased, asymptotically normally and mutually independent, then we can easily show that $\hat{\delta}_{\cdot 1} = \hat{\varphi}_{11}\hat{\delta}_{11} + (1 - \hat{\varphi}_{11})\hat{\delta}_{01}$ is asymptotically normal with mean $\varphi_{11}\delta_{11} + (1 - \varphi_{11})\delta_{01}$ and variance [58, 48],

$$\begin{aligned} Var(\hat{\delta}_{\cdot 1}) = \varphi_{11}^2 Var(\hat{\delta}_{11}) + (1 - \varphi_{11})^2 Var(\hat{\delta}_{01}) + [(\delta_{11} - \delta_{01})^2 + Var(\hat{\delta}_{11}) \\ + Var(\hat{\delta}_{01})]Var(\hat{\varphi}_{11}). \end{aligned} \tag{2.2}$$

Thus in order to estimate $\delta_{\cdot 1}$ and $Var(\hat{\delta}_{\cdot 1})$, we need to estimate φ_{11}, δ_{11}, δ_{01} and their corresponding variances.

Under the random sample design from the CDx IU population for a bridging study, naturally, we assume that n_{11} follow a binomial distribution, $n_{11} \sim \text{Binomial}(n_{\cdot 1}, \varphi_{11})$. Then we have the MLE of φ_{11} and variance as follows:

$$\hat{\varphi}_{11} = \frac{n_{11}}{n_{\cdot 1}} \tag{2.3}$$

$$\widehat{Var(\hat{\varphi}_{11})} = \frac{\hat{\varphi}_{11}(1 - \hat{\varphi}_{11})}{n_{\cdot 1}} \tag{2.4}$$

Provided there is a large enough sample size to secure statistical power, outcome data from the clinical trial of patients that are CTA+ and CDx+ could be used to estimate δ_{11} and $\text{Var}(\hat{\delta}_{11})$, or a subset of the sample could be used. This method does not have the benefit of treatment balance within the CTA and CDx both positive population, with respect to baseline covariates used as stratification factors in the randomization. We evaluate the performance of two approaches to resolve this problem. First approach, we utilizing the doubly robust estimator for causal effect estimations [19]. Second approach, we implement the optimal full matching method to estimate δ_{11} [55].

2.2.2.1 Doubly robust estimatior

We first discuss estimateing δ_{11} based on propensity score methods with the doubly robust estimator. Let Z be the indicator of treatment where $Z = 1$ if treated, and $Z = 0$ if in

control. Columns of $\mathbf{X}$ are covariates measured prior to treatment (baseline). Y_1 and Y_0 are potential outcomes under treatment and control from a causal inference framework, also $Y = ZY_1 + (1-Z)Y_0$ is the observed outcome. For each individual, we observe $(Z_i, X_i, Y_i), i \in \{1, \cdots, n_{11}\}$. The doubly robust estimator for δ_{11} is

$$\hat{\delta}_{11,DR} = n_{11}^{-1} \sum_{i=1} [\frac{Z_i Y_i}{\hat{e}_i} - \frac{Z_i - \hat{e}_i}{\hat{e}_i} m_1(X_i)] - n_{11}^{-1} \sum [\frac{(1-Z_i)Y_i}{1-\hat{e}_i} - \frac{Z_i - \hat{e}_i}{1-\hat{e}_i} m_0(X_i)]$$

[19]. Here $e(X) = P(Z = 1|X)$ is the propensity score, thus the probability of receiving treatment given the observed covariates. Logistic regression with treatment as outcome and observed baseline variables $\mathbf{X}$ as predictors is fitted to estimate $e_i = e(X_i)$. Meanwhile, $m_z(X) = E(Y|Z = z, X)$ is the regression model of outcome for only the group where treatment $Z = z$, $z \in \{0, 1\}$. Logistic regression is fitted for binary outcomes. The doubly robust estimator combines a form of outcome regression with a model for the treatment (i.e., the propensity score) to estimate the causal effect $E(Y_1 - Y_0)$ such that only one of the two models need be correctly specified to obtain an unbiased effect estimator [47].

Two types of variance estimators are considered for $\mathrm{Var}(\hat{\delta}_{11,DR})$, including the sandwich estimator and the bootstrap estimator. From Lunceford and Davidian [47], the sandwich estimator of variance is $n^{-2} \sum \hat{I}_i^2$ with

$$\hat{I}_i = \frac{Z_i Y_i}{\hat{e}_i} - \frac{Z_i - \hat{e}_i}{\hat{e}_i} \hat{m}_1(X_i) - \frac{(1-Z_i)Y_i}{1-\hat{e}_i} - \frac{Z_i - \hat{e}_i}{1-\hat{e}_i} \hat{m}_0(X_i) - \hat{\delta}_{11,DR}.$$

2.2.2.2 Optimal full matching

We also estimate δ_{11} using the optimal full matching method without restrictions on treatment-control balance. Full matching closely matches many more subjects than would matching with a fixed number of controls, and uses as many observations as suitable to be included in a comparison [29]. In a full matching, the sample is divided into a collection of non-overlapping matched sets that are either $1 : c$ or $t : 1$ for the number of treatment patients verses control patients, t and c can be any positive numbers, not fixed for different matched sets. In other words, each matched set consists either of a treated individual and any positive

number of controls, or a single control and any positive number of treated individuals [28]. A matched set assembles treated and control individuals that are comparable with respect to baseline covariates. The problem of finding an optimal full matching is equivalent to a standard problem of finding a minimum cost flow in a certain network, *i.e.* to a standard combinatorial optimization problem for which good algorithms exist [55]. Bootstrapping variance of $\text{Var}(\hat{\delta}_{11,DR})$ is considered for the full matching method.

2.2.3 Simulation settings with sensitivity analysis

Given that we do not have any CTA negative patients enrolled in the targeted design clinical trial, we have no outcome results to estimate δ_{01} or $\text{Var}(\hat{\delta}_{01})$ directly. Instead we conduct a sensitivity analysis suggested by Meijuan Li [43] under the reasonable assumption that δ_{01} is in between 0 and δ_{11}. In other words, we assume the treatment effect for the CTA-&CDx patients is at worse equal to no effect, and at best the same as the treatment effect for CTA and CDx both positive patients. To be more precise, we assume $\hat{\delta}_{01} = c\hat{\delta}_{11}$, for some $c \in [0, 1]$. Also assuming $Var(\hat{\delta}_{01}) = Var(\hat{\delta}_{11})$, we can estimate the drug efficacy in CDx intended use population and its associated variance as follows:

$$\hat{\delta}_{\cdot 1} = [(1-c)\hat{\varphi}_{11} + c]\hat{\delta}_{11}, \tag{2.5}$$

$$\widehat{Var}(\hat{\delta}_{\cdot 1}) = (2\hat{\varphi}_{11}^2 - 2\hat{\varphi}_{11} + 1)\widehat{Var}(\hat{\delta}_{11}) + [(1-c)^2\hat{\delta}_{11}^2 + 2\widehat{Var}(\hat{\delta}_{11})]\widehat{Var}(\hat{\varphi}_{11}) \tag{2.6}$$

[43]. Plugging in the sandwich and bootstrap estimators for $Var(\hat{\delta}_{11})$ gives two estimators for $Var(\hat{\delta}_{\cdot 1})$. We also compare those results to the bootstrap variance of $\hat{\delta}_{\cdot 1}$.

We generated data with a dichotomous treatment Z (50% treated) and a binary true status of the tested biomarker D (30% positive). CTA and CDx are generated to be positive with probability 0.97 when D is positive, and positive with probability 0.07 when D is negative. CTA, Z, D, and normal baseline variables $X = (X_1, X_2)$ are used to generate the missingness of CDx (16.3% missing). The binary potential outcomes (Y_1, Y_0) are simulated

based on D and X with different sets of coefficients. The mean difference between Y_1, Y_0 is -0.36 when D is positive, and -0.04 when D is negative.

We simulate 1000 cohorts of size 1000. For each simulated cohort, any bootstrap is carried out with 100 resamples with replacement. Drug efficacy and variance are estimated in each simulated cohort with all regression models correctly specified using methods discussed in section 2.2. For $\hat{\delta}_{\cdot 1}$, we determine the 95% confidence interval coverage for each method of variance estimation by assessing the proportion of intervals that contained the true value in the 1000 cohorts. Results are presented for when c is $0, c_0$ and 1, where $c_0 = 0.2545$ is the truth for the simulated cohorts.

2.3 Results

All the following numerical computation is performed using RStudio over R version 3.0.2 on a standard MacBook Air with 1.8GHz Intel Core i5 and 4GB 1600MHz DDR3 memory. For optimal full matching, we use the R implementation from the package optmatch [29].

2.3.1 Simulation results

Simulation results are presented in Table 2.2.

The fitted propensity scores mostly overlap for the treated and controlled groups. The doubly robust estimator method performs better than the optimal matching method in all of our simulation setting scenarios. With the doubly robust estimator method, in the scenario of $c = c_0$, the drug efficacy results are unbiased, all estimated standard errors slightly underestimate the true variability of our estimate. And when $c = 1$, $\hat{\delta}_{01} = \hat{\delta}_{11}$, which means we are assuming the treatment effect for the CTA-&CDx patients is equal to the treatment effect for CTA and CDx both positive patients. This assumption is very far from the truth for our simulation setting where $c = c_0 = 0.2545$. Thus as expected, our estimate performs poorly under both methods when $c = 1$. Clearly, a reasonable choice of c is important to achieve unbiased results. Bootstrap confidence intervals for $\hat{\delta}_{\cdot 1}$, in comparison, provided better coverage across all scenarios considered. Thus, we recommend reporting

$\hat{\delta}_{11}$ Method	c	Bias	MSE	SE_s	SE_{b1}	SE_{b2}	95% CI Coverage, %		
							SE_s	SE_{b1}	SE_{b2}
Doubly robust estimator	0	-0.0049	0.0019	0.0412	0.0436	0.0424	93.5	94.1	93.4
	c_0	-0.0186	0.0024	0.0412	0.0436	0.0447	90.3	91.6	92.3
	1	-0.0588	0.0059	0.0412	0.0436	0.0500	67.0	68.7	77.1
Optimal matching	0	-0.0258	0.0030	–	0.0500	0.0480	–	91.5	90.3
	c_0	-0.0404	0.0041	–	0.0490	0.0500	–	84.9	85.7
	1	-0.0833	0.0101	–	0.0490	0.0566	–	57.7	67.8

Table 2.2: Bias, MSE, estimated standard errors and 95% confidence interval coverage for drug efficacy in the CDx+ population ($\hat{\delta}_{\cdot 1}$), under doubly robust estimator and optimal matching methods for δ_{11} estimation.

Abbreviations: MSE, mean squared error; CI, confidence interval; SE_s, standard error from equation (6) by plugging the sandwich estimator for $Var(\hat{\delta}_{11})$; SE_{b1}, standard error from equation (6) by plugging the bootstrap estimator for $Var(\hat{\delta}_{11})$; SE_{b2}, standard error estimated from the 100 bootstrapped estimates of $\delta_{\cdot 1}$. All standard error estimates is an average over 1000 simulated cohorts. $c_0 = 0.2545$ is the truth of c for the simulated cohorts.

estimates using the doubly robust estimator method for δ_{11} and bootstrapped estimates for the standard errors and confidence intervals.

