Disease Prediction based on Functional Connectomes using a Scalable and Spatially-Informed Support Vector Machine

Takanori Watanabe[a,*], Daniel Kessler[c], Clayton Scott[a,b], Michael Angstadt[c], Chandra Sripada[c]

[a] *Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA*
[b] *Department of Statistics, University of Michigan, Ann Arbor, MI, USA*
[c] *Department of Psychiatry, University of Michigan, Ann Arbor, MI, USA*

Abstract

Substantial evidence indicates that major psychiatric disorders are associated with distributed neural dysconnectivity, leading to strong interest in using neuroimaging methods to accurately predict disorder status. In this work, we are specifically interested in a multivariate approach that uses features derived from whole-brain resting state functional connectomes. However, functional connectomes reside in a high dimensional space, which complicates model interpretation and introduces numerous statistical and computational challenges. Traditional feature selection techniques are used to reduce data dimensionality, but are blind to the spatial structure of the connectomes. We propose a regularization framework where the 6-D structure of the functional connectome (defined by pairs of points in 3-D space) is explicitly taken into account via the fused Lasso or the GraphNet regularizer. Our method only restricts the loss function to be convex and margin-based, allowing non-differentiable loss functions such as the hinge-loss to be used. Using the fused Lasso or GraphNet regularizer with the hinge-loss leads to a structured sparse support vector machine (SVM) with embedded feature selection. We introduce a novel efficient optimization algorithm based on augmented Lagrangian and the classical alternating direction method, which can solve both fused Lasso and GraphNet regularized SVM with very little modification. We also demonstrate that the inner subproblems of the algorithm can be solved efficiently in analytic form by coupling the variable splitting strategy with a data augmentation scheme. Experiments on simulated data and resting state scans from a large schizophrenia dataset show that our proposed approach can identify predictive regions that are spatially contiguous in the 6-D "connectome space," offering an additional layer of interpretability that could provide new insights about various disease processes.

Keywords: Classification, feature selection, structured sparsity, resting state fMRI, functional connectivity, support vector machine

1. Introduction

There is substantial interest in establishing neuroimaging-based biomarkers that reliably distinguish individuals with psychiatric disorders from healthy individuals. Towards this end, neuroimaging affords a variety of specific modalities including structural imaging, diffusion tensor imaging (DTI) and tractography, and activation studies under conditions of cognitive challenge (*i.e.*, task-based functional magnetic resonance imaging (fMRI)). In addition, resting state fMRI has emerged

* Corresponding author. 1301 Beal Avenue, 4111 EECS, Ann Arbor, MI, 48109 USA; Tel: +1 734 615 7027
Email addresses: takanori@umich.edu (Takanori Watanabe), kesslerd@umich.edu (Daniel Kessler), clayscot@umich.edu (Clayton Scott), mangstad@med.umich.edu (Michael Angstadt), sripada@umich.edu (Chandra Sripada)

as a mainstream approach that offers robust, sharable, and scalable ability to comprehensively characterize patterns of connections and network architecture of the brain.

Recently a number of groups have demonstrated that substantial quantities of discriminative information regarding psychiatric diseases reside in resting state functional connectomes (Castellanos et al., 2013; Fox and Greicius, 2010). In this article, we define the functional connectomes as the cross-correlation matrix that results from parcellating the brain into hundreds of distinct regions, and computing cross-correlation matrices across time (Varoquaux and Craddock, 2013). Even with relatively coarse parcellation schemes with several hundred regions of interest (ROI), the resulting connectomes encompass hundreds of thousands of connections or more. The massive size of connectomes offers new possibilities, as patterns of connectivity across the entirety of the brain are represented. Nonetheless, the high dimensionality of connectomic data presents critical statistical and computational challenges. In particular, mass univariate strategies that perform separate statistical tests at each edge of the connectome require excessively stringent corrections for multiple comparisons. Multivariate methods are promising, but these require specialized approaches in the context where the number of parameters dominate the number of observations, a setting commonly referred to as the "large p small n problem," denoted $p \gg n$ (Bühlmann and van de Geer, 2011; West, 2003).

In the $p \gg n$ regime, it is important to leverage any potential structure in the data, and sparsity is a natural assumption that arises in many applications (Candes and Wakin, 2008; Fan and Lv, 2010). For example, in the context of connectomics, it is reasonable to believe that only a fraction of the functional connectome is impacted under a specific disorder, an assumption that has been supported in nearly all extant studies (see Castellanos et al. (2013)). Furthermore, when sparsity is coupled with a linear classifier, the nonzero variables can be interpreted as pairs of brain regions that allow reliable discrimination between controls and patients. In other words, sparse linear classifiers have the potential of revealing *connectivity-based biomarkers* that characterize mechanisms of the disease process of interest (Atluri et al., 2013).

The problem of identifying the subset of variables relevant for prediction is called feature selection (Guyon and Elisseeff, 2003; Jain et al., 2000), which can be done in a univariate or a multivariate fashion. In the univariate approach, features are independentally ranked based on their statistical relationship with the target label (*e.g.*, two sample t-test, mutual information), and only the top features are submitted to the classifier. While this method is commonly used (Sripada et al., 2013b; Zeng et al., 2012), it ignores the multivariate nature of fMRI. On the other hand, multivariate approaches such as *recursive feature elimination* (Guyon and Elisseeff, 2003) can be used to capture feature interactions (Craddock et al., 2009; Dai et al., 2012), but these methods are computationally intensive and rely on suboptimal heuristics. However, a more serious shortcoming common to all the methods above is that outside of sparsity, no structural information is taken into account. In particular, we further know that functional connectomes reside in a structured space, defined by pairs of coordinate points in 3-D brain space. Performing prediction and feature selection in a spatially informed manner could potentially allow us to draw more neuroscientifically meaningful conclusions. Fortunately, *regularization methods* allow us to achieve this in a natural and principled way.

Regularization is a classical technique to prevent overfitting (James and Stein, 1961; Tikhonov, 1963), achieved by encoding prior knowledge about the data structure into the estimation problem. Sparsity promoting regularization methods, such as Lasso (Tibshirani, 1996) and Elastic-net (Zou and Hastie, 2005), have the advantage of performing prediction and feature selection jointly (Grosenick et al., 2008; Yamashita et al., 2008); however, they also have the issue of neglecting additional structure the data may have. Recently, there has been strong interest in the machine learning community in designing a convex regularizer that promotes *structured sparsity* (Chen et al., 2012; Mairal et al., 2011; Micchelli et al., 2013), which extends the standard concept of sparsity. Indeed, spatially informed regularizers have been applied successfully in task-based detection, *i.e.*,

decoding, where the goal is to localize in 3-D space the brain regions that become active under an external stimulus (Baldassarre et al., 2012; Gramfort et al., 2013; Grosenick et al., 2013; Jenatton et al., 2012; Michel et al., 2011). Connectomic maps exhibit rich spatial structure, as each connection comes from a pair of localized regions in 3-D space, giving each connection a localization in 6-D space (referred to as "connectome space" hereafter). However, to the best of our knowledge, no framework currently deployed exploits this spatial structure in the functional connectome.

Based on these considerations, the main contributions of this paper are two-fold. First, we propose to explicitly account for the 6-D spatial structure of the functional connectome by using either the fused Lasso (Tibshirani et al., 2005) or the GraphNet regularizer (Grosenick et al., 2013). Second, we introduce a novel scalable algorithm based on the classical alternating direction method (Boyd et al., 2011; Gabay and Mercier, 1976; Glowinski and Marroco, 1975) for solving the nonsmooth, large-scale optimization problem that results from these spatially-informed regularizers. Variable splitting and data augmentation strategies are used to break the problem into simpler subproblems that can be solved efficiently in closed form. The method we propose only restricts the loss function to be convex and margin-based, which allows non-differentiable loss functions such as the hinge-loss to be used. This is important, since using the fused Lasso or the GraphNet regularizer with the hinge-loss function leads to a structured sparse support vector machine (SVM) (Grosenick et al., 2013; Ye and Xie, 2011), where feature selection is *embedded* (Guyon and Elisseeff, 2003), *i.e.*, feature selection is conducted jointly with classification. We demonstrate that the optimization algorithm we introduce can solve both fused Lasso and GraphNet regularized SVM with very little modification. To the best of our knowledge, this is the first application of structured sparse methods in the context of disease prediction using functional connectomes. Additional discussions of technical contributions are reported in Sec. 2.3. We perform experiments on simulated connectomic data and resting state scans from a large schizophrenia dataset to demonstrate that the proposed method identifies predictive regions that are spatially contiguous in the connectome space, offering an additional layer of interpretability that could provide new insights about various disease processes.

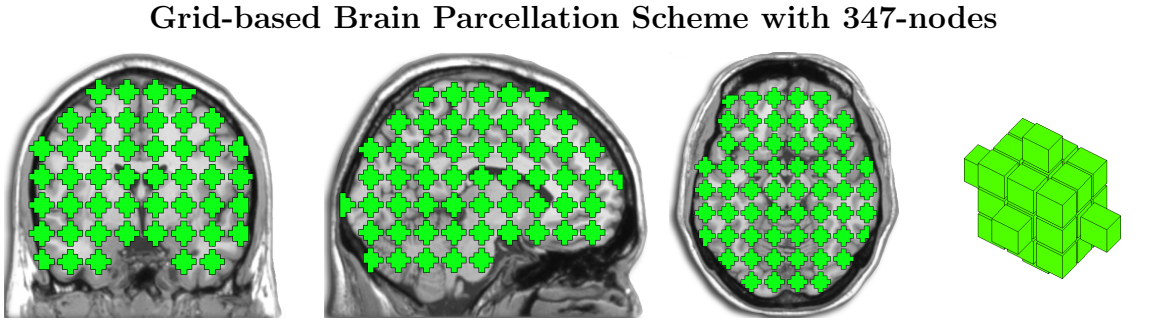
Notation. We let lowercase and uppercase bold letters denote vectors and matrices, respectively. For every positive integer $n \in \mathbb{N}$, we define an index set $[n] := \{1, \ldots, n\}$, and also let $\boldsymbol{I}_n \in \mathbb{R}^{n \times n}$ denote the identity matrix. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times p}$, we let $\boldsymbol{A}^T$ denote its matrix transpose, and $\boldsymbol{A}^H$ denote its Hermitian transpose. Given $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^n$, we invoke the standard notation $\langle \boldsymbol{w}, \boldsymbol{v} \rangle := \sum_{i=1}^n w_i v_i$ to express the inner product in $\mathbb{R}^n$. We also let $\|\boldsymbol{w}\|_p = (\sum_{i=1}^n w_i^p)^{1/p}$ denote the ℓ_p-norm of a vector, $p \geqslant 1$, with the absence of subscript indicating the standard Euclidean norm, $\|\cdot\| = \|\cdot\|_2$.

2. Material and methods

2.1. Defining Functional Connectomes

In this work, we produced a whole-brain resting state functional connectome as follows. First, 347 non-overlapping spherical nodes are placed throughout the entire brain in a regularly-spaced grid pattern, with a spacing of $18 \times 18 \times 18$ mm; each of these nodes represents a pseudo-spherical ROI with a radius of 7.5 mm, which encompasses 33 voxels (the voxel size is $3 \times 3 \times 3$ mm). For a schematic representation of the parcellation scheme, see Fig. 1. Next, for each of these nodes, a single representative time-series is assigned by spatially averaging the BOLD signals falling within the ROI. Then, a cross-correlation matrix is generated by computing Pearson's correlation coefficient between these representative time-series. Finally, a vector $\boldsymbol{x}$ of length $\binom{347}{2} = 60,031$ is obtained by extracting the lower-triangular portion of the cross-correlation matrix. This vector $\boldsymbol{x} \in \mathbb{R}^{60,031}$ represents the whole-brain functional connectome, which serves as the feature vector for disease prediction.

The grid-based scheme for brain parcellation used in this work provides numerous advantages. Of note, this approach has been validated in previous studies (Sripada et al., 2013a, 2014, 2013b).



(a) Coronal (b) Sagittal (c) Axial (d) Node (33 voxels)

Figure 1: Coronal, sagittal, and axial slices depicting the coverage of our brain parcellation scheme along with 3-D rendering of one pseudo-sphereical node. Each contiguous green region represents a pseudo-spherical node representing an ROI containing 33-voxels. Overall, there are 347 non-overlapping nodes placed throughout the entire brain. These nodes are placed on a grid with 18 mm spacing between node centers in the X, Y, and Z dimensions.

Furthermore, the uniformly spaced grid is a good fit with our implementation of fused Lasso and GraphNet, as it provides a natural notion of nearest-neighbor and ordering among the coordinates of the connectome. This property also turns out to be critical for employing our optimization algorithm, which will be discussed in Sec. 2.3. This is in contrast to alternative approaches, such as methods that rely on anatomical (Tzourio-Mazoyer et al., 2002; Zeng et al., 2012) or functional parcellation schemes (Dosenbach et al., 2010). Anatomical parcellations in particular have been shown to yield inferior performance to alternative schemes in the literature (Power et al., 2011). Additionally, grid-based approaches provide scalable density: there is a natural way to increase the spatial resolution of the grid when computational feasibility allows. In particular, to increase node density, one could reduce the inter-node distance and also reduce the node size such that suitable inter-node space remains. This scalable density property turns out to be quite important, as our grid-based scheme is considerably more dense than standard functional parcellations (*e.g.*, Dosenbach et al. (2010); Shirer et al. (2011)) that use as many as several hundred fewer nodes, and thus have tens of thousands fewer connections in the connectome. Finally, the use of our grid-based scheme naturally leaves space between the nodes. While on the surface this may appear to yield incomplete coverage, this is in fact a desirable property to avoid inappropriate inter-node smoothing. This may result as a function of either the point-spread process of fMRI image acquisition or be introduced as a standard preprocessing step. In recognition of these advantages, we have elected to use a grid scheme composed of pseudo-spherical nodes spaced at regular intervals.

One pragmatic advantage of using an *a priori* parcellation scheme as opposed to one that combines parcellation and connectome calculation is that it permits the usage of a grid, and thus yields all the advantages outlined above. Moreover, it allows for easier comparison across studies since an identical (or at least similar) parcellation can be brought to bear on a variety of connectomic investigations. Secondly, while an approach that embeds both parcellation and connectome calculation in a single step may be suitable for recovering a more informative normative connectome, it would not necessarily be appropriate for recovering discriminative information about diseases in the connectome unless features were selected based on their disease-versus-healthy discriminative value. This approach, however, would require nesting parcellation within cross validation and would lead to highly dissimilar classification problems across cross validation folds and present challenges to any sort of inference or aggregation of performance. In light of these challenges, we have elected to use our *a priori* grid-based scheme.

2.2. Statistical learning framework

We now formally introduce the statistical learning framework adopted to perform joint feature selection and disease prediction with spatial information taken into consideration.

2.2.1. Regularized empirical risk minimization and feature selection

In this work, we are interested in the supervised learning problem of linear binary classification. Suppose we are given a set of training data $\{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the input feature vector and $y_i \in \{-1, +1\}$ is the corresponding class label for each $i \in [n]$. In our application, $\boldsymbol{x}_i$ represents functional connectome and y_i indicates the diagnostic status of subject $i \in [n]$, where we adopt the convention of letting $y = +1$ indicate "disorder" and $y = -1$ indicate "healthy" in this article. The goal is to learn a linear decision function $\operatorname{sign}(\langle \boldsymbol{x}, \boldsymbol{w} \rangle)$, parameterized by weight vector $\boldsymbol{w} \in \mathbb{R}^p$, that predicts the label $y \in \{-1, +1\}$ of a new input $\boldsymbol{x} \in \mathbb{R}^p$. A standard approach for estimating $\boldsymbol{w}$ is solving a regularized empirical risk minimization (ERM) problem with the form

$$\underset{\boldsymbol{w} \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle\right) + \lambda \mathcal{R}(\boldsymbol{w}) \,. \tag{1}$$

The first term $\frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle\right)$ corresponds to the *empirical risk* of a margin-based loss function $\ell : \mathbb{R} \to \mathbb{R}_+$ (*e.g.*, hinge, logistic, exponential), which quantifies how well the model fits the data. The second term $\mathcal{R} : \mathbb{R}^p \to \mathbb{R}_+$ is a *regularizer* that curtails overfitting and enforces some kind of structure on the solution by penalizing weight vectors that deviate from the assumed structure. The user-defined regularization parameter $\lambda \geqslant 0$ controls the tradeoff between data fit and regularization. Throughout this work, we assume the loss function and the regularizer to be convex, but not necessarily differentiable. Furthermore, we introduce the following notations

$$\boldsymbol{Y} := \operatorname{diag}\{y_1, \cdots, y_n\}, \quad \boldsymbol{X} := \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}, \quad \boldsymbol{Y}\boldsymbol{X}\boldsymbol{w} = \begin{bmatrix} y_1 \langle \boldsymbol{w}, \boldsymbol{x}_1 \rangle \\ \vdots \\ y_n \langle \boldsymbol{w}, \boldsymbol{x}_n \rangle \end{bmatrix},$$

which allow us to express the empirical risk succinctly by defining a functional $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}_+$ which aggregates the total loss $\mathcal{L}(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}) := \sum_{i=1}^{n} \ell(y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)$.

Regularized ERM (1) has a rich history in statistics and machine learning, and many well known estimators can be recovered from this framework. For example, when the hinge loss $\ell(t) := \max(0, 1 - t)$ is used with the smoothness promoting ℓ_2-regularizer $\|\boldsymbol{w}\|_2^2$, we recover the SVM (Cortes and Vapnik, 1995). However, while smoothness helps prevent overfitting, it is problematic for model interpretation, as all the coefficients from the weight vector contribute to the final prediction function. Automatic feature selection can be done using the ℓ_1-regularizer $\|\boldsymbol{w}\|_1$ known as the Lasso (Tibshirani, 1996), which causes many of the coefficients in $\boldsymbol{w}$ to be exactly zero. Because the prediction function is described by a linear combination between the weight $\boldsymbol{w}$ and the feature vector $\boldsymbol{x}$, we can directly identify and visualize the regions that are relevant for prediction.

While the ℓ_1-regularizer possesses many useful statistical properties, several works have reported poor performance when the features are highly correlated. More precisely, if there are clusters of correlated features, Lasso will select only a single representative feature from each cluster group, ignoring all the other equally predictive features. This leads to a model that is overly sparse and sensitive to data resampling, creating problems for interpretation. To address this issue, Zou and Hastie (2005) proposed to combine the ℓ_1 and ℓ_2 regularizers, leading to the Elastic-net, which has the form $\|\boldsymbol{w}\|_1 + \frac{\gamma}{2\lambda}\|\boldsymbol{w}\|_2^2$, where $\gamma \geqslant 0$ is a second regularization parameter. The ℓ_1-regularizer has the role of encouraging sparsity, whereas the ℓ_2-regularizer has the effect of allowing groups of highly correlated features to enter the model together, leading to a more stable and arguably a

more sensible solution. While Elastic-net addresses part of the limitations of Lasso and has been demonstrated to improve prediction accuracy (Carroll et al., 2009; Ryali et al., 2010), it does not leverage the 6-D structure of connectome space. To address this issue, we employ the fused Lasso and GraphNet (Grosenick et al., 2013).

2.2.2. Spatially informed feature selection and classification via fused Lasso and GraphNet

The original formulation of fused Lasso (Tibshirani et al., 2005) was designed for encoding correlations among successive variables in 1-D data, such as mass spectrometry and comparative genomic hybridization (CGH) data (Ye and Xie, 2011). More specifically, assuming the weight vector $\boldsymbol{w} \in \mathbb{R}^p$ has a natural ordering among its coordinates $j \in [p]$, the regularized ERM problem with the fused Lasso has the following form:

$$\underset{\boldsymbol{w}\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\mathcal{L}(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1 + \gamma \sum_{j=2}^{p} \left| w^{(j)} - w^{(j-1)} \right| , \tag{2}$$

where $w^{(j)}$ indicates the j-th entry of $\boldsymbol{w}$. Like Elastic-net, this regularizer has two components: the first component is the usual sparsity promoting ℓ_1-regularizer, and the second component penalizes the absolute deviation among adjacent coordinates. Together, they have the net effect of promoting sparse and piecewise constant solutions.

