Scalable Machine Learning Methods for Massive Biomedical Data Analysis

by

Takanori Watanabe

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2015

Doctoral Committee:

Associate Professor Clayton D. Scott, Co-Chair
Assistant Professor Chandra S. Sripada, Co-Chair
Professor Jeffrey A. Fessler
Professor Alfred O. Hero III
Professor Charles R. Meyer

To my parents.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Scalable Machine Learning Methods for Massive Biomedical Data Analysis

by

Takanori Watanabe

Chair: Clayton D. Scott

Co-chair: Chandra S. Sripada

Modern data acquisition techniques have enabled biomedical researchers to collect and analyze datasets of substantial size and complexity. The massive size of these datasets allows us to comprehensively study the biological system of interest at an unprecedented level of detail, which may lead to the discovery of clinically relevant *biomarkers*. Nonetheless, the dimensionality of these datasets presents critical computational and statistical challenges, as traditional statistical methods break down when the number of predictors dominates the number of observations, a setting frequently encountered in biomedical data analysis. This difficulty is compounded by the fact that biological data tend to be noisy and often possess complex correlation patterns among the predictors. The central goal of this dissertation is to develop a computationally tractable machine learning framework that allows us to extract scientifically meaningful information from these massive and highly complex biomedical datasets. We motivate the scope of our study by considering two important problems with clinical relevance: (1) uncertainty analysis for biomedical *image registration*, and (2) psychiatric disease prediction based on *functional* *connectomes*, which are high dimensional correlation maps generated from resting state functional MRI.

The first part of the dissertation concerns the problem of analyzing the level of uncertainty involved in biomedical image registration, where *image registration* is the process of finding the spatial transformation that best aligns the coordinates of an image pair. Toward this end, we introduce a data-driven method that allows one to visualize and quantify image registration uncertainty using spatially adaptive confidence regions, and demonstrate that empirical evaluations of the method on 2-D images yield promising results. At the heart of our proposed method is a novel shrinkage-based estimate of the distribution on deformation parameters.

The second part of the dissertation focuses on the supervised learning problem of binary classification, where the goal is to predict the psychiatric disorder status of an individual using functional connectomes derived from resting-state functional MRI. To address the dimensionality of the features, we introduce a regularized empirical risk minimization framework that allows us to encode various structures in the data. Specifically, in contrast to previous methods, our approach explicitly accounts for the 6-D spatial structure of the functional connectomes (defined by pairs of points in 3-D space) by using either the GraphNet, fused Lasso, or the isotropic total variation penalty. Furthermore, we also introduce a multitask extension to this framework, which is suitable when the data are aggregated from multiple imaging institutions. Experiments on both synthetic and real world data reveal that the proposed method can recover results that are more neuroscientifically informative than previous methods while improving predictive performance.

CHAPTER 1

Introduction

With advancing data acquisition technology, high dimensional data have become much more regularly encountered in various areas of biomedical science. For example, advanced microarray technology allows scientists to measure the expression levels of tens of thousands of genes in a single experiment. In addition, modern neuroimaging techniques afford a variety of modalities that produce large-scale measurements that represent different aspects of neuronal activity, such as functional magnetic resonance imaging (fMRI), positron emission tomography (PET), and electroencephalograms (EEG) and magnetoencephalograms (MEG) recordings. The massive size of these data offers new possibilities, as they allow us to comprehensively study the biological system of interest at an unprecedented level of detail, which may lead to the discovery of clinically relevant *biomarkers*[1]. Nonetheless, the dimensionality of these data presents critical computational and statistical challenges, as traditional statistical methods break down when the number of parameters (predictors) dominates the number of observations, a setting frequently encountered in biomedical data analysis. This difficulty is compounded by the fact that biological data often possess complex correlation patterns among the predictors and tend to be noisy for variety of reasons, such as background noise, calibration error in

[1] The word *biomarker* is formally defined by the National Institutes of Health Biomarkers Definitions Working Group as: "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" (Atkinson et al., 2001; Strimbu and Tavel, 2010).

the measurement device, physiological movements (*e.g.*, cardiac and respiratory motion), and other sources of experimental variations. The central goal of this dissertation is to develop a computationally tractable machine learning framework that allows us to extract scientifically meaningful information from these massive and highly complex biomedical data. We motivate the scope of our study by considering two important problems with clinical relevance: (1) uncertainty analysis for biomedical *image registration*, and (2) psychiatric disease prediction based on *functional connectomes*, which are high dimensional correlation maps generated from resting state fMRI.

The remainder of this introductory chapter is organized as follows. First, we will formally present the challenges encountered in high dimensional data analysis, and introduce some of the key tools we utilize to mitigate these problems. Next, we will provide a brief primer on *image registration* and *functional connectomes*, and present the main contributions of our work. Finally, we will conclude this chapter with an outline of the dissertation.

1.1 High Dimensional Challenges

The setup where the number of parameters p greatly exceeds the sample size n is commonly referred to as the "large p small n problem," denoted $p \gg n$ (Bühlmann and van de Geer, 2011; West, 2003). In such setting, classical statistical methods break down in the face of the "curse of dimensionality" (Donoho, 2000; Duda et al., 2000). More concretely, the estimation procedure becomes susceptible to *overfitting*, *i.e.*, the estimated model will perform extremely well on the training data, but will predict poorly on unobserved data. Furthermore, in the $p \gg n$ setup, it is impossible to attain a statistically consistent estimator unless we impose some type of structural assumption on the model (Negahban et al., 2012). This leads us to the notion of *regularization*, a concept that will appear throughout this dissertation.

Regularization is a classical technique to prevent overfitting (James and Stein, 1961;

Tikhonov, 1963), and is achieved by encoding prior knowledge about the data structure into the estimation problem. In fact, many well known estimators from statistics and machine learning are based on solving a *regularized empirical risk minimization* problem (*e.g.*, support vector machine, logistic regression, boosting) that has the following form:

$$\operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^p} \mathcal{L}\left(\boldsymbol{w}\right) + \lambda\mathcal{R}(\boldsymbol{w}) \ . \tag{1.1}$$

The first term $\mathcal{L} : \mathbb{R}^p \to \mathbb{R}_+$ corresponds to the *empirical risk* of some loss function (*e.g.*, square loss, Huber loss, hinge loss), 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 models that deviate from the assumed structure. The user defined regularization parameter $\lambda \geqslant 0$ controls the tradeoff between data fit and regularization. Several different regularizers have been proposed in the literature to promote various forms of structure, such as smoothness (*e.g.*, ridge regression (Hoerl and Kennard, 1970), support vector machine (Cortes and Vapnik, 1995)), sparsity (*e.g.*, Lasso (Tibshirani, 1996), basis pursuit (Chen et al., 2001)), group sparsity (*e.g.*, group Lasso (Yuan and Lin, 2006), latent group Lasso (Obozinski et al., 2011)), low-rank structure (*e.g.*, trace/nuclear norm (Bach, 2008b; Recht et al., 2010)), and sparse covariance and inverse covariance structure (Bien and Tibshirani, 2011; Friedman et al., 2007; Meinshausen and Bühlmann, 2006).

Finally, an equally important aspect of a learning method is its computational tractability, as many statistical learning problems involve solving a numerical optimization problem (*e.g.*, Equation 1.1). In principle, almost all convex optimization problems can be solved with high accuracy using polynomial time interior-point methods. However, these generic solvers are impractical for high dimensional data, since the iteration cost of these methods grows nonlinearly with the problem size p (Boyd and Vandenberghe, 2004; Sra et al., 2012). Furthermore, sparsity promoting regularizers (*e.g.*, Lasso, group Lasso, Elastic-net), which are commonly used in high dimensional statistical inference problems, add to the difficulty by introducing non-differentiability to the objective function. For this reason, first-order optimization methods have generated renewed interest from the statistics and machine learning community, as they are capable of solving large scale and often nonsmooth optimization problems. These methods include conjugate gradient, proximal gradient, projected gradient, and alternating direction methods (Bach et al., 2012; Beck and Teboulle, 2009; Boyd et al., 2011; Nesterov, 2007). The work presented in the dissertation will frequently rely on these types of first-order optimization techniques.

1.2 Biomedical Image Registration and Uncertainty Analysis

Image registration is the process of finding the spatial transformation that maps the homologous image's coordinate space to the reference image's coordinates; Fig. 1.1 provides an example execution of the registration process. Its ability to fuse medical images with complementary information has led to its adoption in a variety of clinical research settings (Hill et al., 2001). For instance, PET and MRI are modalities that are commonly used for surgical planning. On one hand, PET images contain information about cancerous activity within the brain, but do not contain much anatomical structure. On the other hand, MRI images capture anatomical structures in the brain, but provide little physiological information. The variation in the appearance of the anatomy from these modalities can be seen in Fig. 1.2. By registering these images, the cancerous anatomical structures can be localized in a unified coordinate system. Other medical applications of image registration include motion correction, atlas construction, dose estimation, treatment monitoring, radiation therapy, and many more (Hill et al., 2001; Long et al., 2010; Shi et al., 2012; Sotiras et al., 2013).

1.2.1 Background: Elements of Biomedical Image Registration

Image registration is typically cast as an optimization problem, where the goal is to find the transformation that optimizes a user specified similarity measure that quantifies the quality of alignment between the reference image and the transformed homologous image. More formally, given a pair of d-dimensional images $\boldsymbol{f}_{\text{ref}}$ and $\boldsymbol{f}_{\text{hol}}$, image registration aims to solve the following optimization problem:

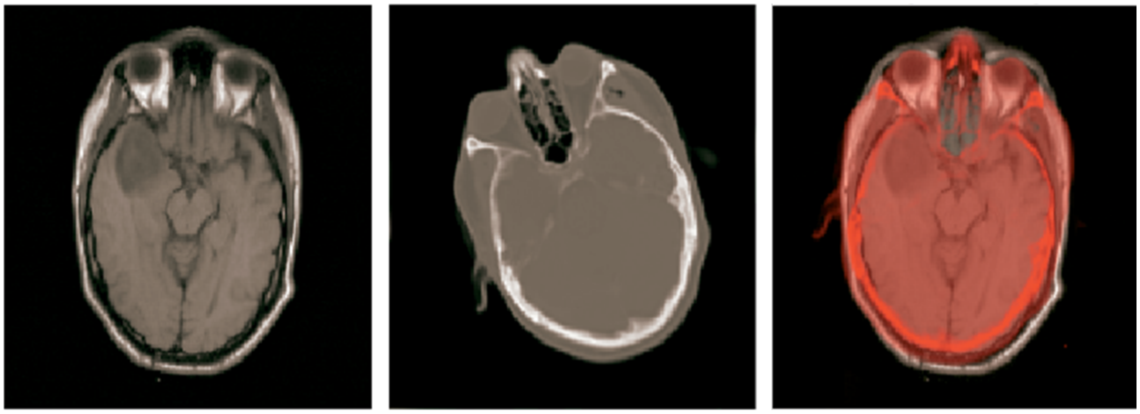
$$\hat{\boldsymbol{T}} = \arg\max_{\boldsymbol{T}} \; \boldsymbol{\Psi}\big(\boldsymbol{f}_{\text{ref}}(\cdot), \boldsymbol{f}_{\text{hol}} \circ \boldsymbol{T}(\cdot)\big), \tag{1.2}$$

where $\boldsymbol{f}_{\text{ref}} : \mathbb{R}^d \rightarrow \mathbb{R}$ and $\boldsymbol{f}_{\text{hol}} : \mathbb{R}^d \rightarrow \mathbb{R}$ are the reference and the homologous image respectively, $\boldsymbol{T} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ denotes the spatial transformation that models the misaligment between the image pair, and $\boldsymbol{\Psi}$ is a user-specified similarity measure that quantifies the quality of the alignment. Importantly, Equation 1.2 illustrates the following three major design components of image registration:

1. the similarity measure $\boldsymbol{\Psi}$,

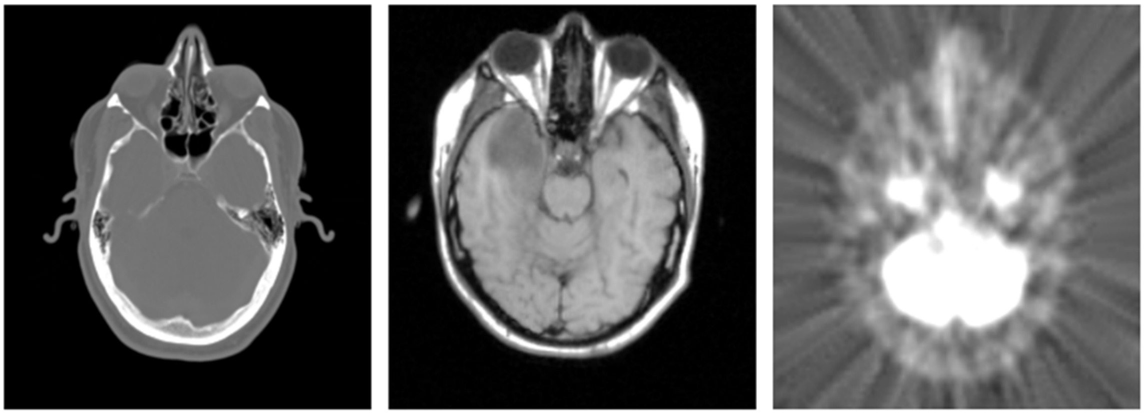
2. the model for the spatial transformation $\boldsymbol{T}$,

3. the optimization algorithm for solving (1.2).



(a) Reference image (b) Homologous image (c) Registered image

Figure 1.1: Example execution of the registration process.

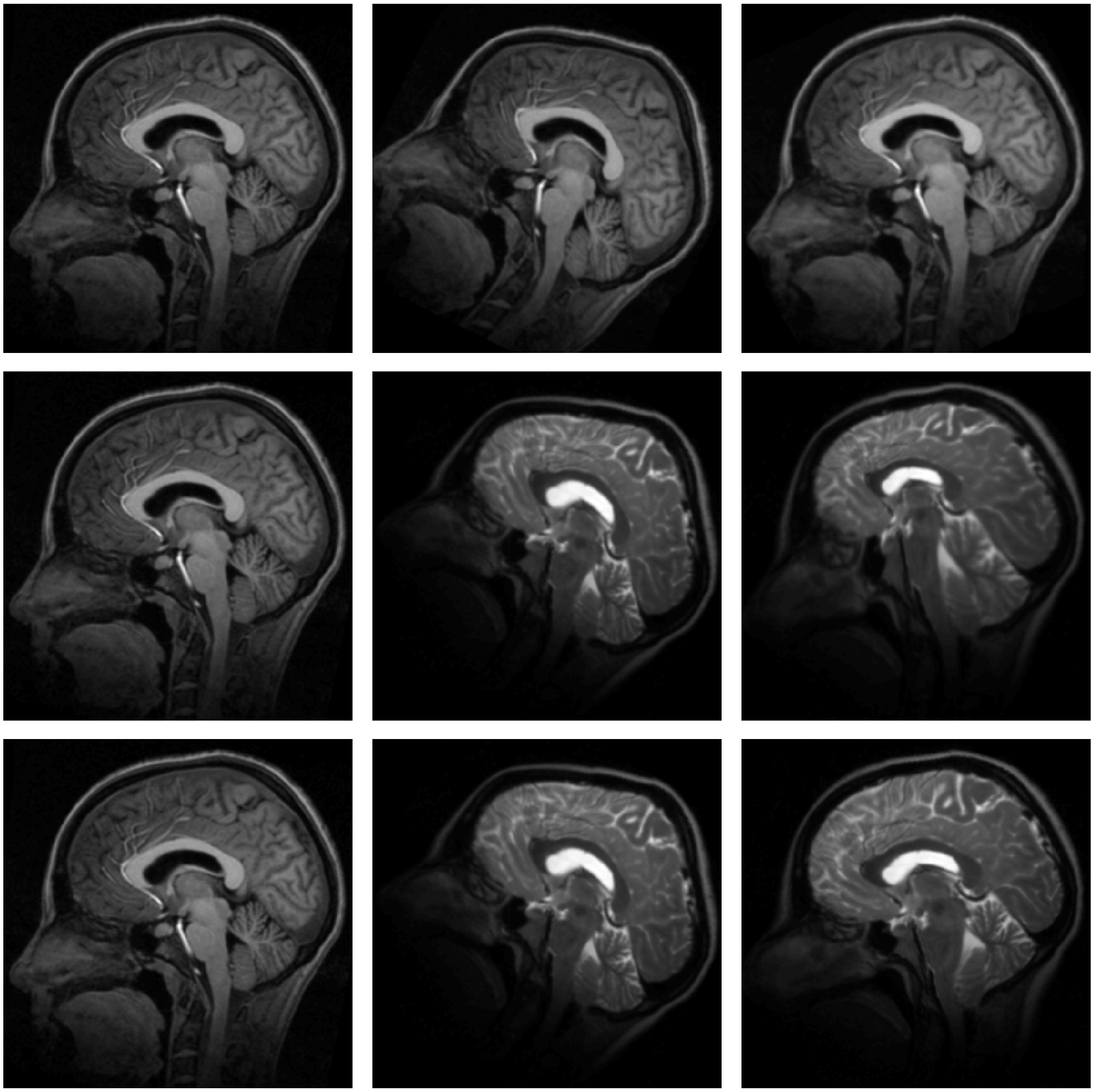


(a) CT (b) MRI (c) PET

Figure 1.2: Brain images acquired from different imaging modalities. Note the variation in the appearance of the anatomy.

Similarity measure (Ψ): The choice of the similarity measure depends on the type of relationship one expects among the pixel (voxel) intensities in the image pair. For example, in the *intramodal* setup, where the images are acquired from the same imaging modality, it is reasonable to assume the intensities of the images to be directly/linearly related. Thus simple similarity measures such as the *sum of squared differences* (SSD) and *Pearson's correlation* are popular choices for this setup. Conversely, in the *intermodal* setup, where the images are acquired from different imaging modalities, the intensities of the two images are no longer directly related, hence SSD and Pearson's correlation become inappropriate. In this case, usually one instead assumes a statistical/probabilistic relationship between the images, and information theoretic measures such as *conditional entropy* and *mutual information* are common choices (Pluim et al., 2003). Fig. 1.3 illustrates how the choice of the similarity measure can have a huge impact on the outcome of a registration algorithm.

Transformation model ($\boldsymbol{T}$): The transformation model describes the type of spatial deformation that is expected between the reference and the homologous image. A parametric approach is commonly adopted for this, where the transformation $\boldsymbol{T}$ is compactly characterized by a parameter vector $\boldsymbol{\theta}$; the size of $\boldsymbol{\theta}$ determines the degrees of freedom (DOF) of the model. The simplest choice is the *rigid* transformation model



(a) Reference (b) Homologous (c) Registered homologous

Figure 1.3: The impact of the choice of similarity measure $\boldsymbol{\Psi}$. **Top row**: successful intramodal registration using SSD. **Middle row**: unsuccessful intermodal registration using SSD; note the misalignment in the corpus callosum, which has a black appearance in the reference image and a white appearance in the homologous image. **Bottom row**: successful intermodal registration using mutual information.

that is characterized by rotation and translation, corresponding to three DOF in 2-D and six DOF in 3-D. While this model is appropriate for describing movements in the hard tissue region, it is not capable of capturing local movements in the soft tissue area (*e.g.*, respiratory and cardiac motion). To model these types of local deformations, *nonrigid* transformation models such as the B-spline and thin-plate spline models are commonly used (Meyer et al., 1997; Rueckert et al., 1999; Unser, 1999). Extensive reviews on nonrigid deformation models can be found in (Holden, 2008; Sotiras et al., 2013). However, the flexibility afforded by the nonrigid model comes at the expense of the size of the parameter vector $\boldsymbol{\theta}$, which can often be on the order of a million. This not only increases computational complexity but also leads to overfitting, which results in a physically unrealistic transformation such as bone-warping. Thus, regularization becomes crucial for stabilizing the estimation procedure, and various regularizers have been introduced in the literature, such as the gradient norm, elastic energy, topology preserving penalty (Chun and Fessler, 2009; Modersitzki, 2004)).

Optimization strategies: As explained earlier, image registration is an optimization problem that aims to find the transformation that best aligns the coordinates of an image pair. Hence the choice of the optimization strategy can have a significant impact on the outcome of the registration algorithm. Iterative gradient based approaches such as gradient descent, conjugate gradient descent, and quasi-Newton methods are frequently used for nonrigid models with high DOF (Holden, 2008; Klein et al., 2007; Sotiras et al., 2013).

1.2.2 Contribution: Registration Uncertainty Analysis using Spatial Confidence Regions

Despite the promises that image registration offers, there are numerous issues that still must be solved before it can be used in the clinical practice. For instance, it is well known that registration accuracy is limited in practice, and the degree of uncertainty varies at different image regions. Such uncertainty arises for variety of reasons, such as the variation in the appearance of the anatomy, measurement noises, deformation model mismatch, local minima, etc. Evaluating this degree of uncertainty is highly non-trivial due to the scarcity of ground-truth data. Understanding the accuracy of a registration result is one of the central themes in modern medical image analysis.

In light of these challenges, in Chapter 2 of the dissertation, we propose a data-driven method that allows one to visualize and quantify the registration uncertainty through spatially adaptive confidence regions. The method applies to any choice of the similarity measure and various parametric transformation models, including high dimensional deformation models such as the B-spline. At the heart of the proposed method is a novel shrinkage-based estimate of the distribution on deformation parameters $\boldsymbol{\theta}$. We present some empirical evaluations of the method in 2-D using images of the lung and liver, and demonstrate that the confidence regions produces promising results.

1.3 Disease Prediction based on Functional Connectomes

The emerging field of *connectomics*, which is the study of the network architecture of the brain, has provided various new insights about neuropsychiatric disorders that are associated with abnormalities in brain connectivity (Biswal et al., 2010; Hagmann, 2005; Sporns et al., 2005). Brain connectivity can be broadly divided into two categories: structural connectivity and functional connectivity. On one hand, "structural connectivity" describes *anatomical connections*, *i.e.*, physical wiring of the brain such as linkages in white matter fiber tracts that can be studied using modalities such as diffusion tensor images (DTI) (Bihan and Johansen-Berg, 2012). On the other hand, "functional connectivity" describes *functional connections* that are typically characterized by the statistical dependencies among the neuronal signals between remote brain regions (Biswal et al., 1995). These brain connectivities are commonly represented as graphs called *structural* and *functional connectomes*, where the nodes represent brain regions and the edges (weighted or binary) represent the structural/functional relationship between the neuronal signals (Bullmore and Sporns, 2009; Smith et al., 2013; Sporns, 2013). Throughout this dissertation, we will focus on *functional connectomes* generated from *resting state* fMRI.

1.3.1 Background: Resting state fMRI and Functional Connectomes

FMRI data consist of a time series of three dimensional volumes imaging the brain, where each 3-D volume encompasses around $10,000 \sim 100,000$ voxels. The univariate time series at each voxel represents a blood oxygen level dependent (BOLD) signal, an indirect measure of neuronal activities in the brain. The imaging process is noninvasive, relatively cheap and accessible, and does not expose subjects to radiation, making fMRI an attractive tool for studying the human brain.

Traditional experiments in the early years of fMRI research involved *task-based studies*, where participants perform a set of tasks during scan time, and the goal is to identify the brain regions associated with the task performance. However, it was later discovered that even in the absence of a cognitive task performance, the BOLD signal follows a synchronized fluctuation pattern at distributed brain regions (Biswal et al., 1995), implying that the brain is functionally connected at rest (Greicius et al., 2003). These temporal correlations between remote brain regions is referred to as *functional connectivity* (Friston, 1994), and *resting state fMRI* has become a vital modality for studying the intrinsic functional architecture of brain networks (Fox and Raichle, 2007; Smith et al., 2013).

A particularly notable tool that made a significant contribution in the development of the field of connectomics is *functional connectome*, which is a correlation map derived from resting state fMRI. More precisely, functional connectomes are constructed by parcellating the brain into multiple distinct regions and computing cross-correlations among the inter-regional BOLD signals (Varoquaux and Craddock, 2013). It is important to note that even with a relatively coarse parcellation scheme with several hundred regions of interest (ROI), the resulting functional connectome will be massive, encompassing hundreds of thousands of connections or more.

A central goal in connectomic research is the identification of an objective, connectivity-based biomarker of psychiatric disorders using functional connectomes. Such discovery would not only substantially extend our knowledge about the network topology of the human brain, but also offers the potential for a machine-based diagnosis system to enter the clinical realm (Atluri et al., 2013). Thus in recent years, machine learning techniques have garnered considerable amount of interests among the neuroimaging community (Pereira et al., 2009; Richiardi et al., 2013). However, many standard "off-the-shelf" machine learning algorithms are not immediately applicable due to the massive size of functional connectomes. Thus, a specialized class of machine learning techniques that are amenable to the dimensionality of functional connectomes is in critical need.

1.3.2 Contribution: Connectome-based Disease Prediction using a Scalable and Spatially-Informed Support Vector Machine

Abundant neurophysiological evidences indicate that major psychiatric disorders such as Alzheimer's disease, Attention Deficit Hyperactive Disorder (ADHD), autism spectrum disorder (ASD), and schizophrenia are associated with altered connectivity in the brain (Bassett and Bullmore, 2009; Castellanos et al., 2013; Dey et al., 2012; Fornito et al., 2012; Fox and Greicius, 2010; Sripada et al., 2014). Thus, there is great interest in developing machine-based methods that reliably distinguish patients from healthy controls using neuroimaging data. In this dissertation, 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 (Castellanos et al., 2013; Craddock et al., 2009; Dai et al., 2012; Sripada et al., 2013b; Zeng et al., 2012).

In Chapter 3, we address these issues by proposing 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 sparse fused Lasso (Tibshirani et al., 2005) or the GraphNet regularizer (Grosenick et al., 2013). Our method only restricts the loss function to be convex and margin-based, allowing non-differentiable loss 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 (Boyd et al., 2011), 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.

1.3.3 Contribution: Multitask Structured Sparse Support Vector Machine for Multisite Connectivity-based Disease Prediction

In response to the significant interest in developing imaging-based methods for diagnosing neuropsychiatric conditions, several data-sharing initiatives have been launched in the neuroimaging field (Biswal et al., 2010; Di Martino et al., 2013; Essen et al., 2012; Mennes et al., 2013; Poldrack et al., 2013; Poline et al., 2012; The ADHD-200 Consortium, 2012; Weiner et al., 2012). Here the datasets are collected across multiple imaging sites throughout the world. While this enables researchers to study the disorders of interest with substantial sample size, it also creates new challenges since the data aggregation process introduces various sources of site-specific heterogeneities.

To address this issue, in Chapter 4 we introduce a multitask structured sparse SVM, an extension to the method introduced in Chapter 3. Specifically, we employ a penalty that accounts for the following two-way structure that exists in a multisite functional connectome dataset: (1) the 6-D *spatial structure* in the functional connectomes captured via either the GraphNet, fused Lasso, or the isotropic total variation penalty, and (2) the *inter-site* structure captured via the multitask ℓ_1/ℓ_2-penalty (Lounici et al., 2009; Obozinski et al., 2010). The potential utility of the proposed method is demonstrated on the multisite ADHD-200 dataset.

1.4 Dissertation Outline

The remainder of the dissertation is organized as follows. In Chapter 2, we introduce a novel data-driven method that allows one to visualize and quantify image registration uncertainty using spatially adaptive confidence regions. In Chapter 3, we present a statistical learning framework for predicting the neuropsychiatric disease status of an individual using *functional connectomes* generated from resting state fMRI. In contrast to previous approaches, the method we present explicitly accounts for the 6-D spatial structure of the data. Chapter 4 presents a multitask extension to the work of Chapter 3, where the imaging sites are treated as the *tasks*. Finally, we conclude in Chapter 5 by providing a summary of the dissertation, and outline directions for future work.

CHAPTER 2

Spatial Confidence Regions for Quantifying and Visualizing Registration Uncertainty

For image registration to be most useful in a clinical setting, it is desirable to know the degree of uncertainty in the returned point-correspondences. In this chapter, we propose a data-driven method that allows one to visualize and quantify the registration uncertainty through spatially adaptive confidence regions. The method applies to any parametric deformation models and to any choice of the similarity criterion. We adopt the B-spline model and the negative sum of squared differences for concreteness. At the heart of the proposed method is a novel shrinkage-based estimate of the distribution on deformation parameters. We present some empirical evaluations of the method in 2-D using images of the lung and liver, and the method generalizes to 3-D.

2.1 Introduction

Image registration is the process of finding the spatial transformation that best aligns the coordinates of an image pair. Its ability to combine physiological and anatomical information has led to its adoption in a variety of clinical settings. However, the registration process is complicated by several factors, such as the variation in the appearance of the

This chapter is based on Watanabe and Scott (2012) anatomy, measurement noises, deformation model mismatch, local minima, etc. Thus, registration accuracy is limited in practice, and the degree of uncertainty varies at different image regions. For image registration to be most useful in clinical practice, it is important to understand its associated uncertainty.

Unfortunately, evaluating the accuracy of a registration result is non-trivial, mainly due to the scarcity of ground-truth data. For rigid-registration, there have been studies where physical landmarks are used to perform error analysis (Fitzpatrick and West, 2001). Statistical performance bounds for simple transformation models have been presented under a Gaussian noise condition (Robinson and Milanfar, 2004; Yetik and Nehorai, 2006). However, it is generally difficult or impractical to extend these methods to nonrigid registration, which limits their applicability since many part of the human anatomy cannot be described by a rigid model.

While characterizing the accuracy of a nonrigid registration algorithm is even more challenging, there have been recent works addressing this issue. Christensen et al. (2006) initiated a project which aims to allow researchers to perform comparative evaluation of nonrigid registration algorithms on brain images. Ruan and Fessler (2008) presented an observation model for image registration that accounts for image noise, and analyzed the performance limit of the model using Cramér-Rao bound analysis. Kybic (2010) used bootstrap resampling to perform multiple registrations on each bootstrap sample, and used the results to compute the statistics of the deformation parameter. Hub et al. (2009) proposed an algorithm and a heuristic measure of local uncertainty to evaluate the fidelity of the registration result. Risholm *et al.* adopted a Bayesian framework in (Risholm et al., 2010), where they proposed a registration uncertainty map based on the inter-quartile range (IQR) of the posterior distribution of the deformation field. Simpson *et al.* also adopted the Bayesian paradigm in (Simpson et al., 2012), where they introduced a probabilistic model that allows inference to take place on both the regularization level and the posterior of the deformation parameters. The mean-field variational Bayesian method was used to approximate the posterior of the deformation parameters, providing an efficient inference scheme.

We view the deformation as a random variable and propose a method that estimates the distribution of the deformation parameters given an image pair and registration algorithm. For illustration purpose, we use the cubic B-spline deformation model and the negative sum of squared differences as the similarity criterion, but the idea is applicable for other forms of parametric model (see Holden (2008) for other possible choices) and intensity-based registration algorithms. The estimated distribution will allow us to simulate realizations of registration errors, which can be used to learn spatial confidence regions. To the best of our knowledge, none of the existing methods view the registration uncertainty through spatial confidence regions represented in the pixel-domain. The confidence regions can be used to create an interactive visual interface that can be used to assess the accuracy of the original registration result. A conceptual depiction of this visual interface is shown in Fig. 2.1. When a user, such as a radiologist, selects a pixel in the reference image, a confidence region appears around the estimated corresponding pixel in the homologous image. If the prespecified confidence level is, say $\gamma = 0.95$, then the actual corresponding point is located within the confidence region with at least 95% probability. The magnitude and the orientation of the confidence region offers an understanding of the geometrical fidelity of the registration result at different spatial locations.

2.2 Method

For clarity, the idea is presented in a 2-D setting, but the method generalizes directly to 3-D.