2.3.2 Application: Real example results

The study under consideration is a ongoing randomized, double-blind, placebo-controlled phase 3 study in subjects with a newly diagnosed subtype of a disease. The primary objective of this study is to evaluate if the drug additional to other treatments prolongs event-free survival among subjects within this subtype. The market-ready CDx is desired for the classification of the subtype of the disease. However, CDx is not available at the onset of enrollment. Instead, the CTA is used to screen all subjects.

As the clinical trial is currently ongoing and blinded, original data is not yet available, thus we simulated data mimicking the setup of the ongoing trial. The binary outcome in our

simulated data is an indicator of death within 3 years. We analyze the data using doubly robust estimator methods for δ_{11}. For 6 choices of $c \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, we estimate $\delta_{\cdot 1}$ with all types of variances mentioned and their corresponding 95% confidence intervals.

c	$\hat{\delta}_{\cdot 1}$	$\mathbf{SE}_s$	$\mathbf{SE}_{b1}$	$\mathbf{SE}_{b2}$	95% **CI**		
					$\mathbf{SE}_s$ **based**	$\mathbf{SE}_{b1}$ **based**	$\mathbf{SE}_{b2}$ **based**
0.0	-0.230	0.035	0.034	0.034	-0.299, -0.162	-0.297, -0.164	-0.296, -0.165
0.2	-0.234	0.035	0.034	0.034	-0.302, -0.165	-0.300, -0.167	-0.300, -0.167
0.4	-0.237	0.035	0.034	0.035	-0.305, -0.169	-0.303, -0.171	-0.305, -0.169
0.6	-0.240	0.035	0.034	0.035	-0.308, -0.172	-0.306, -0.174	-0.309, -0.172
0.8	-0.243	0.035	0.034	0.035	-0.311, -0.176	-0.309, -0.178	-0.313, -0.174
1.0	-0.247	0.035	0.034	0.036	-0.315, -0.179	-0.313, -0.181	-0.317, -0.176

Table 2.3: Estimation (of $\hat{\delta}_{\cdot 1}$), estimated standard errors and 95% confidence intervals for drug efficacy in the CDx+ population, under doubly robust estimator methods for δ_{11} estimation.

Abbreviations: CI, confidence interval; SE_s, standard error from equation (6) by plugging the sandwich estimator for $Var(\hat{\delta}_{11})$; SE_{b1}, standard error from equation (6) by plugging the bootstrap estimator for $Var(\hat{\delta}_{11})$; SE_{b2}, standard error estimated from the standard deviation of 100 bootstrapped estimates of $\delta_{\cdot 1}$.

Results are presented in Table 2.3. Among patients with CTA results, 5% of the them are missing test results of the CDx test. As an example of the results presented, when $c = 0.6$ and we use the bootstrap standard errors, the estimated drug efficacy for the CDx+ population is -0.240 (95% CI: -0.309, -0.172). Indicating that the probability of death within 3 years is 24.0% lower for the treatment group compared to the controls. This result is not abnormal if the true difference in probability is between 17.2% lower to 30.9% lower for the treatment group verses controls.

2.4 Discussion

Using the propensity score method with doubly robust estimation and optimal full matching to solve the covariate imbalance problem in bridging study with missing CDx results, we evaluate the performance of the two methods under a causal inference framework. In the presence of the unknown ratio c, we conduct a sensitivity analysis over c. We showed that estimate of drug efficacy in the CDx positive patients using the doubly robust estimator method for δ_{11} and bootstrapped estimates for the standard errors give good results with a reasonable choice of c. We also analyzed simulated data mimicking the setup of the ongoing trial. Survival outcome with censoring instead of a binary outcome could be explored in future research.

Chapter 3

INCREASING THE STABILITY OF GRAPHICAL MODELING BY USING DATA ACROSS IMAGING SCANS AND STRUCTURAL BRAIN INFORMATION

Abstract

Recent advances in data acquisition and preprocessing have largely improved the reliability of functional magnetic resonance imaging for estimating functional brain networks. This opens doors toward its use for longitudinal mapping of network changes in clinical settings. However, discussion for estimating networks in a longitudinal clinical setting are scarce. Here, we propose an approach that incorporates information from baseline assessment when estimating networks in follow-up data. Using data from the Consortium for Reliability and Reproducibility, we illustrate that our approach produces stable networks across repeated scans of the same individuals. The approach may, therefore, be of particular value to the clinical neurosciences, and we will make code and documentation openly available on Github.

3.1 Introduction

Resting-state fMRI has become a mainstream gateway to brain connectivity networks in the absence of tasks. A popular statistical framework for those networks is Gaussian graphical models. Since these models are typically high-dimensional, that is, the number of nodes rival or even exceeds the number of observations, estimation is often conducted via regularized maximum likelihood-like approaches such as neighborhood selection (MB) [49] and graphical lasso (GLASSO) [2, 18, 70]. The first approach aims at graph reconstruction by aggregating local estimates [49], while the second one is based on a global objective function [2, 18, 70]. Both approaches are now accompanied by a bulk of literature on theory and computation;

we refer to [8, 31] and references therein.

A topic that has become increasingly prominent in the discussion of such graphical models is reproducibility [68]. One aspect of reproducibility is stability, which means similarity among outcomes of a statistical method across similar data sets. fMRI can be obtained multiple times per person. Human brain connectivity has been shown to be reproducible across individuals [13]. Thus network estimations with more stable within-person brain connectivity network edges are preferred. Then an important question is how to leverage data for most stable estimations of brain connectivity networks.

Although some more sophisticated techniques can jointly estimate related Gaussian graphical models from multiple high-dimensional data, stability across these estimated connectivity networks has received limited attention. Understanding the stability of the estimated within-subject connectivity is crucial to the validity of all scientific results directly drawn from brain connectivity networks, as we need to detangle the actual findings from the within-subject variations. More stable within-subject connectivity gives us more confidence comparing similarities and differences between subjects.

In this paper, we extend neighborhood selection to borrow strength across different fMRI scans to reveal more similar within-subject connectivity structures. This joint Gaussian graphical model procedure links the estimation of separate graphical models through a fused penalty as proposed in [4]. Pairwise distances between brain regions are also integrated into the model to leverage the structural information as in Chapter 1 [6]. The advantage of this method is the ability to jointly estimate more stable edges across graphs, which leads to improvements compared to fitting just separate models, as it can borrow information from other related graphs. Thus for each subject, we can inform the brain connectivity structure from latter fMRI scans with previously obtained fMRI scans for more accurate estimation.

Related literature: Abundant literature focus on the joint estimation of multiple precision matrices through varies penalties: [14, 27, 64, 11, 33, 30, 50, 51, 72, 54, 67, 42, 69] A few Bayesian approaches also attempts to jointly estimate multiple precision matrices [53, 46, 12]. Also, [45] propose a Bayesian neighborhood selection method to jointly estimating multiple

brain networks. [4] proposed the penalized neighborhood selection approach with fused lasso to identify differences in Gaussian graphical models with similar structures.

3.2 Method

Starting from the well-established neighborhood selection framework with additional knowledge defined in Chapter 1, for one node $j \in \{1, \ldots, p\}$, the node-wise regression is

$$\hat{\beta}^j \in \underset{\beta \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|X_1^j - X_1^{-j}\beta\|_2^2 + \|\mathbf{r}_{-j}^j \circ \beta\|_1 \right\}, \tag{3.1}$$

where the circle $\circ$ indicates element-wise multiplication. We extend this idea to the problem setting of two data matrices X_1 and $X_2 \in \mathbb{R}^{n \times p}$ by incorporating fused penalties. We aim to leverage the fact that X_1 and X_2 are scans from the same person to encourage the two groups of parameters $\beta_1^j, \beta_2^j \in \mathbb{R}^{p-1}$ to shrink together. For each $j \in \{1, \ldots, p\}$, we consider a joint node-wise regression

$$\hat{\beta}_1^j, \hat{\beta}_2^j \in \underset{\beta_1, \beta_2 \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\{ \frac{1}{2}\|X_1^j - X_1^{-j}\beta_1\|_2^2 + \frac{1}{2}\|X_2^j - X_2^{-j}\beta_2\|_2^2 + \|\mathbf{r}_{-j}^j \circ \beta_1\|_1 + \|\mathbf{r}_{-j}^j \circ \beta_2\|_1 + \gamma^j \|\beta_1 - \beta_2\|_1 \right\}, \tag{3.2}$$

The jointly estimated corresponding regression parameters $\beta_1^j, \beta_2^j \in \mathbb{R}^{p-1}$ for a given node j determines the edges between node j and the other nodes. The vectors $X_1^j, X_2^j \in \mathbb{R}^n$ denote the jth column of the data matrices X_1, X_2, the matrices $X_1^{-j}, X_2^{-j} \in \mathbb{R}^{n \times (p-1)}$ denote X_1, X_2 without the jth column. Tuning parameters are $\mathbf{r}^j \in \mathbb{R}_+^p$ and $\gamma^j \in \mathbb{R}_+$. We set $(\mathbf{r}^j)_i = \bar{r}^j \cdot D_{ij}^3$, and $\gamma^j = \bar{r}^j \cdot \gamma_0^j$. $\mathbf{r}_{-j}^j \in \mathbb{R}^{p-1}$ is $\mathbf{r}^j \in \mathbb{R}^p$ omitting the jth element. Here $\bar{r}^j$ is an overall tuning parameter for the jth node's regression. As customary in high-dimensional statistics, the free parameter $\bar{r}^j$ balances the weights of the data and the structural assumptions (all the penalty terms). γ_0^j balances the ratio between the fuse penalty and other penalties. In practice, $\bar{r}^j$ can be calibrated via 10-fold cross-validation.

In the following, we refer to our method as Geofuse for the integrated geometric information and fused lasso.

3.3 fMRI data

The fMRI dataset we use consists of 30 healthy young adults from the Hangzhou Normal University of the Consortium for Reliability and Reproducibility (CoRR-HNU) dataset [9, 74]. Each participant received ten resting-state fMRI scans across one month, one scan every three days. We utilized all 300 scans in our analysis. Imaging was performed with a GE MR750 3.0 Tesla scanner (GE Medical Systems, Waukesha, WI) with 8-channel standard head coil (EPI: TR/TE = 2000 ms/30 ms, FOV = 220 × 220 mm, flip angle = 90°, matrix = 64 × 64, voxel size = 3.4 × 3.4 × 3.4 mm, slices = 43). A detailed description of the dataset can be found on `http://fcon_1000.projects.nitrc.org/indi/CoRR/html/hnu_1.html`. The preprocessing of the fMRIs was done by Tobias Kaufmann in the same manner as in his paper[38]. Preprocessing details can be found in the method section. We applied the group-wise graph-theory-based 100-parcellation whole-brain atlas defined in [57]. Each preprocessed data contains $n = 295$ consecutive scans over $p = 93$ nodes. An autoregressive integrated moving average model [24, 32] was applied to account for autocorrelation.