The idea of penalizing the deviations among neighboring coefficients can be extended to other situations where there is a natural ordering among the feature coordinates. For instance, the extension of the 1-D fused Lasso (2) for 2-D imaging data is to penalize the *vertical* and *horizontal* difference between pixels; here, the coordinates are described via lexicographical ordering. This type of generalization applies to our 6-D functional connectomes by the virtue of the grid pattern in the nodes, and the ERM formulation reads

$$\underset{\boldsymbol{w}\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\mathcal{L}(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1 + \gamma \sum_{j=1}^{p} \sum_{k\in\mathcal{N}_j} \left| w^{(j)} - w^{(k)} \right| , \tag{3}$$

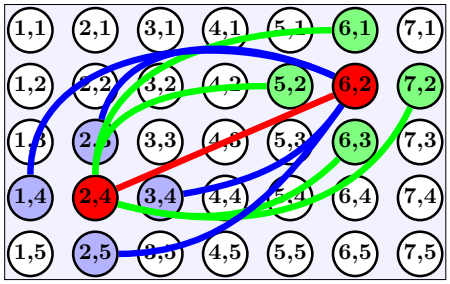
where $\mathcal{N}_j$ is the first-order neighborhood set corresponding to coordinate j in 6-D connectome space. The spatial penalty $\gamma \sum_{j=1}^{p} \sum_{k\in\mathcal{N}_j} \left| w^{(j)} - w^{(k)} \right|$ accounts for the 6-D structure in the connectome by penalizing deviations among *nearest-neighbor* edges, encouraging solutions that are spatially coherent in the connectome space. This type of regularizer is known as an anisotropic total variation (TV) penalty in the image processing community (Wang et al., 2008b), and an analogous isotropic TV penalty was applied by Michel et al. (2011) for the application of 3-D brain decoding.

When the absolute value penalty in the spatial regularizer $\left| w^{(j)} - w^{(k)} \right|$ in (3) is replaced by the squared penalty $\frac{1}{2}(w^{(j)} - w^{(k)})^2$, we recover the GraphNet model proposed by Grosenick et al. (2013):

$$\underset{\boldsymbol{w}\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\mathcal{L}(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1 + \frac{\gamma}{2} \sum_{j=1}^{p} \sum_{k\in\mathcal{N}_j} \left(w^{(j)} - w^{(k)} \right)^2 . \tag{4}$$

GraphNet also promotes spatial contiguity, but instead of promoting sharp piecewise constant patches, it encourages the clusters to appear in smoother form by penalizing the quadratic deviations among the nearest-neighbor edges (*i.e.*, the coordinates of the functional connectome $\boldsymbol{x}$). We emphasize that the optimization algorithm we propose can be used to solve both fused Lasso (3) and GraphNet (4) with very little modification.

To gain a better understanding of the neighborhood set $\mathcal{N}_j$ in the context of our application, let us denote (x, y, z) and (x', y', z') the pair of 3-D points in the brain that define the connectome



$\mathcal{N}_j$ in 4-D Connectome Space

Figure 2: Illustration of the neighborhood structure of the connectome when the nodes reside in 2-D space. The red edge represents coordinate $j = \{(2,4),(6,2)\}$ in 4-D connectome space, and its neighborhood set $\mathcal{N}_j$ is represented by the blue and green edges. This idea extends directly to 6-D connectomes generated from 3-D resting state volumes.

coordinate j. Then, the first-order neighborhood set of j can be written precisely as [1]

$$\mathcal{N}_j := \left\{ \begin{array}{lll} (x \pm 1, y, z, x', y', z'), & (x, y \pm 1, z, x', y', z'), & (x, y, z \pm 1, x', y', z'), \\ (x, y, z, x' \pm 1, y', z'), & (x, y, z, x', y' \pm 1, z'), & (x, y, z, x', y', z' \pm 1) \end{array} \right\} .$$

Fig. 2 provides a pictorial illustration of $\mathcal{N}_j$ in the case of a 4-D connectome, where the nodes reside in 2-D space.

There are multiple reasons why fused Lasso and GraphNet are justified approaches for our problem. For example, fMRI is known to possess high spatio-temporal correlation between neighboring voxels and time points, partly for biological reasons as well as from preprocessing (*e.g.*, spatial smoothing). Consequently, functional connectomes contain rich correlations among nearby coordinates in the connectome space. In addition, there is a neurophysiological basis for why the predictive features are expected to be spatially contiguous rather than being randomly dispersed throughout the brain; this point will be thoroughly discussed in Sec. 4.1. Finally, the spatial coherence that fused Lasso and GraphNet promote helps decrease model complexity and facilitates interpretation.

Letting $\boldsymbol{C} \in \mathbb{R}^{e \times p}$ denote the 6-D *finite differencing matrix* (also known as the *incidence matrix*), the spatial regularization term for both fused Lasso and GraphNet can be written compactly as

$$\|\boldsymbol{C}\boldsymbol{w}\|_q^q = \sum_{j=1}^{p} \sum_{k \in \mathcal{N}_j} |\, w^{(j)} - w^{(k)} \,|^q, \quad q \in \{1, 2\} \ ,$$

where each row in $\boldsymbol{C}$ contains a single $+1$ and a -1 entry, and e represents the total number of adjacent coordinates in the connectome. This allows us to write out the regularized ERM formulation for both fused Lasso (3) and GraphNet (4) in the following unified form:

$$\arg\min_{\boldsymbol{w} \in \mathbb{R}^p} \frac{1}{n} \mathcal{L}(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1 + \frac{\gamma}{q} \|\boldsymbol{C}\boldsymbol{w}\|_q^q \ , \ q \in \{1, 2\} \ . \tag{5}$$

We will focus on this matrix-vector representation hereafter, as it is more intuitive and convenient for analyzing the variable splitting framework in the upcoming section.

[1] If (x, y, z) or (x', y', z') are on the boundary of the brain volume, then neighboring points outside the brain volume are excluded from $\mathcal{N}_j$.

2.3. Optimization

Solving the optimization problem (5) is challenging since the problem size p is large and the three terms in the cost function can each be non-differentiable. To address these challenges, we now introduce a scalable optimization framework based on augmented Lagrangian (AL) methods. In particular, we introduce a variable splitting scheme that converts the unconstrained optimization problem of the form (5) into an equivalent constrained optimization problem, which can be solved efficiently using the alternating direction method of multipliers (ADMM) algorithm (Boyd et al., 2011; Gabay and Mercier, 1976; Glowinski and Marroco, 1975). We demonstrate that by augmenting the weight vector with zero entries at appropriate locations, the inner subproblems associated with ADMM can be solved efficiently in closed form.

2.3.1. Alternating Direction Method of Multipliers

The ADMM algorithm is a powerful algorithm for solving convex optimization problems having the separable structure (Boyd et al., 2011)

$$\underset{\bar{\boldsymbol{x}},\bar{\boldsymbol{y}}}{\text{minimize}}\ \bar{\boldsymbol{f}}(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{g}}(\bar{\boldsymbol{y}}) \quad \text{subject to } \bar{\boldsymbol{A}}\bar{\boldsymbol{x}} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}} = \boldsymbol{0}\ , \tag{6}$$

where $\bar{\boldsymbol{x}} \in \mathbb{R}^{\bar{p}}$ and $\bar{\boldsymbol{y}} \in \mathbb{R}^{\bar{q}}$ are unknown primal variables, $\bar{\boldsymbol{f}} : \mathbb{R}^{\bar{p}} \to \mathbb{R} \cup \{+\infty\}$ and $\bar{\boldsymbol{g}} : \mathbb{R}^{\bar{q}} \to \mathbb{R} \cup \{+\infty\}$ are closed convex functions, and $\bar{\boldsymbol{A}} \in \mathbb{R}^{c\times\bar{p}}$ and $\bar{\boldsymbol{B}} \in \mathbb{R}^{c\times\bar{q}}$ are matrices representing c linear constraints. More specifically, the ADMM algorithm solves for the primal variables in (6) through the following iterative procedure:

$$\bar{\boldsymbol{x}}^{(t+1)} \leftarrow \arg\min_{\bar{\boldsymbol{x}}} \bar{\boldsymbol{f}}(\bar{\boldsymbol{x}}) + \frac{\rho}{2}\left\|\bar{\boldsymbol{A}}\bar{\boldsymbol{x}} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}}^{(t)} + \boldsymbol{u}^{(t)}\right\|^2 \tag{7}$$

$$\bar{\boldsymbol{y}}^{(t+1)} \leftarrow \arg\min_{\bar{\boldsymbol{y}}} \bar{\boldsymbol{g}}(\bar{\boldsymbol{y}}) + \frac{\rho}{2}\left\|\bar{\boldsymbol{A}}\bar{\boldsymbol{x}}^{(t+1)} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}} + \boldsymbol{u}^{(t)}\right\|^2 \tag{8}$$

$$\boldsymbol{u}^{(t+1)} \leftarrow \boldsymbol{u}^{(t)} + \left(\bar{\boldsymbol{A}}\bar{\boldsymbol{x}}^{(t+1)} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}}^{(t+1)}\right)\ , \tag{9}$$

where superscript t denotes the iteration count and $\boldsymbol{u} \in \mathbb{R}^c$ denotes the (scaled) dual variable.

The convergence of the ADMM algorithm has been established in Theorem 1 of Mota et al. (2011), which states that if matrices $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{B}}$ are full column-rank and the problem (6) is solvable (*i.e.*, it has an optimal objective value), the ADMM iterations (7) - (9) converges to the optimal solution. While the AL parameter $\rho > 0$ does not affect the convergence property of ADMM, it can impact its convergence speed. We use the value $\rho = 1$ in all of our implementations.

2.3.2. Variable splitting and data augmentation

The original formulation of our problem (5) does not have the structure of (6). However, we can convert the unconstrained optimization problem (5) into an equivalent constrained optimization problem (6) by introducing auxiliary constraint variables, a method known as *variable splitting* (Afonso et al., 2010). While there are several different ways to introduce the constraint variables, the heart of the strategy is to select a splitting scheme that decouples the problem into more manageable subproblems. For example, one particular splitting strategy we can adopt for problem (5) is

$$\begin{aligned}\underset{\substack{\boldsymbol{w},\boldsymbol{v_1}\\ \boldsymbol{v_2},\boldsymbol{v_3},\boldsymbol{v_4}}}{\text{minimize}}\ & \frac{1}{n}\mathcal{L}(\boldsymbol{v_1}) + \lambda\left\|\boldsymbol{v_2}\right\|_1 + \frac{\gamma}{q}\left\|\boldsymbol{v_3}\right\|_q^q \\ \text{subject to } & \boldsymbol{Y}\boldsymbol{X}\boldsymbol{w} = \boldsymbol{v_1},\ \boldsymbol{w} = \boldsymbol{v_2},\ \boldsymbol{C}\boldsymbol{v_4} = \boldsymbol{v_3},\ \boldsymbol{w} = \boldsymbol{v_4}\ ,\end{aligned} \tag{10}$$

where $\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}, \boldsymbol{v_4}$ are the constraint variables. It is easy to see that problems (5) and (10) are equivalent, and the correspondence with the ADMM formulation (6) is as follows:

$$\bar{\boldsymbol{f}}(\bar{\boldsymbol{x}}) = \frac{\gamma}{q}\left\|\boldsymbol{v_3}\right\|_q^q, \quad \bar{\boldsymbol{g}}(\bar{\boldsymbol{y}}) = \frac{1}{n}\mathcal{L}(\boldsymbol{v_1}) + \lambda\left\|\boldsymbol{v_2}\right\|_1$$

$$\bar{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{YX} & \boldsymbol{0} \\ \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}, \; \bar{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{w} \\ \boldsymbol{v_3} \end{bmatrix}, \bar{\boldsymbol{B}} = \begin{bmatrix} -\boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{C} \\ \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{I} \end{bmatrix}, \bar{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{v_1} \\ \boldsymbol{v_2} \\ \boldsymbol{v_4} \end{bmatrix}. \tag{11}$$

However, there is an issue with this splitting strategy: one of the resulting subproblems from the ADMM algorithm requires us to invert a matrix involving the Laplacian matrix $\boldsymbol{C}^T\boldsymbol{C} \in \mathbb{R}^{p\times p}$, which is prohibitively large. Although this matrix is sparse, it has a distorted structure due to the irregularities in the coordinates of $\boldsymbol{x}$. These irregularities arise from two reasons: (1) the nodes defining the functional connectome $\boldsymbol{x}$ are placed only on the brain, not the entire rectangular field of view (FOV), and (2) $\boldsymbol{x}$ lacks a complete 6-D representation since it only contains the lower-triangular part of the cross-correlation matrix. Fig. 3a displays the Laplacian matrix that results from the 347-node functional connectome defined in Section 2.1, and the distorted structure is clearly visible.

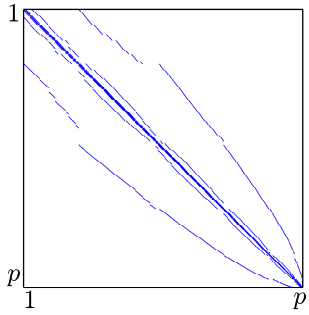
To address this issue, we introduce an *augmentation matrix* $\boldsymbol{A} \in \mathbb{R}^{\tilde{p}\times p}$, whose rows are either the zero vector or an element from the trivial basis $\{\boldsymbol{e}_j \mid j \in [p]\}$, and has the property $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_p$. Furthermore, we define the *augmented weight vector* $\tilde{\boldsymbol{w}} := \boldsymbol{Aw}$, where $\boldsymbol{A}$ rectifies the irregularities in the coordinates of $\boldsymbol{w}$ (and $\boldsymbol{x}$) by padding extra zero entries, accommodating for: (1) the nodes that were not placed in the FOV (*i.e.*, the regions outside the brain), and (2) the diagonal and upper-triangular part of the cross-correlation matrix, which were disposed due to redundancy; further details regarding this augmentation scheme is reported in Appendix A. As a result, we now have a new differencing matrix $\tilde{\boldsymbol{C}} \in \mathbb{R}^{\tilde{e}\times\tilde{p}}$ corresponding to $\tilde{\boldsymbol{w}} \in \mathbb{R}^{\tilde{p}}$, whose Laplacian matrix $\tilde{\boldsymbol{C}}^T\tilde{\boldsymbol{C}} \in \mathbb{R}^{\tilde{p}\times\tilde{p}}$ has a systematic structure, as shown in Fig. 3b. In fact, this matrix has a special structure known as *block-circulant with circulant-blocks* (BCCB), which is critical since the matrix inversion involving $\tilde{\boldsymbol{C}}^T\tilde{\boldsymbol{C}}$ can be computed efficiently in closed form using the fast Fourier transform (FFT) (the utility of this property will be elaborated more in Section 2.3.3). It is important to note that this BCCB structure in the Laplacian matrix arises from the grid structure introduced from the parcellation scheme we adopted for producing the functional connectome.

Finally, by introducing a diagonal masking matrix $\boldsymbol{B} \in \{0,1\}^{\tilde{p}\times\tilde{p}}$, we have $\left\|\boldsymbol{B}\tilde{\boldsymbol{C}}\tilde{\boldsymbol{w}}\right\|_q^q = \left\|\boldsymbol{Cw}\right\|_q^q$ for $\boldsymbol{q} \in \{1,2\}$. Note that this masking strategy was adopted from the recent works of Allison et al. (2013) and Matakos et al. (2013), and has the effect of removing artifacts that are introduced from the data augmentation procedure when computing the $\|\cdot\|_q^q$-norm. This allows us to write out the fused Lasso and GraphNet problem (5) in the following equivalent form:

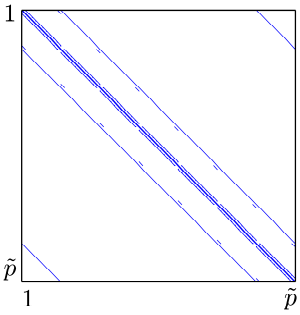
$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{n}\mathcal{L}(\boldsymbol{YXw}) + \lambda\left\|\boldsymbol{w}\right\|_1 + \frac{\gamma}{q}\left\|\boldsymbol{B}\tilde{\boldsymbol{C}}\boldsymbol{Aw}\right\|_q^q, \; q \in \{1,2\}$$

Moreover, this can be converted into a constrained optimization problem

$$\begin{aligned} &\underset{\substack{\boldsymbol{w},\boldsymbol{v_1} \\ \boldsymbol{v_2},\boldsymbol{v_3},\boldsymbol{v_4}}}{\text{minimize}} \; \frac{1}{n}\mathcal{L}(\boldsymbol{v_1}) + \lambda\left\|\boldsymbol{v_2}\right\|_1 + \frac{\gamma}{q}\left\|\boldsymbol{Bv_3}\right\|_q^q \\ &\text{subject to } \boldsymbol{YXw} = \boldsymbol{v_1}, \; \boldsymbol{w} = \boldsymbol{v_2}, \; \tilde{\boldsymbol{C}}\boldsymbol{v_4} = \boldsymbol{v_3}, \; \boldsymbol{Aw} = \boldsymbol{v_4}, \end{aligned} \tag{12}$$



(a) Laplacian matrix: $\boldsymbol{C}^T\boldsymbol{C}$



(b) Augmented Laplacian matrix: $\tilde{\boldsymbol{C}}^T\tilde{\boldsymbol{C}}$

Figure 3: Laplacian matrix corresponding to the original data $\boldsymbol{C}^T\boldsymbol{C}$ and the augmented data $\tilde{\boldsymbol{C}}^T\tilde{\boldsymbol{C}}$, where the rows and columns of these matrices represent the coordinates of the original and augmented functional connectome. Note that the irregularities in the original Laplacian matrix are rectified by data augmentation. The augmented Laplacian matrix has a special structure known as *block-circulant with circulant-blocks* (BCCB), which has important computational advantages that will be exploited in this work.

and the correspondence with the ADMM formulation (6) now becomes:

$$\bar{\boldsymbol{f}}(\bar{\boldsymbol{x}}) = \frac{\gamma}{q}\left\|\boldsymbol{B}\boldsymbol{v_3}\right\|_q^q, \quad \bar{\boldsymbol{g}}(\bar{\boldsymbol{y}}) = \frac{1}{n}\mathcal{L}(\boldsymbol{v_1}) + \lambda\left\|\boldsymbol{v_2}\right\|_1$$

$$\bar{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{Y}\boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{A} & \boldsymbol{0} \end{bmatrix}, \ \bar{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{w} \\ \boldsymbol{v_3} \end{bmatrix}, \bar{\boldsymbol{B}} = \begin{bmatrix} -\boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & -\tilde{\boldsymbol{C}} \\ \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{I} \end{bmatrix}, \bar{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{v_1} \\ \boldsymbol{v_2} \\ \boldsymbol{v_4} \end{bmatrix}. \tag{13}$$

The dual variables corresponding to $\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}$, and $\boldsymbol{v_4}$ are written in block form $\boldsymbol{u} = [\boldsymbol{u_1}^T, \boldsymbol{u_2}^T, \boldsymbol{u_3}^T, \boldsymbol{u_4}^T]^T$. Note that functions $\bar{\boldsymbol{f}}$ and $\bar{\boldsymbol{g}}$ are convex, and matrices $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{B}}$ are full column-rank, so the convergence of the ADMM iterations (7)-(9) is guaranteed (see Theorem 1 in Mota et al. (2011)).