2.2.1 Nonrigid Registration and Deformation Model

When adopting a parametric deformation model, it is common to cast image registration as an optimization problem over a real valued function $\mathbf{\Psi}$, a similarity measure quantifying the quality of the overall registration. Formally, this is written

$$\arg\max_{\boldsymbol{\theta}} \boldsymbol{\Psi}\left(\boldsymbol{f}_{\text{ref}}(\cdot), \boldsymbol{f}_{\text{hol}} \circ \boldsymbol{T}(\cdot\,; \boldsymbol{\theta})\right) , \tag{2.1}$$

where $\boldsymbol{f}_{\text{ref}}, \boldsymbol{f}_{\text{hol}} : \mathbb{R}^2 \rightarrow \mathbb{R}$ are the reference and the homologous images respectively, and $\boldsymbol{T}(\cdot\,; \boldsymbol{\theta}) : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is a transformation parametrized by $\boldsymbol{\theta}$. Letting $\boldsymbol{r} = (x, y)$ denote a pixel location, a nonrigid transformation can be written $T(\boldsymbol{r}; \boldsymbol{\theta}) = \boldsymbol{r} + \boldsymbol{d}(\boldsymbol{r}; \boldsymbol{\theta})$, where $\boldsymbol{d}(\cdot\,; \boldsymbol{\theta})$ is the deformation. To evaluate the value $\boldsymbol{f}_{\text{hol}}(\boldsymbol{T}(\boldsymbol{r}; \boldsymbol{\theta}))$ at non-pixel positions, we use fast B-spline interpolation (Unser et al., 1991, 1993a,b) with a 4-level multiresolution scheme (Unser et al., 1993c). To model the deformation, we adopt the commonly used tensor product of the cubic B-spline basis function β (Kybic and Unser, 2003; Rueckert et al., 1999), where the deformation for each direction $q \in \{x, y\}$ is described independently by parameter coefficients $\{\boldsymbol{\theta}_q\}$ as follows:

$$d_q(\boldsymbol{r}; \boldsymbol{\theta}_q) = \sum_{i,j} \theta_q^{(i,j)}\, \beta\left(\frac{x}{m_x} - i\right) \beta\left(\frac{y}{m_y} - j\right) ,$$

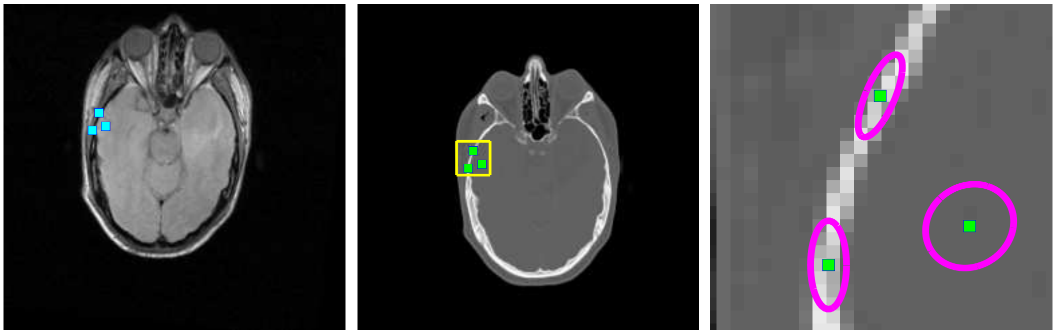

(a) (b) (c)

Figure 2.1: Conceptual illustration of the proposed method. The marks in (a)-(b) are a few point-correspondences estimated by registration. The confidence regions in (c) offer an understanding of the possible registration error for these pixels. We expect the shape of the confidence regions to reflect the local image structure, as demonstrated in (c).

where the B-spline function β has the representation

$$\beta(x) = \begin{cases} \frac{2}{3} - |x|^2 + \frac{|x|^3}{2}, & 0 \leqslant |x| < 1 \\ \frac{(2-|x|)^3}{6}, & 1 \leqslant |x| < 2 \\ 0, & |x| \geqslant 2 \end{cases} .$$

The scale of the deformation is controlled by m_q, which is the knot spacing in the q direction. If K knots are placed on the image, the total dimension of the parameter $\boldsymbol{\theta} = \{\boldsymbol{\theta}_x, \boldsymbol{\theta}_y\}$ is $2K$ since $\boldsymbol{\theta}_x, \boldsymbol{\theta}_y \in \mathbb{R}^K$.

2.2.2 Spatial Confidence Regions

Given the image pair $\boldsymbol{f}_{\text{ref}}$ and $\boldsymbol{f}_{\text{hol}}$, let $\boldsymbol{\Omega}_{\text{ref}} \subset \mathbb{R}^2$ and $\boldsymbol{\Omega}_{\text{hol}} \subset \mathbb{R}^2$ denote the regions of interest in the reference and homologous image respectively. Also, let $\hat{\boldsymbol{\theta}}$ be the deformation coefficients estimated from registration (2.1). We will assume that the underlying ground-truth deformation belongs to the adopted deformation class with deformation parameter $\boldsymbol{\theta}$. Then, the registration error $\boldsymbol{e}$ for pixel $\boldsymbol{r} \in \boldsymbol{\Omega}_{\text{ref}}$ is expressed as

$$\boldsymbol{e}(\boldsymbol{r}) = \big(e_x(\boldsymbol{r}), e_y(\boldsymbol{r})\big) = \boldsymbol{T}(\boldsymbol{r}; \hat{\boldsymbol{\theta}}) - \boldsymbol{T}(\boldsymbol{r}; \boldsymbol{\theta}) \ . \tag{2.2}$$

We will view the true deformation $\boldsymbol{\theta}$ as a random variable, which together with other sources of randomness such as image noise, introduces a distribution on $\boldsymbol{e}(\boldsymbol{r})$ for each $\boldsymbol{r}$. Here, $\boldsymbol{\Omega}_{\text{hol}}$ is the sample space of the registration error $\boldsymbol{e}(\boldsymbol{r})$, and the confidence region $\boldsymbol{\Phi}(\boldsymbol{r}) \subseteq \boldsymbol{\Omega}_{\text{hol}}$ is a set such that

$$\text{Pr}\big\{\boldsymbol{e}(\boldsymbol{r}) \in \boldsymbol{\Phi}(\boldsymbol{r})\big\} \geqslant \gamma,$$

where $\gamma \in [0, 1]$ is a prespecified confidence level. To estimate the spatial confidence regions, we adopt the following two-step process.

First, we estimate the distribution of $\boldsymbol{\theta}$. We assume $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, so the problem reduces to estimating $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. This is a challenging task because there is only a single realization of $\boldsymbol{\theta}$, corresponding to the given reference and homologous images, and this realization is not observed.

Second, given the estimates of $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, we can then simulate approximate realizations of $\boldsymbol{\theta}$, and thereby simulate spatial errors $\boldsymbol{e}(\boldsymbol{r})$. From this it is straight-forward to estimate $\boldsymbol{\Phi}(\boldsymbol{r})$. However, sampling from $\mathcal{N}(\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}})$ is potentially computationally intensive. The total dimension of $\boldsymbol{\theta}$ for the B-spline model is $2K$ in 2-D and $3K$ in 3-D. For a high resolution CT dataset of image size $512 \times 512 \times 480$ with voxel dimensions $1 \times 1 \times 1$ mm^3, B-spline knots placed every 5 mm leads to a dimension on the order of millions. Sampling from a multivariate normal distribution requires a matrix square root of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, but this is clearly prohibitive in both computational cost and memory storage. Therefore it is essential that the estimate $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ have some structure that facilitates efficient sampling.

2.2.3 Estimation of Deformation Distribution

We use the registration result $\hat{\boldsymbol{\theta}}$ as the estimate for $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, and propose the following convex combination for $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} = (1 - \rho)\boldsymbol{\Sigma}_o + \rho \hat{\boldsymbol{\theta}} \hat{\boldsymbol{\theta}}^T \, . \tag{2.3}$$

The first term $\boldsymbol{\Sigma}_o$ is a positive-definite matrix which is an *a priori* baseline we impose on the covariance structure, and the second term is a rank-1 outer product that serves as the data-driven component. The weighting between the two terms is controlled by $\rho \in [0, 1)$. Note that (2.3) has a form of a shrinkage estimator reminiscent of the Ledoit-Wolfe type covariance estimate (Ledoit and Wolf, 2003), but only using the registration result $\hat{\boldsymbol{\theta}}$.

For the baseline covariance $\boldsymbol{\Sigma}_o$, we propose to use a covariance matrix which is motivated from the autoregressive model. Let $\boldsymbol{\Sigma}_{\text{AR}} \in \mathbb{R}_{++}^{K \times K}$ denote the covariance of a first order 2-D autoregressive model, whose entries are given as

$$\boldsymbol{\Sigma}_{\text{AR}}(i,j) = r_x^{|x(i)-x(j)|} r_y^{|y(i)-y(j)|} \,, \qquad 1 \leqslant i,j \leqslant K \,.$$

Here, $|r_x| < 1$ and $|r_y| < 1$ are parameters that control the smoothness between neighboring knots, and $x(i) = \mod(i-1, n_x)$, $y(i) = \lfloor (i-1)/n_x \rfloor$ are the mappings from the lexicographic index i to its corresponding (x,y) coordinate, assuming an $(n_x \times n_y)$ grid of knots. A key property of this dense matrix is that its inverse, or the precision matrix $\boldsymbol{\Theta}_{\text{AR}} = \boldsymbol{\Sigma}_{\text{AR}}^{-1}$, is block-tridiagonal with tridiagonal blocks. Specifically, $\boldsymbol{\Theta}_{\text{AR}}$ has an n_y-by-n_y block matrix structure with each blocks of size $(n_x \times n_x)$, and only the main diagonal and the subdiagonal blocks are non-zero. Furthermore, these non-zero blocks are tridiagonal with the values of the non-zero entries known as a function of r_x and r_y.

Based on $\boldsymbol{\Sigma}_{\text{AR}}$, we propose to use the following baseline covariance $\boldsymbol{\Sigma}_o \in \mathbb{R}_{++}^{2K \times 2K}$ having a 2-by-2 block matrix structure expressed by the Kronecker product:

$$\boldsymbol{\Sigma}_o = \begin{bmatrix} c_x \boldsymbol{\Sigma}_{\text{AR}} & c_{xy} \boldsymbol{\Sigma}_{\text{AR}} \\ c_{xy} \boldsymbol{\Sigma}_{\text{AR}} & c_y \boldsymbol{\Sigma}_{\text{AR}} \end{bmatrix} = \begin{bmatrix} c_x, c_{xy} \\ c_{xy}, c_y \end{bmatrix} \otimes \boldsymbol{\Sigma}_{\text{AR}} \,. \tag{2.4}$$

The coefficients c_x and c_y assign the prior variance level on $\boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_y$, whereas c_{xy} assigns the prior cross-covariance level between $\boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_y$. The only restriction on these values is $(c_x c_y) > c_{xy}^2$, which ensures $\boldsymbol{\Sigma}_o$ is positive-definite. It is important to note that the precision matrix $\boldsymbol{\Theta}_o$ of this baseline covariance is sparse, also having a 2-by-2 block matrix structure

$$\boldsymbol{\Theta}_o = \boldsymbol{\Sigma}_o^{-1} = \begin{bmatrix} c_x, c_{xy} \\ c_{xy}, c_y \end{bmatrix}^{-1} \otimes \boldsymbol{\Sigma}_{\text{AR}}^{-1} = \begin{bmatrix} p_x, p_{xy} \\ p_{xy}, p_y \end{bmatrix} \otimes \boldsymbol{\Theta}_{\text{AR}} \,,$$

where $\{p_x, p_y, p_{xy}\}$ are obtained by inverting the 2×2 coefficient matrix. The sparsity structure of $\boldsymbol{\Theta}_o$ can be interpreted intuitively under a Gaussian graphical model framework. The conditional dependencies between knots are described by the non-zero entries in the matrix, which are represented as edges in an undirected graph. For our model, a knot $\theta_x(i,j)$ has 17 edges, 8 connected to its 8-nearest neighbors and the other 9 connected to the corresponding $\theta_y(i,j)$ knot and its 8-nearest neighbors. Fig. 2.2 provides an illustration of $\mathbf{\Sigma}_o$ and the sparsity structure of its inverse $\mathbf{\Theta}_o$, along with an example realization of B-spline coefficients $\boldsymbol{\theta} = (\boldsymbol{\theta}_x, \boldsymbol{\theta}_y)$.

2.2.4 Efficient Sampling.

We now discuss how the sparsity structure of $\mathbf{\Theta}_o$ can be exploited. Let $\boldsymbol{L}_{\mathbf{\Theta}_{\text{AR}}}$ denote the cholesky factor for $\mathbf{\Theta}_{\text{AR}}$, which can be computed efficiently in $\mathcal{O}(K)$ operations due to its block-tridiagonal with tridiagonal blocks structure (Golub and Van Loan, 1996). Then the Cholesky factor for $\mathbf{\Theta}_o$ can be expressed

$$\boldsymbol{L}_o = \begin{bmatrix} \sqrt{p_x}\boldsymbol{L}_{\mathbf{\Theta}_{\text{AR}}} & \mathbf{0} \\ \frac{p_{xy}}{\sqrt{p_x}}\boldsymbol{L}_{\mathbf{\Theta}_{\text{AR}}} & \sqrt{p_y - \frac{p_{xy}^2}{p_x}}\boldsymbol{L}_{\mathbf{\Theta}_{\text{AR}}} \end{bmatrix} .$$

Defining matrix $\boldsymbol{L}$ as

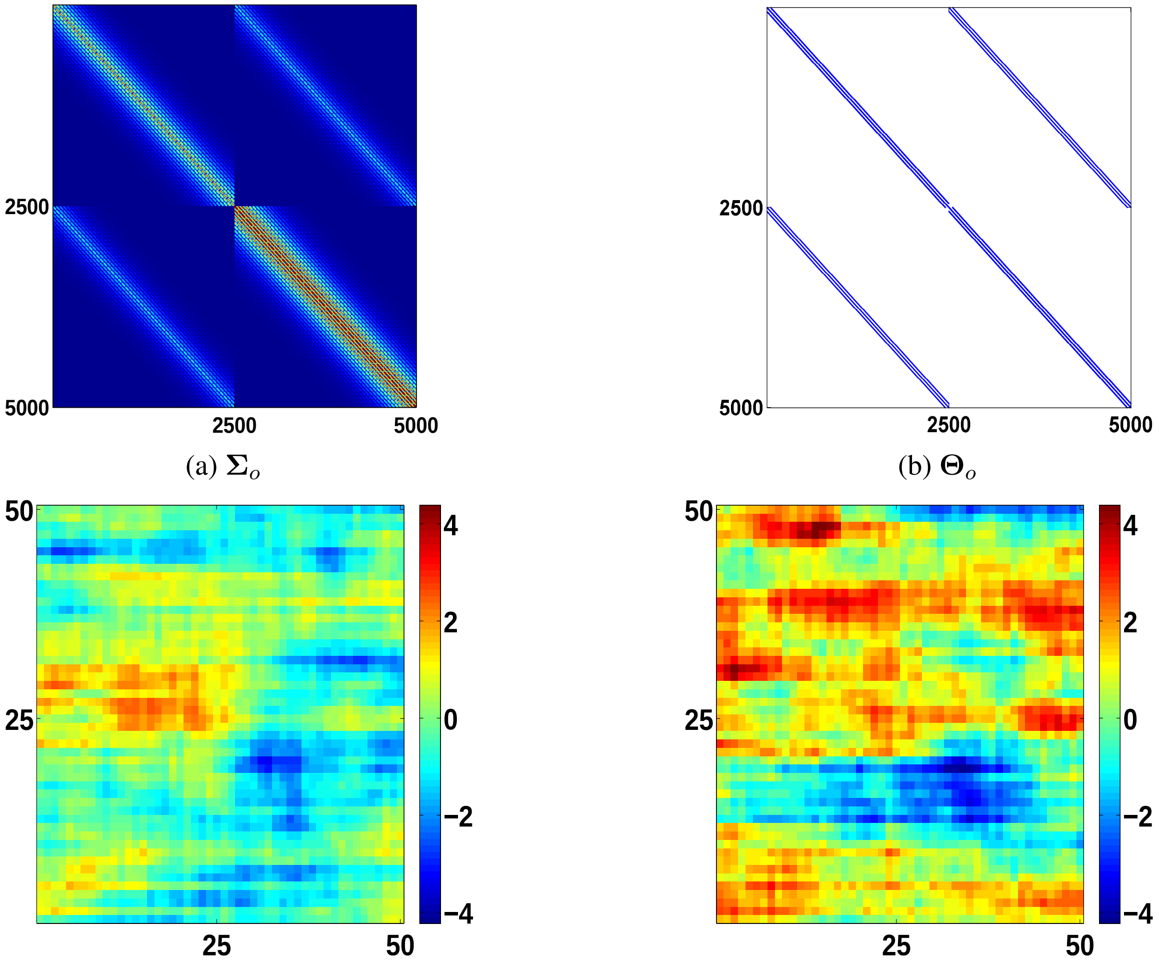
$$\boldsymbol{L} = \sqrt{(1-\rho)}\left[\boldsymbol{L}_o^{-T} + m\left(\frac{\rho}{1-\rho}\right)\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T\boldsymbol{L}_o\right] ,$$

where constants t and m are defined as $t := \left(\frac{\rho}{1-\rho}\right)\hat{\boldsymbol{\theta}}^T\mathbf{\Theta}_o\hat{\boldsymbol{\theta}}$ and $m := \frac{\sqrt{1+t}-1}{t}$, it can be shown that

$$\boldsymbol{L}\boldsymbol{L}^T = (1-\rho)\mathbf{\Sigma}_o + \rho\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T,$$

i.e., $\boldsymbol{L}$ is a matrix square root for $\mathbf{\Sigma}_{\boldsymbol{\theta}}$. Therefore, letting $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, we have

$$\hat{\boldsymbol{\theta}} + \boldsymbol{L}\,\boldsymbol{z} \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, \hat{\mathbf{\Sigma}}_{\boldsymbol{\theta}}),$$



(a) $\mathbf{\Sigma}_o$

(b) $\mathbf{\Theta}_o$

(c) $\boldsymbol{\theta}_x$

(d) $\boldsymbol{\theta}_y$

Figure 2.2: Illustration of the properties of the baseline covariance $\mathbf{\Sigma}_o$. The values used are $(n_x, n_y) = (50, 50)$, $(r_x, r_y) = (0.95, 0.8)$, and $\{c_x, c_y, c_{xy}\} = \{1, 2, 0.5\}$. (a) The baseline covariance $\mathbf{\Sigma}_o$, (b) the sparsity structure of $\mathbf{\Theta}_o = \mathbf{\Sigma}_o^{-1}$, (c)-(d) B-spline coefficients $\boldsymbol{\theta}_x$ and $\boldsymbol{\theta}_y$ obtained from sample $\boldsymbol{\theta} = (\boldsymbol{\theta}_x, \boldsymbol{\theta}_y) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_o)$.

the desired distribution. Furthermore, by invoking the matrix inversion lemma, this matrix-vector product term can be expressed as

$$\boldsymbol{L}\,\boldsymbol{z} = \sqrt{(1-\rho)}\boldsymbol{L}_o^{-T}\boldsymbol{z} + m\left(\frac{\rho}{\sqrt{1-\rho}}\right)\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T\boldsymbol{L}_o\boldsymbol{z}.$$

The first term $\boldsymbol{L}_o^{-T}\boldsymbol{z}$ can be computed in $\mathcal{O}(K)$ operations using backward-substitution and exploiting the sparsity of $\boldsymbol{L}_o$ (Golub and Van Loan, 1996). The second term involves a simple matrix-vector multiplication, thus it can also be computed efficiently.

In summary, we never need to store or directly compute a matrix square root for the dense matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$; we only need to store the sparse precision matrix $\boldsymbol{\Theta}_o$ and compute its cholesky factor $\boldsymbol{L}_o$. Therefore, the sampling procedure scales gracefully to 3-D.

2.2.5 Error Simulations and Spatial Confidence Regions

Using the sampling procedure discussed in the previous section, we can now generate realizations of registration error $\boldsymbol{e}(\boldsymbol{r})$ as follows:

1. Sample $\boldsymbol{\theta}_i \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}})$.
2. Synthesize reference image $\boldsymbol{f}_{\text{ref}}^{(i)}(\boldsymbol{r}) \leftarrow \boldsymbol{f}_{\text{hol}} \circ \boldsymbol{T}(\boldsymbol{r};\boldsymbol{\theta}_i)$.
3. Register $\boldsymbol{f}_{\text{hol}}$ on to $\boldsymbol{f}_{\text{ref}}^{(i)}$ to get estimate $\hat{\boldsymbol{\theta}}_i$.
4. Compute error $\boldsymbol{e}_i(\boldsymbol{r}) = \boldsymbol{T}(\boldsymbol{r};\hat{\boldsymbol{\theta}}_i) - \boldsymbol{T}(\boldsymbol{r};\boldsymbol{\theta}_i)$.

We assume that $\boldsymbol{e}(\boldsymbol{r}) \sim \mathcal{N}\big(\boldsymbol{\mu}_{\boldsymbol{e}}(\boldsymbol{r}), \boldsymbol{\Sigma}_{\boldsymbol{e}}(\boldsymbol{r})\big)$ for all $\boldsymbol{r}$. Then the spatial confidence region associated with pixel $\boldsymbol{r} \in \boldsymbol{\Omega}_{\text{ref}}$ is defined by the ellipsoid

$$\boldsymbol{\Phi}(\boldsymbol{r}) = \left\{\boldsymbol{r}' : \big(\boldsymbol{r}' - \boldsymbol{\mu}_{\boldsymbol{e}}(\boldsymbol{r})\big)^T \boldsymbol{\Sigma}_{\boldsymbol{e}}^{-1}(\boldsymbol{r})\big(\boldsymbol{r}' - \boldsymbol{\mu}_{\boldsymbol{e}}(\boldsymbol{r})\big) < \chi_2^2(1-\gamma)\right\},$$

which is the $100\gamma\%$ level set of the bivariate normal distribution. Under this formulation, confidence region estimation becomes the problem of estimating $\{\boldsymbol{\mu}_{\boldsymbol{e}}(\boldsymbol{r}), \boldsymbol{\Sigma}_{\boldsymbol{e}}(\boldsymbol{r})\}$, the mean and covariance of the registration error at pixel location $\boldsymbol{r}$. We estimate these with the sample mean and covariance based on the simulated errors $\{\boldsymbol{e}_i(\boldsymbol{r})\}$. Algorithm 1 outlines the overall spatial confidence region estimation process.

Note that since we are using $\hat{\boldsymbol{\theta}}$ as the estimate for $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, it is important for the original registration to return an anatomically sensible result (*e.g.*, no bone warping), as severe inaccuracy could negatively impact the quality of the spatial confidence regions.

2.3 Experiments

We now demonstrate an application of the method, and also present preliminary experiments performed in 2-D. For illustration purpose, we used the negative sum of squared differences as the similarity criterion, but other metrics such as mutual information are also appropriate. For optimization, we used the conjugate gradient method, and the line search step size was determined by one step of Newton's method. For image interpolation, we used the popular B-spline model (Unser, 1999). To encourage the estimated deformation to be topology-preserving, we included the penalty term introduced by Chun and Fessler (2009) into the cost function for all experiments.

2.3.1 Application

We first applied the proposed method to two coronal CT slices in the lung region, shown in Fig. 2.3. Both images are size 256×360, and the exhale-frame served as the homologous image while the inhale-frame served as reference. The notable motion in this dataset is the sliding of the diaphragm with respect to the chest wall. Due to the opposing motion fields at this interface, registration uncertainty is expected to be higher around this region. To model the deformation, we used a knot spacing of $(m_x, m_y) = (3, 8)$, resulting in a parameter dimension of $\boldsymbol{\theta} \in \mathbb{R}^{7650}$. A tighter knot spacing was used for m_x since a finer scale of deformation was needed in the x-direction to model the sliding motion at the chest wall. Since the degree of this slide is relatively small for this dataset, the registration result

Algorithm 1 Spatial Confidence Regions Generation

```
**Input:** $\boldsymbol{f}_{\text{ref}}, \boldsymbol{f}_{\text{hol}}$
**Output:** $\{\hat{\boldsymbol{\mu}}_{\boldsymbol{e}}(\boldsymbol{r}), \hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}(\boldsymbol{r})\}$ for all $\boldsymbol{r} \in \boldsymbol{\Omega}_{\text{ref}}$
$\hat{\boldsymbol{\theta}} \leftarrow \arg\max_{\boldsymbol{\theta}'} \boldsymbol{\Psi}\big(\boldsymbol{f}_{\text{ref}}(\cdot), \boldsymbol{f}_{\text{hol}} \circ \boldsymbol{T}(\cdot\,; \boldsymbol{\theta}')\big)$
$\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}}$
$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} \leftarrow (1-\rho)\boldsymbol{\Sigma}_o + \rho\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^T$
**for** $i = 1, \ldots, N$
    sample $\boldsymbol{\theta}_i \leftarrow \mathcal{N}(\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}})$
    generate $\boldsymbol{f}_{\text{ref}}^{(i)}(\boldsymbol{r}) \leftarrow \boldsymbol{f}_{\text{hol}} \circ \boldsymbol{T}(\boldsymbol{r}; \boldsymbol{\theta}_i)$
    register $\hat{\boldsymbol{\theta}}_i \leftarrow \arg\max_{\boldsymbol{\theta}'} \boldsymbol{\Psi}\big(\boldsymbol{f}_{\text{ref}}^{(i)}(\cdot), \boldsymbol{f}_{\text{hol}} \circ \boldsymbol{T}(\cdot\,; \boldsymbol{\theta}')\big)$
    compute $\boldsymbol{e}_i(\boldsymbol{r}) = \boldsymbol{T}(\boldsymbol{r}; \hat{\boldsymbol{\theta}}_i) - \boldsymbol{T}(\boldsymbol{r}; \boldsymbol{\theta}_i)$
**end for**
$\hat{\boldsymbol{\mu}}_{\boldsymbol{e}}(\boldsymbol{r}) \leftarrow \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{e}_i(\boldsymbol{r})$
$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{e}}(\boldsymbol{r}) \leftarrow \frac{1}{N}\sum_{i=1}^{N} \big(\boldsymbol{e}_i(\boldsymbol{r}) - \hat{\boldsymbol{\mu}}_{\boldsymbol{e}}(\boldsymbol{r})\big)\big(\boldsymbol{e}_i(\boldsymbol{r}) - \hat{\boldsymbol{\mu}}_{\boldsymbol{e}}(\boldsymbol{r})\big)^T$
```

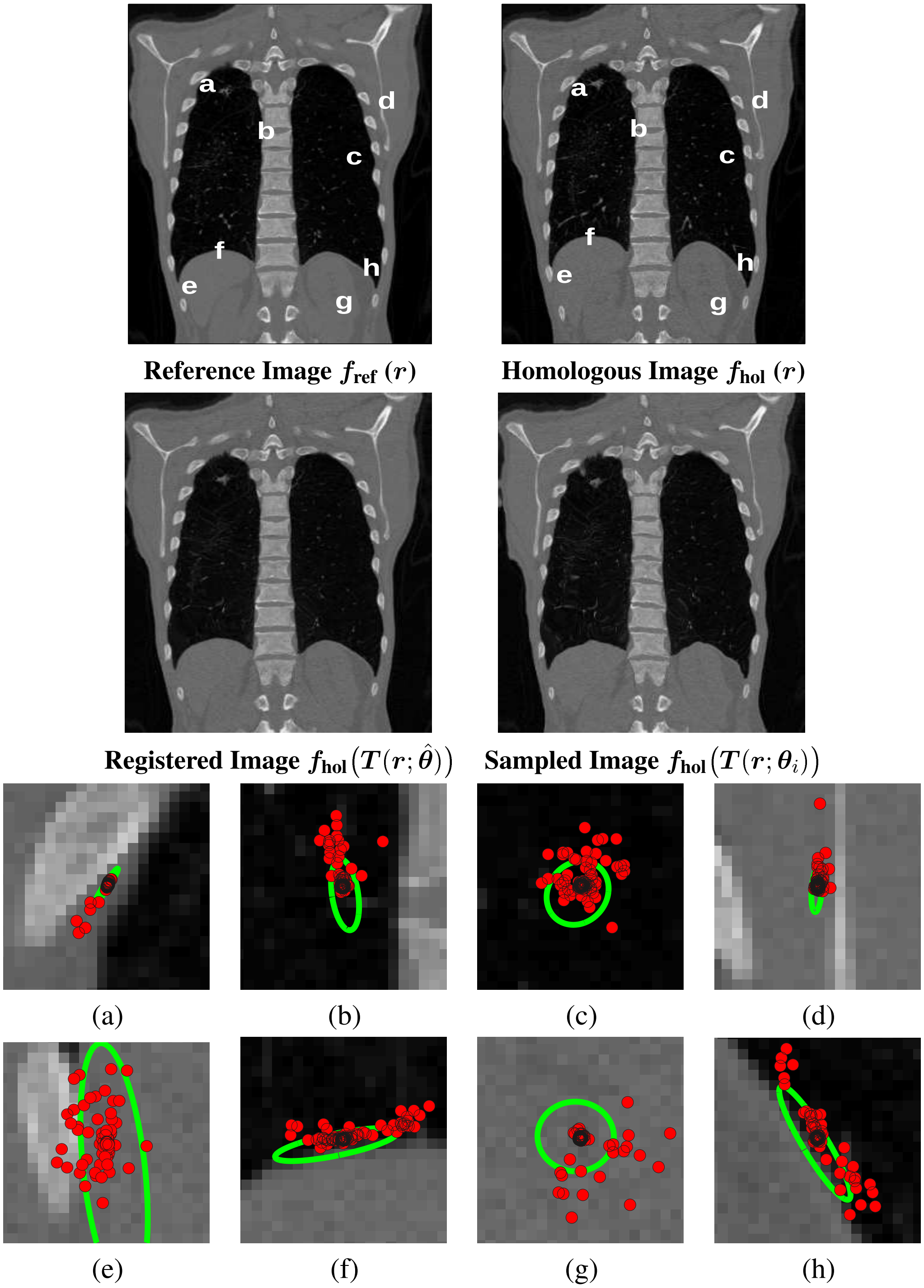
shown in Fig. 2.3 looks reasonably accurate based on visual inspection.

Using $\hat{\boldsymbol{\theta}}$ obtained from registering these images, we used the single-shot mean and covariance estimate and the efficient sampling scheme to obtain 100 new realizations of deformations. For the baseline covariance $\boldsymbol{\Sigma}_o$, we used values of $(r_x, r_y) = (0.9, 0.9)$ and $\{c_x, c_y, c_{xy}\} = \{2, 4, 0.5\}$. A relatively high value for c_y was used since the magnitude of the overall deformation was higher in the y-direction. Finally, $\rho = 0.1$ was used, as it was found to produce sensible deformation samples. One of the synthesized reference images is shown in Fig. 2.3. Following Algorithm 1, we obtained a set of spatial confidence regions $\{\boldsymbol{\Phi}(\boldsymbol{r})\}$ for all $\boldsymbol{r}$ in the region of anatomical interest, using a confidence level of $\gamma = 0.9$. A few of these are displayed in Fig. 2.3 (a)-(h), along with 100 simulated errors. It is important to note how the shapes of these confidence regions reflect the local image structure. The principal major axes of the ellipses are oriented along the edge, indicating higher uncertainty for those directions. The confidence regions for (c) and (g) take on isotropic shapes due to the absence of well-defined image structures. Finally, notice how the confidence region for (e) is quite large, illustrating how difficult it is to accurately register the sliding diaphragm at the chest wall.

2.3.2 Experimental Result

To quantitatively evaluate our method, we manually assigned $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ for the cubic B-spline deformation-generating process. The mean deformation $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ was designed to model the exhale to inhale motion in the abdominal area around the liver region, simulated by a contracting motion field. Manually assigning a sensible ground-truth value for the covariance $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is extremely difficult due to its high dimension and positive-definite constraint. Therefore, we took the shrinkage-based covariance model (2.3) as the ground-truth, using values of $(r_x, r_y) = (0.95, 0.95)$, $\{c_x, c_y, c_{xy}\} = \{2, 3, 0.5\}$, and $\rho = 0.1$. These values imply that the covariance is smooth with moderate level of correlation in the x and y deformations. We sampled a single instance of deformation $\boldsymbol{\theta}$ from this ground-truth distribution, and used it to deform a 2D axial CT slice in the liver region, having image size 512×420. We labeled the original image as the homologous and the deformed image as the reference. This resulting image pair and their difference image are shown in Fig. 2.4. A knot spacing of $(m_x, m_y) = (8, 8)$ was used to define the scale of the ground-truth deformation, resulting in a parameter dimension of $\boldsymbol{\theta} \in \mathbb{R}^{6656}$.