3.4 Stability study

A reliable statistical method should provide similar outcomes across multiple data sets of one person. We thus study stability as a notion of within-subject similarity among outcomes of a method. We use two specific measures: GA and ES. The first measure, GA, assesses the similarity between two graphs $\hat{\mathcal{E}}$ and $\tilde{\mathcal{E}}$ (with corresponding adjacency matrices $\hat{A}$ and $\tilde{A}$) as follows:

$$\mathrm{GA} := 1 - \frac{d_{\mathrm{H}}(\hat{\mathcal{E}}, \tilde{\mathcal{E}})}{|\hat{\mathcal{E}}| + |\tilde{\mathcal{E}}|} = 1 - \frac{\|\hat{A} - \tilde{A}\|_1}{\|\hat{A}\|_1 + \|\tilde{A}\|_1},$$

where Hamming distance between two graphs $\hat{\mathcal{E}}$ and $\tilde{\mathcal{E}}$ is defined by $d_{\mathrm{H}}(\hat{\mathcal{E}}, \tilde{\mathcal{E}}) := \big|\{(i,j) \mid (i,j) \in \hat{\mathcal{E}}, (i,j) \notin \tilde{\mathcal{E}}\} \cup \{(i,j) \mid (i,j) \notin \hat{\mathcal{E}}, (i,j) \in \tilde{\mathcal{E}}\}\big|$. The second measure, ES, we adopt the estimation stability concept of [68]. We compute

$$\mathrm{ES}_v := 1 - \frac{\frac{1}{v}\sum_{k=1}^{v} \|\hat{A}_k - \frac{1}{v}\sum_{l=1}^{v} \hat{A}_l\|_{\mathrm{F}}^2}{\|\frac{1}{v}\sum_{m=1}^{v} \hat{A}_m\|_{\mathrm{F}}^2},$$

where $\|\!| \cdot |\!\|_{\mathrm{F}}$ is the Frobenius norm. This definition of ES_v corresponds to [68, Page 1491], other than our definition is one minus the original one and [68]'s regression accuracy is replaced by a graph recovery accuracy here. Compare also to [44, Page 473].

We compare our approach Geofuse to MB and GLASSO. Specifically for Geofuse, we pair each scan from Scan 2 to Scan 10 with Scan 1 of the same person for joint estimation. We adopt the "and-rule" for neighborhood selection, the overall tuning parameters $\bar{r}^j$ are selected via 5-fold cross-validation while $\gamma_0^j = 1$. For GLASSO and MB, graph estimation is done on each scan separately. We adopt the "and-rule" for neighborhood selection in MB and select the tuning parameters via 10-fold cross-validation. Since the tuning parameter in GLASSO is not amenable to the same cross-validation scheme, GLASSO is calibrated via BIC.

To calculate within-subject stability for one person with GA, we compute the average of all 45 pairwise GA of Scan 2 to 10 from that person. To calculate within-subject stability for one person with ES, we compute ES_9 over the pooled adjacency matrices of Scan 2 to 10 from that subject. For all three methods considered, the values of within-subject GA and ES_9 averaged over all patients are summarized in the left table in Table 3.1.

For comparison, we also estimate the group stability of all 30 subjects, shown in the right table of Table 3.1. For GA each time we pick one scan each from two different people from Scan 2 to 10 to compute GA, then average over all the GA to get a group GA. We compute ES_{270} over the pooled adjacency matrices of Scan 2 to 10 from all 30 subjects.

Both measures rely on the main idea that agreement of estimates arising from two scans of the same person indicates that the method is reliable, while largely disagreeing estimates raise a red flag. Note that for both GA and ES, large values mean high stability. Thus, Geofuse is consistently the most stable method, suggesting that it can leverage multiple fMRI data effectively and make graph estimation more reproducible. We also observe that across all methods, within-person stability is higher than group stability, showing the within-person reproducibility of the human brain connectivity as expected.

Within-subject stability			Group stability		
Method	**GA**	**ES_9**	**Method**	**GA**	**ES_{270}**
Geofuse	0.67	0.59	Geofuse	0.65	0.57
MB	0.39	-0.21	MB	0.33	-0.86
GLASSO	0.59	0.43	GLASSO	0.58	0.28

Table 3.1: Geofuse has higher stability on the fMRI data set than standard methods, indicating that it can leverage multiple fMRI data effectively and make graph estimation more reproducible. All methods have higher within-person stability than group stability, showing the within-person reproducibility of human brain connectivity.

3.5 Brain connectivity results

Having established the stability of our approach, we can now analyze and visualize the fMRI data brain connectivity network estimations. To compare methods, two graphs and their differences (from Scan 1 and Scan 7) for one person are displayed in Figure 3.1. The first row graphs are estimated jointly based on Geofuse. Second and third row graphs are estimated separately. Each column in the figure corresponds to estimations from Scan 1, Scan 7, and their network differences. Each row corresponds to one of the three methods Geofuse, MB, and GLASSO. The plots show anatomical maps, that is, the nodes are positioned according to their 3D coordinates. Edges between close regions (pairwise distances in the lower quartile) are colored orange; all other edges are colored blue. While showing only one person's graphs for illustration, we find two patterns across all subjects: (i) a closer analysis shows that jointly estimated networks from Geofuse are more similar compared to networks estimated separately by other methods. (ii) Geofuse yields more edges between spatially close regions (orange) than the other methods do, the other methods do not exhibit such a preference for short connections. This finding shows that the estimates provided by Geofuse are in agreement with the biological rationale that direct connectivities in the brain are more likely to be between spatially close regions stated in Chapter 1.

To compare among subjects, the within-person averages of the estimated adjacency matrices are calculated:

$$\bar{A}^i := \frac{1}{9}\sum_{j=2}^{10} \hat{A}^i_j \,,$$

where $\hat{A}^i_j$ is the estimate of the adjacency matrix for Subject i ($i \in 1, 2, \ldots, 30$) of Scan j. Corresponding heatmaps are displayed for two subjects in the top row of Figure 3.3. The entries in the heat-map denote the frequencies of the edges within each person. Their differences and the within-person brain graphs averaged over all subjects is displayed in the bottom row.

Our main observation is that the graphs for the two subjects under Geofuse seem to be similar, there is no visible structural variability between the two subjects in either direction (see red and yellow entries on the orange background of the Subject 28 - Subject 7 heatmap in Figure 3.3). In particular, we find the within-person heatmaps to be quite similar to the all-subject averaged heatmap in pattern. The results and patterns we see under Geofuse are not present for GLASSO or MB as shown in 3.5 and 3.7. In clear contrast, the two competing methods' heatmaps for the two subjects and the group average do not present many visible patterns, we can not identify any particular regions of interest from the heatmaps of these two methods.

Through linear regression adjusted for subject id and implementing the Bonferroni correction, we find three observations on other graph metrics: (1) Edge density and max coreness of the graphs differ significantly by subject age (mean: 24 range: 20 - 30) across all the methods we applied. (2) Mean alpha centrality of the graphs differ significantly by subject age under Geofuse, but not under the other two methods. (3) Max edge degree, edge density, and max coreness all differ significantly by scan day (from baseline scan) under Geofuse, but not under the other two methods.

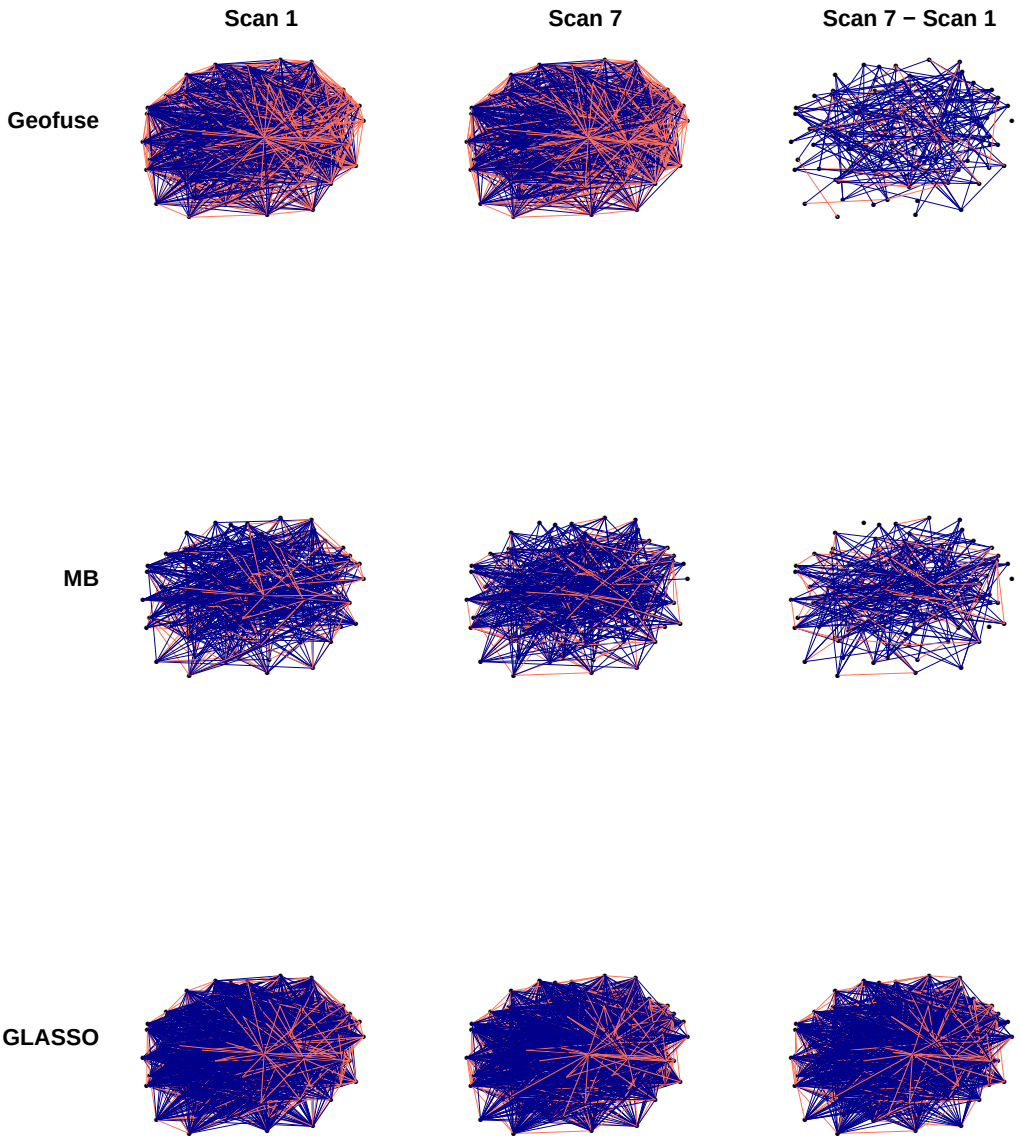


Figure 3.1: Anatomical maps of the estimated brain connectivity networks. First row: one person's example of the two estimated graphs from one model under Geofuse. Second row: graph estimation of corresponding scans from the same person under MB. Third row: graph estimation of corresponding scans from the same person under GLASSO.