2.3.3. ADMM: efficient closed-form updates

With the variable splitting scheme (12) and ADMM formulation (13), the ADMM update for the primal variable $\bar{\boldsymbol{x}}$ (7) decomposes into subproblems

$$\boldsymbol{w}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{w}} \left\{ \left\| \boldsymbol{Y}\boldsymbol{X}\boldsymbol{w} - \left(\boldsymbol{v_1}^{(t)} - \boldsymbol{u_1}^{(t)}\right)\right\|^2 + \left\|\boldsymbol{w} - \left(\boldsymbol{v_2}^{(t)} - \boldsymbol{u_2}^{(t)}\right)\right\|^2 + \left\|\boldsymbol{A}\boldsymbol{w} - \left(\boldsymbol{v_4}^{(t)} - \boldsymbol{u_4}^{(t)}\right)\right\|^2 \right\} \tag{14}$$

$$\boldsymbol{v_3}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{v_3}} \left\{ \frac{\gamma}{q}\left\|\boldsymbol{B}\boldsymbol{v_3}\right\|_q^q + \frac{\rho}{2}\left\|\boldsymbol{v_3} - \left(\tilde{\boldsymbol{C}}\boldsymbol{v_4}^{(t)} - \boldsymbol{u_3}^{(t)}\right)\right\|^2 \right\}, \tag{15}$$

whereas the updates for primal variable $\bar{\boldsymbol{y}}$ (8) are

$$\boldsymbol{v_1}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{v_1}} \left\{ \frac{1}{n}\mathcal{L}(\boldsymbol{v_1}) + \frac{\rho}{2}\left\|\boldsymbol{v_1} - \left(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}^{(t+1)} + \boldsymbol{u_1}^{(t)}\right)\right\|^2 \right\} \tag{16}$$

$$\boldsymbol{v_2}^{(t+1)} \leftarrow \underset{\boldsymbol{v_2}}{\arg\min} \left\{ \lambda \left\| \boldsymbol{v_2} \right\|_1 + \frac{\rho}{2} \left\| \boldsymbol{v_2} - \left(\boldsymbol{w}^{(t+1)} + \boldsymbol{u_2}^{(t)} \right) \right\|^2 \right\} \tag{17}$$

$$\boldsymbol{v_4}^{(t+1)} \leftarrow \underset{\boldsymbol{v_4}}{\arg\min} \left\{ \left\| \widetilde{\boldsymbol{C}} \boldsymbol{v_4} - \left(\boldsymbol{v_3}^{(t+1)} + \boldsymbol{u_3}^{(t)} \right) \right\|^2 + \left\| \boldsymbol{v_4} - \left(\boldsymbol{A} \boldsymbol{w}^{(t+1)} + \boldsymbol{u_4}^{(t)} \right) \right\|^2 \right\} . \tag{18}$$

The update for the dual variable $\boldsymbol{u}$ is a trivial matrix-vector multiplication (9) (see Algorithm 1 line 14-17).

We now demonstrate that the minimization problems (14)-(18) each admits an efficient, closed form solution.

$\boldsymbol{w}$ update. The quadratic minimization problem (14) has the following closed form solution:

$$\boldsymbol{w}^{(t+1)} \leftarrow \left(\boldsymbol{X}^T \boldsymbol{X} + 2\boldsymbol{I}_p \right)^{-1} \left(\boldsymbol{X}^T \boldsymbol{Y}^T [\boldsymbol{v_1}^{(t)} - \boldsymbol{u_1}^{(t)}] + [\boldsymbol{v_2}^{(t)} - \boldsymbol{u_2}^{(t)}] + \boldsymbol{A}^T [\boldsymbol{v_4}^{(t)} - \boldsymbol{u_4}^{(t)}] \right) . \tag{19}$$

Note we used the fact that $\boldsymbol{Y}^T \boldsymbol{Y} = \boldsymbol{I}_n$ and $\boldsymbol{A}^T \boldsymbol{A} = \boldsymbol{I}_p$ to arrive at this expression. Applying update (19) brute force will require an inversion of a $(p \times p)$ matrix, but this can be converted into an $(n \times n)$ inversion problem by invoking the *matrix inversion Lemma*

$$\left(\boldsymbol{X}^T \boldsymbol{X} + 2\boldsymbol{I}_p \right)^{-1} = \frac{1}{2} \boldsymbol{I}_p - \frac{1}{4} \boldsymbol{X}^T \left(\boldsymbol{I}_n + \frac{1}{2} \boldsymbol{X} \boldsymbol{X}^T \right)^{-1} \boldsymbol{X} . \tag{20}$$

In the context of our work, n denotes the number of scanned subjects, which is typically on the order of a few hundred. The matrix $(\boldsymbol{X}^T \boldsymbol{X} + 2\boldsymbol{I}_p)^{-1}$ can be stored in memory if p is small, but the massive dimensionality of the functional connectome in our application dismisses this option. Therefore, we instead precompute the $(p \times n)$ matrix $\boldsymbol{H} := \frac{1}{4} \boldsymbol{X}^T (\boldsymbol{I}_n + \frac{1}{2} \boldsymbol{X} \boldsymbol{X}^T)^{-1}$ in (20), and let

$$\boldsymbol{\varrho}^{(t)} := \boldsymbol{X}^T \boldsymbol{Y}^T [\boldsymbol{v_1}^{(t)} - \boldsymbol{u_1}^{(t)}] + [\boldsymbol{v_2}^{(t)} - \boldsymbol{u_2}^{(t)}] + \boldsymbol{A}^T [\boldsymbol{v_4}^{(t)} - \boldsymbol{u_4}^{(t)}] .$$

This way, the update (19) can be implemented as follows:

$$w^{(t+1)} \leftarrow (\boldsymbol{X}^T \boldsymbol{X} + 2\boldsymbol{I}_p)^{-1} \boldsymbol{\varrho}^{(t)} = \frac{1}{2} \boldsymbol{\varrho}^{(t)} - \boldsymbol{H} \boldsymbol{X} \boldsymbol{\varrho}^{(t)} , \tag{21}$$

which allows us to carry out the $\boldsymbol{w}$-update without having to store a $(p \times p)$ matrix in memory.

$\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ update. The minimization problems (16) and (17) have the form of the (scaled) proximal operator $\mathrm{Prox}_{\tau F} : \mathbb{R}^p \to \mathbb{R}^p$ (Rockafellar and Wets, 1998), defined by

$$\mathrm{Prox}_{\tau F}(\boldsymbol{v}) = \underset{\boldsymbol{u} \in \mathbb{R}^p}{\arg\min} \, \tau F(\boldsymbol{u}) + \frac{1}{2} \left\| \boldsymbol{v} - \boldsymbol{u} \right\|^2 , \quad \tau > 0 , \tag{22}$$

where $F : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is a closed convex function. Using standard subdifferential calculus rules (Borwein and Lewis, 2006), it is straightforward to show that a point $\boldsymbol{u}^* \in \mathbb{R}^p$ solves the minimization in (22) if and only if the condition

$$\boldsymbol{0} \in \partial F(\boldsymbol{u}^*) + (\boldsymbol{u}^* - \boldsymbol{v})/\tau \tag{23}$$

holds. Here, $\partial F(\boldsymbol{u}^*)$ denotes the subdifferential of function F at $\boldsymbol{u}^*$, defined by

$$\partial F(\boldsymbol{u}^*) := \{ \boldsymbol{z} \in \mathbb{R}^p : F(\boldsymbol{u}^*) + \langle \boldsymbol{z}, \boldsymbol{u} - \boldsymbol{u}^* \rangle \leqslant F(\boldsymbol{u}), \ \forall \boldsymbol{u} \in \mathbb{R}^p \} .$$

In addition, both updates (16) and (17) are fully separable across their coordinates, decomposing

into the following sets of elementwise scalar optimization problems:

$$\left[\boldsymbol{v_1}^{(t+1)}\right]_i \leftarrow \mathrm{Prox}_{\frac{\ell}{n\rho}}\left(\left[\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}^{(t+1)} + \boldsymbol{u_1}^{(t)}\right]_i\right), \qquad i \in [n] \tag{24}$$

$$\left[\boldsymbol{v_2}^{(t+1)}\right]_j \leftarrow \mathrm{Prox}_{\frac{\lambda}{\rho}|\cdot|}\left(\left[\boldsymbol{w}^{(t+1)} + \boldsymbol{u_2}^{(t)}\right]_j\right), \qquad j \in [p]\ , \tag{25}$$

where $[\,\cdot\,]_i$ and $[\,\cdot\,]_j$ each index the i-th and j-th element of a vector in $\mathbb{R}^n$ and $\mathbb{R}^p$ respectively. For some margin-based loss functions, their corresponding proximal operator (24) can be derived in closed form using the optimality condition (23). Fig. 4 plots a few commonly used margin-based losses and their corresponding proximal operators, and Table 1 provides their closed form expressions. The choice of the margin-based loss is application dependent, such as whether differentiability is desired or not. The proximal operator of the ℓ_1-norm (17) and the absolute loss function (25) corresponds to the well known *soft-threshold operator* (Donoho, 1995)

$$\mathrm{Soft}_\tau(t) := \begin{cases} t-\tau & \text{if } t > \tau \\ 0 & \text{if } |t| \leqslant \tau \\ t+\tau & \text{if } t < -\tau \end{cases}\ . \tag{26}$$

The absolute loss and the soft-threshold operator are also included in Fig. 4 and Table. 1 for completeness.

$\boldsymbol{v_3}$ *update.* The solution to the minimization problem (15) depends on the choice of $q \in \{1, 2\}$, where $q = 1$ recovers fused Lasso and $q = 2$ recovers GraphNet.

In the fused Lasso case $q = 1$, since the masking matrix $\boldsymbol{B} \in \{0,1\}^{\tilde{p}\times\tilde{p}}$ is diagonal, the update (15) is fully separable. Letting $\boldsymbol{\zeta}^{(t)} := \tilde{\boldsymbol{C}}\boldsymbol{v_4}^{(t)} - \boldsymbol{u_3}^{(t)}$, the minimization problem decouples into a set of scalar minimization problems of the form:

$$\arg\min_{v_k\in\mathbb{R}} \left\{\gamma\, b_k\, |v_k| + \frac{\rho}{2}\left(v_k - \zeta_k^{(t)}\right)^2\right\}\ , \qquad k \in [\tilde{p}] \tag{27}$$

where b_k is the k-th diagonal entry of $\boldsymbol{B}$ and $\zeta_k^{(t)}$ is the k-th entry of $\boldsymbol{\zeta}^{(t)} \in \mathbb{R}^{\tilde{p}}$. On one hand, when $b_k = 0$, the minimizer for problem (27) returns the trivial solution $\zeta_k^{(t)}$. On the other hand, when $b_k = 1$, the minimizer will once again have the form of the proximal operator (22) corresponding to the absolute loss function $|\cdot|$, recovering the soft-threshold operator (26). To summarize, when $q = 1$, the update for $\boldsymbol{v_3}$ (15) can be done efficiently by conducting the following elementwise update for each $k \in [\tilde{p}]$:

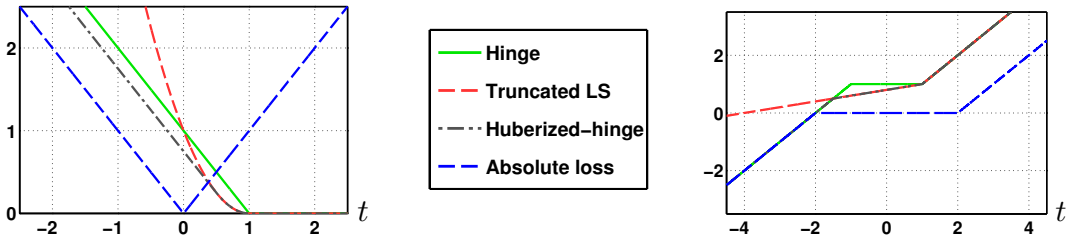
$$\left[\boldsymbol{v_3}^{(t+1)}\right]_k \leftarrow \begin{cases} \mathrm{Soft}_{\gamma/\rho}\left(\left[\tilde{\boldsymbol{C}}\left(\boldsymbol{v_4}^{(t)} - \boldsymbol{u_3}^{(t)}\right)\right]_k\right) & \text{if } \boldsymbol{B}_{k,k} = 1 \\ \left[\tilde{\boldsymbol{C}}\left(\boldsymbol{v_4}^{(t)} - \boldsymbol{u_3}^{(t)}\right)\right]_k & \text{if } \boldsymbol{B}_{k,k} = 0 \end{cases} \tag{28}$$

where $[\cdot]_k$ indexes the k-th element of a vector in $\mathbb{R}^{\tilde{p}}$.

In the GraphNet case $q = 2$, update (15) is a quadratic optimization problem with the closed form solution

$$\boldsymbol{v_3}^{(t+1)} \leftarrow \rho\big(\gamma\boldsymbol{B} + \rho\boldsymbol{I}_{\tilde{p}}\big)^{-1}\tilde{\boldsymbol{C}}(\boldsymbol{v_4}^{(t)} - \boldsymbol{u_3}^{(t)})\ , \tag{29}$$

which is trivial to compute since the matrix $(\gamma\boldsymbol{B} + \rho\boldsymbol{I}_{\tilde{p}})$ is diagonal.



(a) Loss functions $\ell(t)$ (b) Proximal operator $\text{Prox}_{\tau\ell}(t)$

Figure 4: Plots of scalar convex loss functions that are relevant in this work, along with their associated proximal operators. Table 1 provides the closed form expression for these functions. Parameter values of $\tau = 2$ and $\delta = 0.5$ are used in the plot for the proximal operator and the huberized hinge-loss respectively.

	$\ell(t)$	$\text{Prox}_{\tau\ell}(t)$
Hinge	$\max(0, 1-t)$	$\begin{cases} t & \text{if } t > 1 \\ 1 & \text{if } 1-\tau \leqslant t \leqslant 1 \\ t+\tau & \text{if } t < 1-\tau \end{cases}$
Truncated least squares	$\big\{\max(0, 1-t)\big\}^2$	$\begin{cases} t & \text{if } t > 1 \\ \dfrac{t+2\tau}{1+2\tau} & \text{if } t \leqslant 1 \end{cases}$
Huberized hinge (Wang et al., 2008a)	$\begin{cases} 0 & \text{if } t > 1 \\ \dfrac{(1-t)^2}{2\delta} & \text{if } 1-\delta \leqslant t \leqslant 1 \\ 1-t-\frac{\delta}{2} & \text{if } t < 1-\delta \end{cases}$	$\begin{cases} t & \text{if } t > 1 \\ \dfrac{t+\tau/\delta}{1+\tau/\delta} & \text{if } 1-\delta-\tau \leqslant t \leqslant 1 \\ t+\tau & \text{if } t < 1-\delta-\tau \end{cases}$
Absolute loss	$\lvert t\rvert$ (from ℓ_1-regularization)	$\text{Soft}_\tau(t) := \begin{cases} t-\tau & \text{if } t > \tau \\ 0 & \text{if } \lvert t\rvert \leqslant \tau \\ t+\tau & \text{if } t < -\tau \end{cases}$

Table 1: Examples of scalar convex loss functions that are relevant for this work, along with their corresponding proximal operators in closed form.

$\boldsymbol{v_4}$ *update.* The closed form solution to the quadratic optimization problem (18) is

$$\boldsymbol{v_4}^{(t+1)} \leftarrow \left(\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}} + \boldsymbol{I}_{\tilde{p}}\right)^{-1}\left(\widetilde{\boldsymbol{C}}^T[\boldsymbol{v_3}^{(t)} + \boldsymbol{u_3}^{(t)}] + \boldsymbol{A}\boldsymbol{w}^{(t+1)} + \boldsymbol{u_4}^{(t)}\right) . \tag{30}$$

To suppress notations, let us define $\boldsymbol{Q} \in \mathbb{R}^{\tilde{p}\times\tilde{p}}$ and $\boldsymbol{b} \in \mathbb{R}^{\tilde{p}}$, where $\boldsymbol{Q} := \widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}} + \boldsymbol{I}_{\tilde{p}}$ and

$$\boldsymbol{b} := \widetilde{\boldsymbol{C}}^T[\boldsymbol{v_3}^{(t)} + \boldsymbol{u_3}^{(t)}] + \boldsymbol{A}\boldsymbol{w}^{(t+1)} + \boldsymbol{u_4}^{(t)}.$$

As stated earlier, the Laplacian matrix $\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}}$ is block-circulant with circulant-blocks (BCCB), and consequently, the matrix $\boldsymbol{Q}$ is BCCB as well. It is well known that a BCCB matrix can be diagonalized as (Davis, 1979)

$$\boldsymbol{Q} = \boldsymbol{U}^H\boldsymbol{\Lambda}\boldsymbol{U},$$

where $\boldsymbol{U} \in \mathbb{R}^{\tilde{p}\times\tilde{p}}$ is the (6-D) DFT matrix and $\boldsymbol{\Lambda} \in \mathbb{R}^{\tilde{p}\times\tilde{p}}$ is a diagonal matrix containing the (6-D) DFT coefficients of the first column of $\boldsymbol{Q}$. As a result, the update (30) can be carried out efficiently using the (6-D) FFT

$$\boldsymbol{Q}^{-1}\boldsymbol{b} = \left(\boldsymbol{U}^H\boldsymbol{\Lambda}^{-1}\boldsymbol{U}\right)\boldsymbol{b} = \text{ifft}\Big(\text{fft}(\boldsymbol{b}) \circledcirc \boldsymbol{\phi}\Big) , \tag{31}$$

where fft and ifft denote the (6-D) FFT and inverse-FFT operation[2], $\boldsymbol{\phi}$ is a vector containing the diagonal entries of $\boldsymbol{\Lambda}$, and $\oslash$ indicates elementwise division (more precisely, vectors $\boldsymbol{b}$ and $\boldsymbol{\phi}$ are reshaped into 6-D arrays prior to the 6-D FFT and inverse-FFT operations, and the result of these operations is re-vectorized).

AL-based optimization methods that involve this kind of FFT-based inversion have been applied in image processing (Afonso et al., 2010; Allison et al., 2013; Matakos et al., 2013). Problems such as image denoising, reconstruction, and restoration are typically cast as a regularized ERM problem involving the squared loss function. The data augmentation scheme we propose allows us to apply this FFT-based technique with 6-D functional connectomes in the context of binary classification with margin-based loss functions.

Finally, note that the ADMM algorithm was also used to solve the fused Lasso regularized SVM problem in (Ye and Xie, 2011) under a different variable splitting setup. However, their application focuses on 1-D data such as mass spectrometry and array CGH. Consequently, the Laplacian matrix corresponding to their feature vector is tridiagonal with no irregularities present. Furthermore, the variable splitting scheme they propose requires an iterative algorithm to be used for one of the ADMM subproblems. In contrast, the variable splitting scheme and the data augmentation strategy we propose allow the ADMM subproblems to be decoupled in a way that all the updates can be carried out efficiently and non-iteratively in closed form.

Summary: the final algorithm and termination criteria. Algorithm 1 outlines the complete ADMM algorithm for solving both the fused Lasso and GraphNet regularized ERM problem (5), and is guaranteed to converge. In our implementations, all the variables were initialized at zero. The algorithm is terminated when the relative difference between two successive iterates falls below a user-specified threshold:

$$\frac{\left\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)}\right\|}{\left\|\boldsymbol{w}^{(t)}\right\|} \leqslant \varepsilon \,. \tag{32}$$

2.4. Generation of synthetic data: 4-D functional connectomes

To assess the validity of our method, we ran experiments on synthetic 4-D functional connectome data. The data were generated to imitate functional connectomes resulting from a single slice of our grid-based parcellation scheme (see Fig. 1). Specifically, we selected only the nodes that are present at axial slice $z = 18$ in the MNI space; this slice was selected for its substantial X and Y coverage. Fig. 5a provides a schematic representation of the selected nodes.