Next, we generated three classes of spatial confidence regions for this image pair, using confidence levels of $\gamma = 0.9$ and 0.95. The first confidence region $\boldsymbol{\Phi}_1(\boldsymbol{r})$ corresponds to the case where a correct deformation model is used for registration, and the parameter values for the shrinkage-based covariance estimate $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ matches that of the ground truth. The second confidence region $\boldsymbol{\Phi}_2(\boldsymbol{r})$ corresponds to the case where there is a mismatch in the deformation model. Here, we used a fifth-order B-spline function during registration, with a knot spacing of $(m_x, m_y) = (6, 6)$. In addition, we introduced some discrepancies in the



Reference Image $f_{\text{ref}}\,(r)$

Homologous Image $f_{\text{hol}}\,(r)$

Registered Image $f_{\text{hol}}\big(T(r;\hat{\theta})\big)$

Sampled Image $f_{\text{hol}}\big(T(r;\theta_i)\big)$

(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 2.3: The top two rows show the 2-D dataset used in the first experiment, along with the registration result and an image synthesized using one of the sampled deformations. A few of the confidence regions from $r \in \Omega_{\text{ref}}$ are shown in (a)-(h), with the red marks representing 100 realizations of registration error. Note how the confidence regions reflect the local image structure.

	Def. Basis	**Def. Scale**	**Parameter values used for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$**
Conf. Reg. 1 $\boldsymbol{\Phi}_1(\boldsymbol{r})$	Cubic B-spline	$m_x = 8$ $m_y = 8$	$\rho = 0.1, (r_x, r_y) = (0.95, 0.95)$ $\{c_x, c_y, c_{xy}\} = \{2, 3, 0.5\}$
Conf. Reg. 2 $\boldsymbol{\Phi}_2(\boldsymbol{r})$	Fifth order B-spline	$m_x = 6$ $m_y = 6$	$\rho = 0.15, (r_x, r_y) = (0.9, 0.9)$ $\{c_x, c_y, c_{xy}\} = \{2, 2, 0\}$
Conf. Reg. 3 $\boldsymbol{\Phi}_3(\boldsymbol{r})$	Cubic B-spline	$m_x = 8$ $m_y = 8$	$\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}} = \boldsymbol{\mu}_{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ **(Oracle)**

Table 2.1: Spatial Confidence Regions Generated for Validation

parameter values for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$. Finally, the third confidence region $\boldsymbol{\Phi}_3(\boldsymbol{r})$ corresponds to the ideal case, and is constructed for the purpose of comparison. Here, a correct deformation model is used for registration, and the deformations used to train the spatial confidence regions were sampled from the ground-truth $\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ rather than the estimated distribution. The descriptions of these confidence regions are summarized in Table 2.1. All confidence regions were generated using $N = 200$ simulated errors.

To assess the quality of these spatial confidence regions, we evaluated their *coverage rates* by sampling $M = 500$ additional deformations from the ground-truth distribution $\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$. Coverage rate for a given pixel $\boldsymbol{r}$ is defined as the percentage of registration errors that are confined within the confidence region $\boldsymbol{\Phi}(\boldsymbol{r})$, and is written mathematically

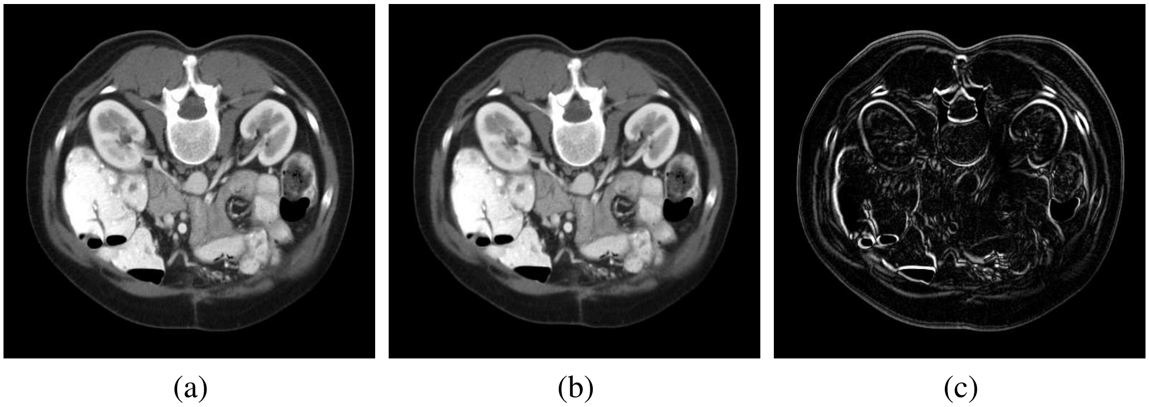


Figure 2.4: The dataset used for validation: (a) the homologous image $\boldsymbol{f}_{\text{hol}}(\boldsymbol{r})$, (b) the reference image $\boldsymbol{f}_{\text{ref}}(\boldsymbol{r}) = \boldsymbol{f}_{\text{hol}}\big(\boldsymbol{T}(\boldsymbol{r}; \boldsymbol{\theta})\big)$ generated by a deformation coefficient sampled from the ground-truth distribution $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, (c) the absolute difference image.

as

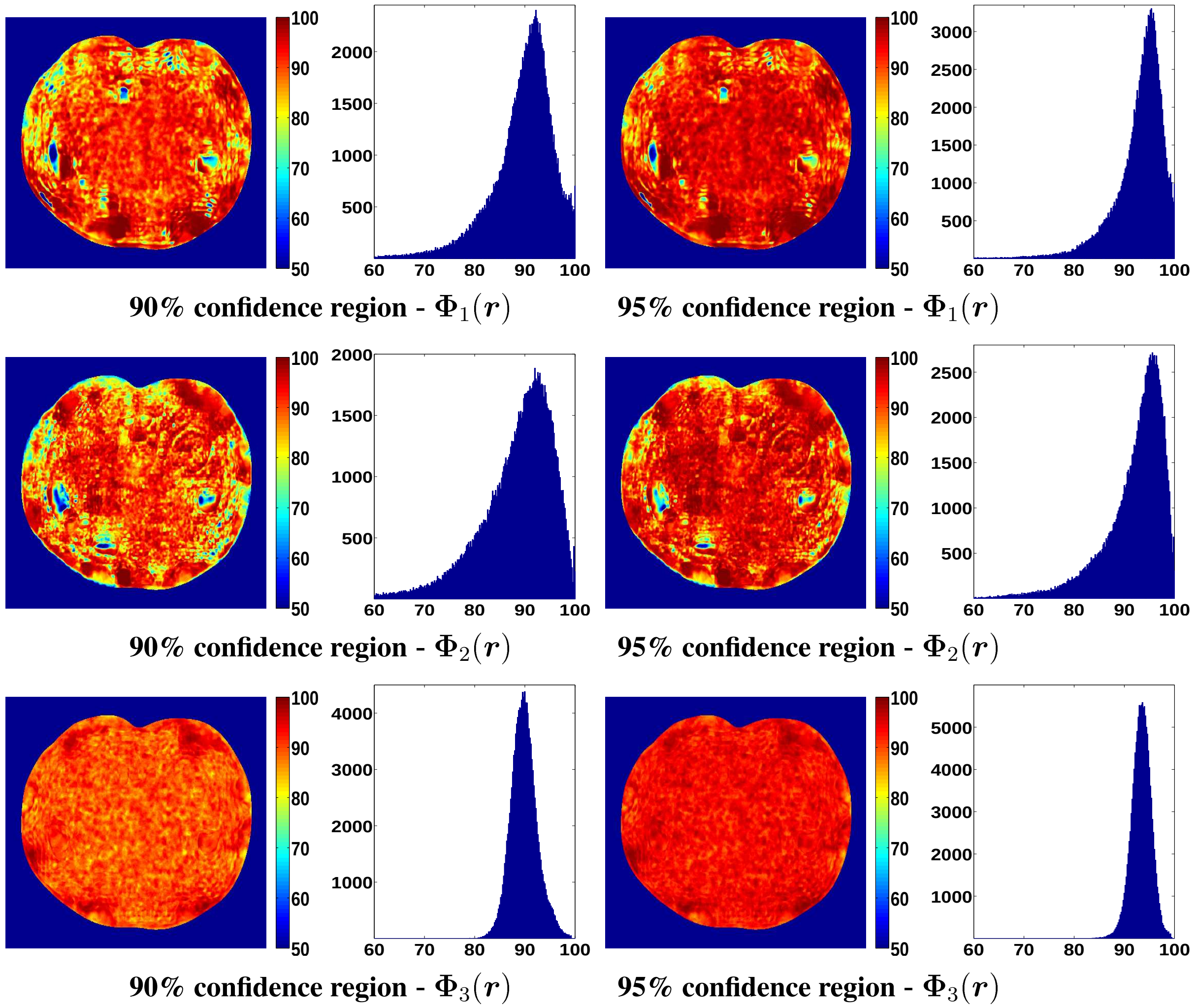
$$\frac{1}{M}\sum_{i=1}^{M}\mathbb{1}\left\{\tilde{\boldsymbol{e}}_i(\boldsymbol{r}) \in \boldsymbol{\Phi}(\boldsymbol{r})\right\} \ , \tag{2.5}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, and $\tilde{\boldsymbol{e}}_i(\boldsymbol{r})$ are registration errors generated from deformations sampled from the ground-truth distribution. We computed the coverage rate for the pixels that are located within the region of anatomy. The resulting coverage rates are rendered as heatmaps and are displayed in Fig. 2.5, along with their corresponding histograms. It can observed that the coverage rates for the first two confidence regions, $\boldsymbol{\Phi}_1(\boldsymbol{r})$ and $\boldsymbol{\Phi}_2(\boldsymbol{r})$, generally come close to the prespecified confidence level γ, although some degree of discrepancy can be observed at some image regions. The third confidence region $\boldsymbol{\Phi}_3(\boldsymbol{r})$ gave the best result as expected; the coverage rate for all pixels comes very close to γ.

In summary, the performance of the spatial confidence regions $\boldsymbol{\Phi}_1(\boldsymbol{r})$ and $\boldsymbol{\Phi}_2(\boldsymbol{r})$ turned out to be reasonably close, having results comparable to the ideal case of $\boldsymbol{\Phi}_3(\boldsymbol{r})$. Although further validation studies are required to obtain a more conclusive finding, this is an encouraging preliminary result.

2.4 Discussion and Conclusion

In this work, we presented a new method to evaluate the accuracy of a registration algorithm using spatially adaptive confidence regions. Preliminary experimental test results in 2-D suggest the confidence regions are effective based on their coverage rates. However, it is important to note that the computational cost of the proposed method is N times the original registration algorithm, since we must register each of the sampled deformations. Depending on the user's choice, this N can be in the order of hundreds to even thousands, with higher values likely to return more reliable confidence regions. We note that the process is easily parallelizable. Furthermore, in application such as surgical planning and radiation therapy, it may not be necessary to have spatial confidence regions for every



90% confidence region - $\Phi_1(r)$

95% confidence region - $\Phi_1(r)$

90% confidence region - $\Phi_2(r)$

95% confidence region - $\Phi_2(r)$

90% confidence region - $\Phi_3(r)$

95% confidence region - $\Phi_3(r)$

Figure 2.5: The coverage rates evaluated for the three classes of spatial confidence regions presented in Table 2.1, displayed in the form of heatmap and histogram. Note that the performances of $\mathbf{\Phi}_1(\boldsymbol{r})$ and $\mathbf{\Phi}_2(\boldsymbol{r})$ are fairly comparable to the ideal confidence region $\mathbf{\Phi}_3(\boldsymbol{r})$, as the coverage rates for many of the pixels come close to the prespecified confidence level γ.

voxel in the image volume. Therefore, after completing the original full 3-D registration, we suggest to run the N registrations only within a subregion where the accuracy of the initial registration must be known. This allows one to obtain spatial confidence regions for these locations at a much more reasonable computational expense.

While the presented work demonstrated promising preliminary results, there are several directions and open questions that remain for future research. For example, the natural next step is to perform more extensive validation studies in 3-D using various similarity criteria and deformation models, and explore a way to quantify the robustness of the method. Furthermore, other choices of *a priori* baseline for the shrinkage-based covariance estimate shall be investigated. It is also important to conduct a simulation study under various noise conditions, as image noise can significantly impact registration accuracy. Finally, it is important to seek a way to incorporate more data into our model to allow a more sophisticated parameter selection to take place.

CHAPTER 3

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

3.1 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 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

This chapter is based on Watanabe et al. (2014a,b) 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[1], 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-

[1] Here we mean *linear* in the correlation values, not the original data.

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., 2012b; 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. 3.4. 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 $\boldsymbol{w}, \boldsymbol{r} \in \mathbb{R}^n$, we invoke the standard notation $\langle \boldsymbol{w}, \boldsymbol{r} \rangle := \sum_{i=1}^n w_i v_i$ 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, $\|\cdot\| = \|\cdot\|_2$.

3.2 Defining Functional Connectomes

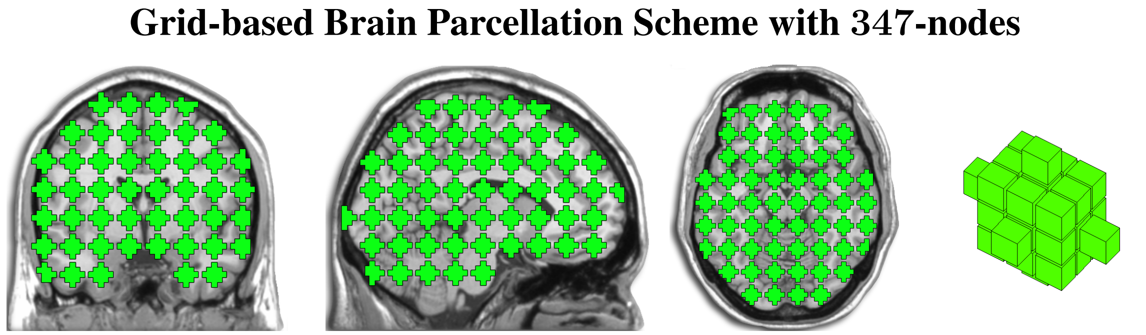
FMRI data consist of a time series of three dimensional volumes imaging the brain, where each 3-D volume encompasses around $10,000 \sim 100,000$ voxels. The univariate time series at each voxel represents a blood oxygen level dependent (BOLD) signal, an indirect measure of neuronal activities in the brain. Traditional experiments in the early years of fMRI research involved *task-based studies*, but after it was discovered that the brain is functionally connected at rest, *resting state* fMRI became a dominant tool for studying the network architecture of the brain. As such, we used the time series from resting state fMRI to generate FC's, which are correlation maps that describe brain connectivity.

More precisely, 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. 3.1. 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). 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. 3.4. 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



(a) Coronal (b) Sagittal (c) Axial (d) 33 voxel node

Figure 3.1: 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.

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.

3.3 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.

3.3.1 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 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

$$\operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell \left(y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle \right) + \lambda \mathcal{R}(\boldsymbol{w}) \,. \tag{3.1}$$

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

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

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 (3.1) 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).

3.3.2 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:

$$\operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^p} \frac{1}{n} \mathcal{L}(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}) + \lambda \left\| \boldsymbol{w} \right\|_1 + \gamma \sum_{j=2}^{p} \left| w^{(j)} - w^{(j-1)} \right| \,, \tag{3.2}$$

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 (3.2) 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

$$\operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^p} \frac{1}{n} \mathcal{L}(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}) + \lambda \left\| \boldsymbol{w} \right\|_1 + \gamma \sum_{j=1}^{p} \sum_{k \in \mathcal{N}_j} \left| \, w^{(j)} - w^{(k)} \right| \,, \tag{3.3}$$

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 $\lvert w^{(j)} - w^{(k)} \rvert$ in (3.3) 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):

$$\operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{n}\mathcal{L}(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}) + \lambda \left\lVert \boldsymbol{w} \right\rVert_1 + \frac{\gamma}{2}\sum_{j=1}^{p}\sum_{k\in\mathcal{N}_j}\left(w^{(j)} - w^{(k)}\right)^2 . \tag{3.4}$$

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 (3.3) and GraphNet (3.4) 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 coordinate j. Then, the first-order neighborhood set of j can be written precisely as[2]

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

Fig. 3.2 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

[2] 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 $\mathcal{N}_j$.

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. 3.7.1. Finally, the spatial coherence that fused Lasso and GraphNet promotes facilitates model 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 (3.3) and GraphNet (3.4) in the following unified form:

$$\underset{\boldsymbol{w} \in \mathbb{R}^p}{\arg\min} \, \frac{1}{n} \mathcal{L}(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1 + \frac{\gamma}{q} \|\boldsymbol{C}\boldsymbol{w}\|_q^q \ , \ q \in \{1, 2\} \ . \tag{3.5}$$

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.

3.4 Optimization

Solving the optimization problem (3.5) 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 (3.5) into an equivalent constrained optimization problem, which can be solved efficiently using the alternating

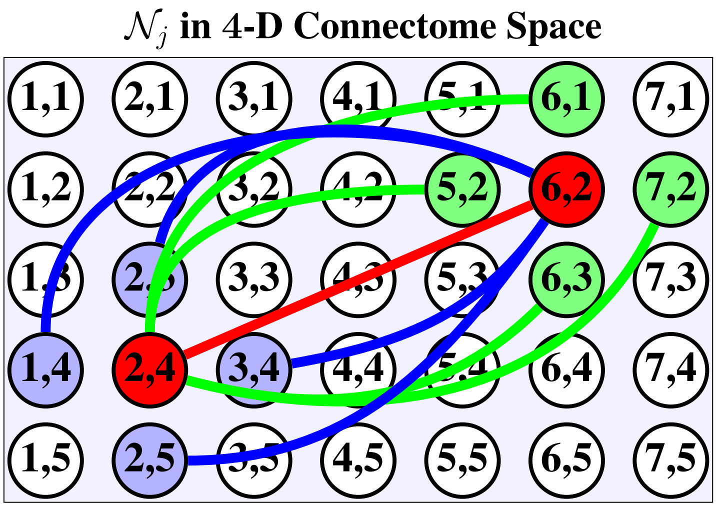


Figure 3.2: Illustration of the neighborhood structure of the connectome when the nodes reside in 2-D space. The red edge represents coordinate $j = \{(2,4),(6,2)\}$ in 4-D connectome space, and its neighborhood set $\mathcal{N}_j$ is represented by the blue and green edges. This idea extends directly to 6-D connectomes generated from 3-D resting state volumes.

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.

3.4.1 Alternating Direction Method of Multipliers

The ADMM algorithm is a powerful algorithm for solving large scale optimization problems. The method was first introduced in the 1970's (Gabay and Mercier, 1976; Glowinski and Marroco, 1975), but has recently generated renewed interest from the statistics and signal processing community, as large-scale datasets became more routinely encountered. We refer the readers to (Boyd et al., 2011) for an extensive review of ADMM.

More precisely, ADMM solves convex optimization problems having the separable structure

$$\min_{\bar{\boldsymbol{x}},\bar{\boldsymbol{y}}} \bar{\boldsymbol{f}}(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{g}}(\bar{\boldsymbol{y}}) \quad \text{subject to } \bar{\boldsymbol{A}}\bar{\boldsymbol{x}} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}} = \boldsymbol{0}\,, \tag{3.6}$$

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. In the classical AL framework, the primal variables are solved by the following iterations

$$\begin{aligned} \left(\bar{\boldsymbol{x}}^{(t+1)}, \bar{\boldsymbol{y}}^{(t+1)}\right) &\leftarrow \arg\min_{\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}} \boldsymbol{L}_{\boldsymbol{\rho}} \left(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}, \bar{\boldsymbol{u}}^{(t)}\right) \\ \bar{\boldsymbol{u}}^{(t+1)} &\leftarrow \bar{\boldsymbol{u}}^{(t)} + \rho \left(\bar{\boldsymbol{A}}\bar{\boldsymbol{x}}^{(t+1)} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}}^{(t+1)}\right), \end{aligned} \tag{3.7}$$

where superscript t denotes the iteration count, and

$$\boldsymbol{L}_{\boldsymbol{\rho}}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}, \bar{\boldsymbol{u}}) := \bar{\boldsymbol{f}}(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{g}}(\bar{\boldsymbol{y}}) + \langle \bar{\boldsymbol{u}}, \bar{\boldsymbol{A}}\bar{\boldsymbol{x}} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}} \rangle + \frac{\rho}{2} \left\| \bar{\boldsymbol{A}}\bar{\boldsymbol{x}} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}} - \bar{\boldsymbol{u}} \right\|^2 \tag{3.8}$$

is the AL function with dual variable $\bar{\boldsymbol{u}} \in \mathbb{R}^c$ and AL parameter $\rho > 0$. In practice, minimizing the AL function jointly over $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ can be challenging. Fortunately, ADMM exploits the separable structure in (3.6) by decomposing the primal variable update in (3.7) into two separate steps

$$\begin{aligned} \bar{\boldsymbol{x}}^{(t+1)} &\leftarrow \arg\min_{\bar{\boldsymbol{x}}} \boldsymbol{L}_{\boldsymbol{\rho}} \left(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}^{(t)}, \bar{\boldsymbol{u}}^{(t)}\right) \\ \bar{\boldsymbol{y}}^{(t+1)} &\leftarrow \arg\min_{\bar{\boldsymbol{y}}} \boldsymbol{L}_{\boldsymbol{\rho}} \left(\bar{\boldsymbol{x}}^{(t+1)}, \bar{\boldsymbol{y}}, \bar{\boldsymbol{u}}^{(t)}\right) \\ \bar{\boldsymbol{u}}^{(t+1)} &\leftarrow \bar{\boldsymbol{u}}^{(t)} + \rho \left(\bar{\boldsymbol{A}}\bar{\boldsymbol{x}}^{(t+1)} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}}^{(t+1)}\right) \ . \end{aligned} \tag{3.9}$$

This *alternating* minimization strategy is especially useful when it is easy to minimize $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ independently over $\boldsymbol{L}_{\boldsymbol{\rho}}$, a situation rather commonly encountered in practice. Note that by completing the square and defining the scaled dual variable $\boldsymbol{u} := \bar{\boldsymbol{u}}/\rho$, the ADMM iterations (3.9) can be written in the following equivalent form:

$$\bar{\boldsymbol{x}}^{(t+1)} \leftarrow \arg\min_{\bar{\boldsymbol{x}}} \bar{\boldsymbol{f}}(\bar{\boldsymbol{x}}) + \frac{\rho}{2} \left\| \bar{\boldsymbol{A}}\bar{\boldsymbol{x}} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}}^{(t)} + \boldsymbol{u}^{(t)} \right\|^2 \tag{3.10}$$

$$\bar{\boldsymbol{y}}^{(t+1)} \leftarrow \arg\min_{\bar{\boldsymbol{y}}} \bar{\boldsymbol{g}}(\bar{\boldsymbol{y}}) + \frac{\rho}{2} \left\| \bar{\boldsymbol{A}}\bar{\boldsymbol{x}}^{(t+1)} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}} + \boldsymbol{u}^{(t)} \right\|^2 \tag{3.11}$$

$$\boldsymbol{u}^{(t+1)} \leftarrow \boldsymbol{u}^{(t)} + \left(\bar{\boldsymbol{A}}\bar{\boldsymbol{x}}^{(t+1)} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}}^{(t+1)}\right) \ . \tag{3.12}$$

Unless otherwise stated, we will focus on this scaled formulation of ADMM, as it is more convenient to work with.

The convergence of the ADMM algorithm has been established by Mota *et al.* in Mota et al. (2011). 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, although this value can be empirically tuned in practice. For completeness and later reference, we now present the theorem providing the sufficient conditions for ADMM to converge.

Theorem 3.1 (Theorem 1 from Mota et al. (2011))**.** *Consider problem* (3.6)*, where* $\bar{\boldsymbol{f}}$ *and* $\bar{\boldsymbol{g}}$ *are convex functions over* $\mathbb{R}^{\bar{p}}$ *and* $\mathbb{R}^{\bar{q}}$ *respectively. Assume the linear constraint matrices* $\bar{\boldsymbol{A}} \in \mathbb{R}^{c\times\bar{p}}$ *and* $\bar{\boldsymbol{B}} \in \mathbb{R}^{c\times\bar{q}}$ *are full column-rank, and also assume problem* (3.6) *is solvable, i.e., it has an optimal objective value. Then the sequence* $\left\{\bar{\boldsymbol{x}}^{(t)}, \bar{\boldsymbol{y}}^{(t)}, \bar{\boldsymbol{u}}^{(t)}\right\}$ *generated by* (3.9) *converges to* $\{\bar{\boldsymbol{x}}^*, \bar{\boldsymbol{y}}^*, \bar{\boldsymbol{u}}^*\}$*, where*

1. $\{\bar{\boldsymbol{x}}^*, \bar{\boldsymbol{y}}^*\}$ *solves* (3.6).

2. $\bar{\boldsymbol{u}}^*$ *solves the dual problem of* (3.6)*:*

$$\max_{\bar{\boldsymbol{u}}} \bar{\boldsymbol{F}}(\bar{\boldsymbol{u}}) + \bar{\boldsymbol{G}}(\bar{\boldsymbol{u}}) \ ,$$

where $\bar{\boldsymbol{F}} := \inf_{\bar{\boldsymbol{x}}} \bar{\boldsymbol{f}}(\bar{\boldsymbol{x}}) + \langle \bar{\boldsymbol{u}}, \bar{\boldsymbol{A}}\bar{\boldsymbol{x}} \rangle$ *and* $\bar{\boldsymbol{G}} := \inf_{\bar{\boldsymbol{y}}} \bar{\boldsymbol{g}}(\bar{\boldsymbol{y}}) + \langle \bar{\boldsymbol{u}}, \bar{\boldsymbol{B}}\bar{\boldsymbol{y}} \rangle$.

3.4.2 Variable splitting and data augmentation

The original formulation of our problem (3.5) does not have the structure of (3.6). However, we can convert the unconstrained optimization problem (3.5) into an equivalent constrained optimization problem (3.6) 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 (3.5) is

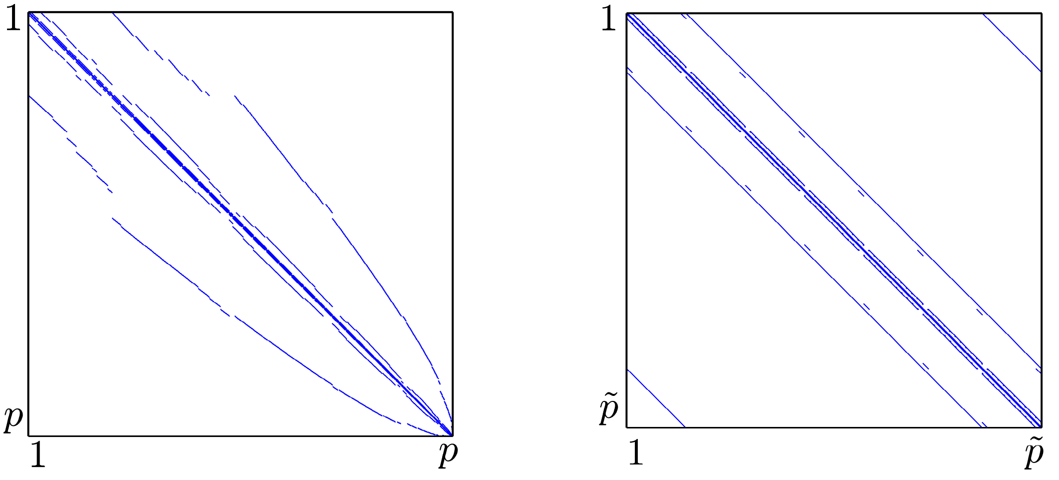
$$\begin{aligned} &\underset{\substack{\boldsymbol{w},\boldsymbol{v_1}\\ \boldsymbol{v_2},\boldsymbol{v_3},\boldsymbol{v_4}}}{\text{minimize}} \frac{1}{n}\mathcal{L}(\boldsymbol{v_1}) + \lambda \left\|\boldsymbol{v_2}\right\|_1 + \frac{\gamma}{q}\left\|\boldsymbol{v_3}\right\|_q^q \\ &\text{subject to } \boldsymbol{YXw} = \boldsymbol{v_1},\ \boldsymbol{w} = \boldsymbol{v_2},\ \boldsymbol{Cv_4} = \boldsymbol{v_3},\ \boldsymbol{w} = \boldsymbol{v_4}\ , \end{aligned} \tag{3.13}$$

where $\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}, \boldsymbol{v_4}$ are the constraint variables. It is easy to see that problems (3.5) and (3.13) are equivalent, and the correspondence with the ADMM formulation (3.6) is as follows:

$$\begin{gathered} \bar{\boldsymbol{f}}(\bar{\boldsymbol{x}}) = \frac{\gamma}{q}\left\|\boldsymbol{v_3}\right\|_q^q, \quad \bar{\boldsymbol{g}}(\bar{\boldsymbol{y}}) = \frac{1}{n}\mathcal{L}(\boldsymbol{v_1}) + \lambda\left\|\boldsymbol{v_2}\right\|_1 \\ \bar{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{YX} & 0 \\ \boldsymbol{I} & 0 \\ 0 & \boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix}, \ \bar{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{w} \\ \boldsymbol{v_3} \end{bmatrix}, \ \bar{\boldsymbol{B}} = \begin{bmatrix} -\boldsymbol{I} & 0 & 0 \\ 0 & -\boldsymbol{I} & 0 \\ 0 & 0 & -\boldsymbol{C} \\ 0 & 0 & -\boldsymbol{I} \end{bmatrix}, \ \bar{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{v_1} \\ \boldsymbol{v_2} \\ \boldsymbol{v_4} \end{bmatrix}. \end{gathered} \tag{3.14}$$

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. 3.3a displays the Laplacian matrix that results from the 347-node functional connectome defined in Section 3.2, 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



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

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

Figure 3.3: Laplacian matrix corresponding to the original data $\boldsymbol{C}^T\boldsymbol{C}$ and the augmented data $\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}}$, 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.

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* $\widetilde{\boldsymbol{w}} := \boldsymbol{A}\boldsymbol{w}$, 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 3.B. As a result, we now have a new differencing matrix $\widetilde{\boldsymbol{C}} \in \mathbb{R}^{\tilde{e}\times\tilde{p}}$ corresponding to $\widetilde{\boldsymbol{w}} \in \mathbb{R}^{\tilde{p}}$, whose Laplacian matrix $\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}} \in \mathbb{R}^{\tilde{p}\times\tilde{p}}$ has a systematic structure, as shown in Fig. 3.3b. In fact, this matrix has a special structure known as *block-circulant with circulant-blocks* (BCCB), which is critical since the matrix inversion involving $\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}}$ can be computed efficiently in closed form using the fast Fourier transform (FFT) (the utility of this property will be elaborated more in Section 3.4.3). 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 $\boldsymbol{B} \in \{0,1\}^{\tilde{e}\times\tilde{e}}$, we have $\|\boldsymbol{B}\widetilde{\boldsymbol{C}}\tilde{\boldsymbol{w}}\|_q^q = \|\boldsymbol{C}\boldsymbol{w}\|_q^q$ 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 (3.5) in the following equivalent form:

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

Moreover, this can be converted into a constrained optimization problem

$$\begin{gathered}\underset{\substack{\boldsymbol{w},\boldsymbol{v_1}\\ \boldsymbol{v_2},\boldsymbol{v_3},\boldsymbol{v_4}}}{\text{minimize}} \frac{1}{n}\mathcal{L}(\boldsymbol{v_1}) + \lambda\left\|\boldsymbol{v_2}\right\|_1 + \frac{\gamma}{q}\left\|\boldsymbol{B}\boldsymbol{v_3}\right\|_q^q \\ \text{subject to } \boldsymbol{Y}\boldsymbol{X}\boldsymbol{w} = \boldsymbol{v_1},\; \boldsymbol{w} = \boldsymbol{v_2},\; \widetilde{\boldsymbol{C}}\boldsymbol{v_4} = \boldsymbol{v_3},\; \boldsymbol{A}\boldsymbol{w} = \boldsymbol{v_4}\;,\end{gathered} \tag{3.15}$$

and the correspondence with the ADMM formulation (3.6) now becomes:

$$\begin{gathered}\bar{\boldsymbol{f}}(\bar{\boldsymbol{x}}) = \frac{\gamma}{q}\left\|\boldsymbol{B}\boldsymbol{v_3}\right\|_q^q,\quad \bar{\boldsymbol{g}}(\bar{\boldsymbol{y}}) = \frac{1}{n}\mathcal{L}(\boldsymbol{v_1}) + \lambda\left\|\boldsymbol{v_2}\right\|_1 \\ \bar{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{Y}\boldsymbol{X} & 0 \\ \boldsymbol{I} & 0 \\ 0 & \boldsymbol{I} \\ \boldsymbol{A} & 0 \end{bmatrix},\; \bar{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{w} \\ \boldsymbol{v_3} \end{bmatrix},\; \bar{\boldsymbol{B}} = \begin{bmatrix} -\boldsymbol{I} & 0 & 0 \\ 0 & -\boldsymbol{I} & 0 \\ 0 & 0 & -\widetilde{\boldsymbol{C}} \\ 0 & 0 & -\boldsymbol{I} \end{bmatrix},\; \bar{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{v_1} \\ \boldsymbol{v_2} \\ \boldsymbol{v_4} \end{bmatrix}.\end{gathered} \tag{3.16}$$

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 (3.10)-(3.12) is guaranteed by Theorem 3.1.