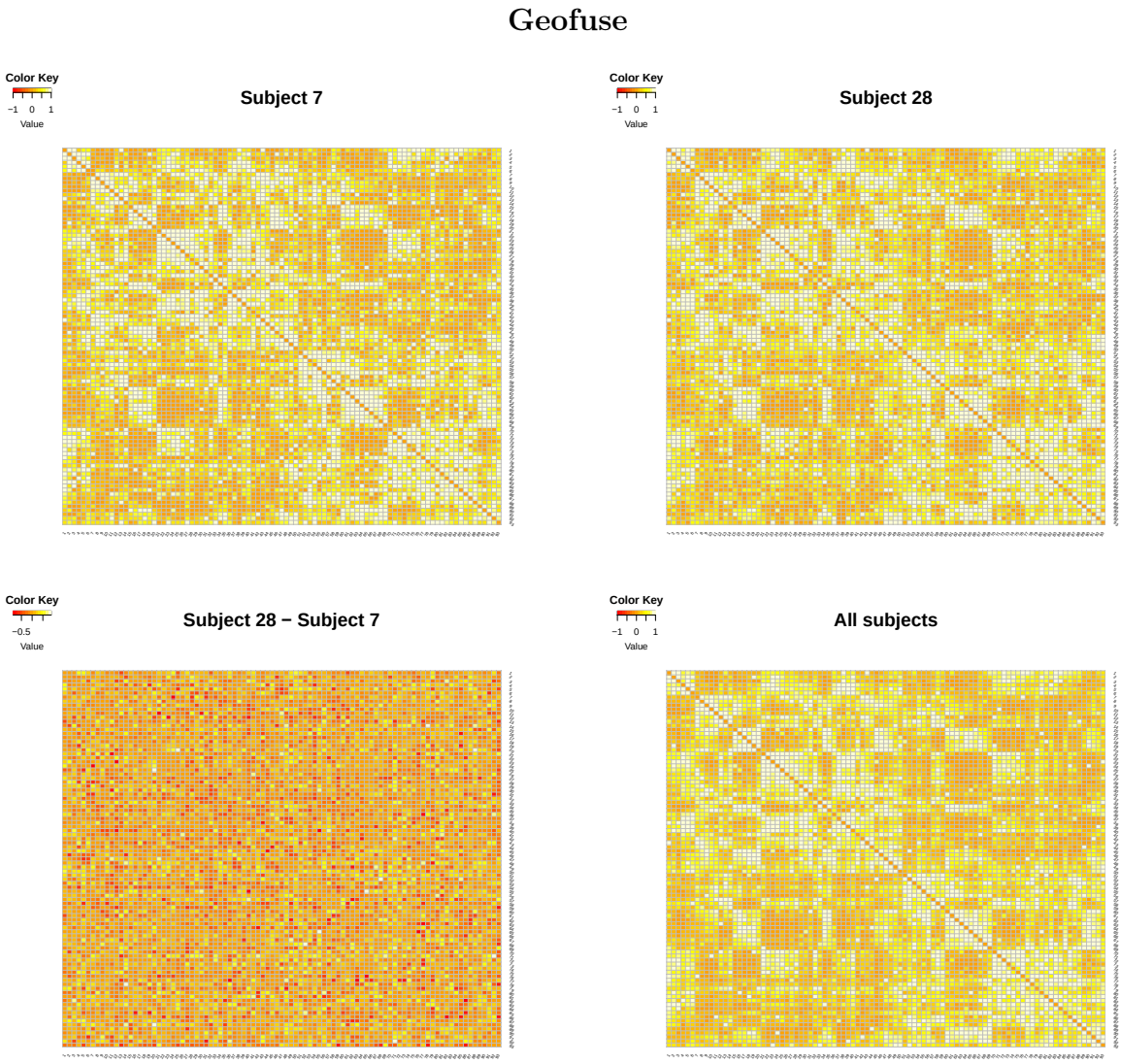


Figure 3.3: Average within-person brain graphs for two subjects under Geofuse. Their differences and the within-person brain graphs averaged over all subjects. The graphs of the two subjects indicate some none structural differences between the two subjects while being quite similar to the group average.

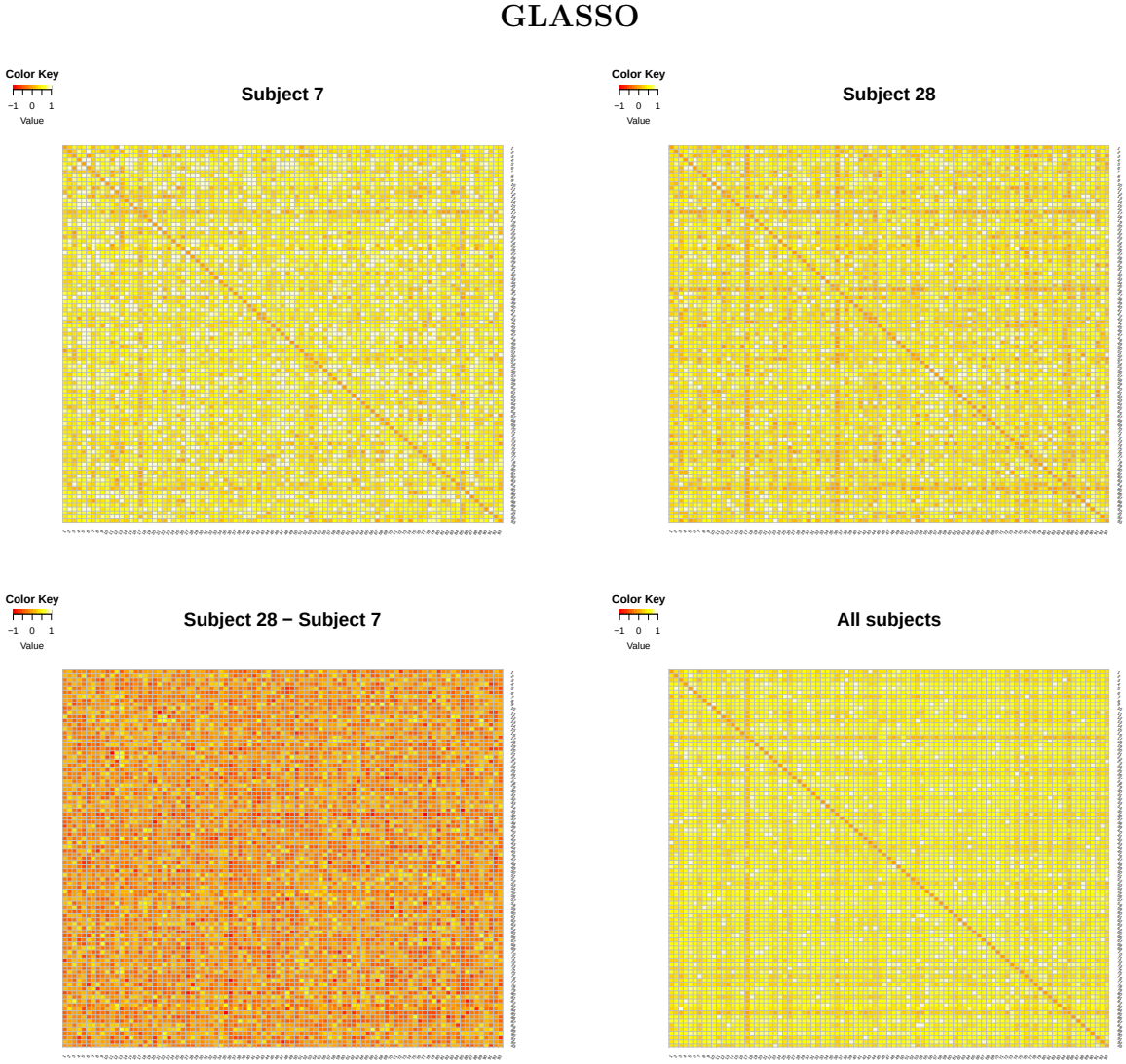


Figure 3.5: Average within-person brain graphs for two subjects under GLASSO. Their differences and the within-person brain graphs averaged over all subjects.

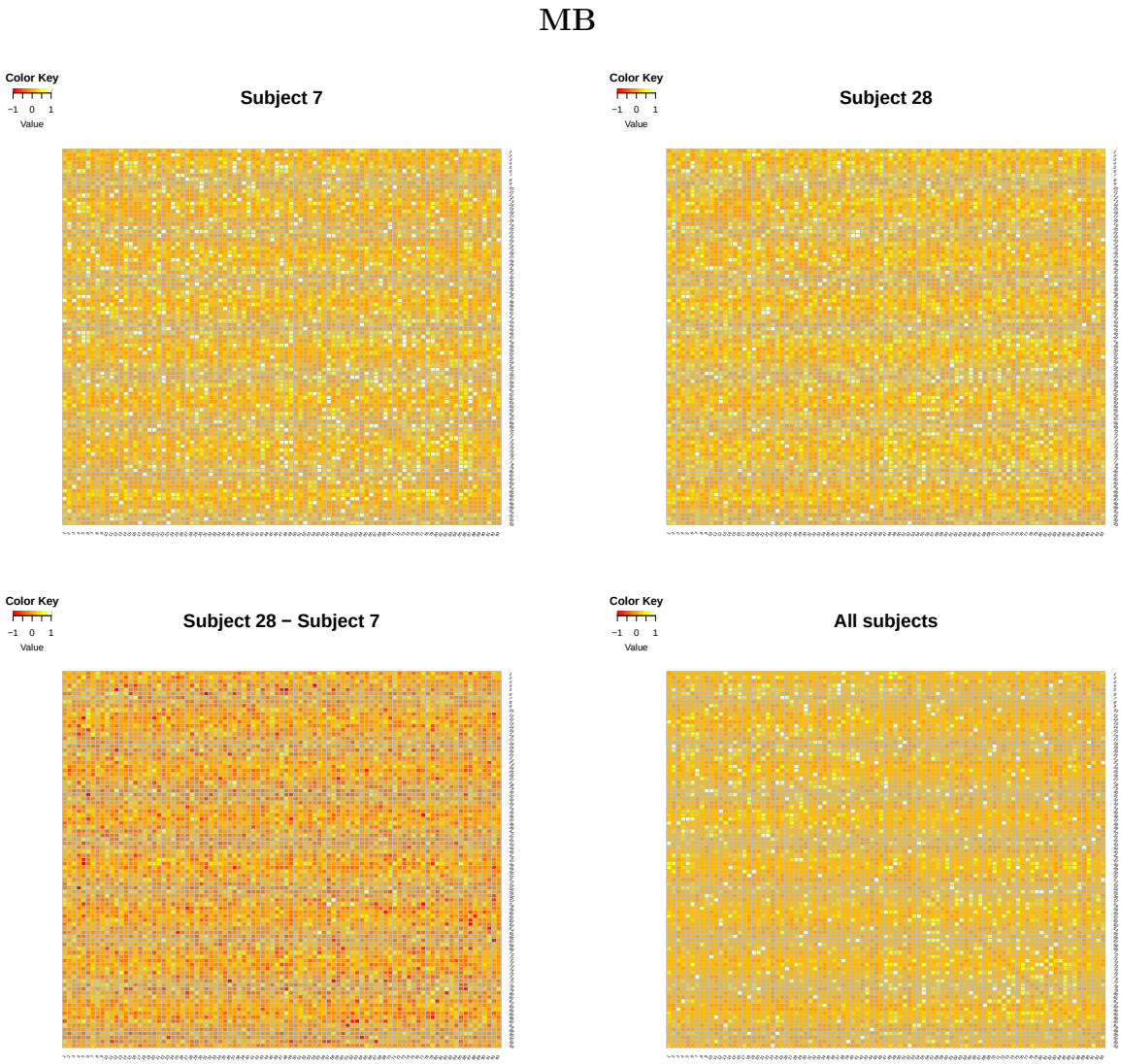


Figure 3.7: Average within-person brain graphs for two subjects under MB. Their differences and the within-person brain graphs averaged over all subjects.