To mimic the *control vs. patient* binary classification setup, we created two classes of functional connectomes sampled from random normal distributions. The mean and the variance for these distributions were assigned using the functional connectomes generated from the real resting state dataset described later in Sec. 2.5. Specifically, we first took the subject-level functional connectomes corresponding to the 67 healthy controls in the dataset, and extracted the entries that represent the edges among the nodes at slice $z = 18$. Since there are 66 nodes within this slice, this gives us $\binom{66}{2} = 2145$ edges for each subjects. Next, we applied Fisher transformation on these edges to map the correlation values to the real line. For each of these transformed edges, we calculated the inter-subject sample mean and sample variance, which we denote by $\{\hat{\mu}(k), \hat{\sigma}^2(k)\}$ with $k \in [2145]$ indexing the edges. Finally, a synthetic subject-level "control class" connectome is realized by sampling edges individually from a set of random normal distributions having the above mean and variance, and

[2]These multidimensional FFT and inverse FFT operations are implemented using `fftn` and `iffn` functions in MATLAB.

Algorithm 1 ADMM for solving fused Lasso ($q = 1$) or GraphNet ($q = 2$)

Initialize primal variables $\boldsymbol{w}, \boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}, \boldsymbol{v_4}$
Initialize dual variables $\boldsymbol{u_1}, \boldsymbol{u_2}, \boldsymbol{u_3}, \boldsymbol{u_4}$
Set $t = 0$, assign $\lambda \geqslant 0$, $\gamma \geqslant 0$
Precompute $\boldsymbol{H} := \frac{1}{4}\boldsymbol{X}^T(\boldsymbol{I}_n + \frac{1}{2}\boldsymbol{X}\boldsymbol{X}^T)^{-1}$
repeat
  $\bar{\boldsymbol{x}}$-update (7)
    $\boldsymbol{w}^{(t+1)} \leftarrow \left(\boldsymbol{X}^T\boldsymbol{X} + 2\boldsymbol{I}_p\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{Y}^T[\boldsymbol{v_1}^{(t)} - \boldsymbol{u_1}^{(t)}] + [\boldsymbol{v_2}^{(t)} - \boldsymbol{u_2}^{(t)}] + \boldsymbol{A}^T[\boldsymbol{v_4}^{(t)} - \boldsymbol{u_4}^{(t)}]\right)$
    ▷ apply update (21)
    $\boldsymbol{v_3}^{(t+1)} \leftarrow \begin{cases} \text{solve using (28)} & \text{if } q = 1 \text{ (fused Lasso)} \\ \text{solve using (29)} & \text{if } q = 2 \text{ (GraphNet)} \end{cases}$
  $\bar{\boldsymbol{y}}$-update (8)
    $\boldsymbol{v_1}^{(t+1)} \leftarrow \text{Prox}_{\frac{\mathcal{L}}{n\rho}}\left(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}^{(t+1)} + \boldsymbol{u_1}^{(t)}\right)$ ▷ apply (24) elementwise
    $\boldsymbol{v_2}^{(t+1)} \leftarrow \text{Soft}_{\lambda/\rho}\left(\boldsymbol{w}^{(t+1)} + \boldsymbol{u_2}^{(t)}\right)$ ▷ apply (25) elementwise
    $\boldsymbol{v_4}^{(t+1)} \leftarrow \left(\tilde{\boldsymbol{C}}^T\tilde{\boldsymbol{C}} + \boldsymbol{I}_{\tilde{p}}\right)^{-1}\left(\tilde{\boldsymbol{C}}^T[\boldsymbol{v_3}^{(t+1)} + \boldsymbol{u_3}^{(t)}] + \boldsymbol{A}\boldsymbol{w}^{(t+1)} + \boldsymbol{u_4}^{(t)}\right)$
    ▷ solve using FFT approach (31)
  $\boldsymbol{u}$-update (9)
    $\boldsymbol{u_1}^{(t+1)} \leftarrow \boldsymbol{u_1}^{(t)} + \boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}^{(t+1)} - \boldsymbol{v_1}^{(t+1)}$
    $\boldsymbol{u_2}^{(t+1)} \leftarrow \boldsymbol{u_2}^{(t)} + \boldsymbol{w}^{(t+1)} - \boldsymbol{v_2}^{(t+1)}$
    $\boldsymbol{u_3}^{(t+1)} \leftarrow \boldsymbol{u_3}^{(t)} + \boldsymbol{v_3}^{(t+1)} - \tilde{\boldsymbol{C}}\boldsymbol{v_4}^{(t+1)}$
    $\boldsymbol{u_4}^{(t+1)} \leftarrow \boldsymbol{u_4}^{(t)} + \boldsymbol{A}\boldsymbol{w}^{(t+1)} - \boldsymbol{v_4}^{(t+1)}$
  $t \leftarrow t + 1$
until stopping criterion is met

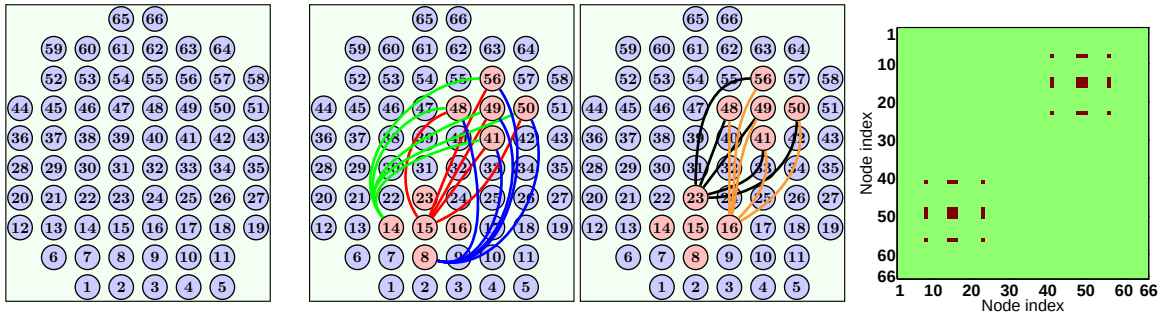
then applying inverse Fisher transformation $\tanh : \mathbb{R} \to (-1, +1)$ on these sampled edges, *i.e.*,

$$\boldsymbol{x} = \left[\tanh\left(x^{(1)}\right), \ldots, \tanh\left(x^{(2145)}\right)\right]^T \text{ where } x^{(k)} \sim \mathcal{N}\left(\hat{\mu}(k), \hat{\sigma}^2(k)\right),\ k \in [2145].$$

Realizations of the "patient class" connectomes are generated in a similar manner, but here we introduced two clusters of *anomalous nodes*, indicated by the red nodes in Fig. 5b. These clusters participate in a disease-specific perturbation, where signal was added to all connections originating in one cluster and terminating in the other. More formally, let $\mathcal{K} \subset [2145]$ denote the index set corresponding to these disease-specific *anomalous edges*, which consist of a complete bipartite graph formed by the anomalous node clusters $\mathcal{C}_1 = \{8, 14, 15, 16, 23\}$ and $\mathcal{C}_2 = \{41, 48, 49, 50, 56\}$, $\mathcal{C}_1, \mathcal{C}_2 \subset [66]$. Under these notations, a synthetic subject-level "patient class" connectome is realized by the following procedure:

$$\boldsymbol{x} = \left[\tanh\left(x^{(1)}\right), \ldots, \tanh\left(x^{(2145)}\right)\right]^T \text{ where } \begin{cases} x^{(k)} \sim \mathcal{N}\Big(\hat{\mu}(k), \hat{\sigma}^2(k)\Big) & \text{if } k \notin \mathcal{K} \\ x^{(k)} \sim \mathcal{N}\Big(\hat{\mu}(k) + d \cdot \hat{\sigma}(k),\ \hat{\sigma}^2(k)\Big) & \text{if } k \in \mathcal{K}\ . \end{cases}$$

In other words, if an edge k is a member of the anomalous edge set $\mathcal{K}$, a non-random signal $d \cdot \hat{\sigma}(k)$



(a) Control (b) Patient (c) Edge support

Figure 5: Schematic representations of the synthetic 4-D functional connectome data generated for the simulation experiments (best viewed in color). (a) Node orientation representing the "control class" connectome, where the blue nodes indicate the normal nodes. (b) Node orientation representing the "patient class" connectome, where there are 25 *anomalous edges* shared among the two *anomalous node* clusters indicated in red (this subfigure is split into two side-by-side figures to improve visibility of the impacted edges). (c) Binary support matrix indicating the locations of the anomalous edges in the connectome space.

is added to the sampled edge-value. Here, d denotes Cohen's effect size (Cohen, 1988), which we set at $d = 0.6$ for our experiments. Overall, since $|\mathcal{C}_1| = |\mathcal{C}_2| = 5$, we have $|\mathcal{K}| = |\mathcal{C}_1| \cdot |\mathcal{C}_2| = 25$, *i.e.*, there are 25 anomalous edges in the patient group; see Fig. 5b for a pictorial illustration of the anomalous edge set $\mathcal{K}$ in the 2-D node space. Fig. 5c presents a binary support matrix indicating the structure of the anomalous edges in the 4-D connectome space, with the locations of the anomalous edges specified by the product set $\mathcal{C}_1 \times \mathcal{C}_2 \subset [66] \times [66]$.

It is important to note that the inclusion of the clusters of anomalous nodes is motivated from the "patchiness assumption" of brain disorders, a view that has been born from multiple task-based and connectivity-based studies; this point will be expounded in finer detail in 4.1. In short, the "patchiness assumption" is the view that major psychiatric disorders manifest in the brain by impacting moderately sized spatially contiguous regions, which is what the clusters of anomalous nodes are intended to mimic in this simulation.

For training the classifiers, we sampled 100 functional connectomes consisting of 50 control samples and 50 patient samples. For evaluating the performance of the classifiers, we sampled 500 additional functional connectomes consisting of 250 control samples and 250 patient samples.

2.5. Real experimental data: schizophrenia resting state dataset

To further assess the utility of the proposed method, we also conducted experiments on real resting state scans.

Participants. We used the Center for Biomedical Research Excellence (COBRE) dataset (`http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html`) made available by the Mind Research Network. The dataset is comprised of 74 typically developing control participants and 71 participants with a DSM-IV-TR diagnosis of schizophrenia. Diagnosis was established by the Structured Clinical Interview for DSM-IV (SCID). Participants were excluded if they had mental retardation, neurological disorder, head trauma, or substance abuse or dependence in the last 12 months. A summary of the participant demographic characteristics is provided in Table 2.

Data Acquisition. A multi-echo MPRAGE (MEMPR) sequence was used with the following parameters: TR/TE/TI = $2530/[1.64, 3.5, 5.36, 7.22, 9.08]/900$ ms, flip angle = 7°, FOV = 256×256 mm, slab thickness = 176 mm, matrix size = $256 \times 256 \times 176$, voxel size = $1 \times 1 \times 1$ mm, number of echoes = 5, pixel bandwidth = 650 Hz, total scan time = 6 minutes. With 5 echoes, the TR and TI

	Healthy Controls				Schizophrenia			
	n	Age	#male	#RH	n	Age	#male	#RH
Pre-exclusion	74	35.8 ± 11.6	51	71	71	38.1 ± 14.0	57	59
Post-exclusion	67	35.2 ± 11.7	46	66	54	35.5 ± 13.1	48	46

Table 2: Demographic characteristics of the participants before and after sample exclusion criteria is applied (RH = right-handed).

time to encode partitions for the MEMPR are similar to that of a conventional MPRAGE, resulting in similar GM/WM/CSF contrast. Resting state data were collected with single-shot full k-space echo-planar imaging (EPI) with ramp sampling correction using the intercomissural line (AC-PC) as a reference (TR: 2 s, TE: 29 ms, matrix size: 64×64, 32 slices, voxel size: $3 \times 3 \times 4mm^3$).

Imaging Sample Selection. Analyses were limited to participants with: (1) MPRAGE anatomical images, with consistent near-full brain coverage (*i.e.*, superior extent included the majority of frontal and parietal cortex and inferior extent included the temporal lobes) with successful registration; (2) complete phenotypic information for main phenotypic variables (diagnosis, age, handedness); (3) mean framewise displacement (FD) within two standard deviations of the sample mean; (4) at least 50% of frames retained after application of framewise censoring for motion ("motion scrubbing"; see below). After applying these sample selection criteria, we analyzed resting state scans from 121 individuals consisting of 67 healthy controls (HC) and 54 schizophrenic subjects (SZ). Demographic characteristics of the post-exclusion sample are shown in Table 2.

Preprocessing. Preprocessing steps were performed using statistical parametric mapping (SPM8; `www.fil.ion.ucl.ac.uk/spm`). Scans were reconstructed, slice-time corrected, realigned to the first scan in the experiment for correction of head motion, and co-registered with the high-resolution T1-weighted image. Normalization was performed using the voxel-based morphometry (VBM) toolbox implemented in SPM8. The high-resolution T1-weighted image was segmented into tissue types, bias-corrected, registered to MNI space, and then normalized using Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra (DARTEL) (Ashburner, 2007). The resulting deformation fields were then applied to the functional images. Smoothing of functional data was performed with an 8 mm^3 kernel.

Connectome generation. Functional connectomes were generated by placing 7.5 mm radius nodes representing ROIs encompassing 33 $3 \times 3 \times 3$ mm voxels in a regular grid spaced at $18 \times 18 \times 18$ mm intervals throughout the brain. Spatially averaged time series were extracted from each of the ROIs. Next, linear detrending was performed, followed by nuisance regression. Regressors included six motion regressors generated from the realignment step, as well as their first derivatives. White matter and cerebrospinal fluid masks were generated from the VBM-based tissue segmentation step noted above, and eroded using the fslmaths program from FSL to eliminate border regions of potentially ambiguous tissue type. The top five principal components of the BOLD time series were extracted from each of the masks and included as regressors in the model – a method that has been demonstrated to effectively remove signals arising from the cardiac and respiratory cycle (Behzadi et al., 2007). The time-series for each ROI was then band-passed filtered in the 0.01 – 0.10 Hz range. Individual frames with excessive head motion were then censored from the time series. Subjects with more than 50% of their frames removed by scrubbing were excluded from further analysis, a threshold justified by simulations conducted by other groups (Fair et al., 2013), as well as by our group. Pearson product-moment correlation coefficients were then calculated pairwise between time courses for each of the 347 ROIs. Standard steps in functional connectivity analysis (removing motion artifacts and nuisance covariates and calculating Pearson's product moment correlations between pairs of nodes)

was performed with `ConnTool`, a functional connectivity analysis package developed by Robert C. Welsh, University of Michigan.

3. Results

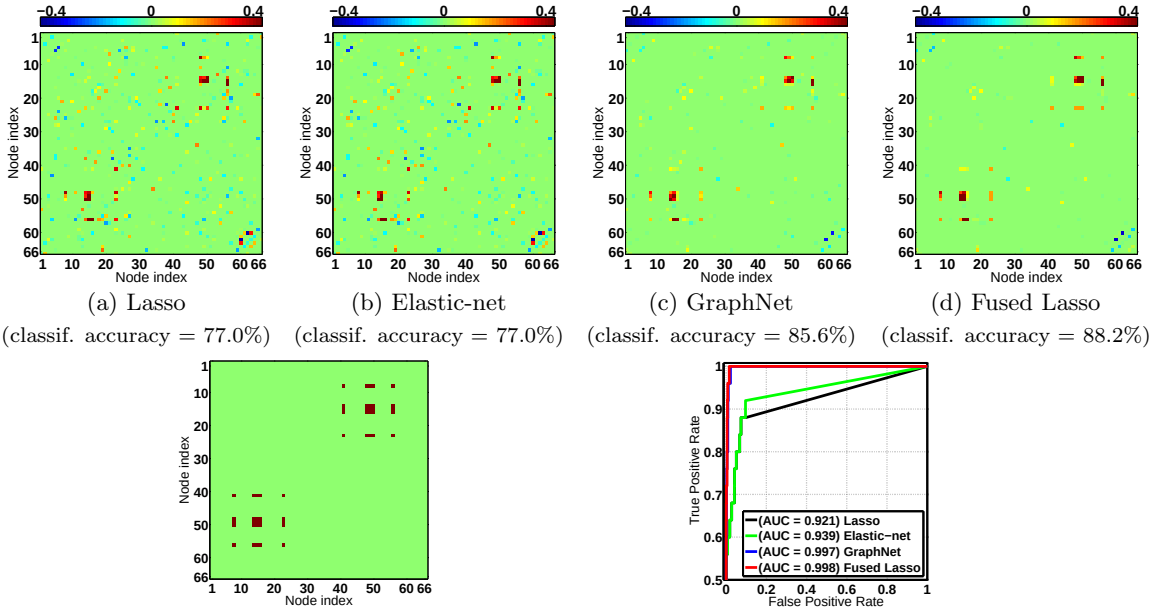
3.1. Results on synthetic functional connectome data

In order to evaluate the validity of our proposed method, we compared the performance of four linear classifiers trained on the synthetic functional connectome data described in Section 2.4, where the training set consists of 100 samples with 50 patients and 50 controls. Specifically, we solved the regularized ERM problem (1) using the hinge-loss and the following four regularizers: Lasso, Elastic-net, GraphNet, and fused Lasso. Lasso and Elastic-net were also solved using ADMM, although the variable splitting scenario and the optimization steps are different from Algorithm 1. The ADMM algorithm for Elastic-net is provided in Appendix B, and the algorithm for Lasso follows directly from Elastic-net by setting $\gamma = 0$. The ADMM algorithm was terminated when the tolerance level (32) fell below $\varepsilon = 4 \times 10^{-3}$ or the algorithm reached 400 iterations. Note that in our experiment, we let $y = +1$ indicate the "patient class" and $y = -1$ indicate the "control class."

With the exception of Lasso, the regularizers we investigated involve two tuning parameters: $\lambda \geqslant 0$ and $\gamma \geqslant 0$. We tuned these regularization parameters by conducting a 5-fold cross-validation on the training set over a two-dimensional grid, and tuned Lasso over a one-dimensional grid. More precisely, the ℓ_1 regularization parameter $\lambda \geqslant 0$ was tuned over the range $\lambda \in \{2^{-11}, 2^{-10.75}, \ldots, 2^{-3.5}\}$ for all four regularizers. The second regularization parameter $\gamma \geqslant 0$ was tuned over the range $\gamma \in \{2^{-16}, 2^{-15.5}, \ldots, 2^{+2}\}$ for Elastic-net and GraphNet and $\gamma \in \{2^{-16}, 2^{-15.5}, \ldots, 2^{-5}\}$ for fused Lasso[3]. The final weight vector estimates are obtained by re-training the classifiers on the entire training set using the regularization parameter values $\{\lambda, \gamma\}$ that yielded the highest 5-fold cross-validation classification accuracy. For visualization, the estimated weight vectors are reshaped into 66×66 symmetric matrices with zeroes on the diagonal (although these are matrices, we will refer to them as "weight vectors" as well), and the classification accuracies are evaluated on a testing set consisting of 500 samples with 250 patients and 250 controls.

The top row of Fig. 6 displays the estimated weight vectors, and the corresponding testing classification accuracies are reported under the subcaptions. Here, the fused Lasso regularized SVM yielded the best classification accuracy at 88.2% using 92 features, followed by 85.6% from GraphNet which used 104 features; Lasso and Elastic-net both achieved 77.0% classification accuracy using 230 and 232 features respectively. However, a perhaps more interesting observation is that fused Lasso and GraphNet were able to recover the structure of the *anomalous edges* much more clearly than Lasso and Elastic-net; this can be seen by comparing the weight vectors estimated by the four regularizers with the support of the anomalous edges displayed in Fig. 6e. While Lasso and Elastic-net yielded weight vector estimates with salt-and-pepper patterns that are difficult to interpret, the weight vector estimates for fused Lasso and GraphNet closely resembles the structure of the *anomalous edges*. To quantify the regularizers' ability to identify the discriminative edges, we generated a receiver operating characteristic (ROC) curve by thresholding the absolute value of the elements of the estimated weight vector. The resulting ROC curve for the four regularizers are plotted in Fig. 6f; we emphasize that this ROC curve summarizes the regularizers' ability to identify the informative edges, and does not represent classification accuracy. From this ROC curve, we see that fused Lasso and GraphNet attain the best performances, achieving a nearly perfect *area under the curve* (AUC) value of 0.998 and 0.997 respectively, whereas the AUC value for Lasso

[3]The grid search region for γ is different for fused Lasso since we observed a clear drop-off in classification performance for any values of γ higher than the range presented. We found this to be true for the real data experiment in Sec. 3.2 as well; see Fig. 7 and Fig. 9.