3.4.3 ADMM: efficient closed-form updates

With the variable splitting scheme (3.15) and ADMM formulation (3.16), the ADMM update for the primal variable $\bar{\boldsymbol{x}}$ (3.10) decomposes into subproblems

$$\boldsymbol{w}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{w}} \left\{ \left\| \boldsymbol{Y}\boldsymbol{X}\boldsymbol{w} - \left(\boldsymbol{v_1}^{(t)} - \boldsymbol{u_1}^{(t)}\right) \right\|^2 + \left\| \boldsymbol{w} - \left(\boldsymbol{v_2}^{(t)} - \boldsymbol{u_2}^{(t)}\right) \right\|^2 \right. \\ \left. + \left\| \boldsymbol{A}\boldsymbol{w} - \left(\boldsymbol{v_4}^{(t)} - \boldsymbol{u_4}^{(t)}\right) \right\|^2 \right\} \tag{3.17}$$

$$\boldsymbol{v_3}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{v_3}} \left\{ \frac{\gamma}{q} \left\| \boldsymbol{B}\boldsymbol{v_3} \right\|_q^q + \frac{\rho}{2} \left\| \boldsymbol{v_3} - \left(\tilde{\boldsymbol{C}}\boldsymbol{v_4}^{(t)} - \boldsymbol{u_3}^{(t)} \right) \right\|^2 \right\}, \tag{3.18}$$

whereas the updates for primal variable $\bar{\boldsymbol{y}}$ (3.11) are

$$\boldsymbol{v_1}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{v_1}} \left\{ \frac{1}{n}\mathcal{L}(\boldsymbol{v_1}) + \frac{\rho}{2} \left\| \boldsymbol{v_1} - \left(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}^{(t+1)} + \boldsymbol{u_1}^{(t)} \right) \right\|^2 \right\} \tag{3.19}$$

$$\boldsymbol{v_2}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{v_2}} \left\{ \lambda \left\| \boldsymbol{v_2} \right\|_1 + \frac{\rho}{2} \left\| \boldsymbol{v_2} - \left(\boldsymbol{w}^{(t+1)} + \boldsymbol{u_2}^{(t)} \right) \right\|^2 \right\} \tag{3.20}$$

$$\boldsymbol{v_4}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{v_4}} \left\{ \left\| \tilde{\boldsymbol{C}}\boldsymbol{v_4} - \left(\boldsymbol{v_3}^{(t+1)} + \boldsymbol{u_3}^{(t)} \right) \right\|^2 + \left\| \boldsymbol{v_4} - \left(\boldsymbol{A}\boldsymbol{w}^{(t+1)} + \boldsymbol{u_4}^{(t)} \right) \right\|^2 \right\}. \tag{3.21}$$

The update for the dual variable $\boldsymbol{u}$ is a trivial matrix-vector multiplication (3.12) (see Algorithm 2 line 14-17).

We now demonstrate that the minimization problems (3.17)-(3.21) each admits an efficient, closed form solution.

$\boldsymbol{w}$ update The quadratic minimization problem (3.17) has the following closed form solution:

$$\boldsymbol{w}^{(t+1)} \leftarrow \left(\boldsymbol{X}^T\boldsymbol{X} + 2\boldsymbol{I}_p \right)^{-1} \left(\boldsymbol{X}^T\boldsymbol{Y}^T[\boldsymbol{v_1}^{(t)} - \boldsymbol{u_1}^{(t)}] + [\boldsymbol{v_2}^{(t)} - \boldsymbol{u_2}^{(t)}] + \boldsymbol{A}^T[\boldsymbol{v_4}^{(t)} - \boldsymbol{u_4}^{(t)}] \right). \tag{3.22}$$

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 (3.22) 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*

$$\left(\boldsymbol{X}^T\boldsymbol{X} + 2\boldsymbol{I}_p\right)^{-1} = \frac{1}{2}\boldsymbol{I}_p - \frac{1}{4}\boldsymbol{X}^T\left(\boldsymbol{I}_n + \frac{1}{2}\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\ . \tag{3.23}$$

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 (3.23), 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 (3.22) can be implemented as follows:

$$w^{(t+1)} \leftarrow (\boldsymbol{X}^T\boldsymbol{X} + 2\boldsymbol{I}_p)^{-1}\boldsymbol{\varrho}^{(t)} = \frac{1}{2}\boldsymbol{\varrho}^{(t)} - \boldsymbol{H}\boldsymbol{X}\boldsymbol{\varrho}^{(t)}\ , \tag{3.24}$$

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 (3.19) and (3.20) 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

$$\mathrm{Prox}_{\tau F}(\boldsymbol{r}) = \operatorname*{arg\,min}_{\boldsymbol{u} \in \mathbb{R}^p} \tau F(\boldsymbol{u}) + \frac{1}{2}\left\|\boldsymbol{r} - \boldsymbol{u}\right\|^2 , \quad \tau > 0\ , \tag{3.25}$$

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 (3.25) if and only if the condition

$$\mathbf{0} \in \partial F(\boldsymbol{u}^*) + (\boldsymbol{u}^* - \boldsymbol{r})/\tau \tag{3.26}$$

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 (3.19) and (3.20) are fully separable across their coordinates, decomposing into the following sets of elementwise scalar optimization problems:

$$\left[\boldsymbol{v_1}^{(t+1)}\right]_i \;\leftarrow\; \operatorname{Prox}_{\frac{\ell}{n\rho}}\left(\left[\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}^{(t+1)} + \boldsymbol{u_1}^{(t)}\right]_i\right), \qquad i \in [n] \tag{3.27}$$

$$\left[\boldsymbol{v_2}^{(t+1)}\right]_j \;\leftarrow\; \operatorname{Prox}_{\frac{\lambda}{\rho}|\cdot|}\left(\left[\boldsymbol{w}^{(t+1)} + \boldsymbol{u_2}^{(t)}\right]_j\right), \qquad j \in [p]\,, \tag{3.28}$$

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 (3.27) can be derived in closed form using the optimality condition (3.26). For example, the proximal operator for the non-differentiable hinge-loss has the expression:

$$\operatorname{Prox}_{\tau\ell}(t) = \begin{cases} t & \text{if } t > 1 \\ 1 & \text{if } 1 - \tau \leqslant t \leqslant 1 \\ t + \tau & \text{if } t < 1 - \tau \end{cases}.$$

If differentiability is desired, one can instead use the *truncated least square* or the *huberized hinge-loss* (Wang et al., 2008a), which both admit closed form proximal operator as well. Fig. 3.4 plots a few commonly used margin-based losses and their corresponding proximal operators, and Table 3.1 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 (3.20) and the absolute loss function (3.28) corresponds to the well known *soft-threshold operator* (Donoho, 1995)

$$\mathrm{Soft}_\tau(t) := \begin{cases} t-\tau & \text{if } t > \tau \\ 0 & \text{if } |t| \leqslant \tau \\ t+\tau & \text{if } t < -\tau \end{cases} . \tag{3.29}$$

The absolute loss and the soft-threshold operator are also included in Fig. 3.4 and Table. 3.1 for completeness.

$\boldsymbol{v_3}$ **update** The solution to the minimization problem (3.18) depends on the choice of $q \in \{1, 2\}$, where $q = 1$ recovers fused Lasso and $q = 2$ recovers GraphNet.

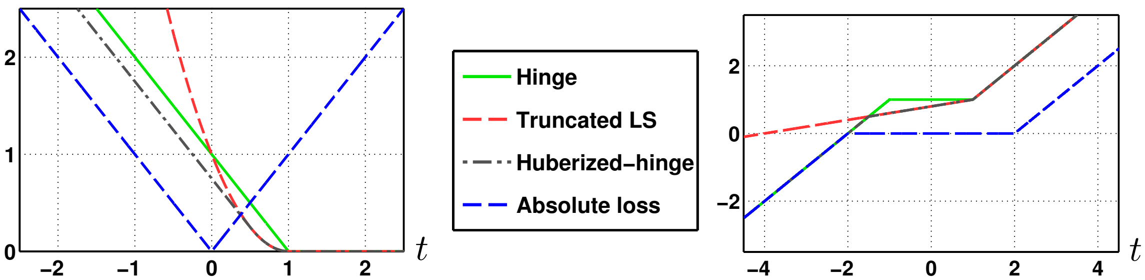
In the fused Lasso case $q = 1$, since the masking matrix $\boldsymbol{B} \in \{0,1\}^{\tilde{e}\times\tilde{e}}$ is diagonal, the update (3.18) is fully separable. Letting $\boldsymbol{\zeta}^{(t)} := \tilde{\boldsymbol{C}}\boldsymbol{v_4}^{(t)} - \boldsymbol{u_3}^{(t)}$, the minimization problem decouples into a set of scalar minimization problems of the form:

$$\operatorname*{arg\,min}_{v_k \in \mathbb{R}} \left\{ \gamma\, b_k\, |v_k| + \frac{\rho}{2}\left(v_k - \zeta_k^{(t)}\right)^2 \right\} \; , \qquad k \in [\tilde{e}] \tag{3.30}$$

where b_k is the k-th diagonal entry of $\boldsymbol{B}$ and $\zeta_k^{(t)}$ is the k-th entry of $\boldsymbol{\zeta}^{(t)} \in \mathbb{R}^{\tilde{e}}$. On one hand, when $b_k = 0$, the minimizer for problem (3.30) 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 (3.25) corresponding to the absolute loss function $|\cdot|$, recovering the soft-threshold operator (3.29). To summarize, when $q = 1$, the update for $\boldsymbol{v_3}$ (3.18) can be done efficiently by conducting the following elementwise update for each $k \in [\tilde{e}]$:

$$\left[\boldsymbol{v_3}^{(t+1)}\right]_k \leftarrow \begin{cases} \mathrm{Soft}_{\gamma/\rho}\left(\left[\tilde{\boldsymbol{C}}\left(\boldsymbol{v_4}^{(t)} - \boldsymbol{u_3}^{(t)}\right)\right]_k\right) & \text{if } \boldsymbol{B}_{k,k} = 1 \\ \left[\tilde{\boldsymbol{C}}\left(\boldsymbol{v_4}^{(t)} - \boldsymbol{u_3}^{(t)}\right)\right]_k & \text{if } \boldsymbol{B}_{k,k} = 0 \end{cases} \tag{3.31}$$

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



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

Figure 3.4: Plots of scalar convex loss functions that are relevant in this work, along with their associated proximal operators. Table 3.1 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.

	$\ell(t)$	$\mathrm{Prox}_{\tau\ell}(t)$
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	$\{\max(0, 1-t)\}^2$	$\begin{cases} t & \text{if } t > 1 \\ \dfrac{t+2\tau}{1+2\tau} & \text{if } t \leqslant 1 \end{cases}$
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)	$\mathrm{Soft}_\tau(t) := \begin{cases} t-\tau & \text{if } t > \tau \\ 0 & \text{if } \lvert t\rvert \leqslant \tau \\ t+\tau & \text{if } t < -\tau \end{cases}$

Table 3.1: Examples of scalar convex loss functions that are relevant for this work, along with their corresponding proximal operators in closed form.

In the GraphNet case $q = 2$, update (3.18) is a quadratic optimization problem with the closed form solution

$$\boldsymbol{v_3}^{(t+1)} \leftarrow \rho\big(\gamma\boldsymbol{B} + \rho\boldsymbol{I}_{\tilde{e}}\big)^{-1}\widetilde{\boldsymbol{C}}\big(\boldsymbol{v_4}^{(t)} - \boldsymbol{u_3}^{(t)}\big)\ , \tag{3.32}$$

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

$\boldsymbol{v}_4$ **update** The closed form solution to the quadratic optimization problem (3.21) is

$$\boldsymbol{v_4}^{(t+1)} \leftarrow \left(\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}} + \boldsymbol{I}_{\tilde{p}}\right)^{-1} \left(\widetilde{\boldsymbol{C}}^T[\boldsymbol{v_3}^{(t)} + \boldsymbol{u_3}^{(t)}] + \boldsymbol{A}\boldsymbol{w}^{(t+1)} + \boldsymbol{u_4}^{(t)}\right) . \tag{3.33}$$

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 (3.33) can be carried out efficiently using the (6-D) FFT

$$\boldsymbol{Q}^{-1}\boldsymbol{b} = \left(\boldsymbol{U}^H\boldsymbol{\Lambda}^{-1}\boldsymbol{U}\right)\boldsymbol{b} = \mathrm{ifft}\Big(\mathrm{fft}(\boldsymbol{b}) \oslash \boldsymbol{\phi}\Big) , \tag{3.34}$$

where fft and ifft denote the (6-D) FFT and inverse-FFT operation[3], $\boldsymbol{\phi}$ 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

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

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 2 outlines the complete ADMM algorithm for solving both the fused Lasso and GraphNet regularized ERM problem (3.5), 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:

$$\frac{\left\|\boldsymbol{w}^{(t+1)} - \boldsymbol{w}^{(t)}\right\|}{\left\|\boldsymbol{w}^{(t)}\right\|} \leqslant \varepsilon \ . \tag{3.35}$$

3.5 Experiment setup

3.5.1 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. 3.1). Specifically, we