BIBLIOGRAPHY

[1] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010.

[2] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res.*, 9(Mar):485–516, 2008.

[3] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[4] Eugene Belilovsky, Gaël Varoquaux, and Matthew B Blaschko. Testing for differences in gaussian graphical models: applications to brain connectivity. In *Advances in Neural Information Processing Systems*, pages 595–603, 2016.

[5] Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 36(2):192–225, 1974.

[6] Yunqi Bu and Johannes Lederer. Integrating additional knowledge into estimation of graphical models. *arXiv preprint arXiv:1704.02739*, 2017.

[7] Yunqi Bu and Xiao-Hua Zhou. Statistical evaluation of drug efficacy for bridging study in companion diagnostic test trials. *J. Biopharm. Statist*, 26(6):1118–1124, 2016.

[8] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[9] Bing Chen, Ting Xu, Changle Zhou, Luoyu Wang, Ning Yang, Ze Wang, Hao-Ming Dong, Zhi Yang, Yu-Feng Zang, Xi-Nian Zuo, and Xu-Chu Weng. Individual variability

and test-retest reliability revealed by ten repeated resting-state brain scans over one month. *PLoS One*, 10(12):e0144963, 2015.

[10] Michaël Chichignoud, Johannes Lederer, and Martin J Wainwright. A practical scheme and fast algorithm to tune the lasso with optimality guarantees. *J. Mach. Learn. Res.*, 17(1):8162–8181, 2016.

[11] Julien Chiquet, Yves Grandvalet, and Christophe Ambroise. Inferring multiple graphical structures. *Statist. Comput.*, 21(4):537–553, 2011.

[12] Giles L Colclough, Mark W Woolrich, Samuel J Harrison, Pedro A Rojas-López, Pedro A Valdes-Sosa, and Stephen M Smith. Multi-subject hierarchical inverse covariance modelling improves estimation of functional brain networks. *NeuroImage*, 178:370–384, 2018.

[13] Jessica S Damoiseaux, Serge ARB Rombouts, Frederik Barkhof, Philip Scheltens, Cornelis J Stam, Stephen M Smith, and Christian F Beckmann. Consistent resting-state networks across healthy subjects. *Proc. Nat. Acad. Sci. India Sect. A*, 103(37):13848–13853, 2006.

[14] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 76(2):373–397, 2014.

[15] Vladimir Filkov. Identifying gene regulatory networks from gene expression data. 2005.

[16] Food and Drug Administration. *Guidance for Industry and Food and Drug Administration Staff - In Vitro Companion Diagnostic Devices*, Aug 6 2014.

[17] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1, 2010.

[18] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[19] Michele J Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *Am. J. Epidemiol.*, 173(7):761–767, 2011.

[20] Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2014.

[21] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, Stephen M. Smith, and David C Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171, 2016.

[22] Rebecca L Gould, Barbaraa Arroyo, Richard G Brown, Adrian M Owen, Ed T Bullmore, and Russell J Howard. Brain mechanisms of successful compensation during learning in alzheimer disease. *Neurology*, 67(6):1011–1017, 2006.

[23] Cheryl L Grady, Maura L Furey, Pietro Pietrini, Barry Horwitz, and Stanley I Rapoport. Altered brain functional connectivity and impaired short-term memory in alzheimer's disease. *Brain*, 124(4):739–756, 2001.

[24] Clive WJ Granger and Michael J Morris. Time series modelling and interpretation. *J. Roy. Statist. Soc. Ser. A*, 139(2):246–257, 1976.

[25] Geoffrey R Grimmett. A theorem about random fields. *Bull. Lond. Math. Soc.*, 5(1):81–84, 1973.

[26] Quanquan Gu, Yuan Cao, Yang Ning, and Han Liu. Local and global inference for high dimensional gaussian copula graphical models. *arXiv preprint arXiv:1502.02347*, 2015.

[27] Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

[28] Ben B Hansen. Full matching in an observational study of coaching for the sat. *J. Amer. Statist. Assoc.*, 99(467):609–618, 2004.

[29] Ben B Hansen and Stephanie O Klopfer. Optimal full matching and related designs via network flows. *J. Comput. Graph. Statist.*, 15(3):609–627, 2006.

[30] Satoshi Hara and Takashi Washio. Common substructure learning of multiple graphical gaussian models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 1–16. Springer, 2011.

[31] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

[32] Larry D Haugh. Checking the independence of two covariance-stationary time series: a univariate residual cross-correlation approach. *J. Amer. Statist. Assoc.*, 71(354):378–385, 1976.

[33] Jean Honorio and Dimitris Samaras. Multi-task learning of gaussian graphical models. In *ICML*, pages 447–454. Citeseer, 2010.

[34] Barry Horwitz, Cheryl L Grady, Nicholas L Schlageter, Ranjan Duara, and Stanley I Rapoport. Intercorrelations of regional cerebral glucose metabolic rates in alzheimer's disease. *Brain Res.*, 407(2):294–306, 1987.

[35] Shuai Huang, Jing Li, Liang Sun, Jieping Ye, Adam Fleisher, Teresa Wu, Kewei Chen, Eric Reiman, and Alzheimer's Disease NeuroImaging Initiative. Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50(3):935–949, 2010.

[36] Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, MA, 2004.

[37] Jana Jankova and Sara Van De Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Stat.*, 9(1):1205–1229, 2015.

[38] Tobias Kaufmann, Dag Alnæs, Nhat Trung Doan, Christine Lycke Brandt, Ole A Andreassen, and Lars T Westlye. Delayed stabilization and individualization in connectome development are related to psychiatric disorders. *Nat. Neurosci.*, 20(4):513, 2017.

[39] Tobias Kaufmann, Dennis van der Meer, Nhat Trung Doan, Emanuel Schwarz, Martina J Lund, Ingrid Agartz, Dag Alnæs, Deanna M Barch, Ramona Baur-Streubel, Alessandro Bertolino, et al. Genetics of brain age suggest an overlap with common brain disorders. *bioRxiv*, page 303164, 2018.

[40] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[41] Johannes Lederer. Graphical models for discrete and continuous data. *arXiv preprint arXiv:1609.05551*, 2016.

[42] Wonyul Lee and Yufeng Liu. Joint estimation of multiple precision matrices with common structures. *J. Mach. Learn. Res.*, 16(1):1035–1062, 2015.

[43] Meijuan Li. Statistical consideration and challenges in bridging study of personalized medicine. *J. Biopharm. Statist.*, 25(3):397–407, 2015.

[44] Chinghway Lim and Bin Yu. Estimation stability with cross-validation (escv). *J. Comput. Graph. Statist.*, 25(2):464–492, 2016.

[45] Zhixiang Lin, Tao Wang, Can Yang, and Hongyu Zhao. On joint estimation of gaussian graphical models for spatial and temporal data. *Biometrics*, 73(3):769–779, 2017.

[46] Joshua Lukemire, Suprateek Kundu, Giuseppe Pagnoni, and Ying Guo. Bayesian joint modeling of multiple brain functional networks. *arXiv preprint arXiv:1708.02123*, 2017.

[47] Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat. Med.*, 23(19):2937–2960, 2004.

[48] David P MacKinnon, Chondra M Lockwood, Jeanne M Hoffman, Stephen G West, and Virgil Sheets. A comparison of methods to test mediation and other intervening variable effects. *Psychol. Methods*, 7(1):83, 2002.

[49] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

[50] Karthik Mohan, Mike Chung, Seungyeop Han, Daniela Witten, Su-In Lee, and Maryam Fazel. Structured learning of gaussian graphical models. In *Advances in neural information processing systems*, pages 620–628, 2012.

[51] Karthik Mohan, Palma London, Maryam Fazel, Daniela Witten, and Su-In Lee. Node-based learning of multiple gaussian graphical models. *J. Mach. Learn. Res.*, 15(1):445–488, 2014.

[52] Trevor Park and George Casella. The bayesian lasso. *J. Amer. Statist. Assoc.*, 103(482):681–686, 2008.

[53] Christine Peterson, Francesco C Stingo, and Marina Vannucci. Bayesian inference of multiple gaussian graphical models. *J. Amer. Statist. Assoc.*, 110(509):159–174, 2015.

[54] Bradley S Price, Charles J Geyer, and Adam J Rothman. Ridge fusion in statistical learning. *J. Comput. Graph. Statist.*, 24(2):439–454, 2015.

[55] Paul R Rosenbaum. A characterization of optimal designs for observational studies. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 53(3):597–610, 1991.

[56] Jeremy A Sabourin, William Valdar, and Andrew B Nobel. A permutation approach for

selecting the penalty parameter in penalized model selection. *Biometrics*, 71(4):1185–1194, 2015.

[57] Xilin Shen, Fuyuze Tokoglu, Xenios Papademetris, and R. Todd Constable. Groupwise whole-brain parcellation from resting-state fmri data for network node identification. *NeuroImage*, 82:403–415, 2013.

[58] Michael E Sobel. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol. Methodol.*, 13:290–312, 1982.

[59] Kaustubh Supekar, Vinod Menon, Daniel Rubin, Mark Musen, and Michael D Greicius. Network analysis of intrinsic functional brain connectivity in alzheimer's disease. *PLoS Comput. Biol.*, 4(6):e1000100, 2008.

[60] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.

[61] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage*, 15(1):273–289, 2002.

[62] Sara A Van De Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electron. J. Stat.*, 3:1360–1392, 2009.

[63] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *Eur. Neuropsychopharmacol.*, 20(8):519–534, 2010.

[64] Gaël Varoquaux, Alexandre Gramfort, Jean-Baptiste Poline, and Bertrand Thirion. Brain covariance selection: better individual functional connectivity models using population prior. In *Advances in neural information processing systems*, pages 2334–2342, 2010.

[65] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1 -constrained quadratic programming (lasso). *IEEE Trans. Inform. Theory*, 55(5):2183–2202, 2009.

[66] Kun Wang, Meng Liang, Liang Wang, Lixia Tian, Xinqing Zhang, Kuncheng Li, and Tianzi Jiang. Altered functional connectivity in early alzheimer's disease: A resting-state fmri study. *Hum. Brain Mapp.*, 28(10):967–978, 2007.

[67] Sen Yang, Zhaosong Lu, Xiaotong Shen, Peter Wonka, and Jieping Ye. Fused multiple graphical lasso. *SIAM J. Optim.*, 25(2):916–943, 2015.

[68] Bin Yu. Stability. *Bernoulli*, 19(4):1484–1500, 2013.