(a) Lasso (classif. accuracy = 77.0%) (b) Elastic-net (classif. accuracy = 77.0%) (c) GraphNet (classif. accuracy = 85.6%) (d) Fused Lasso (classif. accuracy = 88.2%)

(e) Support of the *anomalous edges* (f) ROC (edge identification accuracy)

Figure 6: Simulation experiment result: training set consists of $n = 100$ samples with 50 patients and 50 controls (best viewed in color). (a)-(d) Weight vectors (reshaped into symmetric matrices) estimated from solving the regularized ERM problem (1) using the hinge-loss and four different regularizers. Regularization parameters were tuned via 5-fold cross-validation on the training set, and classification accuracies were evaluated on a testing set consisting of 500 samples with 250 patients and 250 controls. (e) Support matrix indicating the locations of the anomalous edges. (f) ROC curve representing the anomalous edge identification accuracy (not classification accuracy) of the four regularizers.

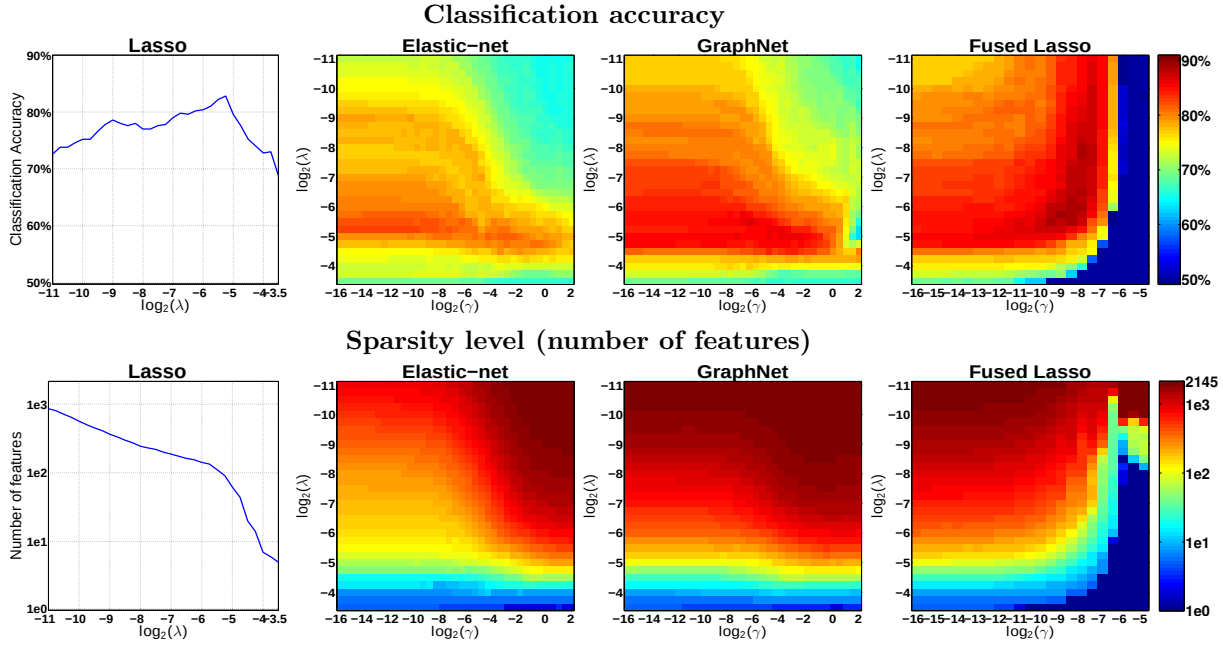


Figure 7: Grid search result for the simulation experiment (best viewed in color). All classifiers were learned using 100 training samples consisting of 50 patients and 50 controls. **Top row**: classification accuracy as a function of the regularization parameters $\{\lambda, \gamma\}$ (evaluated from 500 testing samples consisting of 250 patients and 250 controls). **Bottom row**: the number of features selected as a function of the regularization parameters $\{\lambda, \gamma\}$.

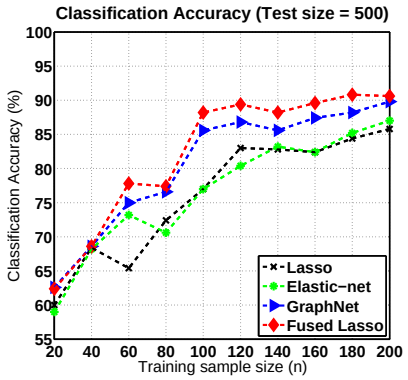


Figure 8: The testing classification accuracy of the different regularizers as a function as a number of training samples n in the simulation experiment. Regularization parameters were tuned via 5-fold cross-validation on the training set. The testing set consists of 500 samples with 250 patients and 250 controls. Table 3 reports the actual numbers.

Testing Classification accuracy (n = training sample size, 500 = test size)

Regularizer	n=20	n=40	n=60	n=80	n=100	n=120	n=140	n=160	n=180	n=200
Lasso	60.0%	68.4%	65.4%	72.4%	77.0%	83.0%	82.8%	82.4%	84.4%	85.8%
Elastic-net	59.7%	68.2%	73.2%	70.6%	77.0%	80.4%	83.2%	82.4%	85.2%	87.0%
GraphNet	**62.6%**	**68.6%**	75.0%	76.6%	85.6%	86.8%	85.6%	87.4%	88.2%	89.8%
Fused Lasso	62.4%	**68.6%**	**77.8%**	**77.4%**	**88.2%**	**89.4%**	**88.2%**	**89.6%**	**90.8%**	**90.6%**

Table 3: The testing classification accuracy of the different regularizers as a function as a number of training samples n in the simulation experiment (the best classification accuracy for each n is denoted in bold font). See Fig. 8 for a plot of this result.

and Elastic-net were 0.921 and 0.939 respectively. In short, Fig. 6a-f demonstrate that fused Lasso and GraphNet not only improved classification accuracy, but also exhibited superior performance in recovering the discriminatory edges with respect to their non-spatially informed counterparts, Lasso and Elastic-net.

In our next analysis, we studied how classification accuracy and sparsity (*i.e.*, number of features selected) behave as a function of the regularization parameters $\{\lambda, \gamma\}$. For this, we conducted a grid search over the same range of λ and γ values presented above, but the classifiers were trained over the entire training set. Classification accuracy was evaluated on the same testing set as the above experiment. The result of the grid search is presented in Fig. 7, where the top row plots the testing classification accuracy and the bottom row plots the number of features selected, both as a function of the regularization parameters $\{\lambda, \gamma\}$.

To further study the performance of our method, we next conducted a *sample complexity analysis* (Gramfort et al., 2011), where we studied how the classification accuracy of the four regularizers behaved as a function of the training sample size n. This was done by repeating our earlier experiment of tuning the regularization parameters via 5-fold cross-validation on the training set, but here we varied the training sample size over the range $n \in \{20, 40, 60, \ldots, 200\}$; the same testing set of size 500 was used throughout for evaluating the classification accuracy. Note the labels are balanced for all datasets, *i.e.*, the training set consists of $n/2$ patients and $n/2$ controls, and similarly the testing set consists of 250 patients and 250 controls. The result of this experiment is reported in Fig. 8 and Table 3. A key observation from this analysis is that the classification accuracy for GraphNet and fused Lasso consistently outperformed Lasso and Elastic-net, which can be attributed to the spatial information injected by these spatially-informed regularizers. Overall, fused Lasso yielded the best classification accuracy.

It is important to note that the inclusion of the anomalous node clusters in the data generating process certainly favors fused Lasso and GraphNet. However, we remind the readers that these anomalous node clusters are not some arbitrary structures we introduced to favor the spatially-informed regularizers, but are motivated from the "patchiness assumption" of brain disorders, a neuroscientific viewpoint which we discuss in detail in Sec. 4.1. The results from the simulation experiments confirm the intuition that if the "patchiness assumption" of brain disorders holds true, spatially-informed classifiers can be a powerful tool for recovering relevant biosignatures.

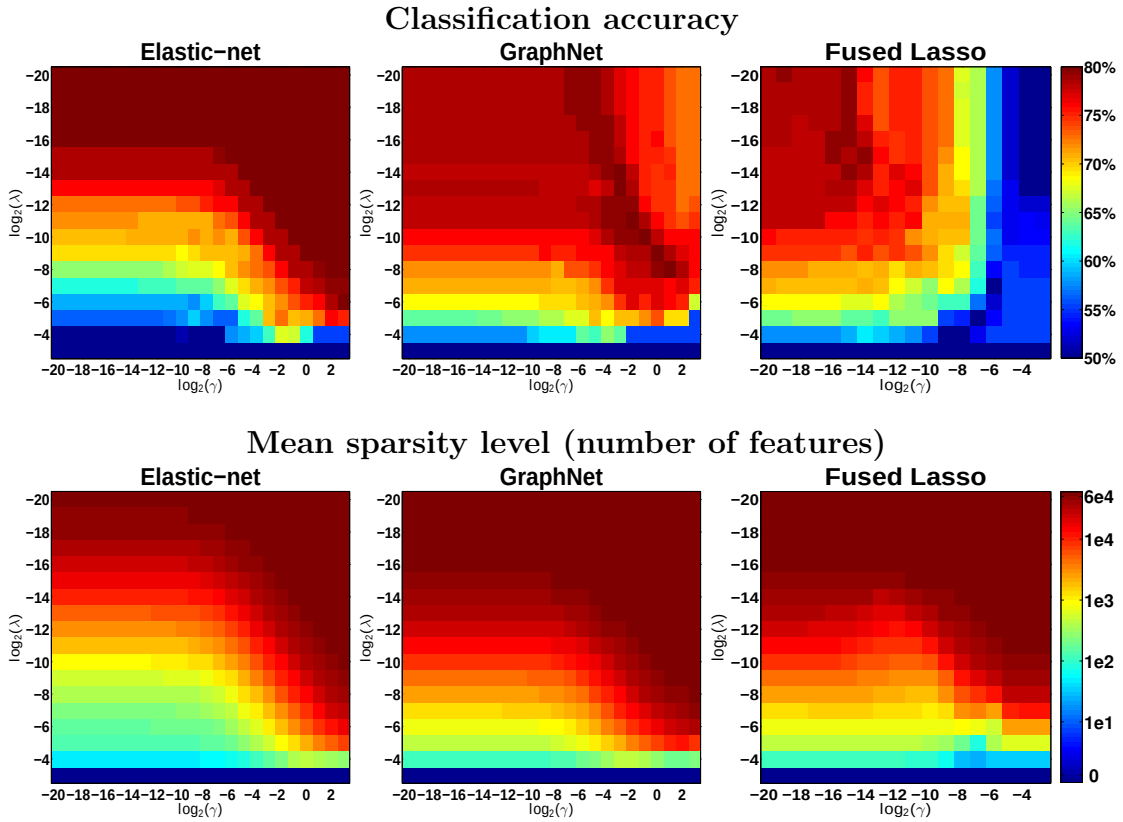
3.2. Results on resting state fMRI data from a schizophrenia dataset

In this experiment, we examined the performance of linear classifiers trained using regularized ERM (1) with the hinge-loss, and three regularizers were subject to comparison: Elastic-net, GraphNet, and fused Lasso. The study involved 121 participants, consisting of 54 schizophrenic subjects (SZ) and 67 healthy controls (HC). We adopt the convention of letting $y = +1$ indicate SZ and $y = -1$ indicate HC subjects. The ADMM algorithm was terminated when the tolerance level (32) fell below $\varepsilon = 4 \times 10^{-3}$ or the algorithm reached 400 iterations. Empirically, we found the algorithm to converge at around 180~300 iterations. For the two regularization parameters, we conducted a two-dimensional grid search: the ℓ_1 regularization parameter $\lambda \geqslant 0$ was searched over the range $\lambda \in \{2^{-20}, 2^{-19}, \cdots, 2^{-3}\}$ for all three regularizers, and the second regularization parameter $\gamma \geqslant 0$ was searched over $\gamma \in \{2^{-20}, 2^{-19}, \cdots, 2^{3}\}$ for Elastic-net and GraphNet and $\gamma \in \{2^{-20}, 2^{-19}, \cdots, 2^{-3}\}$ for fused Lasso. Ten-fold cross-validation to evaluate the generalizability of the classifiers. Furthermore, we analyzed the sparsity level achieved during the grid search by computing the average number of features selected across the cross-validation folds.

The resulting testing classification accuracy and sparsity level for different combinations of $\{\lambda, \gamma\}$ are rendered as heatmaps in Fig. 9. The general trend observed from the grid search is that for all three regularization methods, the classification accuracy improved as more features entered the model. We observed the same trend when using other loss functions as well, specifically the truncated-least squares loss and the huberized-hinge loss (using $\delta = 0.5$) function. Although this behavior may be somewhat surprising, it has been reported that in the $p \gg n$ setting, the unregularized SVM often performs just as well as the best regularized case, and accuracy can degrade when feature pruning takes place (see Ch.18 in Hastie et al. (2009)).

A common practice for choosing the final set of regularization parameters is to select the choice that gives the highest prediction accuracy. Based on the grid search result reported in Fig. 9, one may be tempted to conclude that the prediction models from GraphNet and fused Lasso are not any better than Elastic-net. However, the ultimate goal in our application is the discovery and validation of connectivity-based biomarkers, thus classification accuracy by itself is not sufficient. It is equally important for the prediction model to be interpretable (*e.g.*, sparse) and inform us about the predictive regions residing in the high dimensional connectome space. From the grid search, we found that for all three regularization methods, the classifiers achieved a good balance between accuracy and sparsity when approximately $3,000$ features ($\approx 5\%$) were selected out of $p = 60,031$. More specifically, Elastic-net, GraphNet, and fused Lasso achieved classification accuracies of 73.5%, 70.3%, and 71.9%, using an average of 3076, 3403, and 3140 features across the cross-validation folds. Corresponding regularization parameter values $\{\lambda, \gamma\}$ were: $\{2^{-6}, 2^{-1}\}$, $\{2^{-5}, 2^{-2}\}$, and $\{2^{-9}, 2^{-10}\}$. Therefore, we further analyzed the classifiers obtained from these regularization parameter values.

During cross-validation, we learned a different weight vector for each partitioning of the dataset. In order to obtain a single representative weight vector, we took the approach of Grosenick et al. (2013), computing the elementwise median of the weight vectors across the cross-validation folds. Note that this approach possesses attractive theoretical properties; see Grosenick et al. (2013) and Minsker (2013) for a detailed discussion. For visualization and interpretation, we grouped the indices of these weight vectors according to the network parcellation scheme proposed by Yeo et al. (2011), and augmented this parcellation with subcortical regions and cerebellum derived from the

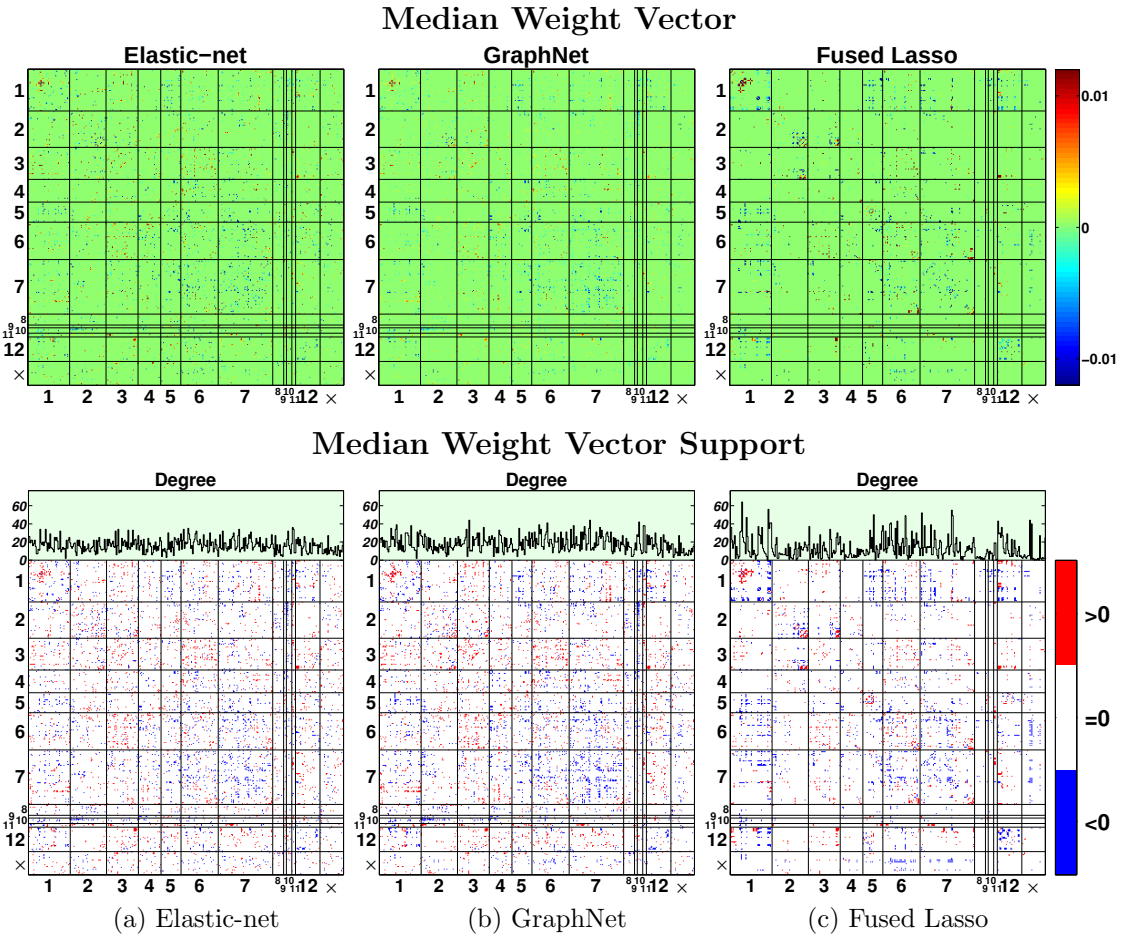


(a) Elastic-net (b) GraphNet (c) Fused Lasso

Figure 9: Grid search result for the real resting state data (best viewed in color). **Top row**: the classification accuracy evaluated from 10-fold cross-validation. **Bottom row**: the average number of features selected across the cross-validation folds. The (x, y)-axis corresponds to the two regularization parameters λ and γ.

parcellation of Tzourio-Mazoyer et al. (2002) (see Table 4); these weight vectors are then reshaped them into 347×347 symmetric matrices with zeroes on the diagonal. Furthermore, we generated trinary representations of these matrices in order to highlight their support structures, where red, blue, and white denotes positive, negative, and zero entries respectively. The resulting matrices are displayed in Fig. 10.