Algorithm 2 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 (3.10)
        $\boldsymbol{w}^{(t+1)} \leftarrow \left(\boldsymbol{X}^T\boldsymbol{X} + 2\boldsymbol{I}_p\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{Y}^T[\boldsymbol{v_1}^{(t)} - \boldsymbol{u_1}^{(t)}] + [\boldsymbol{v_2}^{(t)} - \boldsymbol{u_2}^{(t)}] + \boldsymbol{A}^T[\boldsymbol{v_4}^{(t)} - \boldsymbol{u_4}^{(t)}]\right)$
                                                        ▷ apply update (3.24)
        $\boldsymbol{v_3}^{(t+1)} \leftarrow \begin{cases} \text{solve using (3.31)} & \text{if } q = 1 \text{ (fused Lasso)} \\ \text{solve using (3.32)} & \text{if } q = 2 \text{ (GraphNet)} \end{cases}$
    $\bar{\boldsymbol{y}}$-update (3.11)
        $\boldsymbol{v_1}^{(t+1)} \leftarrow \mathrm{Prox}_{\frac{\mathcal{L}}{n\rho}}\left(\boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}^{(t+1)} + \boldsymbol{u_1}^{(t)}\right)$      ▷ apply (3.27) elementwise
        $\boldsymbol{v_2}^{(t+1)} \leftarrow \mathrm{Soft}_{\lambda/\rho}\left(\boldsymbol{w}^{(t+1)} + \boldsymbol{u_2}^{(t)}\right)$      ▷ apply (3.28) 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 (3.34)
    $\boldsymbol{u}$-update (3.12)
        $\boldsymbol{u_1}^{(t+1)} \leftarrow \boldsymbol{u_1}^{(t)} + \boldsymbol{Y}\boldsymbol{X}\boldsymbol{w}^{(t+1)} - \boldsymbol{v_1}^{(t+1)}$
        $\boldsymbol{u_2}^{(t+1)} \leftarrow \boldsymbol{u_2}^{(t)} + \boldsymbol{w}^{(t+1)} - \boldsymbol{v_2}^{(t+1)}$
        $\boldsymbol{u_3}^{(t+1)} \leftarrow \boldsymbol{u_3}^{(t)} + \boldsymbol{v_3}^{(t+1)} - \tilde{\boldsymbol{C}}\boldsymbol{v_4}^{(t+1)}$
        $\boldsymbol{u_4}^{(t+1)} \leftarrow \boldsymbol{u_4}^{(t)} + \boldsymbol{A}\boldsymbol{w}^{(t+1)} - \boldsymbol{v_4}^{(t+1)}$
    $t \leftarrow t + 1$
until stopping criterion is met
```

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. 3.5a provides a schematic representation of the selected nodes.

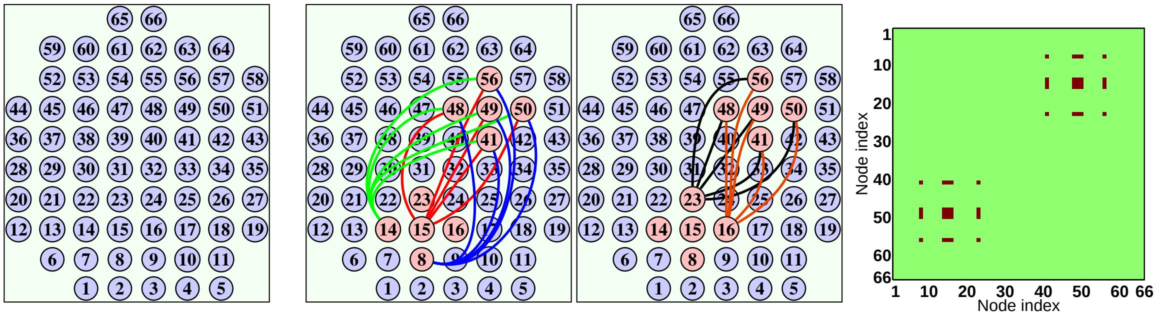
To mimic the *control vs. patient* binary classification setup, we created two classes of functional connectomes sampled from random normal distributions. The mean and the variance for these distributions were assigned using the functional connectomes generated from the real resting state dataset described later in Sec. 3.5.2. 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 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. 3.5b. 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 $d \cdot \hat{\sigma}(k)$ 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



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

Figure 3.5: 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.

Fig. 3.5b for a pictorial illustration of the anomalous edge set $\mathcal{K}$ in the 2-D node space. Fig. 3.5c 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 3.7.1. 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.

3.5.2 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[4] 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 3.2.

Data collection was performed at the Mind Research Network, and funded by a Center of Biomedical Research Excellence (COBRE) grant 5P20RR021938/P20GM103472 from the NIH to Dr. Vince Calhoun. The COBRE data set can also be downloaded from the Collaborative Informatics and Neuroimaging Suite data exchange tool (COINS)[5] (Scott et al., 2011).

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 $= 256 \times 256$ 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 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: $3 \times 3 \times 4mm^3$).

[4] Available at http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html.

[5] Available at http://coins.mrn.org/dx.

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

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 3.2.

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.

In the above preprocessing steps, since the slices of the functional images were not individually registered to the T1-weighted image, there could be motion-induced misalignments among the individual slices. However, we note that the rest scans were collected while the participants are in awake state, and the maximum motion in the data is generally less than the resolution of the voxel. Moreover, since the BOLD signals are spatially averaged over a 15 mm diameter ROI when generating the connectomes, the misalignments among the slices of the functional images will only have a negligible impact on the resulting functional connectomes.

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.

3.6 Results

3.6.1 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 3.5.1, where the training set consists of 100 samples with 50 patients and 50 controls. Specifically, we solved the regularized ERM problem (3.1) 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 2. The ADMM algorithm for Elastic-net is provided in 3.A, and the algorithm for Lasso follows directly from Elastic-net by setting $\gamma = 0$. The ADMM algorithm was terminated when the tolerance level (3.35) 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."

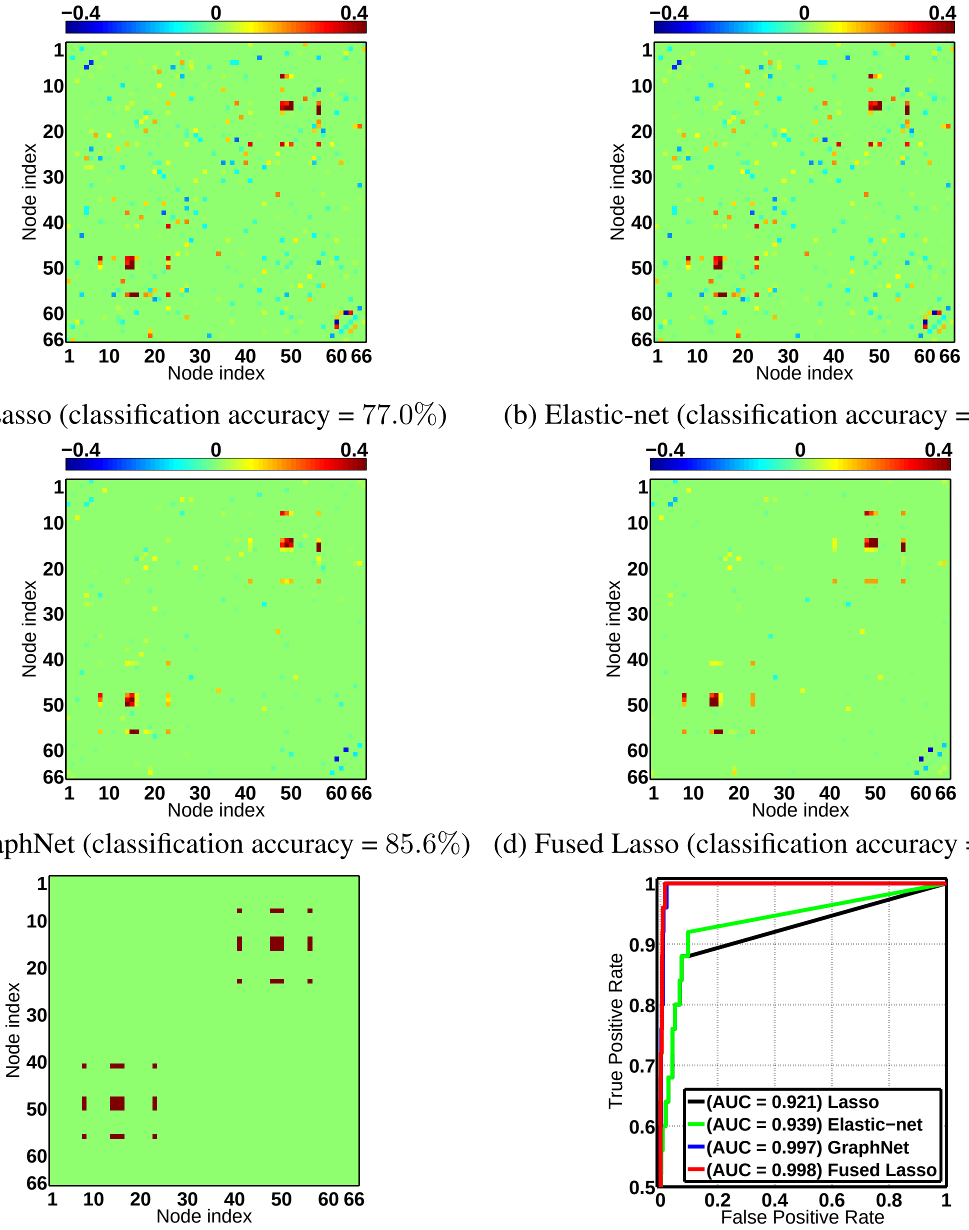
With the exception of Lasso, the regularizers we investigated involve two tuning parameters: $\lambda \geqslant 0$ and $\gamma \geqslant 0$. We tuned these regularization parameters by conducting a 5-fold cross-validation on the training set over a two-dimensional grid, and tuned Lasso over a one-dimensional 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[6]. The final weight vector estimates are obtained by re-training the classifiers on the entire training

[6] 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. 3.6.2 as well; see Fig. 3.7 and Fig. 3.9.

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.

Fig. 3.6a-d 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. 3.6e. 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. 3.6f; 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 and Elastic-net were 0.921 and 0.939 respectively. In short, Fig. 3.6a-f demonstrate that fused Lasso and GraphNet not only



(a) Lasso (classification accuracy = 77.0%)

(b) Elastic-net (classification accuracy = 77.0%)

(c) GraphNet (classification accuracy = 85.6%)

(d) Fused Lasso (classification accuracy = 88.2%)

(e) Support of the *anomalous edges*

(f) ROC (edge identification accuracy)

Figure 3.6: 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 (3.1) 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.

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. 3.7, 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. 3.8 and Table 3.3. 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

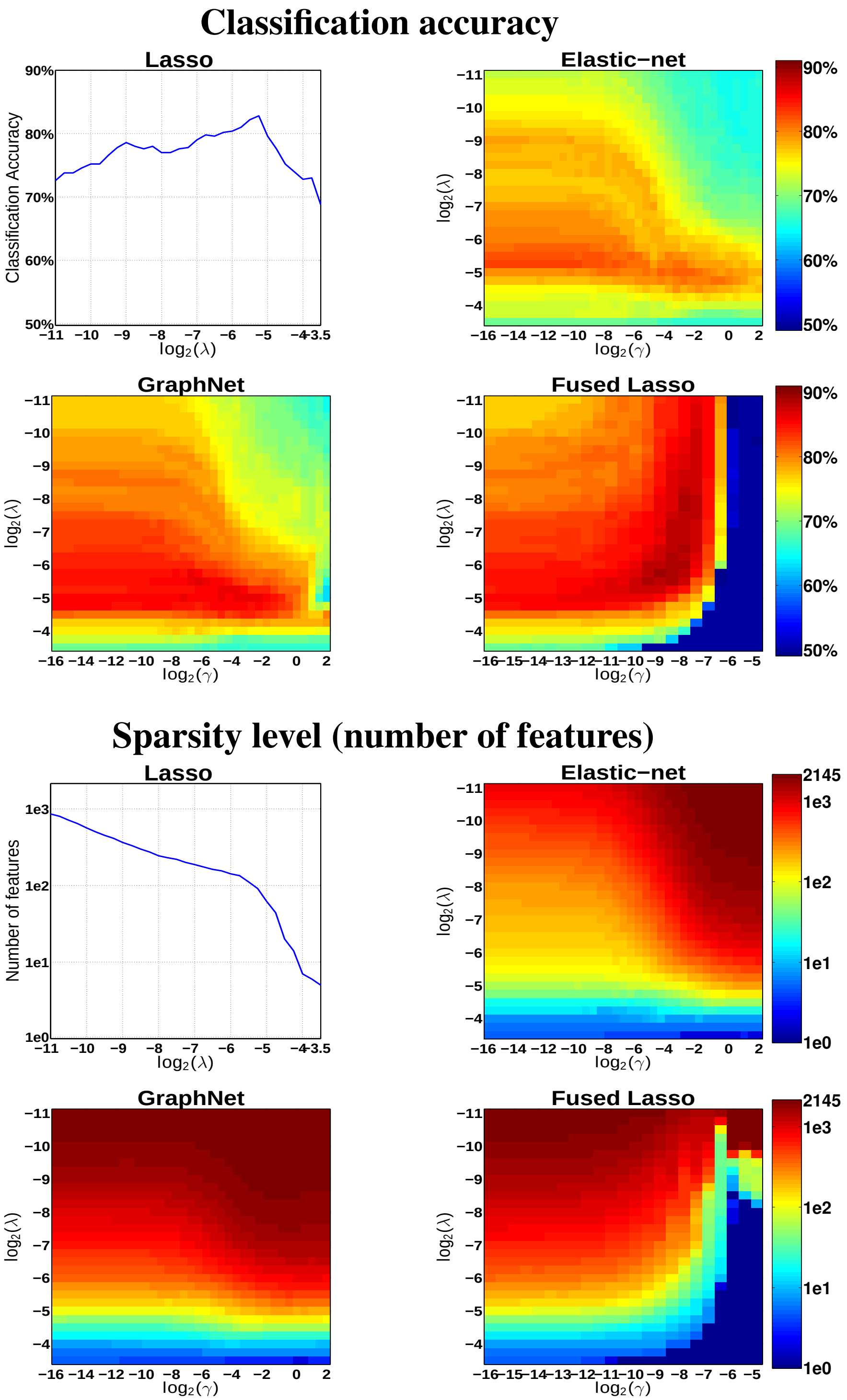


Figure 3.7: Grid search result for the simulation experiment (best viewed in color). All classifiers were learned using 100 training samples consisting of 50 patients and 50 controls. **Top two rows**: classification accuracy as a function of the regularization parameters $\{\lambda, \gamma\}$ (evaluated from 500 testing samples consisting of 250 patients and 250 controls). **Bottom two rows**: the number of features selected as a function of the regularization parameters $\{\lambda, \gamma\}$.

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 3.3: 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. 3.8 for a plot of this result.

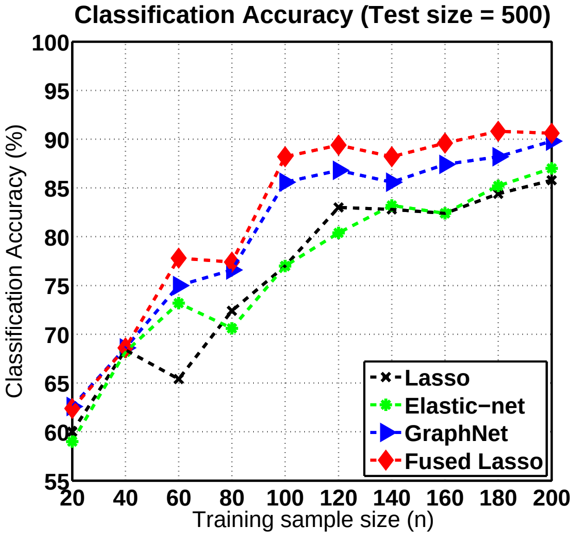


Figure 3.8: The testing classification accuracy of the different regularizers as a function as a number of training samples n in the simulation experiment. Regularization parameters were tuned via 5-fold cross-validation on the training set. The testing set consists of 500 samples with 250 patients and 250 controls. Table 3.3 reports the actual numbers.

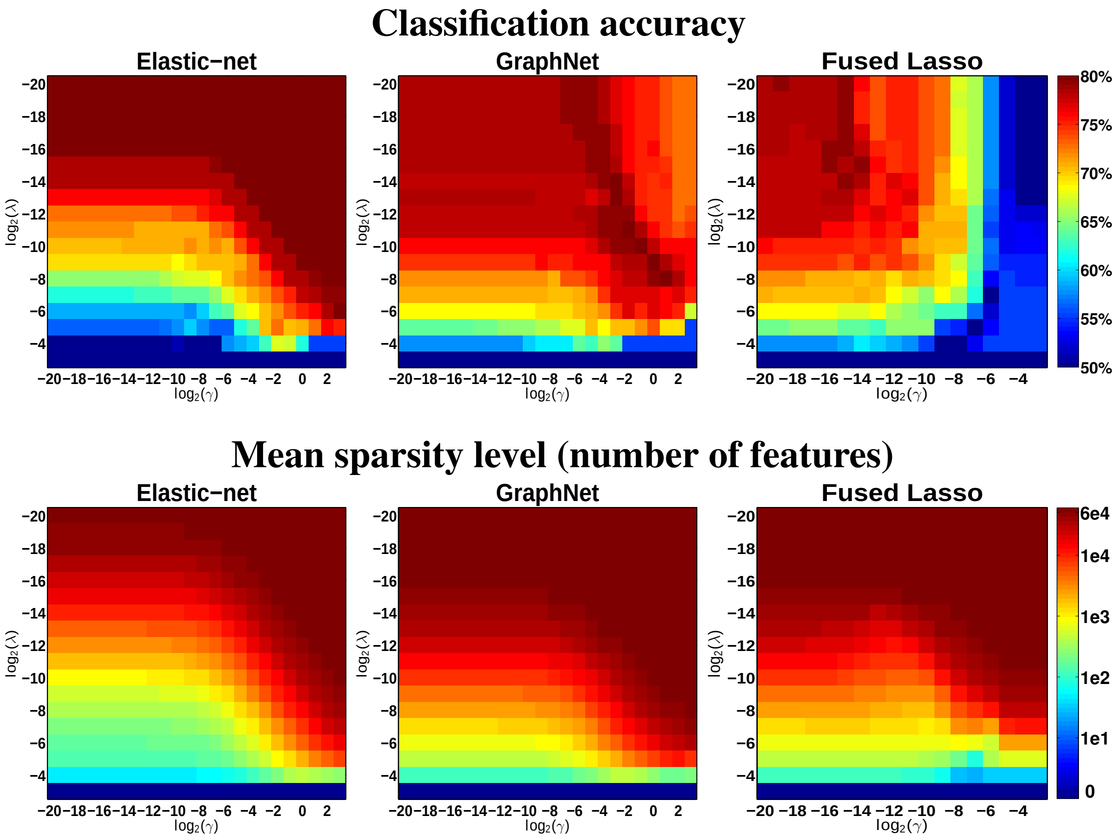
to favor the spatially-informed regularizers, but are motivated from the "patchiness assumption" of brain disorders, a neuroscientific viewpoint which we discuss in detail in Sec. 3.7.1. The results from the simulation experiments confirm the intuition that if the "patchiness assumption" of brain disorders holds true, spatially-informed classifiers can be a powerful tool for recovering relevant biosignatures.

3.6.2 Results on resting state fMRI data from a schizophrenia dataset

In this experiment, we examined the performance of linear classifiers trained using regularized ERM (3.1) 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 (3.35) 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. 3.9. 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



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

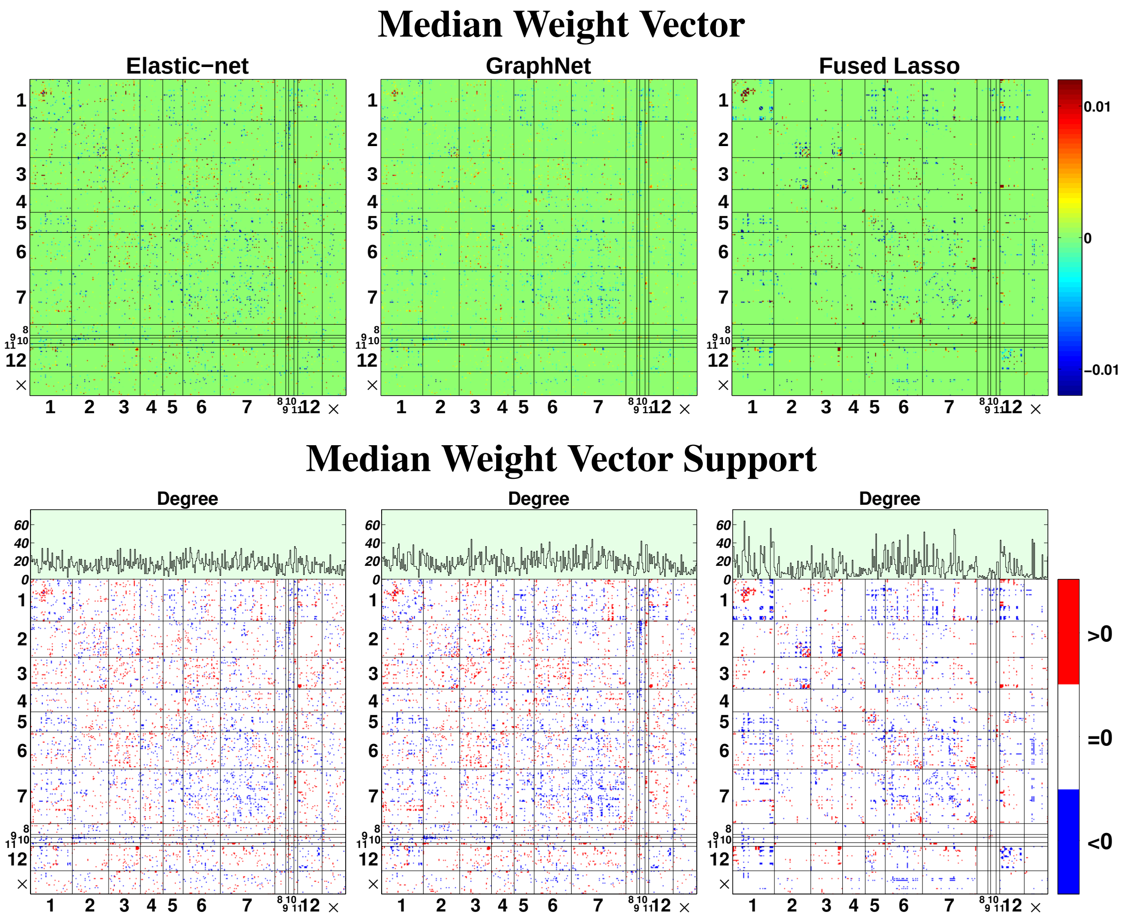
Figure 3.9: Grid search result for the real resting state data (best viewed in color). **Top row**: the classification accuracy evaluated from 10-fold cross-validation. **Bottom row**: the average number of features selected across the cross-validation folds. The (x, y)-axis corresponds to the two regularization parameters λ and γ.

reported in Fig. 3.9, 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. 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 parcellation of Tzourio-Mazoyer et al. (2002) (see Table 3.4); 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. 3.10.

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



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

Figure 3.10: 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 3.4. 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

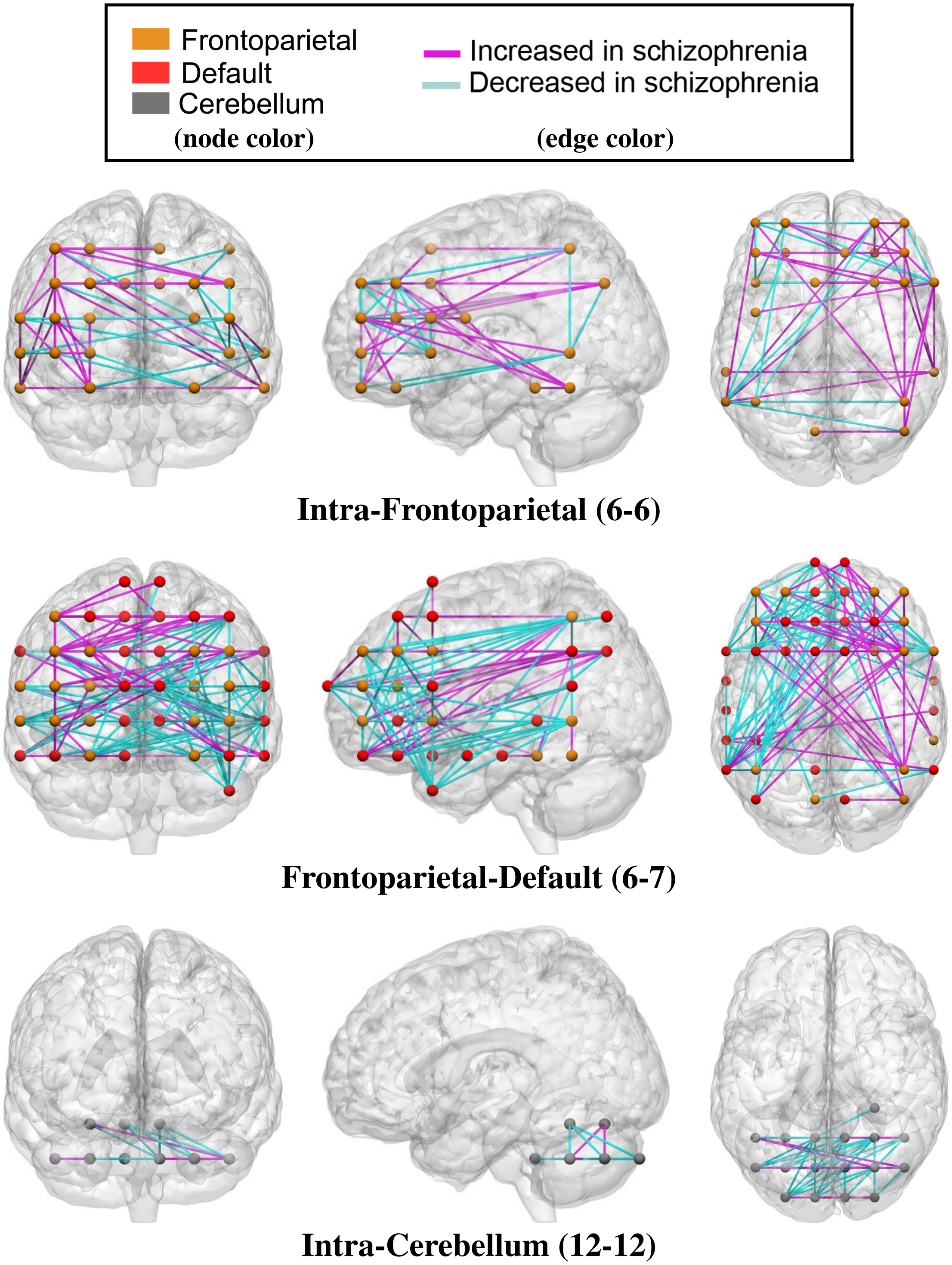
Table 3.4: Network parcellation of the brain proposed by Yeo et al. (2011). In our real resting state fMRI study, the indices of the estimated weight vectors are grouped according to this parcellation scheme; see Fig. 3.10.

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. 3.7.3 of the Discussion). In addition, the weight vector estimate from fused Lasso appears to implicate certain nodes more often in connectivity alterations. In order to emphasize this point, the bottom row in Fig. 3.10 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. 3.11; 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. 3.7). 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. 3.7).

3.6.3 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 3.A, whereas the algorithms for GraphNet



Intra-Frontoparietal (6-6)

Frontoparietal-Default (6-7)

Intra-Cerebellum (12-12)

Figure 3.11: 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.

and fused Lasso are reported in Algorithm 2). 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 $\|\boldsymbol{Cw}\|_q^q,\ q \in \{1,2\}$ is not separable across the coordinates of $\boldsymbol{w}$. To address this difficulty, the variable splitting strategy proposed for GraphNet and fused Lasso (3.15) contains four constraint variables, which is two more than the splitting proposed for Elastic-net (3.36); 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 (3.34), 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[7]. Using the same experimental

[7] 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.

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.

3.7 Discussion

Abundant neurophysiological evidence indicates that major psychiatric disorders are associated with distributed neural dysconnectivity (Konrad and Eickhoff, 2010; Müller et al., 2011; Stephan 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 3.2), 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.

3.7.1 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.

3.7.2 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 3.5.1). 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.

3.7.3 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. 3.10). 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. 3.10 and the 3-D brains shown in Fig. 3.11 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. 3.6.2). 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.

3.7.4 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).

In addition, we used Pearson's correlation coefficient as the measure of dependence between brain regions when constructing the functional connectomes. Since Pearson's correlation can only capture linear dependencies, an interesting future work would be to study the performance of our classifiers when nonlinear dependence measures such as mutual information is used (however, we note that it has been reported that nonlinear measures of dependence may not be necessary for fMRI data; see Hlinka et al. (2011); Richiardi et al. (2013)). Further, the symmetric matrices presented in Fig. 3.10 for interpreting the weight vectors are not positive semidefinite in general, which could limit its interpretability. Hence another important avenue for future exploration would be to modify our method so that these matrices are guaranteed to be positive semidefinite. One immediate way to do this is to convert the ERM problem (3.1) into a matrix optimization problem with a positive semidefinite constraint (Henrion and Malick, 2012).

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 (3.5) 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 (3.28) 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, 2008a; 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. Finally, developing an intuitive and accurate representation of the predictive edges in brain space remains as an open challenge for connectomic studies, as well as devising a performance measure that quantifies both accuracy and interpretability of different classifiers.

3.8 Conclusion

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.

3.A ADMM updates for Elastic-net

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

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

which can be converted into the following equivalent constrained formulation:

$$\operatorname*{minimize}_{\boldsymbol{w},\boldsymbol{v_1}\boldsymbol{v_2}} \frac{1}{n}\mathcal{L}(\boldsymbol{v_1}) + \lambda\left\|\boldsymbol{v_2}\right\|_1 + \frac{\gamma}{2}\left\|\boldsymbol{w}\right\|_2^2 \ \text{ subject to } \boldsymbol{YXw} = \boldsymbol{v_1},\ \boldsymbol{w} = \boldsymbol{v_2}\ . \tag{3.36}$$

With this variable splitting scheme, the correspondence with the ADMM formulation (3.6) 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 $\bar{\boldsymbol{x}}$ (3.10) and $\bar{\boldsymbol{y}}$ (3.11) decomposes into subproblems

$$\boldsymbol{w}^{(t+1)} \leftarrow \operatorname*{arg\,min}_{\boldsymbol{w}} \left\{ \frac{\gamma}{2}\left\|\boldsymbol{w}\right\|^2 + \left\|\boldsymbol{YXw} - \left(\boldsymbol{v_1}^{(t)} - \boldsymbol{u_1}^{(t)}\right)\right\|^2 + \left\|\boldsymbol{w} - \left(\boldsymbol{v_2}^{(t)} - \boldsymbol{u_2}^{(t)}\right)\right\|^2 \right\}$$

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

$$\boldsymbol{v_2}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{v_2}} \left\{ \lambda \left\|\boldsymbol{v_2}\right\|_1 + \frac{\rho}{2}\left\| \boldsymbol{v_2} - \left(\boldsymbol{w}^{(t+1)} + \boldsymbol{u_2}^{(t)}\right)\right\|^2 \right\} \; .$$

The update for $\boldsymbol{w}$ is

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

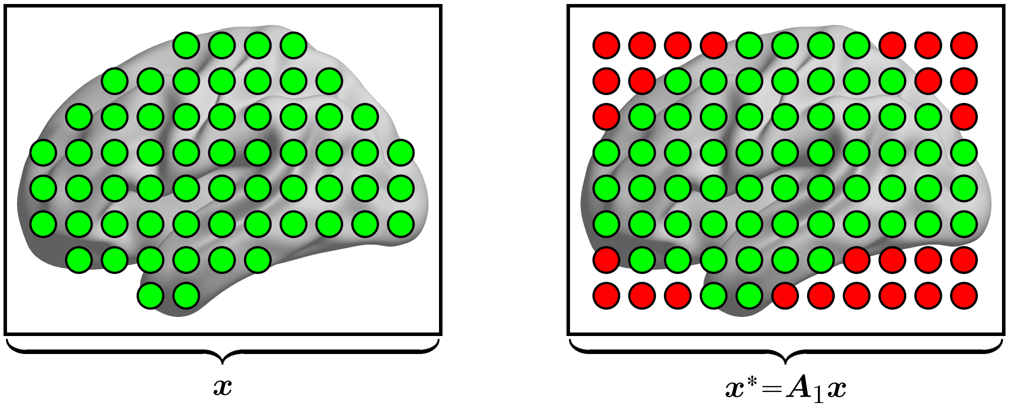
which can be solved efficiently via inversion Lemma (3.23). The update for $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ is identical to (3.19) and (3.20) described in Sec. 3.4.3, which can be solved via coordinate-wise proximal operators (3.27) and (3.28). The dual variable update (3.12) is a trivial matrix-vector multiplication.

3.B Details on the data augmentation scheme

As discussed in Sec. 3.4.2, 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. 3.12). 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. 3.12 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. 3.13). 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. 3.12). 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 3.12: 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.

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

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

(a) Intermediate augmented connectome

(b) Augmented functional connectome

Figure 3.13: 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.

CHAPTER 4

Multisite Disease Classification with Functional Connectomes via Multitask Structured Sparse SVM

4.1 Introduction

There is great interest in identifying neuroimaging biomarkers of psychiatric disorders, such as attention-deficit/hyperactivity disorder (ADHD), autism, Alzheimer's disease, and schizophrenia. Such discovery will not only deeply extend our knowledge about the functional architecture of the brain, but also offers the potential for an objective, machine-based diagnostic system to enter the clinical realm. To this end, multiple data-sharing initiatives have been launched in the neuroimaging field (Poldrack et al., 2013; Poline et al., 2012), including the ADHD-200, Alzheimer's Disease Neuroimaging Initiative (ADNI), Autism Brain Imaging Data Exchange (ABIDE), and Enhanced NKI-Rockland Sample dataset (Di Martino et al., 2013; Mennes et al., 2013; Nooner et al., 2012; The ADHD-200 Consortium, 2012; Weiner et al., 2010). These community-wide collaborative efforts offer unique potential, as they foster reproducible research and allow us to examine the association between diseases and biomarkers with unprecedented sample size.

A significant body of the literature indicates that several major psychiatric disorders are associated with topological alternations in the brain's functional network (Castellanos et al.,

This chapter is based on Watanabe et al. (2014c)

2013; Fox and Greicius, 2010). In particular, functional connectivity, which is measured by the statistical dependencies among the blood oxygenation level dependent (BOLD) signal between remote brain regions (Biswal et al., 1995), have played a critical role in helping us better understand the neurobiological mechanism of various disorders (Fox and Raichle, 2007; Greicius et al., 2003). Motivated by these findings, in this work we are interested in the supervised learning problem of binary classification, where the goal is to predict the diagnostic status of an individual using functional connectomes, which are high dimensional correlation maps derived from resting-state functional magnetic resonance imaging (fMRI) (Varoquaux and Craddock, 2013). However, multisite data present new challenges for this, as the data aggregation process introduces several sources of systematic confounds, such as variability in the scanner quality, image acquisition protocol, subject demographics, and other sources of experimental variations. In order to effectively make use of multisite data, it is important to train the classifiers in a way that accounts for these site-specific heterogeneities. To this end, we propose a classification framework that adopts a multitask learning approach (Argyriou et al., 2008; Caruana, 1997; Obozinski et al., 2010).

The idea behind multitask learning is to *jointly* train multiple tasks in order to improve classification performance, under the assumption that the tasks are related to each other in some sense. Recently, multitask learning methods have been successfully applied in brain decoding (Marquand et al., 2014; Rao et al., 2013), where the *participants* from a multi-subject fMRI study are treated as the tasks. The underlying assumption here is that the brain regions that are activated from a stimulus will share similar patterns across different tasks/subjects. In contrast to these works, the method we propose in this work treats the *sites* from which the resting state fMRI scans are collected as the tasks. In particular, we present *multitask structured sparse support vector machine* (SVM), a multitask extension to the connectome-based disease classification framework introduced in our recent work (Watanabe et al., 2014a). Unlike existing methods, our approach adopts a penalization scheme that accounts for the following two-way structure that exists in a multisite connectomic dataset: (1) the 6-D *spatial structure* in the functional connectomes that arises from pairs of points in 3-D brain space, and (2) the *inter-site* structure captured via the multitask ℓ_1/ℓ_2-penalty, which allows consistent model interpretation to be made by selecting the same set of informative features across sites (Chen et al., 2012a; Obozinski et al., 2010). In addition, to address the large dimensionality of functional connectomes, we introduce a scalable optimization algorithm based on the classical alternating direction method (Boyd et al., 2011; Gabay, 1983; Glowinski and Marroco, 1975).

To demonstrate the utility of our method, we perform experiments on the publicly available ADHD-200 dataset, a multisite dataset that contains resting state scans from seven contributing sites. Our empirical results not only shows that the proposed multitask approach can lead to improvement in classification performance, but also yields interpretable models that have consistent representation of informative features across sites.

Notation We let lowercase and uppercase bold letters denote vectors and matrices, respectively. For every positive integer $n \in \mathbb{N}$, we 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, $\|\cdot\| = \|\cdot\|_2$.

4.2 Material and Methods

4.2.1 Data and Preprocessing

Subjects We used the publicly available ADHD-200 competition data (The ADHD-200 Consortium, 2012), a multisite dataset that contains resting state scans of subjects diagnosed as either typically developing controls (TDC) or with ADHD. The dataset consists of a training set and a validation test set collected across seven sites: New York University Child Study Center, Beijing Normal University, University of Pittsburgh, Oregon Health and Science University, NeuroImage, Washington University at St. Louis, and Kennedy Krieger Institute[1]. Table 4.1 and 4.2 provide a summary of the demographic characteristics for each site in the training and validation test set. Informed consent was provided from all subjects, and study procedures complied with the Human Investigation Review Boards at respective sites. Detailed reporting of phenotypics, assessment protocols, and scanning parameters is available in Fair et al. (2013).

Data acquisition All participants were scanned on 3.0 Tesla scanners. Resting state scans used standard resting-connectivity T2*-weighted echo-planar imaging, whereas the structural scans used standard T1-weighted MPRAGE imaging. All imaging data used are publicly available at the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) (http://fcon_1000.projects.nitrc.org/indi/adhd200).

Image sample selection Analyses were limited to participants with the following: (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, gender, and handedness), although imputation was allowed for missing intelligence quotient (IQ) data (see below for details); (3) full IQ (FIQ) within two standard deviation (SD) of the overall sample mean; (4) mean framewise displacement (FD) within two SD of the overall sample mean.

After applying these sample selection criteria, we analyzed resting state scans from 628 individuals (TDC$=416$, ADHD$=212$) in the training set and 106 subjects (TD$=65$, ADHD$=41$) in the test set. Table 4.1 and 4.2 present the basic demographic characteristics

[1] Participants from Brown site are excluded from our study, as the diagnostic labels for these subjects have not been released.

	Typically Developing Controls				**ADHD**			
Site	n	Age	% Male	IQ	n	Age	% Male	IQ
Pre-exclusion								
KKI	61	10.3 ± 1.3	55.7	111.5 ± 10.3	22	10.2 ± 1.6	54.5	106.0 ± 15.2
NeuroImage	22	17.3 ± 2.6	50.0	111.2	22	17.0 ± 2.8	81.8	111.2
NYU	93	12.1 ± 3.1	45.2	110.7 ± 13.9	116	11.3 ± 2.7	77.6	106.4 ± 14.0
OHSU	41	8.9 ± 1.2	43.9	118.7 ± 12.6	37	8.8 ± 1.0	70.3	108.5 ± 13.9
Peking Univ.	116	11.7 ± 1.7	61.2	118.1 ± 13.3	78	12.4 ± 2.0	91.0	105.4 ± 13.2
Pittsburgh	89	15.1 ± 2.9	51.7	109.8 ± 11.5	— NA —			
Wash. U	59	11.5 ± 3.9	52.5	116.0 ± 14.1	— NA —			
Total	**481**	**12.2 ± 3.3**	**52.6**	**113.8 ± 12.9**	**275**	**11.6 ± 3.0**	**78.9**	**106.7 ± 13.3**
Post-exclusion								
KKI	55	10.4 ± 1.3	56.4	111.1 ± 10.7	19	10.4 ± 1.6	47.4	105.1 ± 14.8
NeuroImage	16	17.1 ± 2.3	50.0	111.2	12	16.6 ± 2.4	91.7	111.2
NYU	78	12.2 ± 3.2	44.9	111.6 ± 11.5	88	11.4 ± 2.8	75.0	108.7 ± 12.9
OHSU	35	9.1 ± 1.2	42.9	117.3 ± 11.9	27	9.0 ± 1.0	77.8	109.2 ± 12.5
Peking Univ.	108	11.8 ± 1.7	62.0	117.3 ± 12.0	66	12.5 ± 2.0	90.9	106.9 ± 12.2
Pittsburgh	78	15.3 ± 2.9	52.6	111.2 ± 10.5	— NA —			
Wash. U	49	11.6 ± 3.9	53.1	115.5 ± 13.6	— NA —			
Total	**419**	**12.3 ± 3.2**	**53.2**	**113.9 ± 11.7**	**212**	**11.6 ± 2.8**	**78.8**	**108.0 ± 12.4**

Table 4.1: Sample characteristics of the participants in the training set, shown both before and after application of exclusion and quality control criteria. Acronyms are: KKI = Kennedy Krieger Institute, NYU = New York University, OHSU = Oregon Health and Science University, Wash. U = Washington University in St. Louis.

of the pre-exclusion and post-exclusion sample for the training set and validation test set, respectively. Of note, for participants lacking a F4 or F2 IQ score, full IQ was estimated by computing the average of the participant's performance and verbal IQ scores. For the NeuroImage site which lacked IQ information in the training set, the mean IQ across the other sites was imputed.

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

	Typically Developing Controls				**ADHD**			
Site	n	Age	% Male	IQ	n	Age	% Male	IQ
Pre-exclusion								
KKI	8	10.6 ± 1.2	87.5	115.0 ± 5.8	3	8.7 ± 0.7	100	110.7 ± 13.3
NeuroImage	14	20.4 ± 3.3	7.1	100.6 ± 14.4	11	17.0 ± 2.2	100	93.3 ± 16.5
NYU	12	11.8 ± 3.0	66.7	114 ± 13.4	29	10.3 ± 2.5	69.0	103 ± 13.6
OHSU	28	9.6 ± 1.3	46.4	113.2 ± 12.8	6	10.1 ± 1.4	66.7	117.0 ± 12.8
Peking Univ.	27	10.2 ± 1.9	48.2	117.2 ± 12.5	24	11.1 ± 2.0	79.2	108.1 ± 12.9
Pittsburgh	5	14.3 ± 0.6	80.0	109.6 ± 15.3	4	15.4 ± 1.4	75.0	103.8 ± 11.0
Wash. U	— NA —				— NA —			
Total	**94**	**12.0 ± 4.2**	**48.9**	**112.7 ± 13.5**	**77**	**11.7 ± 3.2**	**77.9**	**104.7 ± 14.6**
Post-exclusion								
KKI	5	10.3 ± 1.2	80.0	114.8 ± 6.5	3	8.7 ± 0.7	100	110.7 ± 13.3
NeuroImage	11	20.8 ± 2.9	9.1	102.2 ± 15.1	5	17.2 ± 2.7	100	95.6 ± 19.5
NYU	— NA —				2	11.5 ± 3.1	100	115.5 ± 4.9
OHSU	22	9.6 ± 1.3	50.0	112.8 ± 11.1	5	9.9 ± 1.4	60.0	118.0 ± 14.1
Peking Univ.	22	10.3 ± 2.0	40.9	117.0 ± 8.3	22	11.3 ± 2.0	81.8	107.7 ± 9.8
Pittsburgh	5	14.3 ± 0.6	80.0	109.6 ± 15.3	4	15.4 ± 1.4	75.0	103.8 ± 11.0
Wash. U	— NA —				— NA —			
Total	**65**	**12.1 ± 4.5**	**44.6**	**112.3 ± 11.9**	**41**	**12.1 ± 3.1**	**82.9**	**107.7 ± 12.8**

Table 4.2: Sample characteristics of the participants in the validation test set, shown both before and after application of exclusion and quality control criteria.

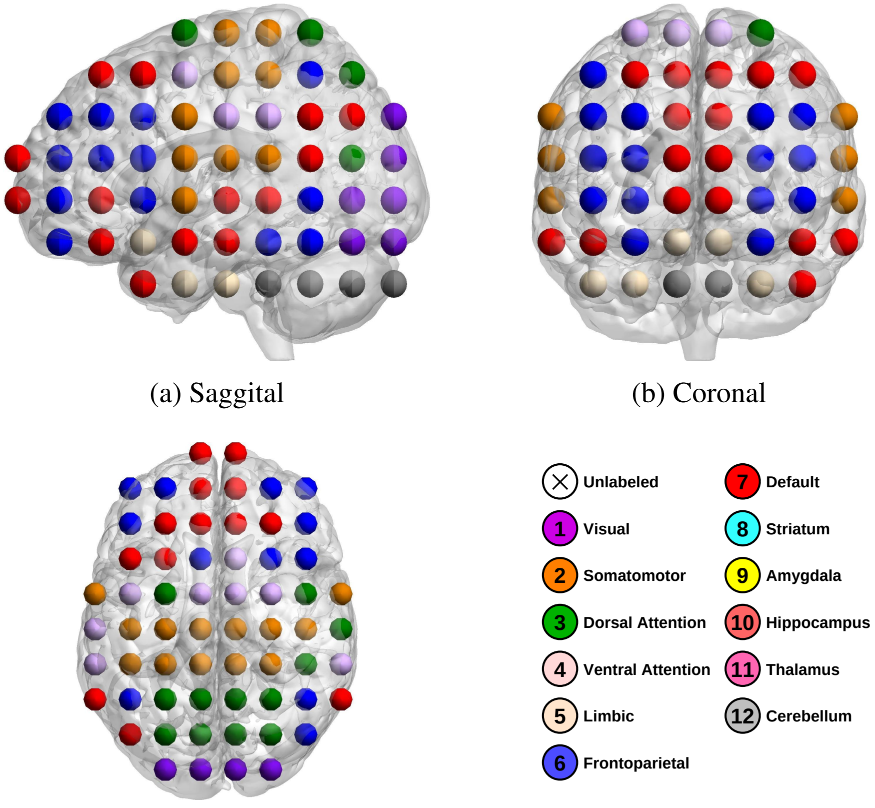
co-registered with the high-resolution T1-weighted image. Normalization was performed using the voxel-based morphometry (VBM) toolbox implemented in SPM8. The high-resolution T1-weighted image was segmented into tissue types, bias-corrected, registered to MNI space, and then normalized using Diffeomorphic 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 To generate the whole-brain resting state functional connectomes, we employed the grid-based parcellation scheme similar to our previous studies (Kessler et al., 2014; Sripada et al., 2014; Watanabe et al., 2014a) (see Watanabe et al. (2014a) for an extensive discussion on the advantages provided by this parcellation scheme). More specifically, we placed 347 non-overlapping nodes throughout the brain, where each of these nodes represents a pseudo-spherical ROI with a radius of 7.5 mm encompassing 33 $3 \times 3 \times 3$ mm voxels in a regular grid spaced at $18 \times 18 \times 18$ mm intervals. Fig. 4.1 provides a pictorial illustration of our parcellation scheme, with the color of the nodes indicating network membership proposed by Yeo et al. (2011).

Spatially averaged time series were next 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. Pearson product-moment correlation coefficients were then calculated pairwise between time courses for each of the 347 ROIs, resulting in a feature vector $\boldsymbol{x}$ of length $\binom{347}{2} = 60,031$ which serves as the feature vector for our disease classification framework.

4.2.2 Supervised Learning and the Multitask Framework

In this work, we propose a penalized empirical risk minimization framework for learning a separate yet related classification model for each site. More formally, suppose we are given K supervised learning tasks, where for each task indexed by $k = 1, \cdots, K$, we are given n_k input and output pairs $(\boldsymbol{x}_1^k, y_1^k), \cdots, (\boldsymbol{x}_{n_k}^k, y_{n_k}^k) \in \mathbb{R}^p \times \{\pm 1\}$. In the context of our work, $\boldsymbol{x}_i^k$ and y_i^k represent the functional connectome and the diagnostic label of the i-



(a) Saggital

(b) Coronal

(c) Axial

(d) Node Labels

Figure 4.1: Sagittal, coronal, and axial slices depicting the coverage of our brain parcellation scheme, where each nodes represents an ROI encompassing 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. The color of the nodes represents the network membership according to the parcellation scheme proposed by Yeo et al. (2011), as outlined in (d).

th subject from the k-th site, respectively. The objective is to simultaneously learn K linear classifiers of the form $f_k(\boldsymbol{x}) = \text{sign}(\langle \boldsymbol{w}^k, \boldsymbol{x} \rangle)$, where $\boldsymbol{w}^1, \ldots, \boldsymbol{w}^K \in \mathbb{R}^p$ are task-specific weight vectors obtained by solving the following optimization problem:

$$\underset{\boldsymbol{w}^1,\ldots,\boldsymbol{w}^K \in \mathbb{R}^p}{\arg\min} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \ell\left(y_i^k \langle \boldsymbol{w}^k, \boldsymbol{x}_i^k \rangle\right) + \mathcal{R}(\boldsymbol{w}^1, \ldots, \boldsymbol{w}^K)\,. \tag{4.1}$$

The first term in (4.1) is the *pooled empirical risk* of a margin-based loss function $\ell : \mathbb{R} \to \mathbb{R}_+$, which quantifies the quality of the model fit across all tasks. Traditional loss functions for classification include hinge, logistic, and exponential loss. In this work, we employ the *hinge-loss* $\ell(t) = \max(1-t, 0)$ from the well known SVM classifier (Cortes and Vapnik, 1995), although other convex margin-based losses can be used as well. The second term $\mathcal{R} : \mathbb{R}^{pK} \to \mathbb{R}_+$ in (4.1) is a penalty function that enforces certain kind of structure on the weight vectors, thereby allowing us to encode prior knowledge about the data. For brevity, let us define a functional $\mathcal{L}(\boldsymbol{Y}^k\boldsymbol{X}^k\boldsymbol{w}^k) := \sum_{i=1}^{n_k} \ell(y_i^k\langle \boldsymbol{w}^k, \boldsymbol{x}_i^k\rangle)$ which aggregates the empirical loss from the k-th task; here $\boldsymbol{X}^k \in \mathbb{R}^{n_k \times p}$ denotes the design matrix for the k-th task and $\boldsymbol{Y}^k \in \{-1, 0, +1\}^{n_k \times n_k}$ is defined as $\boldsymbol{Y}^k := \text{diag}(y_1^k, \dots, y_{n_k}^k)$. Also for conciseness, let $\underline{\boldsymbol{w}} \in \mathbb{R}^{pK}$ denote the vector obtained by stacking the weight vectors $\{\boldsymbol{w}^k\}_{k=1}^K$ together, which lets us rewrite the penalty term in (4.1) as $\mathcal{R}(\underline{\boldsymbol{w}})$.

In a high dimensional setup where the number of features greatly exceeds the sample size, the following ℓ_1-penalty known as the Lasso (Tibshirani, 1996) has been commonly applied in various applications:

$$\mathcal{R}(\underline{\boldsymbol{w}}) = \|\underline{\boldsymbol{w}}\|_1 = \sum_{k=1}^{K}\sum_{j=1}^{p} |w_j^k|\,. \tag{4.2}$$

Lasso possesses the important *variable selection* property, which promotes sparsity by setting many of the weight vector coefficients to zero, which can enhance prediction performance and interpretability by eliminating redundant features that only contribute as noise. Sparsity is also appealing from a connectomic point of view, as it is widely recognized that psychiatric disorders only impact a subset of the brain network, a view that has been validated in many existing studies (Castellanos et al., 2013; Fox and Greicius, 2010). Coupled with a linear classification model (4.1), we can directly interpret the non-zero coefficients of the weight vector as edges that are informative for disease prediction. However, a major drawback of Lasso is that it does not account for any additional structure in the data outside of sparsity. For instance, beyond sparsity, we further know that multisite functional connectome datasets posses the following two structures: (1) the *intra-site* spatial structure that characterizes the geometry of the functional connectomes, and (2) the *inter-site* structure that describes the similarity of the data across imaging sites.

To address this issue, in this work we will focus on convex penalty functions that consist of two parts:

$$\mathcal{R}(\underline{\boldsymbol{w}}) = \gamma \sum_{k=1}^{K} \mathcal{R}_1(\boldsymbol{w}^k) + \lambda \mathcal{R}_2(\underline{\boldsymbol{w}}), \tag{4.3}$$

where $\gamma \geqslant 0$ and $\lambda \geqslant 0$ are hyperparameters. The first penalty $\mathcal{R}_1$ allows us to encode prior knowledge about the *intra-task* structure of the data, *i.e.*, the intrinsic structure in the functional connectome that is independent of its originating site (note how this penalty is separable across the tasks). On the other hand, $\mathcal{R}_2$ is a multitask penalty that allows us incorporate a notion of "task-relatedness" by enforcing some form of mutual dependence among the set of weight vectors $\{\boldsymbol{w}^k\}_{k=1}^K$ across sites. Thus, the objective function we wish to solve has the following form:

$$\underset{\underline{\boldsymbol{w}} \in \mathbb{R}^{Kp}}{\arg\min} \sum_{k=1}^{K} \frac{1}{n_k} \mathcal{L}(\boldsymbol{Y}^k \boldsymbol{X}^k \boldsymbol{w}^k) + \gamma \sum_{k=1}^{K} \mathcal{R}_1(\boldsymbol{w}^k) + \lambda \mathcal{R}_2(\underline{\boldsymbol{w}})\,. \tag{4.4}$$

We next discuss and motivate our choices for $\mathcal{R}_1$ and $\mathcal{R}_2$.

4.2.2.1 Intra-task Structure: 6-D Spatial Penalty

FMRI data are known to exhibit rich spatio-temporal correlation patterns among neighboring voxels and time points. Indeed, several works in the *brain decoding* literature demonstrated that by leveraging these structures, it is possible to enhance the accuracy, interpretability, and stability of the prediction model (Baldassarre et al., 2012; Gramfort et al., 2013; Grosenick et al., 2013; Michel et al., 2011). Motivated by these successes, we recently introduced a single-task penalization framework in the context of connectomics (Watanabe et al., 2014a). In brief, the penalties adopted in Watanabe et al. (2014a) capture the 6-D spatial structure in the functional connectomes, a structure that arises from the fact that the coordinates of connectomes are defined by pairs of points in 3-D brain space. In this work, we propose to utilize and extend these types of penalties for the intra-task $\mathcal{R}_1$ penalty to capture the 6-D spatial structures contained in multisite connectomic data. Of note, this structure reflects the intrinsic geometrical patterns of a functional connectome and is thus independent of its originating site.

More formally, we account for the 6-D spatial structure of the connectomes by employing either the GraphNet (Grosenick et al., 2013), fused Lasso (Tibshirani et al., 2005) (also known as *anisotropic* total variation), or the *isotropic* total variation (TV) penalty (Michel et al., 2011; Wang et al., 2008b). For the case of GraphNet and fused Lasso, the penalty can be expressed compactly using a 6-D finite differencing matrix $\boldsymbol{C} \in \mathbb{R}^{e \times p}$ as follows:

$$\mathcal{R}_1(\boldsymbol{w}^k) = \frac{1}{q} \left\| \boldsymbol{C}\boldsymbol{w}^k \right\|_q^q = \frac{1}{q} \sum_{j=1}^{p} \sum_{m \in \mathcal{N}_j} \left| w_j^k - w_m^k \right|^q = \begin{cases} \text{GraphNet} & \text{if } q = 2 \\ \text{Fused Lasso} & \text{if } q = 1\,. \end{cases}$$

Here $\mathcal{N}_j$ is the first-order nearest-neighbor edge set corresponding to connectome coordinate j, and e indicates the total number of adjacent coordinates in the connectome. The closed form expression for the isotropic TV penalty admits a similar formulation, which is reported in 4.A.

To gain a better understanding of $\mathcal{N}_j$, let us denote (x, y, z) and (x', y', z') the pair of 3-D points in the brain that defines the 6-D connectome coordinate j. Then the first-order neighborhood set of j can be written precisely as:[2]

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

Thus the idea behind the GraphNet, fused Lasso, and isotropic TV is to promote spatial coherence in the weight vectors $\boldsymbol{w}^k$ by penalizing deviations among neighboring edges of the functional connectomes. This allows us to mathematically model our prior knowledge

[2] 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 $\mathcal{N}_j$.

that disease-induced abnormalities manifest in the brain by impacting spatially contiguous regions.

We conclude this section by noting that GraphNet, fused Lasso, and isotropic TV each induces slightly different forms of spatial contiguity. The absolute deviation penalty from fused Lasso encourages the predictive clusters to appear as sharp piecewise constant patches. Likewise, isotropic TV also promotes sharp piecewise constant patches, but also possesses the rotational invariance property that fused Lasso lacks (Michel et al., 2011). Finally, the quadratic penalty from GraphNet encourages the clusters to appear in smoother form.

4.2.2.2 Inter-task Structure: ℓ_1/ℓ_2-Penalty and Group Variable Selection

As mentioned earlier, since we expect the numbers of disease-induced discriminatory edges in the connectomes to be sparse, variable selection is of great importance in terms of both prediction performance and model interpretability. Furthermore, we also expect the connectivity-based biomarkers to be shared across the sites. To formalize this notion of *shared sparsity* pattern, we employ the ℓ_1/ℓ_2-penalty, a multitask penalty that has been widely adopted in various research areas (Chen et al., 2012a; Obozinski et al., 2010).

Specifically, let $\boldsymbol{w}_j = [w_j^1, \ldots, w_j^K]^T \in \mathbb{R}^K$ denote the vector formed by stacking the j-th weight vector coefficients across the K sites. Then the ℓ_1/ℓ_2-penalty is defined as:

$$\mathcal{R}_2(\underline{\boldsymbol{w}}) = \sum_{j=1}^{p} \|w_j\|_2\,, \tag{4.5}$$

i.e., it penalizes the sum of the ℓ_2-norm of $\boldsymbol{w}_j$'s, the vector representing the j-th edge in the connectome. This penalty has the appealing *group variable selection* property (Obozinski et al., 2010), which promotes learning features that are relevant across all sites, thereby simplifying interpretation of the selected features. At the same time, the actual weights associated with a given correlation can vary across site, in contrast to training a single classifier over a pooled dataset. This alleviates the issue of inter-site variability by allowing the amount of influence from a selected edge to vary across site, while the inter-site information are effectively shared to assist the group variable selection process.

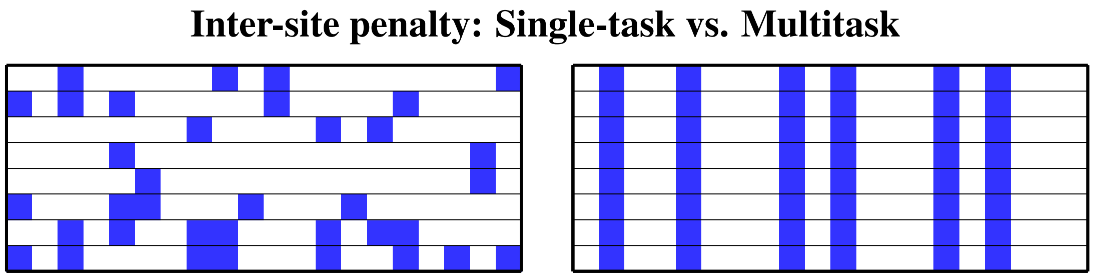
Of note, if we replace the ℓ_2-norm in (4.5) with the ℓ_1-norm, we recover the Lasso penalty (4.2):

$$\mathcal{R}_2(\underline{\boldsymbol{w}}) = \sum_{j=1}^{p} \left\| \boldsymbol{w}_j \right\|_1 = \sum_{k=1}^{K} \left\| \boldsymbol{w}^k \right\|_1 . \tag{4.6}$$

Following Obozinski et al. (2010), we will refer to this penalization scheme as the ℓ_1/ℓ_1-penalty, which is equivalent to a single-task procedure due to the separability of the penalty across the K sites. Fig 4.2 provides an illustration of the type of sparsity pattern one can expect from the single-task ℓ_1/ℓ_1-penalty and the multitask ℓ_1/ℓ_2-penalty.

Recently, Rao et al. (2013) introduced another multitask learning framework for multi-subject fMRI analysis, where the voxels in the fMRI volumes are used as the features for predicting the type of stimulus a subject is processing at different time points (*e.g.*, visual vs. auditory stimulus), and the *subjects* are treated as the tasks. Specifically, they proposed *Sparse Overlapping Sets Lasso (SOS-Lasso)* penalty, which can be viewed as a generalization of the ℓ_1-penalty (4.6) and the ℓ_1/ℓ_2-penalty (4.5). In brief, the SOS-Lasso penalty is motivated by the fact that the fMRI volumes for different individuals can only be crudely aligned during preprocessing, making ℓ_1/ℓ_2-penalty ill-suited for their study as it may potentially select groups of voxels that are misaligned across subjects.

Although the SOS-Lasso is also a valid candidate for the multitask $\mathcal{R}_2$ penalty in (4.4), we have elected to use the ℓ_1/ℓ_2-penalty for our work. The reason for this decision is because when constructing the functional connectomes, the time series of the resting state fMRI volumes are spatially averaged over a 15 mm diameter ROI encompassing 33 voxels (see Sec. 4.2.1). Due to this heavy downsampling, we expect the potential misalignments in the fMRI volumes to only have a negligible impact on the resulting functional connectomes. In addition, the SOS-Lasso introduces another hyperparameter that requires tuning, which creates a heavy computational overhead during cross-validation.



(a) Single-task penalty: $\mathcal{R}_2(\underline{\boldsymbol{w}}) = \sum_{j=1}^{p} \left\| \boldsymbol{w}_j \right\|_1$ (b) Multitask penalty: $\mathcal{R}_2(\underline{\boldsymbol{w}}) = \sum_{j=1}^{p} \left\| \boldsymbol{w}_j \right\|_2$

Figure 4.2: Comparison between the sparsity patterns promoted by the single-task ℓ_1/ℓ_1 and the multitask ℓ_1/ℓ_2 penalty. The rows in the matrices above represent the task-specific weight vectors $\left\{\boldsymbol{w}^k\right\}_{k=1}^{K}$, and the blue entries indicate the non-zero coefficients. Note how the single-task approach yields sparsity patterns that are inconsistent across sites, which can be problematic for interpretation. In contrast, the *group variable selection* property from the multitask approach provides a sparsity pattern that is shared across all sites.

To summarize, the optimization problem we propose to solve in this work has the following formulation:

$$\min_{\underline{\boldsymbol{w}} \in \mathbb{R}^{Kp}} \sum_{k=1}^{K} \frac{1}{n_k} \mathcal{L}(\boldsymbol{Y}^k \boldsymbol{X}^k \boldsymbol{w}^k) + \gamma \sum_{k=1}^{K} \mathcal{R}_1(\boldsymbol{w}^k) + \lambda \sum_{j=1}^{p} \left\| \boldsymbol{w}_j \right\|_2 . \tag{4.7}$$

We employ the hinge-loss for the empirical loss term and use either the GraphNet, fused Lasso, or isotropic TV for the intra-task $\mathcal{R}_1$ penalty, making (4.7) a *multitask structured sparse SVM*, where classification is conducted jointly with group feature selection.

4.2.3 Optimization via Alternating Direction Method

Solving the optimization problem (4.7) is challenging since the problem size $K \cdot p$ is large and the three terms in the cost function can each be non-differentiable. To address these challenges, we introduce an extension to the alternating direction method of multipliers (ADMM) (Boyd et al., 2011; Gabay and Mercier, 1976; Glowinski and Marroco, 1975) algorithm introduced in our earlier work (Watanabe et al., 2014a).

ADMM is a flexible algorithm which iteratively solves problems that have the following separable structure:

$$\min_{\bar{\boldsymbol{x}},\bar{\boldsymbol{y}}} \bar{\boldsymbol{f}}(\bar{\boldsymbol{x}}) + \bar{\boldsymbol{g}}(\bar{\boldsymbol{y}}) \quad \text{subject to } \bar{\boldsymbol{A}}\bar{\boldsymbol{x}} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}} = \boldsymbol{0}\,. \tag{4.8}$$

Here $\bar{\boldsymbol{x}} \in \mathbb{R}^{\bar{p}}$ and $\bar{\boldsymbol{y}} \in \mathbb{R}^{\bar{q}}$ are 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. ADMM exploits the separable structure in (4.8) by applying the following updates:

$$\bar{\boldsymbol{x}}^{(t+1)} \leftarrow \arg\min_{\bar{\boldsymbol{x}}} \bar{\boldsymbol{f}}(\bar{\boldsymbol{x}}) + \frac{\rho}{2} \left\| \bar{\boldsymbol{A}}\bar{\boldsymbol{x}} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}}^{(t)} + \boldsymbol{u}^{(t)} \right\|_2^2 \tag{4.9}$$

$$\bar{\boldsymbol{y}}^{(t+1)} \leftarrow \arg\min_{\bar{\boldsymbol{y}}} \bar{\boldsymbol{g}}(\bar{\boldsymbol{y}}) + \frac{\rho}{2} \left\| \bar{\boldsymbol{A}}\bar{\boldsymbol{x}}^{(t+1)} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}} + \boldsymbol{u}^{(t)} \right\|_2^2 \tag{4.10}$$

$$\boldsymbol{u}^{(t+1)} \leftarrow \boldsymbol{u}^{(t)} + \left(\bar{\boldsymbol{A}}\bar{\boldsymbol{x}}^{(t+1)} + \bar{\boldsymbol{B}}\bar{\boldsymbol{y}}^{(t+1)} \right), \tag{4.11}$$

where t denotes the iteration count, $\boldsymbol{u} \in \mathbb{R}^c$ is the (scaled) dual variable, and $\rho > 0$ is a user defined parameter which we set to $\rho = 1$ in our implementations. The above iterations (4.9)-(4.11) is guaranteed to converge to the optimal solution as long as the constraint matrices $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{B}}$ are full column-rank; see Theorem 1 in Mota et al. (2013). Of note, while the parameter $\rho > 0$ does not affect the convergence of ADMM, it can impact its convergence speed.

4.2.3.1 Variable Splitting and Data Augmentation

Since the objective function for the multitask structured-sparse SVM (4.7) originally has an unconstrained formulation, we use variable splitting techniques (Afonso et al., 2010) to convert it into a constrained problem that is in the canonical ADMM form (4.8). Variable splitting refers to the method of introducing auxiliary constraint variables into an optimization problem, which is particularly useful in an ADMM framework since it allows us to break down an optimization problem into smaller and easier subproblems.

Before we introduce our variable splitting scheme, we note that as it stands, the ADMM algorithm for solving the objective function (4.4) with the GraphNet, fused Lasso, or isotropic TV penalty will require the inversion of the Laplacian matrix $\boldsymbol{C}^T\boldsymbol{C} \in \mathbb{R}^p$, which is prohibitively large. To address this issue, we employ the *data augmentation* + *masking strategy* that was proposed in Watanabe et al. (2014a), which induces a computationally useful structure in the Laplacian matrix. In this section, we will focus on the GraphNet and fused Lasso penalty since these can be succinctly expressed as $\mathcal{R}_1(\boldsymbol{w}^k) = \left\|\boldsymbol{C}\boldsymbol{w}^k\right\|_q^q, q \in \{1, 2\}$. However, the same strategy can be applied for the isotropic TV penalty, and the mathematical detail for this is given in 4.B.

In brief, this strategy introduces an *augmentation matrix* $\boldsymbol{A} \in \mathbb{R}^{\tilde{p}\times p}$, whose rows are either the zero vector or an element from the standard basis $\{\boldsymbol{e}_j\}_{j=1}^p$. Furthermore, this matrix has the property $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_p$, and allows us to define an *augmented weight vector* $\widetilde{\boldsymbol{w}}^k := \boldsymbol{A}\boldsymbol{w}^k$. This results in a new finite differencing matrix $\widetilde{\boldsymbol{C}} \in \mathbb{R}^{\tilde{e}\times\tilde{p}}$ for $\widetilde{\boldsymbol{w}}^k \in \mathbb{R}^{\tilde{p}}$, whose Laplacian matrix $\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}} \in \mathbb{R}^{\tilde{p}\times\tilde{p}}$ has a special structure known as *block-circulant with circulant-blocks* (BCCB), a structure that will be exploited in our ADMM algorithm. Finally, by introducing a diagonal masking matrix $\boldsymbol{B} \in \{0, 1\}^{\tilde{p}\times\tilde{p}}$, we can express the intra-structure spatial penalty in terms of $\widetilde{\boldsymbol{C}}$ and $\widetilde{\boldsymbol{w}}^k$: $\mathcal{R}_1(\boldsymbol{w}^k) = \left\|\boldsymbol{C}\boldsymbol{w}^k\right\|_q^q = \left\|\boldsymbol{B}\widetilde{\boldsymbol{C}}\boldsymbol{A}\boldsymbol{w}^k\right\|_q^q$. We refer the readers to Watanabe et al. (2014a) for additional details regarding this procedure.

In summary, using the augmentation+masking strategy above, we can rewrite the objective for the multitask structured-sparse SVM (4.7) with the GraphNet or fused Lasso as follows (see 4.B for the isotropic TV case):

$$\min_{\underline{\boldsymbol{w}}} \sum_{k=1}^K \frac{1}{n_k}\mathcal{L}(\boldsymbol{Y}^k\boldsymbol{X}^k\boldsymbol{w}^k) + \frac{\gamma}{q}\sum_{k=1}^K \left\|\boldsymbol{B}\widetilde{\boldsymbol{C}}\boldsymbol{A}\boldsymbol{w}^k\right\|_q^q + \lambda\sum_{j=1}^p \left\|\boldsymbol{w}_j\right\|_2 ,$$

which can be converted into the following constrained form:

$$\min_{\{\boldsymbol{w}^k,\boldsymbol{v}_1^k,\boldsymbol{v}_2^k,\boldsymbol{v}_3^k,\boldsymbol{v}_4^k\}_{k=1}^K} \sum_{k=1}^K \frac{1}{n_k}\mathcal{L}(\boldsymbol{v}_1^k) + \frac{\gamma}{q}\sum_{k=1}^K \left\|\boldsymbol{B}\boldsymbol{v}_3^k\right\|_q^q + \lambda\sum_{j=1}^p \left\|\boldsymbol{v}_{2,j}\right\|_2 \tag{4.12}$$

$$\text{subject to } \boldsymbol{Y}^k\boldsymbol{X}^k\boldsymbol{w}^k = \boldsymbol{v}_{\mathbf{1}}^k,\ \boldsymbol{w}^k = \boldsymbol{v}_{\mathbf{2}}^k,\ \widetilde{\boldsymbol{C}}\boldsymbol{v}_{\mathbf{4}}^k = \boldsymbol{v}_{\mathbf{3}}^k,\ \boldsymbol{A}\boldsymbol{w}^k = \boldsymbol{v}_{\mathbf{4}}^k \quad \forall k = 1, \ldots, K.$$

Here, $\{\boldsymbol{v}_{\mathbf{1}}^k, \boldsymbol{v}_{\mathbf{2}}^k, \boldsymbol{v}_{\mathbf{3}}^k, \boldsymbol{v}_{\mathbf{4}}^k\}_{k=1}^K$ are the auxiliary constraint variables introduced from variable splitting. It is straightforward to show that the above two problems are equivalent, and the correspondence with the ADMM formulation (4.8) is given by:

$$\bar{\boldsymbol{f}}(\bar{\boldsymbol{x}}) = \frac{\gamma}{q}\sum_{k=1}^K \left\|\boldsymbol{B}{\boldsymbol{v}_{\mathbf{3}}}^k\right\|_q^q, \quad \bar{\boldsymbol{g}}(\bar{\boldsymbol{y}}) = \sum_{k=1}^K \frac{1}{n_k}\mathcal{L}({\boldsymbol{v}_{\mathbf{1}}}^k) + \lambda\sum_{j=1}^p \left\|\boldsymbol{v}_{2,j}\right\|_2$$

$$\bar{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{Y}\boldsymbol{X} & \mathbf{0} \\ \boldsymbol{I}_{Kp} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{K\tilde{e}} \\ \boldsymbol{I}_K \otimes \boldsymbol{A} & \mathbf{0} \end{bmatrix}, \quad \bar{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{w}^1 \\ \vdots \\ \boldsymbol{w}^K \\ {\boldsymbol{v}_{\mathbf{3}}}^1 \\ \vdots \\ {\boldsymbol{v}_{\mathbf{3}}}^K \end{bmatrix}, \quad \bar{\boldsymbol{B}} = \begin{bmatrix} -\boldsymbol{I}_n & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\boldsymbol{I}_{Kp} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{I}_K \otimes \widetilde{\boldsymbol{C}} \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{I}_{K\tilde{p}} \end{bmatrix}, \quad \bar{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{v}_1^1 \\ \vdots \\ \boldsymbol{v}_1^K \\ \boldsymbol{v}_2^1 \\ \vdots \\ \boldsymbol{v}_2^K \\ \boldsymbol{v}_4^1 \\ \vdots \\ \boldsymbol{v}_4^K \end{bmatrix}. \tag{4.13}$$

where "⊗" represents the Kronecker product, and we define

$$\boldsymbol{Y}\boldsymbol{X} = \begin{bmatrix} \boldsymbol{Y}^1\boldsymbol{X}^1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Y}^2\boldsymbol{X}^2 & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{Y}^K\boldsymbol{X}^K \end{bmatrix} \in \mathbb{R}^{n \times Kp}, \qquad n = \sum_{k=1}^K n_k.$$

Note that the constraint matrices $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{B}}$ are both full column rank, so the convergence of the ADMM algorithm is guaranteed (see Theorem 1 in Mota et al. (2013)).

4.2.3.2 ADMM – Analytical Updates

Under the variable splitting scheme (4.12), the ADMM update for $\bar{\boldsymbol{x}}$ (4.9) decomposes into the following subproblems:

$$\operatorname*{arg\,min}_{\boldsymbol{w}^k} \left\| \boldsymbol{Y}^k \boldsymbol{X}^k \boldsymbol{w}^k - \boldsymbol{v}_{\mathbf{1}}^k + \boldsymbol{u}_{\mathbf{1}}^k \right\|^2 + \left\| \boldsymbol{w}^k - \boldsymbol{v}_{\mathbf{2}}^k + \boldsymbol{u}_{\mathbf{2}}^k \right\|^2 + \left\| \boldsymbol{A}\boldsymbol{w}^k - \boldsymbol{v}_{\mathbf{4}}^k + \boldsymbol{u}_{\mathbf{4}}^k \right\|^2 \tag{4.14}$$

$$\operatorname*{arg\,min}_{\boldsymbol{v}_{\mathbf{3}}^k} \frac{\gamma}{q} \left\| \boldsymbol{B}\boldsymbol{v}_{\mathbf{3}}^k \right\|_q^q + \frac{\rho}{2} \left\| \boldsymbol{v}_{\mathbf{3}}^k - \left(\widetilde{\boldsymbol{C}} \boldsymbol{v}_{\mathbf{4}}^k - \boldsymbol{u}_{\mathbf{3}}^k \right) \right\|^2, \qquad k = 1, \ldots, K, \tag{4.15}$$

where as update for $\bar{\boldsymbol{y}}$ (4.10) decomposes to:

$$\operatorname*{arg\,min}_{\boldsymbol{v}_{\mathbf{1}}^k} \frac{1}{n_k} \mathcal{L}\left(\boldsymbol{v}_{\mathbf{1}}^k \right) + \frac{\rho}{2} \left\| \boldsymbol{v}_{\mathbf{1}}^k - \left(\boldsymbol{Y}^k \boldsymbol{X}^k \boldsymbol{w}^k + \boldsymbol{u}_{\mathbf{1}}^k \right) \right\|_2 = \mathrm{Prox}_{\ell/(n_k \cdot \rho)} \left(\boldsymbol{Y}^k \boldsymbol{X}^k \boldsymbol{w}^k + \boldsymbol{u}_{\mathbf{1}}^k \right) \tag{4.16}$$

$$\operatorname*{arg\,min}_{\boldsymbol{v}_{\mathbf{4}}^k} \left\| \widetilde{\boldsymbol{C}} \boldsymbol{v}_{\mathbf{4}}^k - \left(\boldsymbol{v}_{\mathbf{3}}^k + \boldsymbol{u}_{\mathbf{3}}^k \right) \right\|^2 + \left\| \boldsymbol{v}_{\mathbf{4}}^k - \left(\boldsymbol{A}\boldsymbol{w}^k + \boldsymbol{u}_{\mathbf{4}}^k \right) \right\|^2, \qquad k = 1, \cdots, K \tag{4.17}$$

$$\operatorname*{arg\,min}_{\boldsymbol{v}_{2,j}} \lambda \left\| \boldsymbol{v}_{2,j} \right\|_2 + \frac{\rho}{2} \left\| \boldsymbol{v}_{2,j} - \left(\boldsymbol{w}_j + \boldsymbol{u}_{2,j} \right) \right\|^2 = \mathrm{vsoft}_{\lambda/\rho} \left(\boldsymbol{w}_j + \boldsymbol{u}_{2,j} \right), \qquad j = 1, \cdots, p \tag{4.18}$$

The close form solutions for these are summarized in Algorithm 3, which outlines the complete ADMM algorithm. We note that the update for the isotropic TV only differs in (4.15), corresponding to Line 7 of Algorithm 3; see 4.B. We now demonstrate that the above updates all admit closed form solutions that can be computed efficiently.

$\bar{\boldsymbol{x}}$-update The $\boldsymbol{w}^k$ update (4.14) corresponds to a quadratic minimization problem, and its solution can be obtained by setting the gradient of the cost function to zero, giving us the solution

$$\left((\boldsymbol{X}^k)^T \boldsymbol{X}^k + 2\boldsymbol{I}_p \right)^{-1} \left((\boldsymbol{Y}^k \boldsymbol{X}^k)^T \left(\boldsymbol{v}_{\mathbf{1}}^k - \boldsymbol{u}_{\mathbf{1}}^k \right) \left(\boldsymbol{v}_{\mathbf{2}}^k - \boldsymbol{u}_{\mathbf{2}}^k \right) + \boldsymbol{A}^T \left(\boldsymbol{v}_{\mathbf{4}}^k - \boldsymbol{u}_{\mathbf{4}}^k \right) \right).$$

This update can be converted into a much simpler $(n_k \times n_k)$ inversion problem using the *matrix inversion Lemma*, where n_k indicates the number of subjects from the k-th site.

The solution to the $\boldsymbol{v}_3^k$ update (4.15) depends on the choice of $q \in \{1, 2\}$. When $q = 2$, the solution for (4.15) represents the update for GraphNet, which is given by:

$$\boldsymbol{v}_3^k \leftarrow \rho(\gamma \boldsymbol{B} + \rho \boldsymbol{I}_{\tilde{e}})^{-1} \widetilde{\boldsymbol{C}}(\boldsymbol{v}_4^k - \boldsymbol{u}_3^k).$$

This is easy to compute since the matrix $(\gamma \boldsymbol{B} + \rho \boldsymbol{I}_{\tilde{e}})$ is diagonal. On the other hand, $q = 1$ corresponds to fused Lasso, giving rise to the following elementwise update:

$$\left[\boldsymbol{v}_3^k\right]_s \leftarrow \begin{cases} \text{Soft}_{\gamma/\rho}\left(\left[\widetilde{\boldsymbol{C}}(\boldsymbol{v}_4^k - \boldsymbol{u}_3^k)\right]_s\right) & \text{if } \boldsymbol{B}_{s,s} = 1 \\ \left[\widetilde{\boldsymbol{C}}(\boldsymbol{v}_4^k - \boldsymbol{u}_3^k)\right]_s & \text{if } \boldsymbol{B}_{s,s} = 0, \end{cases} \tag{4.19}$$

where $\text{Soft}_\tau(t) := \max(1 - \frac{\tau}{|t|}, 0) \cdot t$ denotes the *scalar* soft-threshold operator and $[\cdot]_s$ indexes the s-th element of a vector.

$\bar{\boldsymbol{y}}$-update The $\text{Prox}_{\tau\ell}(\cdot)$ in the $\boldsymbol{v}_1^k$ update (4.16) represents the proximal operator (Combettes and Pesquet, 2011) of the hinge-loss $\ell(t) = (1 - t)_+$ given by:

$$\text{Prox}_{\tau\ell}(t) := \begin{cases} t & \text{if } t > 1 \\ 1 & \text{if } 1 - \tau \leqslant t \leqslant 1 \\ t + \tau & \text{if } t < 1 - \tau, \end{cases} \tag{4.20}$$

The closed form solution for the $\boldsymbol{v}_4^k$-update (4.17) is:

$$\boldsymbol{v}_4^k \leftarrow \left(\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}} + \boldsymbol{I}_{\tilde{p}}\right)^{-1}\left(\widetilde{\boldsymbol{C}}^T[\boldsymbol{v}_3^k + \boldsymbol{u}_3^k] + \boldsymbol{A}\boldsymbol{w}^k + \boldsymbol{u}_4^k\right).$$

Since the augmented Laplacian matrix $\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}}$ has a BCCB structure, it can be diagonalized as $\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}} = \boldsymbol{U}^H\boldsymbol{\Lambda}\boldsymbol{U}$ (Davis, 1979; Gray, 2005), where $\boldsymbol{U}$ is the 6-D discrete Fourier

Algorithm 3 ADMM for Multitask Structured Sparse SVM