[69] Donghyeon Yu, Sang Han Lee, Johan Lim, Guanghua Xiao, Richard C Craddock, and Bharat B Biswal. Fused lasso regression for identifying differential correlations in brain connectome graphs. *Stat. Anal. Data Min.*, 11(5):203–226, 2018.

[70] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[71] Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res*, 7(Nov):2541–2563, 2006.

[72] Sihai Dave Zhao, T Tony Cai, and Hongzhe Li. Direct estimation of differential networks. *Biometrika*, 101(2):253–268, 2014.

[73] Hui Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.

[74] Xi-Nian Zuo, Jeffrey S Anderson, Pierre Bellec, Rasmus M Birn, Bharat B Biswal, Janusch Blautzik, John CS Breitner, Randy L Buckner, Vince D Calhoun, F. Xavier Castellanos, et al. An open science resource for establishing reliability and reproducibility in functional connectomics. *Sci. Data*, 1:140049, 2014.

Appendix A

R CODE FOR GRAPH ESTIMATION WITH SI

See also https://github.com/LedererLab.

```
library(glmnet) # Required for cv.glmnet.

LambdaGrid <- function( lambdaMax, numLambda, factor ) {
  # Construct an adaptive tuning parameter grid ranging between factor*lambdaMax
  #  and factor*(lambdaMax/numLambda) with grid increments of 1 / i.
  #
  # Args:
  #   lambdaMax: Largest tuning parameter.
  #   numLambda: Number of elements that should be in the grid.
  #   factor: Scaling factor that will be applied to to all elements in the grid.
  #
  # Returns:
  #   Vector of tuning parameters of size numLambda.
  #
  # Raises:
  #   Error if numLambda is zero or negative.

  if( numLambda <= 0 ) {
    stop( 'Number of grid elements must be greater than 0.' )
  }

  lambdas <- vector( mode="list", length = numLambda )

  lambdas <- lapply( seq_along(lambdas),
                     function( y, n, i ) { factor*lambdaMax/i },
                     y = lambdas,
                     n = names(lambdas) )

  return( unlist(lambdas) )

}

```

```
BetaCol ← function( X, D, i ) {
  # Populate a column of the adjacency matrix ( Beta ) when constructing a
  # connectivity network.
  #
  # Args:
  #   X: An n x p design matrix.
  #   D: A p x p matrix containing pairwise distances between regions
  #   i: Index of the column being updated.
  #
  # Returns:
  #   Vector corresponding to the i'th column of the adjacency matrix.
  #
  # Raises:
  #   Error if requested index i is larger than number of columns of X.

  n ← dim(X)[1]
  p ← dim(X)[2]

  if( i > n ) {
    stop( "Requested index ", i,
          " is larger than numbers of columns in design matrix ", p )
  }

  y ← X[,i]
  predictors ← scale( X[,-i] )

  # Get distances of the i'th region to other regions.
  Di ← D[-i,i]

  factor ← sum( Di^3 )/( p - 1 )
  lambdaMax ← max( abs( crossprod( predictors, y ) ) ) / n

  lambdas ← LambdaGrid( lambdaMax, 1000, factor )

  # Calibrate the tuning parameter and select
  # estimator accordingly.
  model ← cv.glmnet( x = predictors,
                     y = y,
                     lambda = lambdas,
                     intercept = TRUE,
                     penalty.factor = Di^3 )
```

```
74
75    beta ← array( data = 0, dim = p )
76
77    # Construct a vector using elements from the model, except for the
          i'th element
78    # which is set to zero.
79    # The model generated by cv.glmnet has the intercept as the first
          term, which
80    # needs to be removed before populating beta.
81    beta[-i] ← coef( model )[-1,]
82
83    return( beta )
84
85  }
86
87  GraphEstimation ← function( X, D, reproducible = TRUE ) {
88    # Estimate a connectivity network ( adjacency matrix ) using the SI
          method.
89    #
90    # Args:
91    #   X: An n x p design matrix.
92    #   D: A p x p matrix containing pairwise distances between regions
          .
93    #   reproducible : Set to TRUE to make results reproducible between
          runs.
94    #
95    # Returns:
96    #   Adjacency matrix of dimension p x p.
97    #
98    # Raises:
99    #   Error if either X or D contain any non-numeric entries.
100   #   Error if matrix dimensions are incorrect.
101   #   Error if matrices are 1 x 1 or smaller.
102   #   Error if either X or D contain missing ( NA ) values.
103
104   if( !is.numeric(X) ) {
105     stop( "The design matrix must contain all numeric values." )
106   }
107
108   if( !is.numeric(D) ) {
109     stop( "The distance matrix must contain all numeric values." )
110   }
111
112   n ← dim(X)[1] # number of samples
```

```
113   p ← dim(X)[2] # number of regions
114
115   if( ( dim(D)[1] != dim(D)[2] ) || ( dim(D)[1] != p ) ) {
116     stop( "Matrix dimensions are incorrect. Design matrix is ",
117     dim(X)[1], "x", dim(X)[2] , " and Distance matrix is ",
118     dim(D)[1], "x", dim(D)[2] )
119   }
120
121   if( ( n ≤ 1 ) || ( p ≤ 1 ) ) {
122     stop( "All matrix dimensions must be greater than 1 x 1" )
123   }
124
125   missing_X ← sum( is.na(X) )
126   missing_D ← sum( is.na(D) )
127
128   if( missing_X > 0 ) {
129     stop( "Design matrix contained ", missing_X, " missing values." )
130   }
131
132   if( missing_D > 0 ) {
133     stop( "Distance matrix contained ", missing_D, " missing values."
          )
134   }
135
136   if( reproducible ) {
137     set.seed(42)
138   }
139
140   betaMatrix ← matrix( data = NA, nrow = p, ncol = p )
141
142   for ( i in 1 : p ) {
143     betaMatrix[,i] ← BetaCol( X, D, i )
144   }
145
146   # Compute matrix of non-zero entries in the coefficient matrix.
147   adjacencyMatrix ← sign( abs( betaMatrix ) )
148
149   # Apply the 'and' rule.
150   return( adjacencyMatrix * t( adjacencyMatrix ) )
151
152 }
```

Appendix B

TECHNICAL PROOF

We now prove Theorem 1.3.1. For this, we first recall some notation: β_*^j is the jth regression target vector, X_*^{-j} corresponds to X^{-j} on the support of β_*^j denoted by $S_*^j := \text{supp}(\beta_*^j)$ and X_{-*}^{-j} to the remaining parts of X^{-j}, and ε^j is the noise vector of the jth regression. For ease of notation, we also denote the size of the jth support by $s_*^j := |S_*^j|$.

We now establish the primal-dual construction.

lemma B.1 (Primal-Dual Witness Construction)**.** *Assume that $X_*^{-j\top} X_*^{-j}$ is invertible for each $j \in \{1, \ldots, p\}$. Let $\mathbf{r}_{-j}^j = (\mathbf{r}_*^{j\top}, \mathbf{r}_{-*}^{j\top})^\top \in \mathbb{R}^{p-1}$ be a fixed, positive vector for any $j \in \{1, \ldots, p\}$. Here, $\mathbf{r}_*^j$ corresponds to the sub-vector of $\mathbf{r}_{-j}^j$ on the support of β_*^j and $\mathbf{r}_{-*}^j$ corresponds to the rest of $\mathbf{r}_{-j}^j$. Define vectors $\hat{\alpha}^j = (\hat{\alpha}_*^{j\top}, \hat{\alpha}_{-*}^{j\top})^\top \in \mathbb{R}^{p-1}$ (primal vector) and $\hat{z}^j = (\hat{z}_*^{j\top}, \hat{z}_{-*}^{j\top})^\top \in \mathbb{R}^{p-1}$ (dual vector) as follows:*
Primal construction: define $\hat{\alpha}_*^j \in \boldsymbol{R}^{s_*^j}$ such that

$$\hat{\alpha}_*^j \in \underset{\beta \in \boldsymbol{R}^{s_*^j}}{\text{argmin}} \left\{ \|X^j - X_*^{-j}\beta\|_2^2 + \|\mathbf{r}_*^j \circ \beta\|_1 \right\},$$

where the circle $\circ$ indicates element-wise multiplication, and set $\hat{\alpha}_{-*}^j := \mathbf{0}_{p-1-s_*^j}$.
Dual construction: define $\hat{z}^j$ such that

$$-2X^{-j\top}\left(X^j - X^{-j}\hat{\alpha}^j\right) + \hat{z}^j = \mathbf{0}_{p-1}.$$

Now, the following holds:

$$(i)\ \hat{z}_*^j \in \partial \|\mathbf{r}_*^j \circ \hat{\alpha}_*^j\|_1,$$

and

$$(ii)\ \|\hat{z}_{-*}^j\|_d < 1 \quad \Rightarrow \quad \hat{\alpha}^j \in \mathbb{R}^{p-1} \text{ is the unique lasso solution.}$$

$\|\hat{z}^j_{-*}\|_d$ above is the dual norm of the weighted norm $\|\mathbf{r}^j_{-*} \circ \hat{\alpha}^j_{-*}\|_1$. In general, the dual norm for a weighted norm $\|w \circ \beta\|_1$ is denoted by $\|\beta\|_d := \|\frac{1}{w} \circ \beta\|_\infty$.

Proof of lemma B.1. We can summarize the properties of the sub-vectors in the following four relations.

1. The vector $\hat{\alpha}^j_* \in \boldsymbol{R}^{s^j_*}$ satisfies

$$\hat{\alpha}^j_* \in \operatorname*{argmin}_{\beta \in \boldsymbol{R}^{s^j_*}} \left\{ \|X^j - X^{-j}_* \beta\|_2^2 + \|\mathbf{r}^j_* \circ \beta\|_1 \right\};$$

2. It holds that $\hat{\alpha}^j_{-*} := \mathbf{0}_{p-1-s^j_*}$;

3. The vector $\hat{z}^j_* \in \boldsymbol{R}^{s^j_*}$ satisfies

$$-2X^{-j\top}_* (X^j - X^{-j}_* \hat{\alpha}^j_*) + \hat{z}^j_* = \mathbf{0}_{s^j_*};$$

4. The vector $\hat{z}^j \in \boldsymbol{R}^{p-1}$ satisfies

$$-2X^{-j\top}(X^j - X^{-j}\hat{\alpha}^j) + \begin{pmatrix} \hat{z}^j_* \\ \hat{z}^j_{-*} \end{pmatrix} = \mathbf{0}_{p-1}.$$

Part 1: duality on S^j_* We first show result (i), that is,

$$\hat{z}^j_* \in \partial \|\mathbf{r}^j_* \circ \hat{\alpha}^j_*\|_1.$$

By 1. and the KKT conditions for the oracle lasso (with design matrix X^{-j}_*), there is a $\kappa \in \partial \|\mathbf{r}^j_* \circ \hat{\alpha}^j_*\|_1 \subset \boldsymbol{R}^{s^j_*}$ such that

$$-2X^{-j\top}_* (X^j - X^{-j}_* \hat{\alpha}^j_*) + \kappa = \mathbf{0}_{s^j_*}.$$

By 3., we conclude that $\hat{z}^j_* = \kappa$, so that $\hat{z}^j_* \in \partial\|\mathbf{r}^j_* \circ \hat{\alpha}^j_*\|_1$ as desired.