From these figures, one can observe that Elastic-net yields solutions that are scattered throughout the connectome space, which can be problematic for interpretation. In contrast, the weight vector returned from GraphNet has a much smoother structure, demonstrating the impact of the smooth spatial penalty; this is arguably a far more sensible structure from a biological standpoint. Finally, the weight vector from fused Lasso reveals systematic sparsity patterns with multiple contiguous clusters present, indicating that the predictive regions are compactly localized in the connectome space (*e.g.*, see the rich connectivity patterns present in the intra-visual and intra-cerebellum network). It is noteworthy the fused Lasso not only appears to identify more densely packed patches of abnormalities in certain regions, it also generates large areas of relative sparsity (*e.g.*, see somatomotor network interconnections with other networks, and the nodes that fall outside the augmented Yeo parcellation scheme, which are labeled "×"). These areas are more sparse in the fused Lasso map, and this appears to be consistent with existing knowledge of connectivity alterations in schizophrenia (see Sec. 4.3 of the Discussion). In addition, the weight vector estimate from fused Lasso appears



(a) Elastic-net (b) GraphNet (c) Fused Lasso

Figure 10: Weight vectors (reshaped into symmetric matrices) generated by computing the elementwise median of the estimated weight vectors across the cross-validation folds (best viewed in color). The rows and columns of these matrices are grouped according to the network parcellation scheme proposed by Yeo et al. (2011), which is reported in Table 4. The top row displays the heatmap of the estimated weight vectors, whereas the bottom row displays their support structures, with red, blue, and white indicating positive, negative, and zero entries respectively. In order to highlight the structure of the estimated weight vectors, the bottom row further plots the degree of the nodes, *i.e.*, the number of connections a node makes with the rest of the network.

Network Membership Table (× is "unlabeled")			
1. Visual	2. Somatomotor	3. Dorsal Attention	4. Ventral Attention
5. Limbic	6. Frontoparietal	7. Default	8. Striatum
9. Amygdala	10. Hippocampus	11. Thalamus	12. Cerebellum

Table 4: Network parcellation of the brain proposed by Yeo et al. (2011). In our real resting state fMRI study, the indices of the estimated weight vectors are grouped according to this parcellation scheme; see Fig. 10.

to implicate certain nodes more often in connectivity alterations. In order to emphasize this point, the bottom row in Fig. 10 also plots the degree of the nodes, *i.e.*, the number of connections a node makes with the rest of the nodes (this is another example of "spatial contiguity" in the 6-D connectome space).

Finally, in order to convey the regional distribution of the edges recovered by fused Lasso, we rendered implicated edges on canonical 3-D brains (Fig. 11; these figures were generated with the

BrainNet Viewer, http://www.nitrc.org/projects/bnv/). We focus on the three sets of network-to-network connections, intra-frontoparietal, frontoparietal-default, and intra-cerebellum, as these three networks have particularly extensive evidence of their involvement in schizophrenia (see Discussion in Sec. 4). It is noteworthy that lateral prefrontal cortex, an important region in frontoparietal network, is well represented in the fused Lasso map. Edges involving this region represent 39.3% of the intra-frontoparietal connections and 43.6% of the frontoparietal-default network connections. This finding is consistent with previous studies of schizophrenia that emphasize the importance of this region (see Discussion in Sec. 4).

3.3. Computational considerations

It is important to note that the benefit of spatial regularization comes with higher computational expense. To illustrate this point, we ran the ADMM algorithms for Elastic-net, GraphNet, and fused Lasso for 1000 iterations on the full resting state dataset using regularization parameter values $\{\lambda, \gamma\} = \{2^{-15}, 2^{-15}\}$ and compared their computation times (the algorithm for Elastic-net is reported in Appendix B, whereas the algorithms for GraphNet and fused Lasso are reported in Algorithm 1). This timing experiment was implemented in MATLAB version 7.13.0 on a desktop PC with Intel quad-core 3.40 GHz CPU and 12 GB RAM. The total computation times for Elastic-net, GraphNet, and fused Lasso were 17.04 seconds, 96.07 seconds, and 112.45 seconds respectively. The increase in computation time for GraphNet and fused Lasso stems from the fact that unlike the ℓ_2-penalty in Elastic-net, the spatial penalty $\|\boldsymbol{C}\boldsymbol{w}\|_q^q$, $q \in \{1, 2\}$ is not separable across the coordinates of $\boldsymbol{w}$. To address this difficulty, the variable splitting strategy proposed for GraphNet and fused Lasso (12) contains four constraint variables, which is two more than the splitting proposed for Elastic-net (B.1); as a consequence, the ADMM algorithms for GraphNet and fused Lasso contain two additional subproblems. Furthermore, the computational bottlenecks of the ADMM algorithms for GraphNet and fused Lasso are the 6-D FFT and inverse-FFT operations (31), which are not conducted for the Elastic-net. Therefore, if achieving high classification accuracy is the central goal, then Elastic-net would be the most sensible and practical choice, as it yields good classification accuracy and is by far the fastest among the three regularization methods we studied.

Finally, in order to assess the practical utility of our proposed algorithm with respect to existing methods, we conducted another timing experiment using the ADMM algorithm proposed by Ye and Xie (2011), which also solves fused Lasso regularized SVM. It is important to note that the variable splitting scheme they employ is different from the one we introduce, and consequently, their method requires the following matrix inversion problem to be solved for one of the ADMM updates:

$$\boldsymbol{w}^{(t+1)} \leftarrow \left(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{C}^T\boldsymbol{C} + \boldsymbol{I}_p\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{Y}^T[\boldsymbol{v_1}^{(t)} - \boldsymbol{u_1}^{(t)}] + [\boldsymbol{v_2}^{(t)} - \boldsymbol{u_2}^{(t)}] + \boldsymbol{C}^T[\boldsymbol{v_3}^{(t)} - \boldsymbol{u_3}^{(t)}]\right).$$

As suggested in Ye and Xie (2011), we applied the conjugate gradient algorithm to numerically solve this large scale matrix inversion problem[4]. Using the same experimental protocol as our first timing experiment, we ran Ye and Xie's algorithm for 1000 iterations on the full resting state dataset, which resulted in a total computation time of 331.36 seconds, which is nearly three times longer than the algorithm we proposed. This illustrates the practical benefit of our proposed variable splitting and data augmentation scheme, which allows all the ADMM updates to be solved analytically.

4. Discussion

Abundant neurophysiological evidence indicates that major psychiatric disorders are associated with distributed neural dysconnectivity (Konrad and Eickhoff, 2010; Müller et al., 2011; Stephan

[4]The conjugate gradient algorithm was ran until either the ℓ_2-norm of the residual fell below 1×10^{-3} or the algorithm reached 60 iterations.

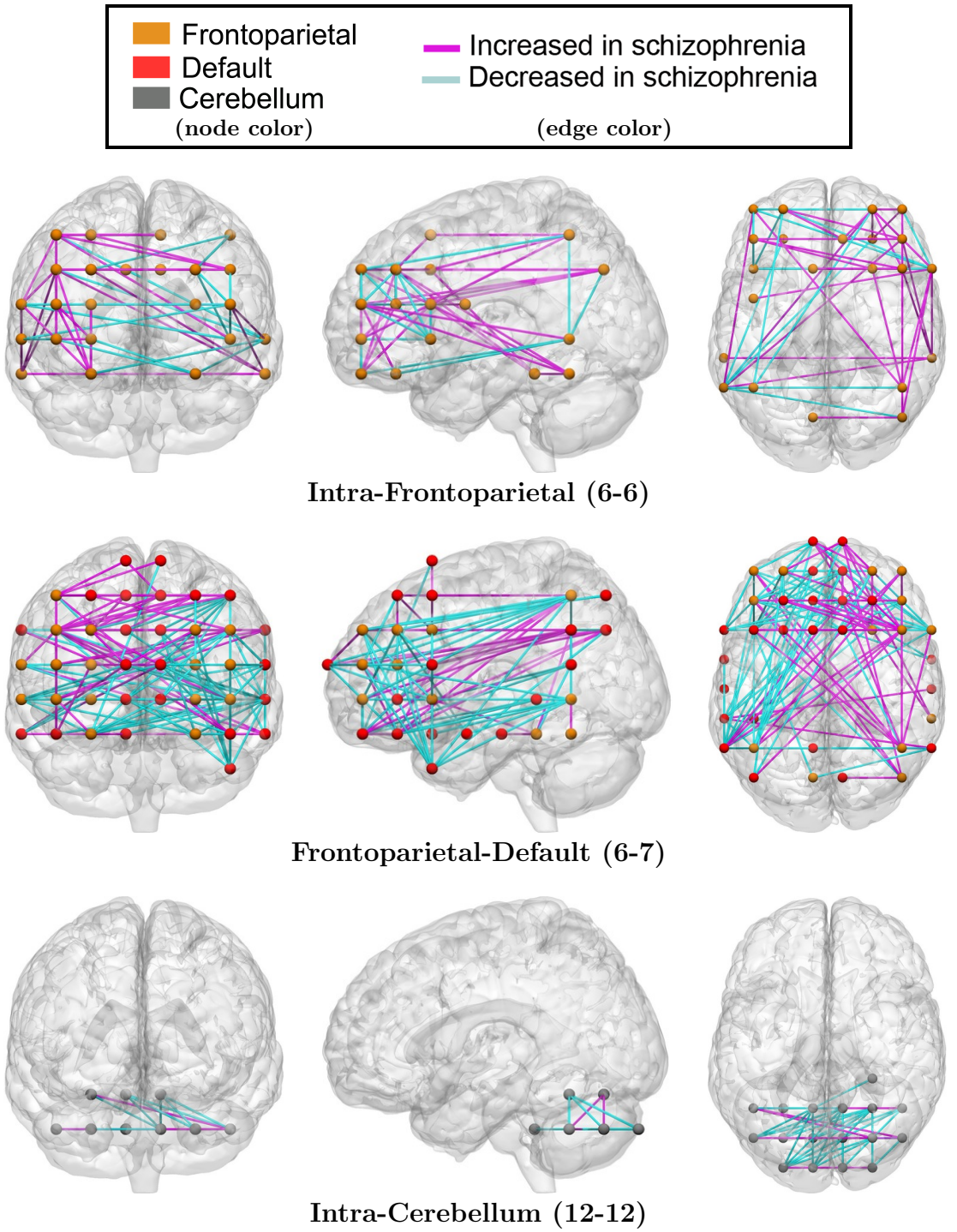


Figure 11: Nonzero edge values of the median weight vector generated from the fused Lasso regularized SVM. For three sets of network-to-network connections, we rendered abnormal connections separately on anterior, sagittal, and axial views of a canonical brain. Notice the prominent involvement of lateral prefrontal regions in connections within frontoparietal network and in connections between frontoparietal network and default network.

et al., 2006). Thus, there is strong interest in using neuroimaging methods to establish connectivity-based biomarkers that accurately predict disorder status (Cohen et al., 2011; Klöppel et al., 2012; Sundermann et al., 2013). Multivariate methods that use whole-brain functional connectomes are particularly promising since they comprehensively look at the network structure of the entire brain (Castellanos et al., 2013; Fornito et al., 2012), but the massive size of connectomes requires some form of dimensionality reduction.

In this work, we developed and deployed a multivariate approach based on the SVM (Cortes and Vapnik, 1995) and regularization methods that leverage the 6-D spatial structure of the functional connectome, namely the fused Lasso (Tibshirani et al., 2005) and the GraphNet regularizer (Grosenick et al., 2013). In addition, we introduced a novel and scalable algorithm based on the classical alternating direction method (Boyd et al., 2011; Gabay and Mercier, 1976; Glowinski and Marroco, 1975) for solving the nonsmooth, large-scale optimization problem that results from the structured sparse SVM. Note that most existing multivariate methods in the literature rely on some form of *a priori* feature selection or feature extraction (*e.g.*, principal component analysis, locally linear embedding) before invoking some "off the shelf" classifier (*e.g.*, nearest-neighbor, SVM, linear discriminant analysis) (Castellanos et al., 2013). In contrast, our feature selection method is not only spatially informed, but is also *embedded* (Guyon and Elisseeff, 2003), meaning that feature selection is conducted together with model fitting. This type of joint feature selection and classification has been rarely applied in the disease prediction framework with functional connectomes.

We used a grid-based parcellation scheme for producing whole-brain resting state functional connectomes (see Section 2.1), and this has two advantages. First, it endows a natural ordering and a notion of nearest neighbors among the coordinates of functional connectomes, which is important when defining the neighborhood set for fused Lasso and GraphNet (one may consider predefining an arbitrary graph structured neighborhood set, but we prefer an approach that enforces little *a priori* assumption on the structure of the predictive regions). Second, the finite differencing matrix corresponding to this (augmented) functional connectome has a special structure that allows efficient FFT-based matrix inversion to be applied (this structure is absent when a functional or an anatomical based parcellation scheme is adopted). When this property is used in tandem with variable splitting, the inner subproblems associated with the proposed ADMM algorithm admit closed form solutions that can be carried out efficiently and non-iteratively.

Using a simulation method and a large real-world schizophrenia dataset, we demonstrate that the proposed spatially-informed regularization methods can achieve accurate disease prediction with superior interpretability of discriminative features. To the best of our knowledge, this is the first application of structured sparse methods in the context of disease prediction using functional connectomes.

4.1. Rationale behind spatial regularization

The rationale for using the fused Lasso and GraphNet regularizer can be better appreciated by considering the "patchiness assumption" – the view that major psychiatric diseases manifest in the brain by impacting moderately-sized (*e.g.*, $1,000$ mm^3 to $30,000$ mm^3) spatially contiguous neural regions. This assumption has been repeatedly born out across different imaging modalities. In structural studies and task-based activation studies, theorists have consistently identified mid-sized blobs in maps of differences between patients and controls (Dickstein et al., 2006; Glahn et al., 2005; Wright et al., 2000). In studies of functional connectivity, the patchiness assumption has found clear support. The vast majority of previous connectivity studies are seed-based; they create maps of connectivity with a single or a handful of discrete seeds, and compare these maps between patients and controls. These studies nearly always report connectivity between patients and controls is altered at one or more discrete medium-sized blobs, similar to structural studies and activation-based studies (Etkin and Wager, 2007; van den Heuvel and Pol, 2010; Konrad and Eickhoff, 2010).

In addition to actual findings from previous connectivity studies, the patchiness assumption is justified by careful examination of the hypotheses proposed by theorists. It is exceedingly common for theorists to state their hypotheses in terms of altered connectivity between two discrete regions or discrete sets of regions. For example, based on hypofrontality models of auditory hallucinations in schizophrenia, Lawrie and colleagues (Lawrie et al., 2002) predicted that individuals with schizophrenia would exhibit decreased connectivity between dorsal lateral prefrontal cortex (DLPFC; Brodman's areas 9 and 10), involved in top-down control, and superior temporal gyrus (STG), which is involved in auditory processing. Both DLPFC and STG are large structures, and they encompass roughly a dozen nodes each in our grid-based parcellation. If Lawrie and colleagues' conjecture is correct, then we should observe alterations in connectivity between the large set of connections that link the nodes that fall within the respective brain structures. Moreover, Lawrie and colleagues' hypothesis implies that the predicted changes will be relatively discrete and localized to connections linking these two regions. For example, the finding of salt and pepper changes throughout the connectome would of course not support their conjecture. Moreover, their hypothesis predicts that even regions that are relatively close to dorsal lateral prefrontal cortex, for example precentral gyrus, involved in motor processing, do not change their connectivity with STG – the connectivity changes they predict are relatively localized and discrete.

In addition to hypotheses about region-to-region abnormalities, the patchiness assumption is also evident in recent network models of mental disorders. In recent years, theorists have recognized that the human brain is organized into large-scale networks that operate as cohesive functional units (Bressler and Menon, 2010; Laird et al., 2011; Yeo et al., 2011). Each individual network is composed of a set of discrete regions, and each region itself encompasses multiple nodes given a standard, suitably dense parcellation scheme (such as our grid-based scheme). Concurrent with the rise of this network understanding of neural organization, theorists have proposed models in which psychiatric disorders are seen to involve perturbations in the interrelationships between individual pairs of network, where the remainder of the network interrelationships remain essentially unaffected (Lynall et al., 2010; Menon, 2011; Tu et al., 2013). If these network models of disease are correct, then using functional connectivity methods, we should discover that in a psychiatric disease that is proposed to affect the interrelationship between network A and network B, the set of regions that make up network A change their relationship with the set of regions in network B. The regions that abut the regions in networks A and B are, by hypothesis, not proposed to alter their connectivity. In connectomic space, this pattern would be represented as patchy changes in the sets of connections linking the blobs of contiguous nodes that represent networks A and B, with the remainder of the connectome remaining largely unaffected.

In sum, actual results from structural, task-based, and connectivity studies suggest the patchiness assumption is reasonable, while close examination of the form of the hypotheses routinely made by psychiatric researchers suggests the assumption underlies theorists' conjectures about disease processes. If these claims are correct, then this provides a powerful rationale for both the fused Lasso and GraphNet penalty. Fused Lasso penalizes abrupt discontinuities, favoring the detection of piecewise constant patches in noisy contexts. Similarly, GraphNet also promotes spatial contiguity, but encourages the clusters to appear in smoother form. Given that there is a solid basis for expecting that the disease discriminative patterns in functional connectomes will consist of spatially contiguous patches, rather than consisting of salt-and-pepper patterns randomly dispersed throughout the brain, then fused Lasso and GraphNet are well very positioned to uncover these patchy discriminative signatures. In addition, the spatial coherence promoted by these spatially-informed regularizers helps decrease model complexity and facilitates interpretation.

4.2. Simulation study and interpretability of results

The analytic intuitions discussed above were confirmed in our simulation study. Here, we imposed "patchiness" in the ground truth by introducing clusters of *anomalous nodes* in the synthetic

functional connectomes that represent the patient group (see Section 2.4). For comparison, we learned SVM classifiers from the training data using the hinge-loss and one of the following regularizers: Lasso, Elastic-net, GraphNet, and fused Lasso. Our results indicate that fused Lasso and GraphNet not only improved classification accuracy, but also exhibited superior performance in recovering the discriminatory edges with respect to their non-spatially informed counterparts, Lasso and Elastic-net.

4.3. Application: classifying healthy controls vs. schizophrenic subjects

Our results indicate that at similar sparsity level, the classification accuracy with Elastic-net, GraphNet, and fused Lasso are comparable. However, studying the structure of the learned weight vectors reveals the key advantage of GraphNet and fused Lasso: they facilitate interpretation by promoting sparsity patterns that are spatially contiguous in the connectome space. Fused Lasso recovers highly systematic sparsity patterns with multiple spatially contiguous clusters, including nodes with diffuse connectivity profiles, which is one manifestation of the "patchiness assumption" discussed earlier. On the other hand, the smooth sparsity structure that GraphNet recovers is biologically more sensible than the salt-and-pepper like structure yielded by the Elastic-net. These decreases in model complexity come without sacrificing prediction accuracy, which fits well with the principle of *Occam's razor* – given multiple equally predictive models, the simplest choice should be selected.

Finally, additional evidence that fused Lasso recovered more interpretable discriminative features for the schizophrenia dataset comes from comparing visualizations of the respective weight vectors from the three regularizers (see Fig. 10). The map of the fused Lasso support shows more prominent and clearly localized alterations in connectivity involving frontoparietal network, default network, and cerebellum, among other regions. These networks also exhibited increased node degree, indicating diffuse connectivity alterations with other networks. Interestingly, these networks are among the most commonly implicated in schizophrenia. Frontoparietal network, which has multiple important hubs in prefrontal cortex, is involved in executive processing and cognitive control (Cole et al., 2013), and has been shown to exhibit abnormal activation (see Minzenberg et al. (2009) for a quantitative meta-analysis) and connectivity (Repovs et al. (2011); Tu et al. (2013); see Fornito et al. (2012) for a review) in schizophrenia. Fused Lasso also recovered altered connectivity between frontoparietal network and default mode network, an important brain network involved in autobiographical memory and internally generated mental simulations (Buckner et al., 2008; Raichle et al., 2001). The weight vectors shown in Fig. 10 and the 3-D brains shown in Fig. 11 evidence a substantial number of aberrant connections between frontoparietal network and default network, with a predominance of reduced connectivity in schizophrenia. Frontoparietal network and default network become more interconnected throughout childhood and adolescence (Anderson et al., 2011; Fair et al., 2007), which might reflect development of top-down cognitive control by frontoparietal regions over default network. Reduced connectivity between these two networks is among the most commonly observed findings in connectivity research in schizophrenia (Jafri et al., 2008; Repovs et al., 2011; Woodward et al., 2011; Zhou et al., 2007a,b), and has been proposed to reflect disruptions and/or delays in normal trajectories of maturation (Repovs et al., 2011). It is also noteworthy that a sizable portion of the aberrant connection within frontoparietal cortex and between frontoparietal network and default network involved dorsal lateral prefrontal cortex (see results in Sec. 3.2). This region is perhaps the most frequently described as being abnormal in schizophrenia (Bunney and Bunney, 2000; Callicott et al., 2000; Zhou et al., 2007a). A third network highlighted by fused Lasso is cerebellum, which is featured in the influential 'cognitive dysmetria' hypothesis of schizophrenia (Andreasen et al., 1998). Abnormalities in cerebellum have been found in post-mortem (Weinberger et al., 1980), structural (Wassink et al., 1999), and functional connectivity studies (Mamah et al., 2013).