```
Initialize primal variables $\{\boldsymbol{w}^k, \boldsymbol{v}_{\boldsymbol{1}}^k, \boldsymbol{v}_{\boldsymbol{2}}^k, \boldsymbol{v}_{\boldsymbol{3}}^k, \boldsymbol{v}_{\boldsymbol{4}}^k\}_{k=1}^K$
Initialize dual variables $\{\boldsymbol{u}_{\boldsymbol{1}}^k, \boldsymbol{u}_{\boldsymbol{2}}^k, \boldsymbol{u}_{\boldsymbol{3}}^k, \boldsymbol{u}_{\boldsymbol{4}}^k\}_{k=1}^K$
Assign hyperparameters $\lambda, \gamma \geqslant 0$
repeat
    for $k = 1, \dots, K$
        $\boldsymbol{w}^k \leftarrow \left((\boldsymbol{X}^k)^T \boldsymbol{X}^k + 2\boldsymbol{I}_p\right)^{-1} \left((\boldsymbol{Y}^k \boldsymbol{X}^k)^T (\boldsymbol{v}_{\boldsymbol{1}}^k - \boldsymbol{u}_{\boldsymbol{1}}^k) (\boldsymbol{v}_{\boldsymbol{2}}^k - \boldsymbol{u}_{\boldsymbol{2}}^k) + \boldsymbol{A}^T (\boldsymbol{v}_{\boldsymbol{4}}^k - \boldsymbol{u}_{\boldsymbol{4}}^k)\right)$
                                                  ▷ solve using matrix inversion Lemma
        $\boldsymbol{v}_{\boldsymbol{3}}^k \leftarrow \begin{cases} \text{apply Equation (4.19)} & \text{if using fused Lasso} \\ \rho(\gamma \boldsymbol{B} + \rho \boldsymbol{I})^{-1} \widetilde{\boldsymbol{C}}(\boldsymbol{v}_{\boldsymbol{4}}^k - \boldsymbol{u}_{\boldsymbol{3}}^k) & \text{if using GraphNet} \\ \text{See 4.B} & \text{if using isotropic TV} \end{cases}$
        $\boldsymbol{v}_{\boldsymbol{1}}^k \leftarrow \operatorname{Prox}_{\ell/(\rho n_k)} \left(\boldsymbol{Y}^k \boldsymbol{X}^k \boldsymbol{w}^k + \boldsymbol{u}_{\boldsymbol{1}}^k\right)$      ▷ Apply (4.20) elementwise
        $\boldsymbol{v}_{\boldsymbol{4}}^k \leftarrow \left(\widetilde{\boldsymbol{C}}^T \widetilde{\boldsymbol{C}} + \boldsymbol{I}_{\tilde{p}}\right)^{-1} \left(\widetilde{\boldsymbol{C}}^T [\boldsymbol{v}_{\boldsymbol{3}}^k + \boldsymbol{u}_{\boldsymbol{3}}^k] + \boldsymbol{A}\boldsymbol{w}^k + \boldsymbol{u}_{\boldsymbol{4}}^k\right)$      ▷ solve using FFT
    end for