Part 2: unique lasso solution We now show that

$$(ii)\ \|\hat{z}^j_{-*}\|_d < 1 \quad \Rightarrow \quad \hat{\alpha}^j \in \mathbb{R}^{p-1} \text{ is the unique lasso solution.}$$

We proceed in four steps.

Step 1: lasso solution. We first show that if $\|\hat{z}^j_{-*}\|_d \leq 1$, the pair $(\hat{\alpha}^j, \hat{z}^j)$ constructed above is a primal-dual solution pair of the KKT conditions for the lasso, that is, $\hat{z}^j \in \partial\|\mathbf{r}^j_{-j} \circ \hat{\alpha}^j\|_1$ and

$$-2X^{-j\top}(X^j - X^{-j}\hat{\alpha}^j) + \hat{z}^j = \mathbf{0}_{p-1}\,.$$

According to Part 1, $\hat{z}^j_* \in \partial\|\mathbf{r}^j_* \circ \hat{\alpha}^j_*\|_1$. In view of the assumption $\|\hat{z}^j_{-*}\|_d < 1 \leq 1$ and 2., it then holds that $\hat{z}^j \in \partial\|\mathbf{r}^j_{-j} \circ \hat{\alpha}^j\|_1$. Moreover, by 4.,

$$-2X^{-j\top}(X^j - X^{-j}\hat{\alpha}^j) + \hat{z}^j = \mathbf{0}_{p-1}\,.$$

This implies that indeed $(\hat{\alpha}^j, \hat{z}^j)$ is a primal-dual solution pair of the KKT conditions of the lasso.

Step 2: support size. We now show that if $\|\hat{z}^j_{-*}\|_d < 1$, it holds that $\mathrm{supp}(\hat{\beta}^j) \subset S^j_*$ for all $\hat{\beta}^j \in \mathbb{R}^{p-1}$ that satisfy

$$\hat{\beta}^j \in \operatorname*{argmin}_{\beta \in \mathbb{R}^{p-1}} \left\{\|X^j - X^{-j}\beta\|_2^2 + \|\mathbf{r}^j_{-j} \circ \beta\|_1\right\}.$$

For this, assume that $\hat{\beta}^j$ is a lasso solution. Then, $\hat{\beta}^j$ minimizes the above objective function. However, by Step 1, $(\hat{\alpha}^j, \hat{z}^j)$ is a feasible primal-dual pair for the X^{-j}-lasso, and thus, also $\hat{\alpha}^j$ minimizes the above objective function. Hence,

$$\|X^j - X^{-j}\hat{\beta}^j\|_2^2 + \|\mathbf{r}^j_{-j} \circ \hat{\beta}^j\|_1 = \|X^j - X^{-j}\hat{\alpha}^j\|_2^2 + \|\mathbf{r}^j_{-j} \circ \hat{\alpha}^j\|_1\,.$$

Since, by Step 1, $\hat{z}^j \in \partial\|\mathbf{r}^j_{-j} \circ \hat{\alpha}^j\|_1$, it holds that $\|\mathbf{r}^j_{-j} \circ \hat{\alpha}^j\|_1 = \langle \hat{z}^j, \hat{\alpha}^j \rangle$. Indeed, for all $i \in \{1, \ldots, p\}$,

$$\hat{z}^j_i \hat{\alpha}^j_i = \begin{cases} \mathrm{sign}(\hat{\alpha}^j_i)(\mathbf{r}^j_{-j})_i \hat{\alpha}^j_i = (\mathbf{r}^j_{-j})_i |\hat{\alpha}^j_i| & \text{if } i \in \mathrm{supp}(\hat{\alpha}^j) \\ \hat{z}^j_i \cdot 0 = 0 = (\mathbf{r}^j_{-j})_i |\hat{\alpha}^j_i| & \text{if } i \notin \mathrm{supp}(\hat{\alpha}^j) \end{cases}.$$

Plugging this into the previous display yields

$$\|X^j - X^{-j}\hat{\beta}^j\|_2^2 + \|\mathbf{r}^j_{-j} \circ \hat{\beta}^j\|_1 \;=\; \|X^j - X^{-j}\hat{\alpha}^j\|_2^2 + \langle \hat{z}^j, \hat{\alpha}^j \rangle \,.$$

We can now subtract $\langle \hat{z}^j, \hat{\beta}^j \rangle$ on both sides to obtain

$$\|X^j - X^{-j}\hat{\beta}^j\|_2^2 + \|\mathbf{r}^j_{-j} \circ \hat{\beta}^j\|_1 - \langle \hat{z}^j, \hat{\beta}^j \rangle \;=\; \|X^j - X^{-j}\hat{\alpha}^j\|_2^2 + \langle \hat{z}^j, (\hat{\alpha}^j - \hat{\beta}^j) \rangle \,.$$

By 4. above, it holds that $\hat{z}^j = 2X^{-j\top}(X^j - X^{-j}\hat{\alpha}^j)$, so that we can further deduce

$$\|X^j - X^{-j}\hat{\beta}^j\|_2^2 + \|\mathbf{r}^j_{-j} \circ \hat{\beta}^j\|_1 - \langle \hat{z}^j, \hat{\beta}^j \rangle \;=\; \|X^j - X^{-j}\hat{\alpha}^j\|_2^2 + \langle 2X^{-j\top}(X^j - X^{-j}\hat{\alpha}^j), \hat{\alpha}^j - \hat{\beta}^j \rangle \,.$$

Rearranging this yields

$$\|\mathbf{r}^j_{-j} \circ \hat{\beta}^j\|_1 - \langle \hat{z}^j, \hat{\beta}^j \rangle \;=\; \|X^j - X^{-j}\hat{\alpha}^j\|_2^2 + \langle -2X^{-j\top}(X^j - X^{-j}\hat{\alpha}^j), \hat{\beta}^j - \hat{\alpha}^j \rangle - \|X^j - X^{-j}\hat{\beta}^j\|_2^2 \,.$$

Setting $f : \beta \to \|X^j - X^{-j}\beta\|_2^2$ makes the right-hand side equal to

$$f(\hat{\alpha}^j) + \langle f'(\hat{\alpha}^j), \hat{\beta}^j - \hat{\alpha}^j \rangle - f(\hat{\beta}^j) \,.$$

However, since f is convex, it holds that

$$f(\hat{\beta}^j) \;\geq\; f(\hat{\alpha}^j) + \langle f'(\hat{\alpha}^j), \hat{\beta}^j - \hat{\alpha}^j \rangle \,.$$

Collecting terms yields

$$\|\mathbf{r}^j_{-j} \circ \hat{\beta}^j\|_1 - \langle \hat{z}^j, \hat{\beta}^j \rangle \;\leq\; 0 \,.$$

Hence,

$$\|\mathbf{r}^j_{-j} \circ \hat{\beta}^j\|_1 \;\leq\; \langle \hat{z}^j, \hat{\beta}^j \rangle \,.$$

Recall that by Step 1, $\hat{z}^j \in \partial \|\mathbf{r}^j_{-j} \circ \hat{\alpha}^j\|_1$, so that $\|\hat{z}^j\|_d \leq 1$. Hence,

$$\langle \hat{z}^j, \hat{\beta}^j \rangle \;=\; \langle \hat{z}^j_*, \hat{\beta}^j_* \rangle + \langle \hat{z}^j_{-*}, \hat{\beta}^j_{-*} \rangle \;\leq\; \|\mathbf{r}^j_* \circ \hat{\beta}^j_*\|_1 + \langle \hat{z}^j_{-*}, \hat{\beta}^j_{-*} \rangle \,,$$

so that

$$\|\mathbf{r}^j_{-*} \circ \hat{\beta}^j_{-*}\|_1 \;\leq\; \langle \hat{z}^j_{-*}, \hat{\beta}^j_{-*} \rangle \,.$$

Now, in view of the strict dual feasibility condition $\|\hat{z}^j_{-*}\|_d < 1$, we have

$$(\hat{z}^j_{-*})_i \cdot (\hat{\beta}^j_{-*})_i \begin{cases} < (\mathbf{r}^j_{-*})_i |(\hat{\beta}^j_{-*})_i| & \text{if } i \in \operatorname{supp}(\hat{\beta}^j_{-*}) \\ = 0 = |(\hat{\beta}^j_{-*})_i| & \text{if } i \notin \operatorname{supp}(\hat{\beta}^j_{-*}) \,. \end{cases}$$

Together with the above display, this means that $(\hat{\beta}^j)_i = 0$ for all $i \notin \operatorname{supp}(\hat{\alpha}^j)$. This concludes the proof of Step 2.

Step 3: uniqueness. We now show that $\hat{\beta}^j = \hat{\alpha}^j$ for all $\hat{\beta}^j \in \mathbb{R}^{p-1}$ that satisfy

$$\hat{\beta}^j \in \operatorname*{argmin}_{\beta \in \mathbb{R}^{p-1}} \{\|X^j - X^{-j}\beta\|_2^2 + \|\mathbf{r}^j_{-j} \circ \beta\|_1\} \,.$$

From Step 2, we deduce that $\hat{\beta}^j = (\hat{\beta}^{j\top}_*, \mathbf{0})^\top$, so that

$$\hat{\beta}^j_* \in \operatorname*{argmin}_{\beta \in \boldsymbol{R}^{s^j_*}} \{\|X^j - X^{-j}_* \beta\|_2^2 + \|\mathbf{r}^j_* \circ \beta\|_1\} \,.$$

However, since the minimal eigenvalue of $\frac{(X^{-j}_*)^\top X^{-j}_*}{s^j_*}$ is larger than zero according to the assumptions of lemma B.1, this problem has a unique solution. □

With lemma B.1 proven, we can now establish a result on support recovery and ℓ_∞-loss. To this end, we assume an irrepresentability condition, that is, we assume that $X^{-j\top}_* X^{-j}_*$ is invertible and for a constant $b_j > 0$, it holds that

$$\|\mathbf{r}^j_* \circ (X^{-j\top}_* X^{-j}_*)^{-1} X^{-j\top}_* X^i\|_1 / (\mathbf{r}^j)_i \leq 1 - b_j$$

for all $i \in \{k \in \{1, \ldots, p\} \mid k \neq j \text{ and } (k, j) \notin \mathcal{E}\}$.

lemma B.2 (Support Recovery Under Irrepresentability Condition)**.** *Assume that the Irrepresentability Condition is satisfied with constant $b_j > 0$, $j \in \{1, \ldots, p\}$. Moreover, assume that*

$$1 > 2\|X^{-j\top}_{-*}(\mathrm{I}_n - X^{-j}_*(X^{-j\top}_* X^{-j}_*)^{-1} X^{-j\top}_*)\varepsilon^j\|_d / b_j \,,$$

and

$$1 > n\|(X^{-j\top}_* X^{-j}_*)^{-1} X^{-j\top}_* \varepsilon^j\|_d \,.$$

Then,

$$(i) \quad \mathrm{supp}(\hat{\beta}^j) \subset S_*^j\,,$$

and

$$(ii) \quad \|(\beta_*^j)_* - \hat{\beta}_*^j\|_d \leq 1/n + \|(X_*^{-j\top} X_*^{-j}/n)^{-1}\hat{z}_*^j\|_d/(2n)\,.$$