Fused Lasso also tended to generate more sparsity in regions of the connectome that are not associated with schizophrenia pathology. For example, connectivity abnormalities in somatomotor

network, and in particular its interconnections with attention network and frontoparietal network, have as far as we know not been described in previous schizophrenia connectivity studies. The same is true of the nodes that fell outside the Yeo parcellation augmented with subcortical regions and cerebellum. These too have not been associated with schizophrenia pathology and tended to be sparser with fused Lasso. Overall, fused Lasso appeared to identify regions known from prior research to be involved in schizophrenia and appeared to generate more sparsity outside of these regions, providing some corroboration for the interpretability of fused Lasso findings.

4.4. Future Directions

While the spatially-informed disease prediction framework we introduced is capable of yielding predictive and highly interpretable results, there are several open questions that remain for future investigation. For example, with little modification, the variable splitting and the data augmentation procedure we introduced should be applicable to the isotropic TV penalty, which also promotes spatial contiguity (Wang et al., 2008b). This is important because on one hand, fused Lasso lacks the rotational invariance property of the isotropic TV penalty, whereas on the other hand, isotropic TV penalty is known to introduce artifacts at corner structured regions (Birkholz, 2011; Grasmair and Lenzen, 2010). Therefore, fused Lasso and isotropic TV penalty can both potentially be problematic for connectomic investigations, and a thorough comparison between these two penalties with our functional connectome data would be an important direction for future investigation. In addition, there are multiple works that have introduced a framework for achieving structured sparsity by coupling the isotropic TV penalty with the differentiable logistic loss function (Baldassarre et al., 2012; Gramfort et al., 2013; Michel et al., 2011). Although our method has the advantage that it can handle non-differentiable loss functions and hence the SVM, the algorithm employed in the above works enjoy a faster rate of convergence than the ADMM algorithm we employ (Beck and Teboulle, 2009; He and Yuan, 2012). Investigating ways to accelerate our proposed ADMM algorithm will be important for future work (Deng and Yin, 2012; Goldstein et al., 2012).

There are several other interesting extensions that remain for future research as well. First, functional and anatomical parcellations (which lack a grid structure and hence the BCCB structure) are often used in connectomic investigations. Future work should extend our methodology so the ADMM subproblems can be solved efficiently in analytic form even when a irregularly structured parcellation scheme is used (although the ADMM algorithm proposed by Ye and Xie (2011) is applicable in this setup, their approach requires an iterative update to be used to numerically solve one of the ADMM subproblems). Furthermore, with the emergence of various data sharing projects in the neuroimaging community such as Autism Brain Imaging Data Exchange (ABIDE) (Di Martino et al., 2013), ADHD-200 (The ADHD-200 Consortium, 2012), 1000 Functional Connectomes Project, and the International Neuroimaging Data-sharing Initiative (INDI) (Mennes et al., 2013), there is a need for a principled framework to handle the heterogeneity introduced by aggregating the data from multiple imaging centers. Toward this end, we are seeking ways to combine the currently presented spatial regularization scheme and multi-task learning (Caruana, 1997), where the tasks correspond to the imaging centers from which the resting state scans originate. One particular approach we have in mind for this is to replace the ℓ_1-regularizer in the objective function (5) with the ℓ_1/ℓ_2 mixed-norm regularizer (Gramfort et al., 2012; Lounici et al., 2009), which encourages the weight vectors across the different tasks to share similar sparsity patterns (a structure often referred to as block-sparsity). Our proposed ADMM algorithm can easily be modified to handle this change, as this simply amounts to replacing the scalar soft-threshold operator for the $\boldsymbol{v_2}$ update (25) with the vector soft-threshold operator (see Gramfort et al. (2012)). Finally, a more sophisticated approach for parameter tuning is needed, ideally a model selection strategy that provides statistical guarantees (Cawley and Talbot, 2010). Resampling-based approaches (Bach, 2008; Varoquaux et al., 2012) such as stability selection (Meinshausen and Bühlmann, 2010) may be considered, albeit these methods can be computationally demanding in high dimension.

5. Conclusions

In this work, we introduced a regularized ERM framework that explicitly accounts for the 6-D spatial structure in the connectome via the fused Lasso and the GraphNet regularizer. We demonstrate that our method recovers sparse and highly interpretable patterns across the connectome while maintaining predictive power, and thus could generate new insights into how psychiatric disorders impact brain networks.

Acknowledgments

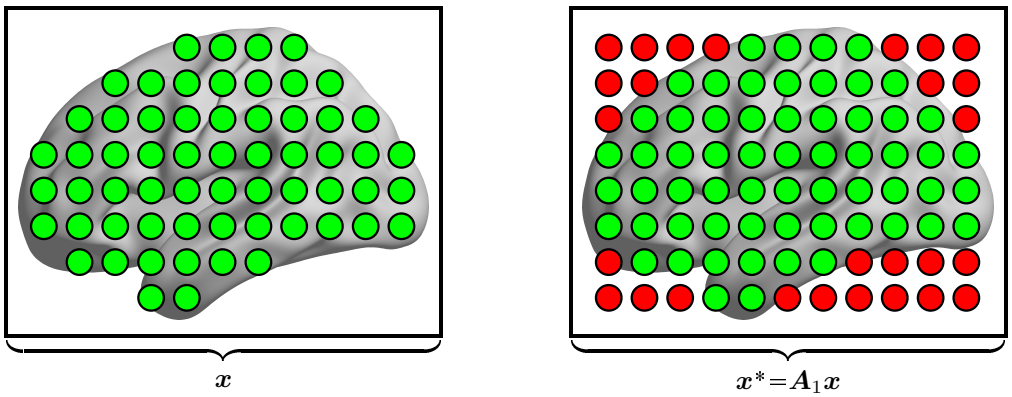
T. Watanabe and C. Scott's research was supported by NIH grant P01CA087634 and by NSF Grant CCF 1217880. C. Sripada's research was supported by NIH grant K23-AA-020297, Center for Computational Medicine Pilot Grant, and the John Templeton Foundation. The authors would like to thank A. Hero and J. Fessler, University of Michigan, for the valuable discussions and their insightful feedbacks. The authors would also like to thank Robert C. Welsh, University of Michigan, for providing us with `ConnTool`, a functional connectivity analysis package used to generate the functional connectome data.

Appendix A. Details on the data augmentation scheme

As discussed in Sec. 2.3.2, the augmentation matrix $\boldsymbol{A} \in \mathbb{R}^{\tilde{p} \times p}$ aims to rectify the irregularities in the Laplacian matrix $\boldsymbol{C}^T\boldsymbol{C}$. To gain a better understanding about $\boldsymbol{A}$, it is best to think of it as a concatenation of two matrices, $\boldsymbol{A} = \boldsymbol{A}_2\boldsymbol{A}_1$. We refer to $\boldsymbol{A}_1 \in \mathbb{R}^{p^* \times p}$ and $\boldsymbol{A}_2 \in \mathbb{R}^{\tilde{p} \times p^*}$ as the *first level* and the *second level* augmentation matrix respectively.

Role of $\boldsymbol{A}_1$. The first source of irregularities is that the nodes defining the functional connectome $\boldsymbol{x} \in \mathbb{R}^p$ are placed only on the brain, not the entire rectangular FOV. As a consequence, $\boldsymbol{x}$ only contains edges among the nodes placed on the support of the brain (represented by the green nodes in Fig. A.1). To fix these irregularities, $\boldsymbol{A}_1$ pads extra zero entries on $\boldsymbol{x}$ to create an *intermediate* augmented connectome $\boldsymbol{x}^* = \boldsymbol{A}_1\boldsymbol{x}$, where $\boldsymbol{x}^* \in \mathbb{R}^{p^*}$. Here, $\boldsymbol{x}^*$ can be treated as if the nodes were placed throughout the entire rectangular FOV; the red nodes in Fig. A.1 represent a set of *ghost nodes* that were not originally present. The coordinates of $\boldsymbol{x}^*$ contain all possible edges between the *ghost nodes* and the original set of nodes, where the edges connected with the *ghost nodes* have zero values.

Role of $\boldsymbol{A}_2$. The second source of irregularities is that $\boldsymbol{x}$ (and $\boldsymbol{x}^*$) lack a complete 6-D representation since it only contains the lower-triangular part of the cross-correlation matrix. Consequently, the coordinates of $\boldsymbol{x}^*$ lack symmetry, as their entries only contain edges for the following set of 6-D coordinate points: $\{(\boldsymbol{r}_j, \boldsymbol{r}_k) \mid j > k\}$, where $\boldsymbol{r}_j = (x_j, y_j, z_j)$ and $\boldsymbol{r}_k = (x_k, y_k, z_k)$ are the 3-D locations of the node-pairs defining the edges. Matrix $\boldsymbol{A}_2$ fixes this asymmetry by padding zero entries to fill in for the 6-D coordinate points $\{(\boldsymbol{r}_j, \boldsymbol{r}_k) \mid j \leqslant k\}$, which correspond to the diagonal and the upper-triangular entries in the cross-correlation matrix that were disposed due to redundancy (see Fig. A.2). Applying $\boldsymbol{A}_2$ on $\boldsymbol{x}^* = \boldsymbol{A}_1\boldsymbol{x}$ provides the desired augmented functional connectome $\tilde{\boldsymbol{x}} = \boldsymbol{A}_2\boldsymbol{x}^* = \boldsymbol{A}\boldsymbol{x}$, and similarly the augmented weight vector $\tilde{\boldsymbol{w}} = \boldsymbol{A}\boldsymbol{w}$. Here, $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{w}}$ contain the full set of 6-D coordinate points $\{(\boldsymbol{r}_j, \boldsymbol{r}_k) \mid j, k \in [d]\}$, where d is the total number of nodes on the rectangular FOV including the *ghost nodes* (*i.e.*, both the green and the red nodes in Fig. A.1). Note that dimension $\tilde{p}$ of the augmented functional connectome is $\tilde{p} = d^2$, and the total number of adjacent coordinates $\tilde{e}$ in this augmented 6-D connectome space is $\tilde{e} = 6\tilde{p}$.



(a) Original Functional connectome

(b) Intermediate augmented connectome

Figure A.1: The effect of the first level augmentation matrix $\boldsymbol{A}_1$. **Left:** the original functional connectome $\boldsymbol{x}$ only contains edges between the nodes placed on the support of the brain (represented by the green nodes). **Right:** $\boldsymbol{A}_1$ pads extra zero entries on $\boldsymbol{x}$ to create the intermediate augmented connectome $\boldsymbol{x}^*$. Here, $\boldsymbol{x}^*$ can be treated as if the nodes were placed throughout the entire rectangular FOV (the red bubbles represent nodes that are outside the brain support), as its entries contain all possible edges between the green and red nodes; the edges that connect with the red nodes all have zero values.

$$\boldsymbol{x}^* = \left[\begin{array}{c} \boldsymbol{x}^*(\boldsymbol{r}_2,\boldsymbol{r}_1) \\ \boldsymbol{x}^*(\boldsymbol{r}_3,\boldsymbol{r}_1) \\ \vdots \\ \boldsymbol{x}^*(\boldsymbol{r}_d,\boldsymbol{r}_1) \\ \hline \boldsymbol{x}^*(\boldsymbol{r}_3,\boldsymbol{r}_2) \\ \boldsymbol{x}^*(\boldsymbol{r}_4,\boldsymbol{r}_2) \\ \vdots \\ \boldsymbol{x}^*(\boldsymbol{r}_d,\boldsymbol{r}_2) \\ \hline \vdots \\ \hline \boldsymbol{x}^*(\boldsymbol{r}_d,\boldsymbol{r}_{d-1}) \end{array}\right], \qquad \tilde{\boldsymbol{x}} = \boldsymbol{A}_2\boldsymbol{x}^* = \left[\begin{array}{c} \tilde{\boldsymbol{x}}(\boldsymbol{r}_1,\boldsymbol{r}_1) \\ \tilde{\boldsymbol{x}}(\boldsymbol{r}_2,\boldsymbol{r}_1) \\ \tilde{\boldsymbol{x}}(\boldsymbol{r}_3,\boldsymbol{r}_1) \\ \vdots \\ \tilde{\boldsymbol{x}}(\boldsymbol{r}_d,\boldsymbol{r}_1) \\ \hline \tilde{\boldsymbol{x}}(\boldsymbol{r}_1,\boldsymbol{r}_2) \\ \tilde{\boldsymbol{x}}(\boldsymbol{r}_2,\boldsymbol{r}_2) \\ \tilde{\boldsymbol{x}}(\boldsymbol{r}_3,\boldsymbol{r}_2) \\ \vdots \\ \tilde{\boldsymbol{x}}(\boldsymbol{r}_d,\boldsymbol{r}_2) \\ \hline \vdots \\ \hline \tilde{\boldsymbol{x}}(\boldsymbol{r}_d,\boldsymbol{r}_d) \end{array}\right] = \left[\begin{array}{c} 0 \\ \boldsymbol{x}^*(\boldsymbol{r}_2,\boldsymbol{r}_1) \\ \boldsymbol{x}^*(\boldsymbol{r}_3,\boldsymbol{r}_1) \\ \vdots \\ \boldsymbol{x}^*(\boldsymbol{r}_d,\boldsymbol{r}_1) \\ \hline 0 \\ 0 \\ \boldsymbol{x}^*(\boldsymbol{r}_3,\boldsymbol{r}_2) \\ \vdots \\ \boldsymbol{x}^*(\boldsymbol{r}_d,\boldsymbol{r}_2) \\ \hline \vdots \\ \hline 0 \end{array}\right]$$

(a) Intermediate augmented connectome

(b) Augmented functional connectome

Figure A.2: The effect of the second level augmentation matrix $\boldsymbol{A}_2$. The entries of $\boldsymbol{x}^*$ represent edges localized by 6-D coordinate points $\{(\boldsymbol{r}_j,\boldsymbol{r}_k) \mid j > k\}$, where $\boldsymbol{r}_j = (x_j, y_j, z_j)$ and $\boldsymbol{r}_k = (x_k, y_k, z_k)$ are the 3-D locations of the node pairs defining the edges. $\boldsymbol{A}_2$ fixes the asymmetry in the coordinates of $\boldsymbol{x}^*$ by padding zero entries to accommodate for the 6-D coordinate points $\{(\boldsymbol{r}_j,\boldsymbol{r}_k) \mid j \leqslant k\}$; these are the diagonal and the upper-triangular entries in the cross-correlation matrix that were disposed for redundancy.

Appendix B. ADMM updates for Elastic-net

The unconstrained formulation of the Elastic-net regularized ERM problem reads

$$\underset{\boldsymbol{w}\in\mathbb{R}^p}{\arg\min}\ \frac{1}{n}\mathcal{L}(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}) + \lambda\left\|\boldsymbol{w}\right\|_1 + \frac{\gamma}{2}\left\|\boldsymbol{w}\right\|_2^2\ ,$$

which can be converted into the following equivalent constrained formulation:

$$\underset{\boldsymbol{w},\boldsymbol{v_1}\boldsymbol{v_2}}{\text{minimize}}\, \frac{1}{n}\mathcal{L}(\boldsymbol{v_1}) + \lambda \left\|\boldsymbol{v_2}\right\|_1 + \frac{\gamma}{2}\left\|\boldsymbol{w}\right\|_2^2 \;\; \text{subject to } \boldsymbol{YXw} = \boldsymbol{v_1},\; \boldsymbol{w} = \boldsymbol{v_2}\;. \tag{B.1}$$

With this variable splitting scheme, the correspondence with the ADMM formulation (6) is

$$\bar{\boldsymbol{f}}(\bar{\boldsymbol{x}}) = \tfrac{\gamma}{2}\left\|\boldsymbol{w}\right\|_2^2, \quad \bar{\boldsymbol{g}}(\bar{\boldsymbol{y}}) = \frac{1}{n}\mathcal{L}(\boldsymbol{v_1}) + \lambda\left\|\boldsymbol{v_2}\right\|_1$$

$$\bar{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{YX} \\ \boldsymbol{I} \end{bmatrix}, \quad \bar{\boldsymbol{x}} = \boldsymbol{w}, \quad \bar{\boldsymbol{B}} = -\boldsymbol{I}, \quad \bar{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{v_1} \\ \boldsymbol{v_2} \end{bmatrix}\;.$$

and the ADMM updates for $\bar{\boldsymbol{x}}$ (7) and $\bar{\boldsymbol{y}}$ (8) decomposes into subproblems

$$\begin{aligned}
\boldsymbol{w}^{(t+1)} &\leftarrow \arg\min_{\boldsymbol{w}} \left\{ \frac{\gamma}{2}\left\|\boldsymbol{w}\right\|^2 + \left\|\boldsymbol{YXw} - \left(\boldsymbol{v_1}^{(t)} - \boldsymbol{u_1}^{(t)}\right)\right\|^2 + \left\|\boldsymbol{w} - \left(\boldsymbol{v_2}^{(t)} - \boldsymbol{u_2}^{(t)}\right)\right\|^2 \right\} \\
\boldsymbol{v_1}^{(t+1)} &\leftarrow \arg\min_{\boldsymbol{v_1}} \left\{ \frac{1}{n}\mathcal{L}(\boldsymbol{v_1}) + \frac{\rho}{2}\left\|\boldsymbol{v_1} - \left(\boldsymbol{YX}\boldsymbol{w}^{(t+1)} + \boldsymbol{u_1}^{(t)}\right)\right\|^2 \right\} \\
\boldsymbol{v_2}^{(t+1)} &\leftarrow \arg\min_{\boldsymbol{v_2}} \left\{ \lambda\left\|\boldsymbol{v_2}\right\|_1 + \frac{\rho}{2}\left\|\boldsymbol{v_2} - \left(\boldsymbol{w}^{(t+1)} + \boldsymbol{u_2}^{(t)}\right)\right\|^2 \right\}\;.
\end{aligned}$$

The update for $\boldsymbol{w}$ is

$$\boldsymbol{w}^{(t+1)} \leftarrow \left(\rho\boldsymbol{X}^T\boldsymbol{X} + [\gamma+\rho]\boldsymbol{I}_p\right)^{-1}\left(\rho\boldsymbol{X}^T\boldsymbol{Y}^T[\boldsymbol{v_1}^{(t)} - \boldsymbol{u_1}^{(t)}] + \rho[\boldsymbol{v_2}^{(t)} - \boldsymbol{u_2}^{(t)}]\right)$$

which can be solved efficiently via inversion Lemma (20). The update for $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ is identical to (16) and (17) described in Sec. 2.3.3, which can be solved via coordinate-wise proximal operators (24) and (25). The dual variable update (9) is a trivial matrix-vector multiplication.