    for $j = 1, \dots, p$
        $\boldsymbol{v}_{\boldsymbol{2}j} \leftarrow \operatorname{vsoft}_{\lambda/\rho} (\boldsymbol{w}_j + \boldsymbol{u}_{2,j})$      ▷ $\operatorname{vsoft}_\tau(\boldsymbol{t}) := \max(1 - \frac{\tau}{\|\boldsymbol{t}\|_2}, 0)\, \boldsymbol{t}, \;\; \boldsymbol{t} \in \mathbb{R}^K$
    end for

    for $k = 1, \dots, K$                                        ▷ dual variable update
        $\boldsymbol{u}_{\boldsymbol{1}}^k \leftarrow \boldsymbol{u}_{\boldsymbol{1}}^k + \boldsymbol{Y}^k \boldsymbol{X}^k \boldsymbol{w}^k - \boldsymbol{v}_{\boldsymbol{1}}^k$
        $\boldsymbol{u}_{\boldsymbol{2}}^k \leftarrow \boldsymbol{u}_{\boldsymbol{2}}^k + \boldsymbol{w}^k - \boldsymbol{v}_{\boldsymbol{2}}^k$
        $\boldsymbol{u}_{\boldsymbol{3}}^k \leftarrow \boldsymbol{u}_{\boldsymbol{3}}^k + \boldsymbol{v}_{\boldsymbol{3}}^k - \widetilde{\boldsymbol{C}} \boldsymbol{v}_{\boldsymbol{4}}^k$
        $\boldsymbol{u}_{\boldsymbol{4}}^k \leftarrow \boldsymbol{u}_{\boldsymbol{4}}^k + \boldsymbol{A}\boldsymbol{w}^k - \boldsymbol{v}_{\boldsymbol{4}}^k$
    end for