Proof of lemma B.2. Step 1: Under the assumptions stated in the lemma, it holds for any primal-dual pair $(\hat{\alpha}^j, \hat{z}^j) \in \mathbb{R}^{p-1} \times \mathbb{R}^{p-1}$ constructed as in lemma B.1 that

$$\|\hat{z}_{-*}^j\|_d < 1\,.$$

To show this, we want to rewrite 4. in the construction in lemma B.1, solve for $\hat{z}_{-*}^j$, and then show that this norm is smaller than 1. We first invoke the model to rewrite 4. of the construction in lemma B.1 as

$$-2X^{-j\top}(X^{-j}\beta_*^j + \varepsilon^j - X^{-j}\hat{\alpha}^j) + \begin{pmatrix} \hat{z}_*^j \\ \hat{z}_{-*}^j \end{pmatrix} = \mathbf{0}_{p-1}\,.$$

Rearranging, we find

$$-2X^{-j\top}\left(X^{-j}\begin{pmatrix}(\beta_*^j)_* \\ \mathbf{0}\end{pmatrix} - X^{-j}\begin{pmatrix}\hat{\alpha}_*^j \\ \mathbf{0}\end{pmatrix}\right) - 2X^{-j\top}\varepsilon^j + \begin{pmatrix} \hat{z}_*^j \\ \hat{z}_{-*}^j \end{pmatrix} = \mathbf{0}_{p-1}\,.$$

We can write this in block matrix form according to

$$-2\begin{pmatrix} X_*^{-j\top}X_*^{-j} & X_*^{-j\top}X_{-*}^{-j} \\ X_{-*}^{-j\top}X_*^{-j} & X_{-*}^{-j\top}X_{-*}^{-j} \end{pmatrix}\begin{pmatrix}(\beta_*^j)_* - \hat{\alpha}_*^j \\ \mathbf{0}\end{pmatrix} - 2\begin{pmatrix}X_*^{-j\top}\varepsilon^j \\ X_{-*}^{-j\top}\varepsilon^j\end{pmatrix} + \begin{pmatrix} \hat{z}_*^j \\ \hat{z}_{-*}^j \end{pmatrix} = \mathbf{0}_{p-1}\,. \tag{B.1}$$

We now solve this equation for $\hat{z}_{-*}^j$ via

$$-2X_{-*}^{-j\top}X_*^{-j}((\beta_*^j)_* - \hat{\alpha}_*^j) - 2X_{-*}^{-j\top}\varepsilon^j + \hat{z}_{-*}^j = \mathbf{0}_{p-1-s_*^j}\,,$$

and hence, by rearranging,

$$\hat{z}_{-*}^j = X_{-*}^{-j\top}X_*^{-j}(2((\beta_*^j)_* - \hat{\alpha}_*^j)) + 2X_{-*}^{-j\top}\varepsilon^j\,. \tag{B.2}$$

We now want to solve Equation (B.1) for $2((\beta_*^j)_* - \hat{\alpha}_*^j)$ and plug this in. We find

$$-2X_*^{-j\top}X_*^{-j}((\beta_*^j)_* - \hat{\alpha}_*^j) - 2X_*^{-j\top}\varepsilon^j + \hat{z}_*^j = \mathbf{0}_{s_*^j}\,,$$

and hence, by rearranging,

$$2X_*^{-j\top}X_*^{-j}((\beta_*^j)_* - \hat{\alpha}_*^j) \;=\; -2X_*^{-j\top}\varepsilon^j + \hat{z}_*^j\,.$$

Since the matrix $X_*^{-j\top}X_*^{-j}$ is invertible by the Irrepresentable Condition, we can solve this equation for $2((\beta_*^j)_* - \hat{\alpha}_*^j)$ and find

$$2((\beta_*^j)_* - \hat{\alpha}_*^j) \;=\; -2(X_*^{-j\top}X_*^{-j})^{-1}X_*^{-j\top}\varepsilon^j + (X_*^{-j\top}X_*^{-j})^{-1}\hat{z}_*^j\,. \tag{B.3}$$

Combining this equation with Equation (B.2) yields

$$\begin{aligned}\hat{z}_{-*}^j \;=\; &-2X_{-*}^{-j\top}X_*^{-j}(X_*^{-j\top}X_*^{-j})^{-1}X_*^{-j\top}\varepsilon^j\\ &+X_{-*}^{-j\top}X_*^{-j}(X_*^{-j\top}X_*^{-j})^{-1}\hat{z}_*^j + 2X_{-*}^{-j\top}\varepsilon^j\\ =\;&2X_{-*}^{-j\top}(\mathrm{I}_n - X_*^{-j}(X_*^{-j\top}X_*^{-j})^{-1}X_*^{-j\top})\varepsilon^j + X_{-*}^{-j\top}X_*^{-j}(X_*^{-j\top}X_*^{-j})^{-1}\hat{z}_*^j\,.\end{aligned}$$

The assumption of lemma B.2 yields for the norm of the first part

$$2\|X_{-*}^{-j\top}(\mathrm{I}_n - X_*^{-j}(X_*^{-j\top}X_*^{-j})^{-1}X_*^{-j\top})\varepsilon^j\|_d \;<\; b_j\,.$$

For the norm of the second term, we find

$$\begin{aligned}&\|X_{-*}^{-j\top}X_*^{-j}(X_*^{-j\top}X_*^{-j})^{-1}\hat{z}_*^j\|_d\\ =&\max_{\substack{i\notin S_*^j\\ i\neq j}}|X^{i\top}X_*^{-j}(X_*^{-j\top}X_*^{-j})^{-1}\hat{z}_*^j|/(\mathbf{r}^j)_i && \text{(definition of } \|\cdot\|_d)\\ =&\max_{\substack{i\notin S_*^j\\ i\neq j}}|(X_*^{-j}(X_*^{-j\top}X_*^{-j})^{-1}\hat{z}_*^j)^\top X^i|/(\mathbf{r}^j)_i && \text{(transpose property)}\\ =&\max_{\substack{i\notin S_*^j\\ i\neq j}}|\hat{z}_*^{j\top}(X_*^{-j\top}X_*^{-j})^{-1}X_*^{-j\top}X^i|/(\mathbf{r}^j)_i && \text{(transpose property)}\\ \leq&\max_{\substack{i\notin S_*^j\\ i\neq j}}\|\hat{z}_*^j\|_d\cdot\|\mathbf{r}_*^j\circ(X_*^{-j\top}X_*^{-j})^{-1}X_*^{-j\top}X^i\|_1/(\mathbf{r}^j)_i && \text{(Hölder's inequality)}\\ \leq&\max_{\substack{i\notin S_*^j\\ i\neq j}}\|\mathbf{r}_*^j\circ(X_*^{-j\top}X_*^{-j})^{-1}X_*^{-j\top}X^i\|_1/(\mathbf{r}^j)_i && (\|\hat{z}_*^j\|_d\leq 1)\\ \leq&\,1-b_j\,. && \text{(Irrepresentable Cond.)}\end{aligned}$$

Collecting terms gives

$$\|\hat{z}^j_{-*}\|_d \,<\, b_j + (1 - b_j) \,=\, 1\,,$$

Thus as proven in lemma B.1 Step 2, we now have $\text{supp}(\hat{\beta}^j) \subset S^j_*$.

Step 2: Under the assumptions stated in the lemma, it holds that

$$\|(\beta^j_*)_* - \hat{\beta}^j_*\|_d \,\leq\, 1/n + 1/n\|(X^{-j\top}_* X^{-j}_*/n)^{-1}\hat{z}^j_*\|_d/2\,.$$

To show this, we deduce as in Step 1, see Equation (B.2), that

$$(\beta^j_*)_* - \hat{\beta}^j_* \,=\, -(X^{-j\top}_* X^{-j}_*)^{-1}X^{-j\top}_*\varepsilon^j + (X^{-j\top}_* X^{-j}_*)^{-1}\hat{z}^j_*/2\,.$$

Taking dual-norms on both sides yields

$$\begin{aligned}
&\|(\beta^j_*)_* - \hat{\beta}^j_*\|_d \\
=&\,\| - (X^{-j\top}_* X^{-j}_*)^{-1}X^{-j\top}_*\varepsilon^j + (X^{-j\top}_* X^{-j}_*)^{-1}\hat{z}^j_*/2\|_d \\
\leq&\,\|(X^{-j\top}_* X^{-j}_*)^{-1}X^{-j\top}_*\varepsilon^j\|_d + \|(X^{-j\top}_* X^{-j}_*)^{-1}\hat{z}^j_*/2\|_d && \text{(triangle inequality)} \\
\leq&\,1/n + \|(X^{-j\top}_* X^{-j}_*)^{-1}\hat{z}^j_*/2\|_d && \text{(lemma B.2 assum.)} \\
=&\,1/n + 1/n\cdot n\|(X^{-j\top}_* X^{-j}_*)^{-1}\hat{z}^j_*/2\|_d && (1 = 1/n\cdot n) \\
=&\,1/n + \|(X^{-j\top}_* X^{-j}_*/n)^{-1}\hat{z}^j_*\|_d/(2n) && \text{(linearity of norm)}
\end{aligned}$$

as desired. □

Combining the inequality of lemma B.2 (ii) and the beta-min condition yields for all $i \in \{1, \ldots, s^j_*\}$

$$\frac{1}{(\mathbf{r}^j_*)_i}|((\beta^j_*)_* - \hat{\beta}^j_*)_i| \,\leq\, 1/n + \|(X^{-j\top}_* X^{-j}_*/n)^{-1}\hat{z}^j_*\|_d/(2n) \,<\, \frac{1}{(\mathbf{r}^j_*)_i}|((\beta^j_*)_*)_i|\,.$$

Thus, $S^j_* \subset \text{supp}(\hat{\beta}^j)$. Together with $\text{supp}(\hat{\beta}^j) \subset S^j_*$ from lemma B.2 (i), this implies $\text{supp}(\hat{\beta}^j) = S^j_*$, that is, SI provides exact graph recovery,

$$\widehat{\mathcal{E}} \,=\, \mathcal{E}\,.$$

This completes the proof of Theorem 1.3.1.

Appendix C

DETAILS ON IMAGING AND PREPROCESSING DATA

The fMRI data in the paper is collected from outpatients at the Department of Neurology at Beijing Hospital from April 2012 through December 2013. Imaging was performed with Philips Achieva 3.0 T with 16-channel standard head coil (FFE-EPI: TR/TE = 3000 ms/35 ms, FOV = 220 mm×220 mm, flip angle = 90°, matrix = 64 × 64, slice thickness = 44 mm, slice gap = 0, slices = 34). The preprocessing of the fMRIs was done with `SPM8` in `Matlab`. Steps include: slice timing correction, realignment, spatial normalization, and averaging within volume. An autoregressive integrated moving average model [24, 32] was applied to account for autocorrelation. The preprocessed data and a detailed description of the preprocessing pipeline can be found on our GitHub page `https://github.com/LedererLab/fMRI`.