M. Afonso, J. Bioucas-Dias, M.A.T. Figueiredo, Fast image recovery using variable splitting and constrained optimization, IEEE Trans. Image Proc. 19 (2010).

M. Allison, S. Ramani, J. Fessler, Accelerated regularized estimation of MR coil sensitivities using Augmented Lagrangian methods, IEEE Trans. Med. Imaging 32 (2013).

J.S. Anderson, M.A. Ferguson, M. Lopez-Larson, D. Yurgelun-Todd, Connectivity gradients between the default mode and attention control networks., Brain Connectivity 1 (2011).

N.C. Andreasen, S. Paradiso, D.S. O'Leary, "Cognitive dysmetria" as an integrative theory of schizophrenia: A dysfunction in cortical-subcortical-cerebellar circuitry?, Schizophr. Bull. 24 (1998).

J. Ashburner, A fast diffeomorphic image registration algorithm, NeuroImage 38 (2007).

G. Atluri, K. Padmanabhan, G. Fang, M. Steinbach, J.R. Petrella, K. Lim, A.M. III, N.F. Samatova, P.M. Doraiswamy, V. Kumar, Complex biomarker discovery in neuroimaging data: Finding a needle in a haystack, Neuoimage: Clin. 3 (2013).

F.R. Bach, Bolasso: model consistent Lasso estimation through the bootstrap., Proc. Int. Conf. Mach. Learn. (2008).

L. Baldassarre, J. Mourao-Miranda, M. Pontil, Structured sparsity models for brain decoding from fMRI data, Workshop on Pattern Recognition and NeuroImaging (2012).

A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM J. Img. Sci. 2 (2009).

Y. Behzadi, K. Restom, J. Liau, T.T. Liu, A component based noise correction method (CompCor) for BOLD and perfusion based fMRI, NeuroImage 37 (2007).

H. Birkholz, A unifying approach to isotropic and anisotropic total variation denoising models, J. Comp. Appl. Mathematics 235 (2011).

J.M. Borwein, A.S. Lewis, Convex Analysis and Nonlinear Optimization: Theory and Examples, Springer Verlag, 2006.

S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (2011).

S.L. Bressler, V. Menon, Large-scale brain networks in cognition: emerging methods and principles, Trends Cogn. Sci. 14 (2010).

R.L. Buckner, J.R. Andrews-Hanna, D.L. Schacter, The brain's default network, Ann. N.Y. Acad. Sci. 1124 (2008).

P. Bühlmann, S. van de Geer, Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer series in statistics, Springer, 2011.

W.E. Bunney, B.G. Bunney, Evidence for a compromised dorsolateral prefrontal cortical parallel circuit in schizophrenia, Brain Research Reviews 31 (2000).

J.H. Callicott, A. Bertolino, V.S. Mattay, F.J. Langheim, J. Duyn, R. Coppola, T.E. Goldberg, D.R. Weinberger, Physiological dysfunction of the dorsolateral prefrontal cortex in schizophrenia revisited, Schizophr. Res. 10 (2000).

E. Candes, M. Wakin, An introduction to compressive sampling, IEEE Trans. Signal Proc. Magazine 25 (2008).

M.K. Carroll, G.A. Cecchi, I. Rish, R. Garg, A.R. Rao, Prediction and interpretation of distributed neural activity with sparse models, NeuroImage 44 (2009).

R. Caruana, Multitask learning, Mach. Learn. 28 (1997).

F.X. Castellanos, A.D. Martino, R.C. Craddock, A.D. Mehta, M.P. Milham, Clinical applications of the functional connectome, NeuroImage 80 (2013).

G.C. Cawley, N.L. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, J. Mach. Learn. Res. 11 (2010).

X. Chen, Q. Lin, S. Kim, J.G. Carbonell, E.P. Xing, Smoothing proximal gradient method for general structured sparse regression, Ann. Appl. Stat. 6 (2012).

J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2 ed., Routledge, 1988.

J.R. Cohen, R.F. Asarnow, F.W. Sabb, R.M. Bilder, S.Y. Bookheimer, B.J. Knowlton, R.A. Poldrack, Decoding continuous behavioral variables from neuroimaging data, Front. Neurosci. 5 (2011).

M.W. Cole, J.R. Reynolds, J.D. Power, G. Repovs, A. Anticevic, T.S. Braver, Multi-task connectivity reveals flexible hubs for adaptive task control, Nature Neurosci. 16 (2013).

C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995).

R.C. Craddock, P.E. Holtzheimer, X.P. Hu, H.S. Mayberg, Disease state prediction from resting state functional connectivity, Magn. Reson. Med. 62 (2009).

D. Dai, J. Wang, J. Hua, H. He, Classification of ADHD children through multimodal magnetic resonance imaging, Front. Neurosci. 6 (2012).

P. Davis, Circulant Matrices, Wiley, 1979.

W. Deng, W. Yin, On the global and linear convergence of the generalized alternating direction method of multipliers, Rice CAAM technical report TR12-14 (2012).

A. Di Martino, C.G. Yan, Q. Li, E. Denio, F.X. Castellanos, K. Alaerts, J.S. Anderson, M. Assaf, S.Y. Bookheimer, M. Dapretto, B. Deen, S. Delmonte, I. Dinstein, B. Ertl-Wagner, D.A. Fair, L. Gallagher, D.P. Kennedy, C.L. Keown, C. Keysers, J.E. Lainhart, C. Lord, B. Luna, V. Menon, N.J. Minshew, C.S. Monk, S. Mueller, R.A. Müeller, M.B. Nebel, J.T. Nigg, K. O'Hearn, K.A. Pelphrey, S.J. Peltier, J.D. Rudie, S. Sunaert, M. Thioux, J.M. Tyszka, L.Q. Uddin, J.S. Verhoeven, N. Wenderoth, J.L. Wiggins, S.H. Mostofsky, M.P. Milham, The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism., Mol. Psychiatry (2013).

S.G. Dickstein, K. Bannon, F. Xavier Castellanos, M.P. Milham, The neural correlates of attention deficit hyperactivity disorder: an ALE meta-analysis, J. Child Psychol. and Psychiatry 47 (2006).

D. Donoho, De-noising by soft-thresholding, IEEE Trans. Inf. Theory 41 (1995).

N.U.F. Dosenbach, B. Nardos, A.L. Cohen, D.A. Fair, J.D. Power, J.A. Church, S.M. Nelson, G.S. Wig, A.C. Vogel, C.N. Lessov-Schlaggar, K.A. Barnes, J.W. Dubis, E. Feczko, R.S. Coalson, J.R. Pruett, D.M. Barch, S.E. Petersen, B.L. Schlaggar, Prediction of individual brain maturity using fMRI, Science 329 (2010).

E. Etkin, T. Wager, Functional neuroimaging of anxiety: a meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia, Am. J. Pshychiatry 164 (2007).

D. Fair, J.T. Nigg, S. Iyer, D. Bathula, K.L. Mills, N.U. Dosenbach, B.L. Schlaggar, M. Mennes, D. Gutman, S. Bangaru, J.K. Buitelaar, D.P. Dickstein, A. Di Martino, D.N. Kennedy, C. Kelly, B. Luna, J.B. Schweitzer, K. Velanova, Y.F. Wang, S.H. Mostofsky, F.X. Castellanos, M.P. Milham, Distinct neural signatures detected for ADHD subtypes after controlling for micro-movements in resting state functional connectivity MRI data, Front. Syst. Neurosci. 6 (2013).

D.A. Fair, N.U.F. Dosenbach, J.A. Church, A.L. Cohen, S. Brahmbhatt, F.M. Miezin, D.M. Barch, M.E. Raichle, S.E. Petersen, B.L. Schlaggar, Development of distinct control networks through segregation and integration, Proc. Natl. Acad. Sci. 104 (2007).

J. Fan, J. Lv, A selective overview of variable selection in high dimensional feature space, Stat. Sinica 20 (2010).

A. Fornito, A. Zalesky, C. Pantelis, E.T. Bullmore, Schizophrenia, neuroimaging and connectomics, NeuroImage 62 (2012).

M.D. Fox, M. Greicius, Clinical applications of resting state functional connectivity, Front. Syst. Neurosci. 4 (2010).

D. Gabay, B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, Comput. Math. Appl. 2 (1976).

D.C. Glahn, J.D. Ragland, A. Abramoff, J. Barrett, A.R. Laird, C.E. Bearden, D.I. Velligan, Beyond hypofrontality: A quantitative meta-analysis of functional neuroimaging studies of working memory in schizophrenia, Hum. Brain Mapp. 25 (2005).

R. Glowinski, A. Marroco, Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires, Revue Française d'Automatique, Informatique, et Recherche Opérationelle 9 (1975).

T. Goldstein, B. O'Donoghue, S. Setzer, Fast alternating direction optimization methods, CAM report (2012).

A. Gramfort, M. Kowalski, M. Hämäläinen, Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods, Physics in Medicine and Biology 57 (2012).

A. Gramfort, B. Thirion, G. Varoquaux, Identifying predictive regions from fMRI with TV-L1 prior, Workshop on Pattern Recognition and NeuroImaging (2013).

A. Gramfort, G. Varoquaux, B. Thirion, Beyond brain reading: Randomized sparsity and clustering to simultaneously predict and identify, Machine Learning and Interpretation in Neuroimaging (2011).

M. Grasmair, F. Lenzen, Anisotropic total variation filtering, Appl. Mathematics and Optimization 62 (2010).

L. Grosenick, S. Greer, B. Knutson, Interpretable classifiers for fMRI improve prediction of purchases, IEEE Trans. Neural Syst. Rehabil. Eng. 16 (2008).

L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, J.E. Taylor, Interpretable whole-brain prediction analysis with GraphNet, NeuroImage 72 (2013).

I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003).

T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2 ed., Springer, 2009.

B. He, X. Yuan, On the O(1/n) Convergence Rate of the Douglas-Rachford Alternating Direction Method., SIAM J. Numer. Anal. 50 (2012).

M. van den Heuvel, H.H. Pol, Exploring the brain network: A review on resting-state fMRI functional connectivity, Eur. Neuropsychopharmacol. 20 (2010).

M.J. Jafri, G.D. Pearlson, M. Stevens, V.D. Calhoun, A method for functional network connectivity among spatially independent resting-state components in schizophrenia, NeuroImage 39 (2008).

A. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000).

W. James, J. Stein, Estimation with quadratic loss, Proc. Third Berkeley Symp. Math. Stat. and Probab. (1961).

R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, E. Eger, F. Bach, B. Thirion, Multi-scale mining of fMRI data with hierarchical structured sparsity, SIAM J. Imaging Sci. 5 (2012).

S. Klöppel, A. Abdulkadir, C.R.J. Jr., N. Koutsouleris, J. Mourao-Miranda, P. Vemuri, Diagnostic neuroimaging across diseases, NeuroImage 61 (2012).

K. Konrad, S.B. Eickhoff, Is the ADHD brain wired differently? a review on structural and functional connectivity in attention deficit hyperactivity disorder, Hum. Brain Mapp. 31 (2010).

A.R. Laird, P.M. Fox, S.B. Eickhoff, J.A. Turner, K.L. Ray, D.R. McKay, D.C. Glahn, C.F. Beckmann, S.M. Smith, P.T. Fox, Behavioral interpretations of intrinsic connectivity networks., J. Cogn. Neurosci. 23 (2011).

S.M. Lawrie, C. Buechel, H.C. Whalley, C.D. Frith, K.J. Friston, E.C. Johnstone, Reduced frontotemporal functional connectivity in schizophrenia associated with auditory hallucinations, Biol. Psychiatry 51 (2002).

K. Lounici, M. Pontil, A.B. Tsybakov, S.A. van de Geer, Taking advantage of sparsity in multi-task learning, Conf. Learn. Theory (2009).

M.E. Lynall, D.S. Bassett, R. Kerwin, P.J. McKenna, M. Kitzbichler, U. Muller, E. Bullmore, Functional Connectivity and Brain Networks in Schizophrenia, J. Neurosci. 30 (2010).

J. Mairal, R. Jenatton, G. Obozinski, F. Bach, Convex and network flow optimization for structured sparsity, J. Mach. Learn. Res. 12 (2011).

D. Mamah, D.M. Barch, G. Repovs, Resting state functional connectivity of five neural networks in bipolar disorder and schizophrenia, J. Affect. Disord. 150 (2013).

A. Matakos, S. Ramani, J. Fessler, Accelerated edge-preserving image restoration without boundary artifacts, IEEE Trans. Image Proc. 22 (2013).

N. Meinshausen, P. Bühlmann, Stability selection, J. R. Stat. Soc. Ser. B Stat. Methodol. 72 (2010).

M. Mennes, B.B. Biswal, F.X. Castellanos, M.P. Milham, Making data sharing work: The FCP/INDI experience, NeuroImage 82 (2013).

V. Menon, Large-scale brain networks and psychopathology: a unifying triple network model, Trends Cogn. Sci. 15 (2011).

C.A. Micchelli, J.M. Morales, M. Pontil, Regularizers for structured sparsity, Adv. Comput. Math. 38 (2013).

V. Michel, A. Gramfort, G. Varoquaux, E. Eger, B. Thirion, Total variation regularization for fMRI-based prediction of behavior, IEEE Trans. Med. Imaging 30 (2011).

S. Minsker, Geometric median and robust estimation in banach spaces, Preprint, arXiv:1308.1334 (2013).

M. Minzenberg, A. Laird, S. Thelen, C. Carter, D. Glahn, Meta-analysis of 41 functional neuroimaging studies of executive function in schizophrenia, Arch. Gen. Psychiatry 66 (2009).

J. Mota, J. Xavier, P. Aguiar, M. Püschel, A proof of convergence for the alternating direction method of multipliers applied to polyhedral-constrained functions, Preprint, arXiv:1112.2295 (2011).

R.A. Müller, P. Shih, B. Keehn, J.R. Deyoe, K.M. Leyden, D.K. Shukla, Underconnected, but how? a survey of functional connectivity MRI studies in autism spectrum disorders, Cereb. Cortex 21 (2011).

J.D. Power, A.L. Cohen, S.M. Nelson, G.S. Wig, K.A. Barnes, J.A. Church, A.C. Vogel, T.O. Laumann, F.M. Miezin, B.L. Schlaggar, S.E. Petersen, Functional network organization of the human brain, Neuron 72 (2011).

M.E. Raichle, A.M. MacLeod, A.Z. Snyder, W.J. Powers, D.A. Gusnard, G.L. Shulman, A default mode of brain function, Proc. Natl. Acad. Sci. 98 (2001).

G. Repovs, J. Csernansky, D. Barch, Brain network connectivity in individuals with schizophrenia and their siblings, Biol. Psychiatry 69 (2011).

R.T. Rockafellar, R.J.B. Wets, Variational Analysis, Springer, 1998.

S. Ryali, K. Supekar, D.A. Abrams, V. Menon, Sparse logistic regression for whole-brain classification of fMRI data, NeuroImage 51 (2010).

W.R. Shirer, S. Ryali, E. Rykhlevskaia, V. Menon, M.D. Greicius, Decoding subject-driven cognitive states with whole-brain connectivity patterns, Cerebral Cortex (2011).

C. Sripada, M. Angstadt, D. Kessler, K.L. Phan, I. Liberzon, G.W. Evans, R. Welsh, P. Kim, J.E. Swain, Volitional regulation of emotions produces distributed alterations in connectivity between visual, attention control, and default networks, NeuroImage (2013a).

C. Sripada, D. Kessler, Y. Fang, K. Kumar, M. Angstadt, Disrupted network architecture of the resting brain in attention-deficit/hyperactivity disorder, Human Brain Mapping (accepted) (2014).

C. Sripada, D. Kessler, R. Welsh, M. Angstadt, I. Liberzon, K.L. Phan, C. Scott, Distributed effects of methylphenidate on the network structure of the resting brain: A connectomic pattern classification analysis, NeuroImage 81 (2013b).

K.E. Stephan, T. Baldeweg, K.J. Friston, Synaptic plasticity and dysconnection in schizophrenia, Biol. Psychiatry 59 (2006).

B. Sundermann, D. Herr, W. Schwindt, B. Pfleiderer, Multivariate classification of blood oxygen level-dependent fMRI data with diagnostic intention: A clinical perspective, Am. J. Neuroradiol. (2013).

The ADHD-200 Consortium, The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience, Front. Syst. Neurosci. 6 (2012).

R. Tibshirani, Regression shrinkage and selection via the Lasso, J. Roy. Stat. Soc. Ser. B 58 (1996).

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused Lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2005).

A. Tikhonov, Solution of incorrectly formulated problems and the regularization method, Soviet Math. Doklady 4 (1963).

P.C. Tu, Y.C. Lee, Y.S. Chen, C.T. Li, T.P. Su, Schizophrenia and the brain's control network: Aberrant within- and between-network connectivity of the frontoparietal network in schizophrenia, Schizophr. Res. 147 (2013).

N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, Automated Anatomical Labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain, NeuroImage 15 (2002).

G. Varoquaux, R.C. Craddock, Learning and comparing functional connectomes across subjects, NeuroImage 80 (2013).

G. Varoquaux, A. Gramfort, B. Thirion, Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering, Proc. Int. Conf. Mach. Learn. (2012).

L. Wang, J. Zhu, H. Zou, Hybrid huberized support vector machines for microarray classification and gene selection, Bioinforma. 24 (2008a).

Y. Wang, J. Yang, W. Yin, Y. Zhang, A new alternating minimization algorithm for total variation image reconstruction, SIAM J. Imaging Sci. 1 (2008b).

T. Wassink, N. Andreasen, P. Nopoulos, M. Flaum, Cerebellar morphology as a predictor of symptom and psychosocial outcome in schizophrenia, Biol. Psychiatry 45 (1999).

D. Weinberger, J. Kleinman, D. Luchins, L. Bigelow, R. Wyatt, Cerebellar pathology in schizophrenia: A controlled postmortem study, Am. J. Psychiatry 137 (1980).

M. West, Bayesian factor regression models in the "Large p, Small n" paradigm, Bayesian Stat. 7 (2003).

N.D. Woodward, B. Rogers, S. Heckers, Functional resting-state networks are differentially affected in schizophrenia, Schizophrenia Research 130 (2011).

I.C. Wright, S. Rabe-Hesketh, P.W. Woodruff, A.S. David, R.M. Murray, R.M. Murray, E.T. Bullmore, Meta-analysis of regional brain volumes in schizophrenia, Am. J. Psychiatry 157 (2000).

O. Yamashita, M. Sato, T. Yoshioka, F. Tong, Y. Kamitani, Sparse estimation automatically selects voxels relevant for the decoding of fmri activity patterns, NeuroImage 42 (2008).

G.B. Ye, X. Xie, Split bregman method for large scale fused Lasso, Comput. Stat. Data Anal. 55 (2011).

B. Yeo, F. Krienen, J. Sepulcre, M. Sabuncu, D. Lashkari, M. Hollinshead, J. Roffman, J. Smoller, L. Zöllei, J. Polimeni, B. Fischl, H. Liu, R. Buckner, The organization of the human cerebral cortex estimated by intrinsic functional connectivity., J. Neurophysiol. 106 (2011).

L. Zeng, H. Shen, L. Liu, L. Wang, B. Li, P. Fang, Z. Zhou, Y. Li, D. Hu, Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis., Brain 135 (2012).

Y. Zhou, M. Liang, T. Jiang, L. Tian, Y. Liu, Z. Liu, H. Liu, F. Kuang, Functional dysconnectivity of the dorsolateral prefrontal cortex in first-episode schizophrenia using resting-state fMRI, Neuroscience Letters 417 (2007a).

Y. Zhou, M. Liang, L. Tian, K. Wang, Y. Hao, H. Liu, Z. Liu, T. Jiang, Functional disintegration in paranoid schizophrenia using resting-state fMRI, Schizophr. Res. 97 (2007b).

H. Zou, T. Hastie, Regularization and variable selection via the Elastic Net, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2005).