until stopping criterion is met
```

transform (DFT) matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix containing the 6-D DFT coefficients of the first column of $\widetilde{\boldsymbol{C}}^T\widetilde{\boldsymbol{C}}$. Thus, the $\boldsymbol{v}_4^k$ update can be implemented efficiently using fast Fourier transform (FFT).

Finally, the solution for the $\boldsymbol{v}_{2,j}$ update (4.18) is given in terms of the *vector* soft-threshold operator: $\mathrm{vsoft}_\tau(\boldsymbol{t}) := \max\left(1 - \frac{\tau}{\|\boldsymbol{t}\|_2}, 0\right) \cdot \boldsymbol{t}$, where $\boldsymbol{t} \in \mathbb{R}^K$. We conclude this section by noting that if the ℓ_1/ℓ_2-penalty in (4.7) is replaced with the ℓ_1/ℓ_1-penalty (4.6), the $\boldsymbol{v}_{2,j}$ update will be replaced by the *scalar* soft-threshold operator, thus recovering the ADMM algorithm for the single-task version of the structured sparse SVM proposed in Watanabe et al. (2014a).

4.3 Results

4.3.1 Experimental Setup

To assess the validity of the proposed method, we compared the performance of various SVM-based classifiers using resting-state functional connectomes derived from the ADHD-200 dataset (see Sec. 4.2.1 for details on preprocessing). For the intra-task penalty $\mathcal{R}_1$, we compared four different regularization schemes: Elastic-net (Chen et al., 2012a; Zou and Hastie, 2005) with $\mathcal{R}_1(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2$, GraphNet, fused Lasso, and istropic TV. For the inter-task penalty $\mathcal{R}_2$, we compared three different approaches:

1. **Pooled** ℓ_1: a single classifier is trained on the entire ADHD-200 dataset (thus we have $\mathcal{R}_2(\underline{\boldsymbol{w}}) = \|\underline{\boldsymbol{w}}\|_1$ with $\underline{\boldsymbol{w}} \in \mathbb{R}^p$ as $K = 1$).

2. **Single-task** ℓ_1/ℓ_1: equivalent to training separately across sites due to the separability of the penalty ($\mathcal{R}_2(\underline{\boldsymbol{w}}) = \sum_{j=1}^p \|\boldsymbol{w}_j\|_1 = \sum_{k=1}^K \|\boldsymbol{w}^k\|_1$).

3. **Multitask** ℓ_1/ℓ_2: *jointly* train the classifiers by solving (4.4).

The regularization parameters $\{\lambda, \gamma\}$ are tuned by conducting a 5-fold cross-validation on the training set over the following two-dimensional grid: $\lambda, \gamma \in \{2^{-13}, 2^{-12}, \ldots, 2^{-3}\}$. The final weight vector estimate is obtained by re-training the classifiers on the entire training set using the $\{\lambda, \gamma\}$ values that maximized the cross-validation classification accuracy; for validation, we predicted the labels of the test set subjects using this weight vector. All methods were solved using ADMM with the algorithm terminated when the condition $\left\|\underline{\boldsymbol{w}}^{\text{new}} - \underline{\boldsymbol{w}}^{\text{old}}\right\|_2 \leqslant 5{\cdot}10^{-3} \times \left\|\underline{\boldsymbol{w}}^{\text{old}}\right\|_2$ was met or the iteration count reached 400.

To evaluate the quality of the classifiers, we analyzed the following set of performance measures for both the 5-fold cross-validation and the validation test set results:

- Classification accuracy (ACC)
- Area under the ROC curve (AUC)
- Balanced score rate (BSR) $= \dfrac{(\text{sensitivity} + \text{specificity})}{2}$
- P-value (PVAL) computed from an one-sided binomial test.
- Sparsity level (SP%) $= 100 \cdot \dfrac{|\#\text{ non-zero features}|}{pK}$
- Stability score (Stab.) $= \dfrac{1}{M(M-1)} \sum_{i \neq j} O_{ij}$ (see (4.21) for precise definition).

The *AUC* and *BSR* are analyzed since classification accuracy by itself can be misleading when the dataset labels are imbalanced (ACC, AUC, and BSR are averaged across the tasks); the ROC curves are constructed by varying the threshold of the classifiers. Classifier performance on the test set was compared to random guessing via a binomial test based on a binomial distribution $B(p, n)$ with $p = 0.5$ and $n = 109$ samples, with *PVAL* evaluated via an one-sided binomial test (Heinzle et al., 2012; Sripada et al., 2013b); the alternative approach of permutation test was not pursued due to its severe computational cost. *Sparsity level* is simply the fraction of features selected in the final model. Finally, *stability score* is a measure introduced in (Rasmussen et al., 2012) which quantifies the stability of the features selected across the cross-validation folds (Baldassarre et al., 2012; Rondina et al., 2014). More precisely, letting S_i and S_j denote the support of the weight vector estimated in the i-th and j-th split of an M-fold cross-validation procedure, we define:

$$O_{ij} := \frac{|S_i \cap S_j| - E_i}{|S_i|}, \qquad E_i := \frac{|S_i|^2}{pK}, \qquad i, j \in 1, \cdots, M. \tag{4.21}$$

Here O_{ij} measures the degree overlap between S_i and S_j, and E_i is a heuristic correction factor introduced in (Rasmussen et al., 2012), and the final *stability score* is obtained by averaging O_{ij} across all cross-validation folds.

4.3.2 Results and Discussion

Table 4.3 presents the classification results from the 5-fold cross-validation and validation on the test-set, and Fig. 4.3 displays the corresponding ROC curves. In addition, Fig. 4.4 and Fig. 4.5 present the classification accuracy and the mean sparsity level obtained at different combinations of $\{\lambda, \gamma\}$ during cross-validation. These results demonstrate that training a single classifier via the "pooling" approach yields the worst performance in terms of accuracy, AUC, and BSR, suggesting that blindly aggregating the datasets across different sites can be problematic for accurate disease classification. Comparison between the single-task and the multitask approaches shows that the ℓ_1/ℓ_2-penalized approach yields superior performance in terms of AUC, although no striking difference can be observed in terms of accuracy and BSR.

In addition to the performance gain, the set of weight vector estimates $\{\hat{\boldsymbol{w}}^k\}_{k=1}^K \in \mathbb{R}^p$ from the multitask approach all share a common support of length p due to the *group variable selection* property of the ℓ_1/ℓ_2-penalty (Chen et al., 2012a; Obozinski et al., 2010). This is invaluable for interpretation, as the selected features can be viewed as edges that are informative across all sites. For visualization, we grouped the indices of this support according to the network parcellation scheme proposed by Yeo et al. (2011), and augmented this parcellation with subcortical regions and cerebellum derived from the parcellation of Tzourio-Mazoyer et al. (2002) (see Table 4.4); this support vector is then

Table 4.3: The classification results from the 5-fold cross-validation and the validation test-set.

	CV (628 subjects)				**Test-set (106 subjects)**				
	ACC	**AUC**	**BSR**	**Stab.**	**ACC**	**AUC**	**BSR**	**PVAL**	**SP%**
Elastic-net (ℓ_1)	.689	.687	.630	.277	.557	.617	.476	.143	2.54%
GraphNet (ℓ_1)	.704	.708	.631	.253	.594	.608	.494	.032	28.88%
Fused Lasso (ℓ_1)	.688	.720	.586	.059	.632	.592	.530	.004	64.85%
TV (ℓ_1)	.701	.715	.620	.005	.623	.608	.521	.007	90.32%
Elastic-net (ℓ_1/ℓ_1)	.709	.752	.649	.276	.623	.609	.530	.007	0.28%
GraphNet (ℓ_1/ℓ_1)	.713	.750	.652	.165	.642	.613	.573	.002	67.14%
Fused Lasso (ℓ_1/ℓ_1)	.715	.750	.659	.329	.632	.634	.547	.004	1.30%
TV (ℓ_1/ℓ_1)	.718	.753	.661	.345	.642	.654	.550	.002	1.61%
Elastic-net (ℓ_1/ℓ_2)	.720	.754	.657	.217	.651	.645	.556	.001	0.25%
GraphNet (ℓ_1/ℓ_2)	.720	.766	.657	.320	.642	.668	.546	.002	1.03%
Fused Lasso (ℓ_1/ℓ_2)	.718	.766	.653	.315	.642	.673	.546	.002	0.79%
TV (ℓ_1/ℓ_2)	.720	.766	.658	.316	.642	.672	.546	.002	0.80%

reshaped them into 347×347 symmetric matrix with zeroes on the diagonal. The resulting support matrices for the Elastic-net+ℓ_1/ℓ_2 and the fused Lasso+ℓ_1/ℓ_2-penalized SVM are presented in Fig. 4.6 (results for GraphNet+ℓ_1/ℓ_2 and isotropic TV+ℓ_1/ℓ_2 were very similar to fused Lasso+ℓ_1/ℓ_2). An interesting observation here is that the support structure from the fused Lasso and ℓ_1/ℓ_2-penalized SVM shows concentrated connectivity patterns in the intra-frontoparietal (6-6) and the intra-default network (7-7) regions; Fig. 4.6 provides a brain space representation of these connections (figures generated with the BrainNet Viewer, `http://www.nitrc.org/projects/bnv/`). These network regions are frequently reported to exhibit disrupted connectivity patterns in resting state studies of ADHD (Castellanos and Proal, 2012; Sripada et al., 2014), although the accuracies obtained from our classifiers are not at the level where the selected features can be interpreted as reliable ADHD biosignatures.

We note that most of the accuracies reported on the validation test-set in

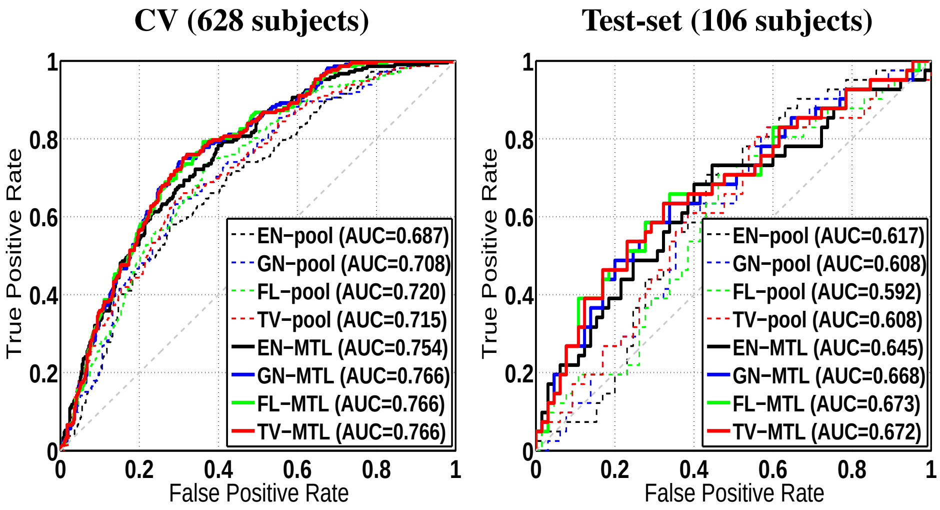


Figure 4.3: The ROC curves obtained by varying the threshold of the classifiers in Table 4.3 classifiers' ROC. The ROC curves for the single-task ℓ_1/ℓ_1-case are omitted to improve curve visibility. (EN = Elastic-net, GN = GraphNet, FL = fused Lasso, TV = isotropic total variation).

Table 4.3 exceeded the highest result from the actual ADHD-200 competition,which was 61.54% (The ADHD-200 Consortium, 2012). However, there are two major caveats: (1) the results in this work cannot be directly compared with the official competition results due to the subject screening procedure we applied on the test set (the criteria such as the FD-based one is important for avoiding confounds from excessive head motion), and (2) the participants in the actual competition were required to predict the labels of 26 subjects from the Brown site, despite the fact that no training data were provided from this site, thereby making it harder to predict the labels for these subjects. The second caveat also implies that most MTL methods, including the ℓ_1/ℓ_2-penalty employed in this work, cannot be applied since there are no means to train a weight vector for a task whose data are not provided. An alternative approach such as *transfer learning* (Pan and Yang, 2010) may be considered for future work. Finally, although the ℓ_1/ℓ_2-penalty facilitates interpretation by selecting the same set of features across sites, it does not ensure the *sign* of the selected features to be consistent, preventing us from interpreting the direction of the selected edges. Future work should extend our methodology so that the sign of the selected edges are guaranteed to be consistent across sites. One possible approach for this is to introduce vectors $\boldsymbol{w}_j^+, \boldsymbol{w}_j^- \in \mathbb{R}^K$, and make the substitution

$$\boldsymbol{w}_j = \boldsymbol{w}_j^+ - \boldsymbol{w}_j^-, \qquad \boldsymbol{w}_j^+ \geqslant 0,\ \boldsymbol{w}_j^- \geqslant 0.$$

Then we can adopt the following inter-task penalty:

$$\begin{aligned} \mathcal{R}_2(\underline{\boldsymbol{w}}) &= \sum_{j=1}^{p} \left(\left\| \boldsymbol{w}_j^+ \right\|_2 + \left\| \boldsymbol{w}_j^- \right\|_2 \right) + \sum_{j=1}^{p} \left(\left\| \boldsymbol{w}_j^+ \right\|_1 + \left\| \boldsymbol{w}_j^- \right\|_1 \right) \\ \text{s.t. } & \boldsymbol{w}_j^+, \boldsymbol{w}_j^- \geqslant \mathbf{0}. \end{aligned} \tag{4.22}$$

The first summation term in (4.22) promotes group sparsity through the ℓ_1/ℓ_2-penalty, and the second summation term promotes the sign of the selected features to be consistent across sites by discouraging $\boldsymbol{w}_j^+$ and $\boldsymbol{w}_j^-$ from both being positive at the same time through the ℓ_1-penalty.

4.4 Conclusion

We presented a multitask structured sparse SVM, a multitask extension to the connectome-based disease classification method introduced our earlier work in Chapter 3, where the imaging sites are treated as *tasks*. Experimental results on the multisite ADHD-200 dataset suggest that the multitask approach using the ℓ_1/ℓ_2-penalty can provide improvement in classification performance over the naive *pooling approach*, where a single classifier is trained on the entire multisite dataset, an approach predominantly adopted in the original ADHD-200 competition. In addition, the mulitask ℓ_1/ℓ_2-penalty achieved higher AUC scores than the single-task ℓ_1/ℓ_1-penalty, and the *group variable selection* property of the multitask approach gives a more interpretable model by selecting the same set of features across sites, which can be visualized compactly in brain space.

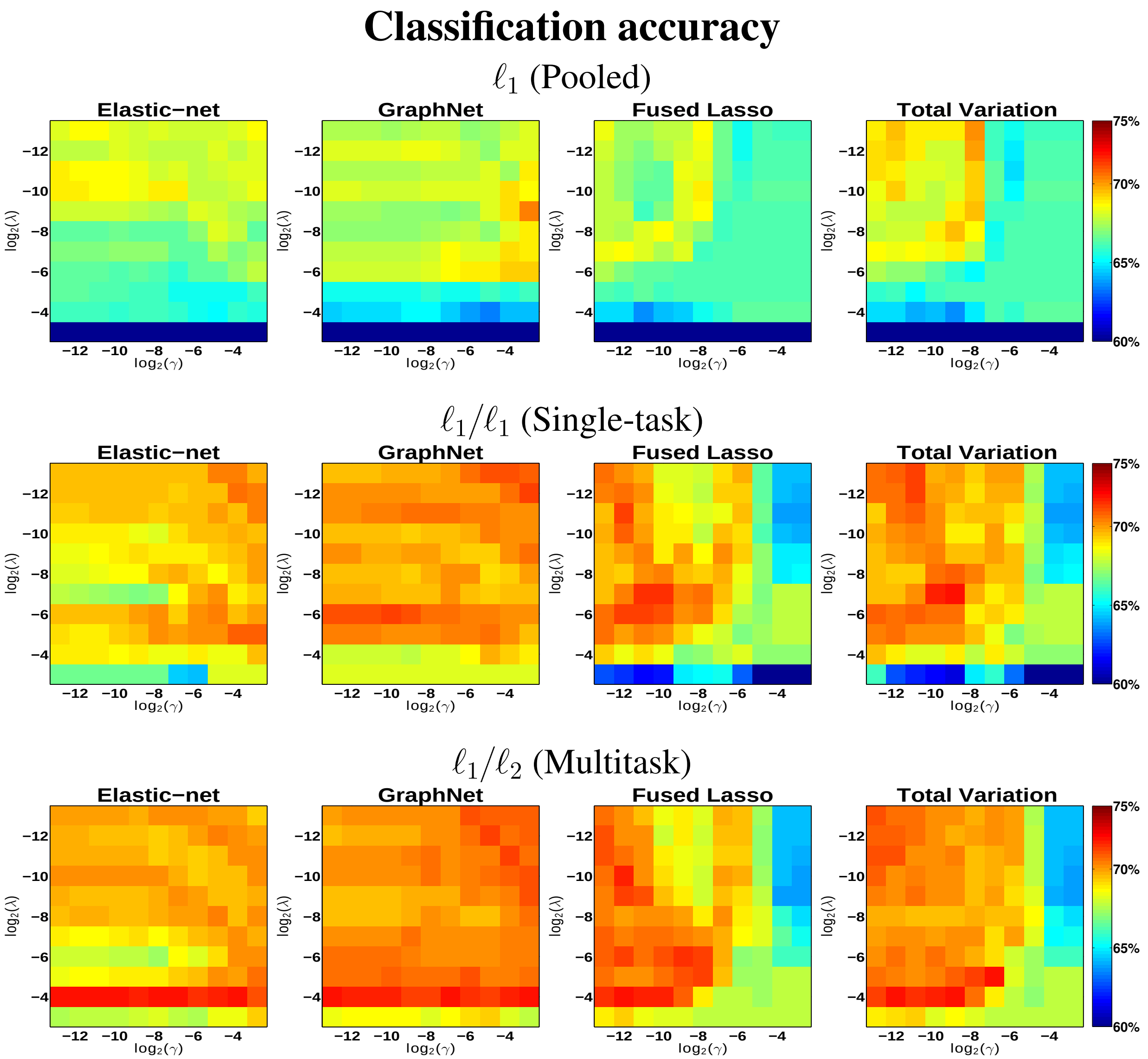


Figure 4.4: Classification accuracy evaluated from 5-fold cross-validation (best viewed in color). The (x, y)-axis corresponds to the two regularization parameters λ and γ.

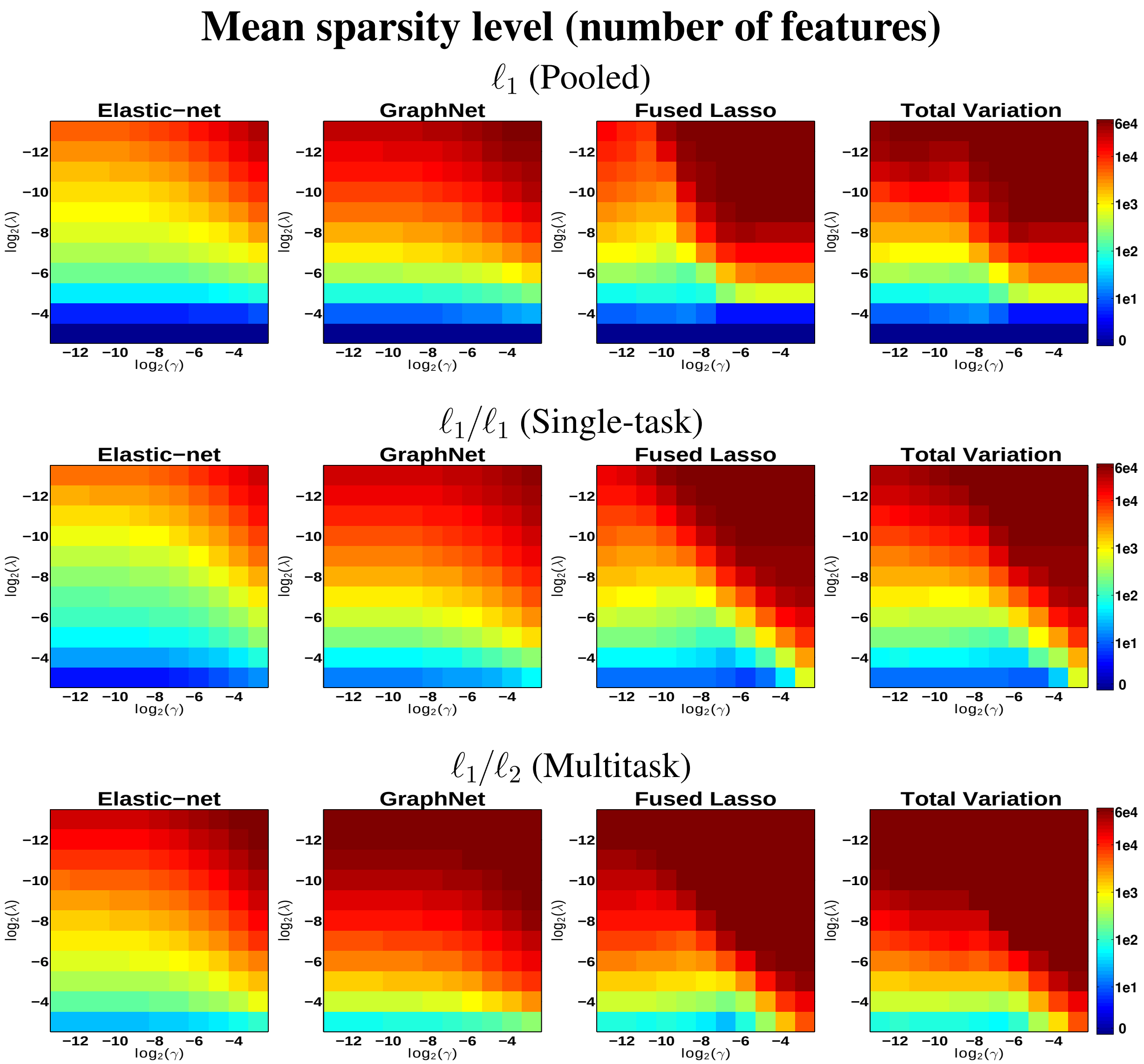
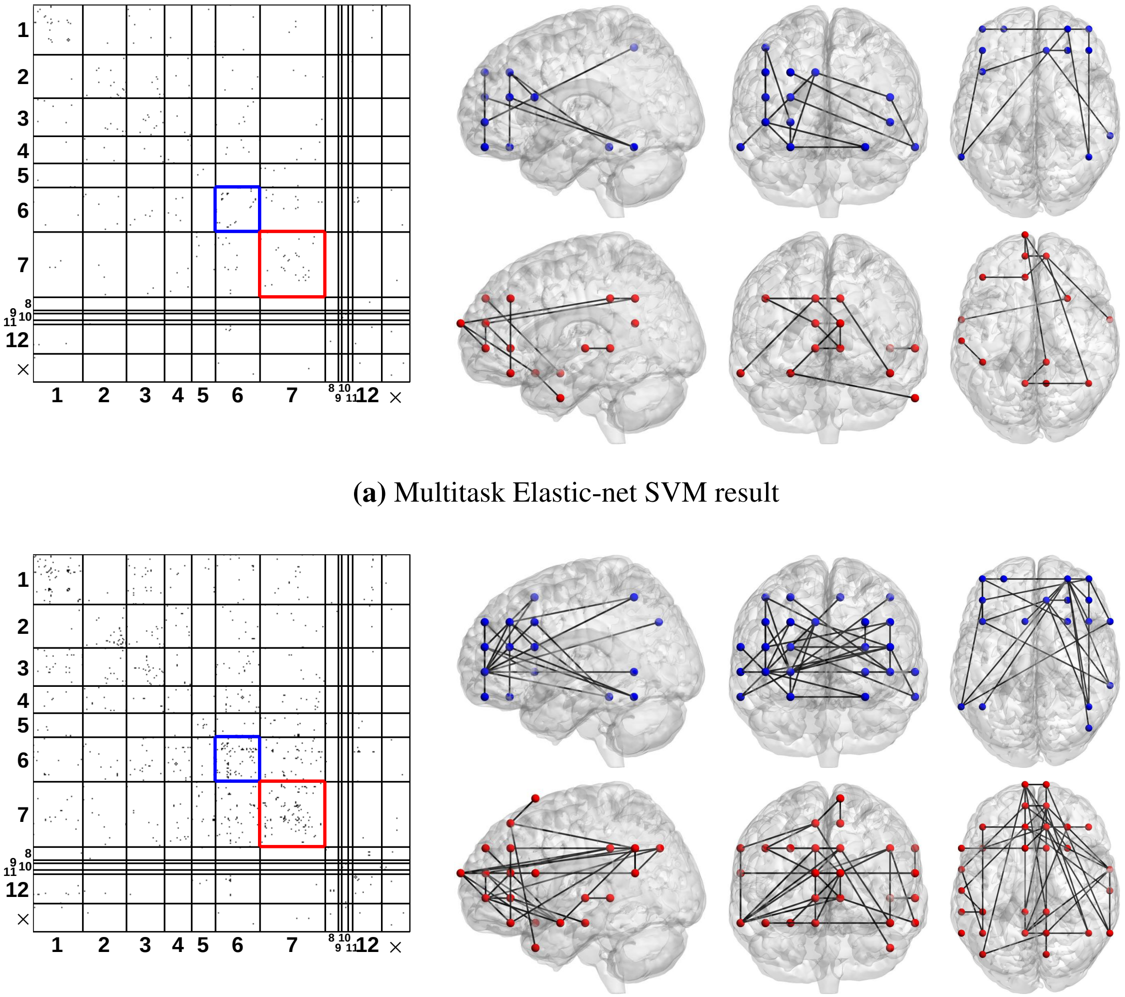


Figure 4.5: Average number of features selected across the cross-validation folds (best viewed in color). The (x, y)-axis corresponds to the two regularization parameters λ and γ.

Table 4.4: Network parcellation scheme of the brain proposed by Yeo et al. (2011).

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



(a) Multitask Elastic-net SVM result

(b) Multitask Fused Lasso SVM result

Figure 4.6: Weight vectors estimated from the Elastic-net+ℓ_1/ℓ_2 and fused Lasso+ℓ_1/ℓ_2-penalized SVM. **Left:** support matrices of the selected features (rows/cols grouped by network membership). **Right:** brain space representation of the selected edges in the intra-frontoparietal (6-6: blue) and the intra-default network (7-7: red).

4.A The expression for the isotropic total variation penalty

Let $\boldsymbol{D}_j \in \mathbb{R}^{6\times p}$ denote the 6-D *discrete gradient operator* of $\boldsymbol{w}$ at coordinate $j \in \{1,\ldots,p\}$. That is, letting (x_j, y_j, z_j) and (x'_j, y'_j, z'_j) denote the pair of 3-D points in the brain that defines a connection w_j, we have:

$$\boldsymbol{D}_j\boldsymbol{w} = \begin{bmatrix} (\nabla_x \boldsymbol{w})_j \\ (\nabla_y \boldsymbol{w})_j \\ (\nabla_z \boldsymbol{w})_j \\ (\nabla_{x'} \boldsymbol{w})_j \\ (\nabla_{y'} \boldsymbol{w})_j \\ (\nabla_{z'} \boldsymbol{w})_j \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}(x_j,y_j,z_j,x'_j,y'_j,z'_j) - \boldsymbol{w}(x_j-1,y_j,z_j,x'_j,y'_j,z'_j) \\ \boldsymbol{w}(x_j,y_j,z_j,x'_j,y'_j,z'_j) - \boldsymbol{w}(x_j,y_j-1,z_j,x'_j,y'_j,z'_j) \\ \boldsymbol{w}(x_j,y_j,z_j,x'_j,y'_j,z'_j) - \boldsymbol{w}(x_j,y_j,z_j-1,x'_j,y'_j,z'_j) \\ \boldsymbol{w}(x_j,y_j,z_j,x'_j,y'_j,z'_j) - \boldsymbol{w}(x_j,y_j,z_j,x'_j-1,y'_j,z'_j) \\ \boldsymbol{w}(x_j,y_j,z_j,x'_j,y'_j,z'_j) - \boldsymbol{w}(x_j,y_j,z_j,x'_j,y'_j-1,z'_j) \\ \boldsymbol{w}(x_j,y_j,z_j,x'_j,y'_j,z'_j) - \boldsymbol{w}(x_j,y_j,z_j,x'_j,y'_j,z'_j-1). \end{bmatrix} \in \mathbb{R}^6.$$

Then the 6-D isotropic TV penalty can be expressed as

$$\mathcal{R}_1(\boldsymbol{w}) = \sum_{j=1}^{p} \|\boldsymbol{D}_j\boldsymbol{w}\|_2\,, \tag{4.23}$$

which is a rotationally invariant counterpart of the fused Lasso penalty. Note that if the ℓ_2-norm in (4.23) is replaced with the ℓ_1-norm, we recover fused Lasso, also known as the *anisotropic* TV penalty. We further note that $\sum_{j=1}^{p} \boldsymbol{D}_j^T\boldsymbol{D}_j = \boldsymbol{C}^T\boldsymbol{C}$.

Thus, the multitask structured-sparse SVM formulation for the isotropic TV penalty can be written as:

$$\min_{\underline{\boldsymbol{w}} \in \mathbb{R}^{Kp}} \sum_{k=1}^{K} \frac{1}{n_k} \mathcal{L}(\boldsymbol{Y}^k\boldsymbol{X}^k\boldsymbol{w}^k) + \gamma \sum_{k=1}^{K}\sum_{j=1}^{p} \left\|\boldsymbol{D}_j\boldsymbol{w}^k\right\|_2 + \lambda \sum_{j=1}^{p} \|\boldsymbol{w}_j\|_2\,. \tag{4.24}$$

4.B Details on the ADMM update for the Isotropic Total Variation Penalty

Let $\widetilde{\boldsymbol{D}}_j \in \mathbb{R}^{6\times\tilde{p}},\ j = 1,\dots,\tilde{p}$, denote the 6-D discrete gradient operator corresponding to the augmented weight vector $\widetilde{\boldsymbol{w}} = \boldsymbol{A}\boldsymbol{w} \in \mathbb{R}^{\tilde{p}}$. Furthermore, let $\boldsymbol{B}_j \in \{0,1\}^{6\times 6},\ j = 1,\dots,\tilde{p}$ denote a collection diagonal masking matrix that ensures the isotropic TV remains unaffected by the augmentation scheme:

$$\sum_{j=1}^{p} \|\boldsymbol{D}_j\boldsymbol{w}\|_2 = \sum_{j=1}^{\tilde{p}} \left\|\boldsymbol{B}_j\widetilde{\boldsymbol{D}}_j\boldsymbol{A}\boldsymbol{w}\right\|_2 .$$

Then we can rewrite (4.24) as:

$$\min_{\underline{\boldsymbol{w}}\in\mathbb{R}^{Kp}} \sum_{k=1}^{K} \frac{1}{n_k}\mathcal{L}(\boldsymbol{Y}^k\boldsymbol{X}^k\boldsymbol{w}^k) + \gamma\sum_{k=1}^{K}\sum_{j=1}^{\tilde{p}} \left\|\boldsymbol{B}_j\widetilde{\boldsymbol{D}}_j\boldsymbol{A}\boldsymbol{w}^k\right\|_2 + \lambda\sum_{j=1}^{p}\|\boldsymbol{w}_j\|_2 ,$$

which can be converted into the following equivalent formulation via variable splitting:

$$\min_{\{\boldsymbol{w}^k,\boldsymbol{v}_{\mathbf{1}}^k,\boldsymbol{v}_{\mathbf{2}}^k,\boldsymbol{v}_{\mathbf{3}}^k,\boldsymbol{v}_{\mathbf{4}}^k\}_{k=1}^{K}} \sum_{k=1}^{K}\frac{1}{n_k}\mathcal{L}(\boldsymbol{v}_{\mathbf{1}}^k) + \gamma\sum_{k=1}^{K}\sum_{j=1}^{\tilde{p}}\|\boldsymbol{B}\boldsymbol{v}_{3,j}^k\|_2 + \lambda\sum_{j=1}^{p}\|\boldsymbol{v}_{2,j}\|_2 \tag{4.25}$$

$$\text{subject to } \boldsymbol{Y}^k\boldsymbol{X}^k\boldsymbol{w}^k = \boldsymbol{v}_{\mathbf{1}}^k,\ \boldsymbol{w}^k = \boldsymbol{v}_{\mathbf{2}}^k,\ \underbrace{\{\widetilde{\boldsymbol{D}}_j\boldsymbol{v}_{4,j}^k = \boldsymbol{v}_{3,j}^k\}_{j=1}^{\tilde{p}}}_{\widetilde{\boldsymbol{C}}\boldsymbol{v}_{\mathbf{4}}^k = \boldsymbol{v}_{\mathbf{3}}^k},\quad \boldsymbol{A}\boldsymbol{w}^k = \boldsymbol{v}_{\mathbf{4}}^k \quad \forall k = 1,\dots,K.$$

Applying the standard ADMM iterations (4.9)-(4.11) results in a nearly identical algorithm with the GraphNet and fused Lasso case, except the ADMM update (4.15) gets replaced by the following:

$$\operatorname*{arg\,min}_{\boldsymbol{v}_{3,j}^k} \gamma\|\boldsymbol{B}\boldsymbol{v}_{3,j}^k\|_2 + \frac{\rho}{2}\left\|\boldsymbol{v}_{3,j}^k - \left(\widetilde{\boldsymbol{D}}_j\widetilde{\boldsymbol{w}}_j - \boldsymbol{u}_{3,j}\right)\right\|_2^2 \quad j = 1,\dots,\tilde{p},\ k = 1,\dots,K.$$

Since $\boldsymbol{B}_j$ is a diagonal masking matrix, this further decomposes into the following subproblems with known closed form solutions:

$$\left[\boldsymbol{v}_{3,j}^k\right]_{\mathcal{I}_j} \leftarrow \underset{\left[\boldsymbol{v}_{3,j}^k\right]_{\mathcal{I}_j}}{\arg\min}\, \gamma \left\| \left[\boldsymbol{v}_{3,j}^k\right]_{\mathcal{I}_j} \right\|_2 + \frac{\rho}{2} \left\| \left[\boldsymbol{v}_{3,j}^k - \left(\widetilde{\boldsymbol{D}}_j \widetilde{\boldsymbol{w}}_j - \boldsymbol{u}_{3,j}\right)\right]_{\mathcal{I}_j} \right\|_2^2 = \mathrm{vsoft}_{\gamma/\rho}\left(\left[\widetilde{\boldsymbol{D}}_j \widetilde{\boldsymbol{w}}_j - \boldsymbol{u}_{3,j}\right]_{\mathcal{I}_j}\right)$$

$$\left[\boldsymbol{v}_{3,j}^k\right]_{\mathcal{I}_j^c} \leftarrow \underset{\left[\boldsymbol{v}_{3,j}^k\right]_{\mathcal{I}_j^c}}{\arg\min} \left\| \left[\boldsymbol{v}_{3,j}^k - \left(\widetilde{\boldsymbol{D}}_j \widetilde{\boldsymbol{w}}_j - \boldsymbol{u}_{3,j}\right)\right]_{\mathcal{I}_j^c} \right\|_2^2 = \left[\widetilde{\boldsymbol{D}}_j \widetilde{\boldsymbol{w}}_j - \boldsymbol{u}_{3,j}\right]_{\mathcal{I}_j^c}.$$

Here $\mathcal{I}_j \subseteq \{1, \cdots, 6\}$ is an index set that indicates the location of the nonzero diagonal entry in $\boldsymbol{B}_j$ with $\mathcal{I}_j^c$ representing its complement. Finally, $\left[\,\cdot\,\right]_{\mathcal{I}_j}$ and $\left[\,\cdot\,\right]_{\mathcal{I}_j^c}$ denote the subset of a vector indexed by $\mathcal{I}_j$ and $\mathcal{I}_j^c$, respectively. For example, if $\mathbf{z} \in \mathbb{R}^6$ and $\boldsymbol{B}_j = \mathrm{diag}(1, 0, 1, 1, 1, 0)$, then we have:

$$\mathcal{I}_j = \{1, 3, 4, 5\}, \quad \mathcal{I}_j^c = \{2, 6\}, \quad \mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \end{bmatrix}, \qquad \left[\mathbf{z}\right]_{\mathcal{I}_j} = \begin{bmatrix} z_1 \\ z_3 \\ z_4 \\ z_5 \end{bmatrix}, \qquad \left[\mathbf{z}\right]_{\mathcal{I}_j^c} = \begin{bmatrix} z_2 \\ z_6 \end{bmatrix}.$$

CHAPTER 5

Conclusion and Future Work

5.1 Summary of Contributions

The central theme of this dissertation was to devise a computationally tractable machine learning method that allows us to extract scientifically meaningful information from massive and highly complex biomedical data, despite being limited in sample size. To this end, we presented innovations in two areas of biomedical science that are of substantial clinical interest: (1) biomedical image registration and (2) psychiatric disease prediction based on functional connectomes.

Chapter 2 highlights our first major contribution, where we tackled the challenging problem of quantitatively evaluating the accuracy of an image registration result. In particular, we introduced a novel data-driven method that allows one to visualize and quantify registration uncertainty using spatially adaptive confidence regions. A vital component to our proposed method is a shrinkage-based estimate of the distribution on deformation parameters. This estimate allows us to simulate realizations of registration errors, which can then be used as training data for learning spatial confidence regions. Experimental results in 2-D suggest that the confidence regions are effective based on their empirical *coverage rates*.

Chapter 3 and 4 were devoted to the topic of *connectomics*, which is the study of brain connectivity. The goal here was to establish a multivariate method that allows us to predict the diagnostic status of an individual using whole-brain functional connectomes derived from resting state fMRI. As opposed to previous approaches which are generally blind to the spatial structure of the data, the method we introduce in Chapter 3 explicitly accounts for the 6-D structure in the connectome via spatially-informed regularizers, namely the fused Lasso and the GraphNet penalty. To solve the resulting nonsmooth and high dimensional optimization problem, we introduced a scalable algorithm based on the alternating direction method, and showed that the inner subproblems of the algorithm can be solved efficiently in analytical form by coupling the variable splitting strategy with a data augmentation scheme. Chapter 4 extends these ideas to a setting where the data are collected from multiple imaging sites. In brief, rather than training a single classifier over a pooled dataset, we proposed to simultaneously learn an individual classifier for each site by adopting a multitask learning framework, where the *sites* are treated as the *tasks*. Experiments on large real-world schizophrenia and ADHD dataset demonstrated that our methods generate accurate disease prediction with superior interpretability of discriminative features, and thus could provide new insights into how psychiatric disorders impact brain network topology.

5.2 Future Directions

Machine learning methods are increasingly being applied in various areas of biomedical science, and several promising results have been produced in the field of connectomics. However, we are far from achieving the goal of identifying a robust, universally accepted connectivity-based biomarker that accurately reflects the underlying neurobiological mechanism of the disease process of interest. For instance, while the multitask learning approach introduced in Chapter 4 produced superior results on the multisite ADHD-200 dataset, the classification accuracies are far from the level where the selected features can be interpreted as reliable ADHD biosignatures. Such result corroborates the fact that multisite data are highly complex and diverse, and it remains to be seen whether there are better ways to handle the numerous sources of inter-site heterogeneities. Moreover, the interpretability of the features selected from the multitask approach in Chapter 4 is heavily limited since the *sign* of the groups of features do not necessarily agree across imaging sites. Thus, it is important to investigate ways to extend our method so that the groups of features selected are consistent across sites.

Another interesting direction for future research is to investigate ways to integrate *multimodal fusion* techniques into our connectome-based disease prediction framework. In particular, there is a recent trend in neuroimaging research to combine multiple image modalities for multivariate pattern analysis (Uludag and Roebroeck, 2014; Zhu et al., 2014), where the idea is to enhance prediction performance by leveraging the complementary information available from different modalities. For example, Alzheimer's disease and mild cognitive impairment are known to be related with symptoms such as brain atrophy and neuro-metabolic alterations, which can be measured from modalities such as structural MRI, PET, and cerebrospinal fluid. Recent researches demonstrated that when classifying patients with Alzheimer's disease from healthy controls, the prediction performance can be substantially improved by training over these modalities (Liu et al., 2014; Zhang and Shen, 2012; Zhang et al., 2011). In the context of connectomics, it would be interesting to see if combining other modalities such as EEG and structural connectomes (typically constructed from DTI) can improve prediction performance and give more precise estimates of connectivity-based biomarkers.

As an overall remark, there are several remaining questions that still must be addressed before an automated neuroimaging-based diagnostic system to enter the clinical realm, and new statistical modeling techniques are in critical need.

BIBLIOGRAPHY

BIBLIOGRAPHY

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

M. Allison, S. Ramani, J. Fessler, Accelerated regularized estimation of MR coil sensitivities using Augmented Lagrangian methods, Medical Imaging, IEEE Transactions on 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).

A. Argyriou, T. Evgeniou, M. Pontil, Convex multi-task feature learning, Mach. Learn. 73 (2008).

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

A.J. Atkinson, W.A. Colburn, V.G. Degruttola, D.L. Demets, G.J. Downing, D.F. Hoth, J.A. Oates, C.C. Peck, R.T. Schooley, B.A. Spilker, J. Woodcock, S.L. Zeger, Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework, Clinical Pharmacology & Therapeutics 69 (2001).

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, NeuroImage: Clinical 3 (2013).

F. Bach, R. Jenatton, J. Mairal, G. Obozinski, Optimization with sparsity-inducing penalties, Found. Trends Mach. Learn. 4 (2012).

F.R. Bach, Bolasso: model consistent Lasso estimation through the bootstrap., in: Proc. 25th Int. Conf. Machine Learning (ICML), 2008, pp. 33–40.

F.R. Bach, Consistency of trace norm minimization, J. Mach. Learn. Res. 9 (2008b).

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

D.S. Bassett, E.T. Bullmore, Human brain networks in health and disease., Current opinion in neurology 22 (2009).

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).

J. Bien, R.J. Tibshirani, Sparse estimation of a covariance matrix, Biometrika 98 (2011).

D.L. Bihan, H. Johansen-Berg, Diffusion MRI at 25: Exploring brain tissue structure and function, NeuroImage 61 (2012).

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

B. Biswal, F. Zerrin Yetkin, V.M. Haughton, J.S. Hyde, Functional connectivity in the motor cortex of resting human brain using echo-planar MRI, Magnetic Resonance in Medicine 34 (1995).

B.B. Biswal, M. Mennes, X.N. Zuo, S. Gohel, C. Kelly, S.M. Smith, C.F. Beckmann, J.S. Adelstein, R.L. Buckner, S. Colcombe, A.M. Dogonowski, M. Ernst, D. Fair, M. Hampson, M.J. Hoptman, J.S. Hyde, V.J. Kiviniemi, R. KÃűtter, S.J. Li, C.P. Lin, M.J. Lowe, C. Mackay, D.J. Madden, K.H. Madsen, D.S. Margulies, H.S. Mayberg, K. McMahon, C.S. Monk, S.H. Mostofsky, B.J. Nagel, J.J. Pekar, S.J. Peltier, S.E. Petersen, V. Riedl, S.A.R.B. Rombouts, B. Rypma, B.L. Schlaggar, S. Schmidt, R.D. Seidler, G.J. Siegle, C. Sorg, G.J. Teng, J. Veijola, A. Villringer, M. Walter, L. Wang, X.C. Weng, S. Whitfield-Gabrieli, P. Williamson, C. Windischberger, Y.F. Zang, H.Y. Zhang, F.X. Castellanos, M.P. Milham, Toward discovery science of human brain function, Proceedings of the National Academy of Sciences 107 (2010).

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. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

S.L. Bressler, V. Menon, Large-scale brain networks in cognition: emerging methods and principles, Trends in Cognitive Sciences 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.

E. Bullmore, O. Sporns, Complex brain networks: graph theoretical analysis of structural and functional systems, Nature Reviews Neuroscience 10 (2009).

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, Signal Processing Magazine, IEEE 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, Machine Learning 28 (1997).

F. Castellanos, E. Proal, Large-scale brain systems in ADHD: beyond the prefrontal-striatal model, Trends Cogn. Sci. 16 (2012).

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).

S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM Review 43 (2001).

X. Chen, J. He, R. Lawrence, J.G. Carbonell, Adaptive multi-task sparse learning with an application to fMRI study, in: SDM, pp. 212–223.

X. Chen, Q. Lin, S. Kim, J.G. Carbonell, E.P. Xing, Smoothing proximal gradient method for general structured sparse regression, The Annals of Applied Statistics 6 (2012b).

G. Christensen, X. Geng, J. Kuhl, J. Bruss, T. Grabowski, I. Pirwani, M. Vannier, J. Allen, H. Damasio, Introduction to the non-rigid image registration evaluation project (NIREP), in: Biomedical Image Registration, volume 4057, 2006, pp. 128–135.

S.Y. Chun, J.A. Fessler, A simple regularizer for B-spline nonrigid image registration that encourages local invertibility, IEEE J. Sel. Top. Sig. Proc. 3 (2009), special Issue on Digital Image Processing Techniques for Oncology.

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, Frontiers in Neuroscience 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).

P. Combettes, J.C. Pesquet, Proximal splitting methods in signal processing, in: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer Optimization and Its Applications, Springer New York, 2011, pp. 185–212.

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, Magnetic Resonance in Medicine 62 (2009).

D. Dai, J. Wang, J. Hua, H. He, Classification of ADHD children through multimodal magnetic resonance imaging, Frontiers in Systems Neuroscience 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).

S. Dey, A.R. Rao, M. Shah, Exploiting the brain's network structure in identifying ADHD, Frontiers in Systems Neuroscience 6 (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., Molecular 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, Journal of Child Psychology and Psychiatry 47 (2006).

D. Donoho, De-noising by soft-thresholding, Information Theory, IEEE Transactions on 41 (1995).

D.L. Donoho, High-dimensional data analysis: the curses and blessings of dimensionality, in: Aide-Memoire of a Lecture at AMS Conference on Math Challenges of the 21st Century, 2000.

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).

R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification (2Nd Edition), Wiley-Interscience, 2000.

D.V. Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. Curtiss, S.D. Penna, D. Feinberg, M. Glasser, N. Harel, A. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S. Petersen, F. Prior, B. Schlaggar, S. Smith, A. Snyder, J. Xu, E. Yacoub, The human connectome project: A data acquisition perspective, NeuroImage 62 (2012).

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, Frontiers in Systems Neuroscience 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, Statist. Sinica 20 (2010).

J.M. Fitzpatrick, J.B. West, The distribution of target registration error in rigid-body, point-based registration, IEEE Trans. Med. Imaging 20 (2001).

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, Frontiers in Systems Neuroscience 4 (2010).

M.D. Fox, M.E. Raichle, Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging, Nature Reviews Neuroscience 8 (2007).

J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9 (2007).

K.J. Friston, Functional and effective connectivity in neuroimaging: a synthesis, Hum Brain Mapp 2 (1994).

D. Gabay, Applications of the method of multipliers to variational inequalities, in: M. Fortin, R. Glowinski (Eds.), Augmented Lagrangian Methods: Applications to the Solution of Boundary Value Problems, North-Holland, Amsterdam, 1983, pp. 299–340.

D. Gabay, B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation, Computers & Mathematics with Applications 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, Human Brain Mapping 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).

G.H. Golub, C.F. Van Loan, Matrix computations, 3 ed., Johns Hopkins University Press, 1996.

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).

R.M. Gray, Toeplitz and circulant matrices: a review, Commun. Inf. Theory 2 (2005).

M.D. Greicius, B. Krasnow, A.L. Reiss, V. Menon, Functional connectivity in the resting brain: A network analysis of the default mode hypothesis, Proceedings of the National Academy of Sciences 100 (2003).

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, Journal of Machine Learning Research 3 (2003).

P. Hagmann, From Diffusion MRI to Brain Connectomics, Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne, 2005.

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).

J. Heinzle, M.A. Wenzel, J.D. Haynes, Visuomotor functional network topology predicts upcoming tasks, J. Neurosci. 32 (2012).

D. Henrion, J. Malick, Projection methods in conic optimization, in: Handbook on Semidefinite, Conic and Polynomial Optimization, volume 166 of *International Series in Operations Research & Management Science*, Springer US, 2012, pp. 565–600.

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

D.L.G. Hill, P.G. Batchelor, M. Holden, D.J. Hawkes, Medical image registration, Phys. Med. Biol. 46 (2001).

J. Hlinka, M. Palus, M. Vejmelka, D. Mantini, M. Corbetta, Functional connectivity in resting-state fMRI: Is linear correlation sufficient?, NeuroImage 54 (2011).

A.E. Hoerl, R.W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics 12 (1970).

M. Holden, A review of geometric transformations for nonrigid body registration, IEEE Trans. Med. Imag. 27 (2008).

M. Hub, M.L. Kessler, C.P. Karger, A stochastic approach to estimate the uncertainty involved in B-spline image registration, IEEE Trans. Med. Imaging 28 (2009).

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, Multiscale mining of fMRI data with hierarchical structured sparsity, SIAM Journal on Imaging Sciences 5 (2012).

D. Kessler, M. Angstadt, R. Welsh, Y. Fang, C. Sripada, Modality spanning deficits in attention-deficit/hyperactivity disorder in functional networks, gray matter, and white matter, The Journal of Neuroscience (accepted) (2014).

S. Klein, M. Staring, J.P.W. Pluim, Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines, Image Processing, IEEE Transactions on 16 (2007).

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, Human Brain Mapping 31 (2010).

J. Kybic, Bootstrap resampling for image registration uncertainty estimation without ground truth, IEEE Transactions on Image Processing 19 (2010).

J. Kybic, M. Unser, Fast parametric elastic image registration, IEEE Trans. Im. Proc. 12 (2003).

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. Cognitive Neuroscience 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).

O. Ledoit, M. Wolf, Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, Journal of Empirical Finance 10 (2003).

F. Liu, C.Y. Wee, H. Chen, D. Shen, Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification, NeuroImage 84 (2014).

Y. Long, J.A. Fessler, J.M. Balter, Accuracy estimation for projection-to-volume targeting during rotational therapy: A feasibility study, Medical Physics 37 (2010).

K. Lounici, M. Pontil, A.B. Tsybakov, S.A. van de Geer, Taking advantage of sparsity in multi-task learning, in: COLT.

M.E. Lynall, D.S. Bassett, R. Kerwin, P.J. McKenna, M. Kitzbichler, U. Muller, E. Bullmore, Functional connectivity and brain networks in schizophrenia, The Journal of Neuroscience 30 (2010).

J. Mairal, R. Jenatton, G. Obozinski, F. Bach, Convex and network flow optimization for structured sparsity, Journal of Machine Learning Research 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. Marquand, M. Brammer, S. Williams, O. Doyle, Bayesian multi-task learning for decoding multi-subject neuroimaging data, NeuroImage 92 (2014).

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

N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the lasso, Ann. Statist. 34 (2006).

N. Meinshausen, P. Bühlmann, Stability selection, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 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 in Cognitive Sciences 15 (2011).

C.R. Meyer, J.L. Boes, B. Kim, P.H. Bland, K.R. Zasadny, P.V. Kison, K. Koral, K.A. Frey, R.L. Wahl, Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using affine and thin-plate spline warped geometric deformations., Medical Image Analysis (1997).

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, Medical Imaging, IEEE Transactions on 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. Modersitzki, Numerical Methods for Image Registration, Oxford University Press, 2004.

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).

J. Mota, J. Xavier, P. Aguiar, M. Puschel, D-ADMM: A communication-efficient distributed algorithm for separable optimization, Signal Processing, IEEE Transactions on 61 (2013).

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, Cerebral Cortex 21 (2011).

S. Negahban, P.D. Ravikumar, M.J. Wainwright, B. Yu, A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, Statistical Science 27 (2012).

Y. Nesterov, Gradient methods for minimizing composite objective function, CORE Discussion Papers 2007076, Université Catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007.

K.B. Nooner, S. Colcombe, R. Tobe, M. Mennes, M. Benedict, A. Moreno, L. Panek, S. Brown, S. Zavitz, Q. Li, S. Sikka, D. Gutman, S. Bangaru, R.T. Schlachter, S. Kamiel, A. Anwar, C. Hinz, M. Kaplan, A. Rachlin, S. Adelsberg, B. Cheung, R. Khanuja, C. Yan, C. Craddock, V. Calhoun, W. Courtney, M. King, D. Wood, C. Cox, C. Kelly, A. DiMartino, E. Petkova, P. Reiss, N. Duan, D. Thompsen, B. Biswal, B. Coffey, M. Hoptman, D.C. Javitt, N. Pomara, J. Sidtis, H. Koplewicz, F.X. Castellanos, B. Leventhal, M. Milham, The nki-rockland sample: A model for accelerating the pace of discovery science in psychiatry, Frontiers in Neuroscience 6 (2012).

G. Obozinski, L. Jacob, J.P. Vert, Group Lasso with overlaps: the Latent Group Lasso approach, Preprint, arXiv:1110.0413 (2011).

G. Obozinski, B. Taskar, M.I. Jordan, Joint covariate selection and joint subspace selection for multiple classification problems, Statistics and Computing 20 (2010).

S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2010).

F. Pereira, T. Mitchell, M. Botvinick, Machine learning classifiers and fMRI: A tutorial overview, NeuroImage 45 (2009).

J.P.W. Pluim, J.B.A. Maintz, M.A. Viergever, Mutual-information-based registration of medical images: a survey, IEEE Trans. Med. Imag. 22 (2003).

R.A. Poldrack, D.M. Barch, J.P. Mitchell, T.D. Wager, A.D. Wagner, J.T. Devlin, C. Cumba, O. Koyejo, M.P. Milham, Toward open sharing of task-based fMRI data: the OpenfMRI project, Front Neuroinform 7 (2013).

J.B. Poline, J.L. Breeze, S.S. Ghosh, K. Gorgolewski, Y.O. Halchenko, M. Hanke, K.G. Helmer, D.S. Marcus, R.A. Poldrack, Y. Schwartz, J. Ashburner, D.N. Kennedy, Data sharing in neuroimaging research, Front. Neuroinformatics 6 (2012).

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).

N.S. Rao, C.R. Cox, R.D. Nowak, T.T. Rogers, Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis., NIPS (2013).

P.M. Rasmussen, L.K. Hansen, K.H. Madsen, N.W. Churchill, S.C. Strother, Model sparsity and brain pattern interpretation of classification models in neuroimaging, Pattern Recognition 45 (2012).

B. Recht, M. Fazel, P.A. Parrilo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, SIAM Review 52 (2010).

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

J. Richiardi, S. Achard, H. Bunke, D. Van De Ville, Machine learning with brain graphs: Predictive modeling approaches for functional imaging in systems neuroscience, Signal Processing Magazine, IEEE 30 (2013).

P. Risholm, S. Pieper, E. Samset, W.M.W. III, Summarizing and visualizing uncertainty in non-rigid registration, in: MICCAI (2), pp. 554–561.

M.D. Robinson, P. Milanfar, Fundamental performance limits in image registration, IEEE Transactions on Image Processing 13 (2004).

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

J. Rondina, T. Hahn, L. de Oliveira, A. Marquand, T. Dresler, T. Leitner, A. Fallgatter, J. Shawe-Taylor, J. Mourao-Miranda, SCoRS–a method based on stability for feature selection and mapping in neuroimaging, IEEE Trans. Med. Imaging 33 (2014).

D. Ruan, J.A. Fessler, Fundamental performance analysis in image registration problems: Cramér-Rao bound and its variations, Technical Report 386, Communications and Signal Processing Laboratory, Dept. of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, 2008.

D. Rueckert, L.I. Sonoda, C. Hayes, D.L.G. Hill, M.O. Leach, D.J. Hawkes, Nonrigid registration using free-form deformations: application to breast MR images, IEEE Trans. Med. Imag. 18 (1999).

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

A. Scott, W. Courtney, D. Wood, R. De la Garza, S. Lane, R. Wang, J.A. Turner, M. King, J. Roberts, V.D. Calhoun, COINS: An innovative informatics and neuroimaging tool suite built for large heterogeneous datasets, Frontiers in Neuroinformatics 5 (2011).

W. Shi, X. Zhuang, H. Wang, S. Duckett, D. Luong, C. Tobon-Gomez, K. Tung, P. Edwards, K. Rhode, R. Razavi, S. Ourselin, D. Rueckert, A comprehensive cardiac motion estimation framework using both untagged and 3-d tagged mr images based on nonrigid registration, Medical Imaging, IEEE Transactions on 31 (2012).

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).

I.J. Simpson, J.A. Schnabel, A.R. Groves, J.L. Andersson, M.W. Woolrich, Probabilistic inference of regularisation in non-rigid registration, NeuroImage 59 (2012).

S.M. Smith, C.F. Beckmann, J. Andersson, E.J. Auerbach, J. Bijsterbosch, G. Douaud, E. Duff, D.A. Feinberg, L. Griffanti, M.P. Harms, M. Kelly, T. Laumann, K.L. Miller, S. Moeller, S. Petersen, J. Power, G. Salimi-Khorshidi, A.Z. Snyder, A.T. Vu, M.W. Woolrich, J. Xu, E. Yacoub, K. UÄ§urbil, D.C.V. Essen, M.F. Glasser, Resting-state fMRI in the human connectome project, NeuroImage 80 (2013).

A. Sotiras, C. Davatzikos, N. Paragios, Deformable medical image registration: A survey, Medical Imaging, IEEE Transactions on 32 (2013).

O. Sporns, The human connectome: Origins and challenges, NeuroImage 80 (2013).

O. Sporns, G. Tononi, R. Kötter, The human connectome: A structural description of the human brain, PLoS Computational Biology 1 (2005).

S. Sra, S. Nowozin, S.J. Wright, Optimization for Machine Learning, MIT Press, 2012.

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, R.C. Welsh, K. Prem Kumar, M. Angstadt, Disrupted network architecture of the resting brain in attention-deficit/hyperactivity disorder, Human Brain Mapping 35 (2014).

C.S. 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, Biological Psychiatry 59 (2006).

K. Strimbu, J.A. Tavel, What are Biomarkers?, Current Opinion in HIV and AIDS 5 (2010).

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

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. Statist. 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, in: Soviet Math. Doklady, volume 4, pp. 1035–1038.

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, Schizophrenia Research 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).

K. Uludag, A. Roebroeck, General overview on the merits of multimodal neuroimaging data fusion, NeuroImage 102, Part 1 (2014).

M. Unser, Splines: A perfect fit for signal and image processing, IEEE Sig. Proc. Mag. 16 (1999).

M. Unser, A. Aldroubi, M. Eden, Fast B-spline transforms for continuous image representation and interpolation, Pattern Analysis and Machine Intelligence, IEEE Transactions on 13 (1991).

M. Unser, A. Aldroubi, M. Eden, B-spline signal processing. I. theory, Signal Processing, IEEE Transactions on 41 (1993a).

M. Unser, A. Aldroubi, M. Eden, B-spline signal processing. II. Efficiency design and applications, Signal Processing, IEEE Transactions on 41 (1993b).

M. Unser, A. Aldroubi, M. Eden, The L2-polynomial spline pyramid, Pattern Analysis and Machine Intelligence, IEEE Transactions on 15 (1993c).

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, ICML 2012.

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

Y. Wang, J. Yang, W. Yin, Y. Zhang, A new alternating minimization algorithm for total variation image reconstruction, SIAM J. Img. 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).

T. Watanabe, D. Kessler, C. Scott, M. Angstadt, C. Sripada, Disease prediction based on functional connectomes using a scalable and spatially-informed support vector machine, NeuroImage 96 (2014a).

T. Watanabe, C. Scott, Spatial confidence regions for quantifying and visualizing registration uncertainty, Biomedical Image Registration 7359 (2012).

T. Watanabe, C. Scott, D. Kessler, M. Angstadt, C. Sripada, Scalable fused Lasso SVM for connectome-based disease prediction, in: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 5989–5993.

T. Watanabe, C. Scott, D. Kessler, C. Sripada, Multisite disease classification with functional connectomes via multitask structured sparse SVM, in: Workshop on Sparsity Techniques in Medical Imaging.

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

M.W. Weiner, P.S. Aisen, J. Jack, Clifford R., W.J. Jagust, J.Q. Trojanowski, L. Shaw, A.J. Saykin, J.C. Morris, N. Cairns, L.A. Beckett, A. Toga, R. Green, S. Walter, H. Soares, P. Snyder, E. Siemers, W. Potter, P.E. Cole, M. Schmidt, The Alzheimer's Disease Neuroimaging Initiative: Progress report and future plans, Alzheimer's & Dementia: The Journal of the Alzheimer's Association 6 (2010).

M.W. Weiner, D.P. Veitch, P.S. Aisen, L.A. Beckett, N.J. Cairns, R.C. Green, D. Harvey, C.R. Jack, W. Jagust, E. Liu, J.C. Morris, R.C. Petersen, A.J. Saykin, M.E. Schmidt, L. Shaw, J.A. Siuciak, H. Soares, A.W. Toga, J.Q. Trojanowski, The Alzheimer's disease neuroimaging initiative: A review of papers published since its inception, Alzheimer's & Dementia 8 (2012).

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, Computational Statistics and Data Analysis 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., Journal of Neurophysiology 106 (2011).

I.S. Yetik, A. Nehorai, Performance bounds on image registration, IEEE Transactions on Signal Processing 54 (2006).

M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc. Ser. B Stat. Methodol. 68 (2006).

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).

D. Zhang, D. Shen, Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease, NeuroImage 59 (2012).

D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, Multimodal classification of Alzheimer's disease and mild cognitive impairment, NeuroImage 55 (2011).

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).

D. Zhu, T. Zhang, X. Jiang, X. Hu, H. Chen, N. Yang, J. Lv, J. Han, L. Guo, T. Liu, Fusing DTI and fMRI data: A survey of methods and applications, NeuroImage 102, Part 1 (2014).

H. Zou, T. Hastie, Regularization and variable selection via the Elastic Net, Journal of the Royal Statistical Society, Series B 67 (2005).