# Machine Learning Techniques for High-dimensional Neuroimaging Data Analysis

by

Xin Yang

A Dissertation Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy in Computational Sciences

Middle Tennessee State University

August 2016

Dissertation Committee:

Dr. Don Hong, Chair

Dr. Qiang Wu, Co-Chair

Dr. Hyrum D. Carroll

Dr. Cen Li

Dr. John Wallin

**I dedicate this research to my mother. I love you, Mom.**

## ACKNOWLEDGMENTS

The past four years at Middle Tennessee State University were challenging but a very fruitful time of my life. There are many people, whom I deeply appreciate, and without whom this work would never have been finished.

First and foremost is my advisor Dr. Don Hong, who provided invaluable insights, advice, immeasurable amount of patience and guidance throughout my research and career.

I was fortunate to work with Dr. Qiang Wu from whom I learned diverse statistical and mathematical skills. I want to thank him for bearing with me when I was asking so many questions during the discussions. I am also grateful to Dr. Lixin Shen from Syracuse University for valuable discussions related to this study.

I would like to express my gratitude to Dr. John Wallin, director of the Computational Science PhD program, for his guidance and support during my graduate studies. I would also like to thank my dissertation committee members Dr. Cen Li and Dr. Hyrum Carroll for their valuable discussions and suggestions for my dissertation writing. I would like to thank Alan Parker who helped to polish my dissertation writing.

Most of all, I would like to thank my family, especially my mom, Xiaoping Qian, my dad Jinxiu Yang, my sister Yang Yang, and my husband Ning Zhang, for always being supportive and encouraging. Without their boundless love, this dissertation would not have been possible.

## ABSTRACT

In the past two decades, neuroimaging has become the most commonly used imaging technique for the study of human brain, which has given us insights about the complex neural characteristics of the human brain and also provided helpful information for the diagnosis of various diseases.

However, the analysis of neuroimaing data is extremely complex, requiring the use of sophisticated techniques from acquiring raw data to image processing and statistical analysis. The purpose of this dissertation is to provide accurate and efficient machine learning models for neuroimaging data analysis. In this dissertation, we will focus on the study of two neuroimaging techniques: functional MRI data and MRI data.

Functional magnetic resonance imaging (fMRI) has become one of the most widely used techniques in investigating human brain function over the past two decades. However, the analysis of fMRI data is extremely complex due to its difficulties in big data processing. Hence, efficient and accurate machine learning models are necessary to interpret fMRI data by incorporating both spatial and temporal information. We will investigate a class of spatial multitask learning models which incorporates spatial information of each task's 2-dimensional neighborhood. Simulation and real application results show satisfactory performance from spatial multitask learning algorithms.

As Magnetic Resonance Imaging (MRI) has matured, a large number of researchers have studied Alzheimer's disease (AD) image data. Many high-dimensional classification methods use structural MRI brain images for classification between AD and healthy individuals. As computer computation power has improved, neural networks have been widely applied in Alzheimer's disease diagnosis. However, the first layer of this method is based on individual brain voxel, which means neural networks learn each voxel individually without considering the brain spatial information. This

method may lose some important information since the neighbor effect is ignored. Because the voxel of the brain is not isolated, in reality some brain area has an extremely close relationship. To overcome the shortcomings of the spatial correlation problem, we proposed a new technique called spatial regularization neural network (SRNN), which incorporates spatial information provided by each voxel's 3-dimensional neighbor voxels. It is successfully applied in real applications.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1 Overview

The human brain has almost 100 billion neurons with more than trillion's connections. Despite our technological advancements, we have still been unable to unlock the mysteries of the human brain. As a result, we still can not prevent or cure brain disorders such as Alzheimer's disease, autism, stroke, and so on. In 2013, president Barak Obama of the United States announced a bold new research investment: BRAIN Initiative. This bold new research is to revolutionize our understanding of the human brain and uncover its mysteries so that we can find new ways to prevent and cure brain disorders. Since the BRAIN Initiative has been announced, many academic institutions and scientists have answered this call and made significant contributions and progress. Some universities have invested large sums of money to buy neuroimaging scanners in order to collect human brain data, enabling data scientists to use mathematical models to analyze this data mathematically.

Among all these research efforts, neuroimaging has become the most commonly used imaging technique for the study of human brain, which has given us insights about the complex neural characteristics of the human brain and also provided helpful information for the diagnosis of many diseases. However, the analysis of neuroimaging data is extremely complex, requiring the use of sophisticated techniques from acquiring raw data to image processing and statistical analysis. As a result, it is often said that "neuroscience is data rich yet theory-poor." Therefore, the aim of this dissertation is to provide theories and specific models for the neuroimaging data

analysis. In this dissertation, we will focus on the study of machine learning techniques for high-dimensional neuroimaging data, especially to provide insight into the complex brain activities of the fMRI data and also Alzheimer's disease classification of MRI data.

This dissertation is organized as follows: Chapter 1 provides the background information of neuroimaging technique, such as fMRI and MRI, and presents the basic steps of image preprocessing. Chapter 2 presents a novel spatial regularization multitask learning framework for fMRI data analysis. A class of spatial multitask learning models: MTLRidge, MTLLasso, MTLEN was proposed. Simulation and real application are used to verify their performance. In Chapter 3, we will propose a spatial regularization approach for neural network and apply it to structural MRI Alzheimer's disease classification. Chapter 4 summarizes this dissertation and discusses future work to be done in this field of research.

## 1.2 Neuroimaging Techniques

Currently, two major neuroimaging techniques are available: the anatomical technique and the functional technique.

The anatomical technique is used to track normal and abnormal development of the human brain in both a healthy condition and disease condition. Moreover, the anatomical technique is also combined with the functional technique to track brain activity. The earliest brain imaging technique was computed tomography (CT). Now, CT has been largely replaced by the more powerful magnetic resonance imaging (MRI) technique [1]. MRI provides high quality images of brain structure.

The functional technique has become dominant in cognitive neuroscience because it can track the neural activity associated with the corresponding ability to perform a

particular cognitive task. There are a variety of noninvasive functional neuroimaging techniques. They are divided into two categories: The first one directly measures electrical activity associated with neuronal activity, such as electroencephalography (EEG) and magnetoencephalography (MEG). The second one indirectly measures neuronal activity by measuring changes in the local oxygenation of blood, such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) [1].

These techniques enable researchers to study the anatomical structure and metabolic function of the human brain throughout the life span, in both sickness and health. This will help uncover the mysteries of the human brain so that we can find new ways to prevent and treat brain disorders. In this dissertation, we have applied our proposed methodology to both fMRI and MRI data.

### 1.2.1 MRI

Magnetic resonance imaging (MRI) is an imaging technique to produce high quality images of the human body, which is primarily used for the human brain [2]. MRI provides good contrast between the different soft tissues of the body based on the facts that the human body is largely composed of water molecules, and MRI is based on Nuclear magnetic resonance (NMR) of hydrogen protons [3].

One very important feature of MRI is that it can generate different contrast characteristics images, such as T1 weighted MRI, T2 weighted MRI. T1 is when the machine only looks at the longitudinal movement of protons, so T1 images are usually used to look at anatomical information of healthy people. T2 is the transverse movement of protons, which is usually used to track the pathology of the patient. MRI has a wide range of application in medical diagnosis, because most of the dis-

ease tissue tends to have higher water molecules than healthy individuals [4]. As a result, MRI has been used as a matured medical diagnosis tool in hospitals and clinics.

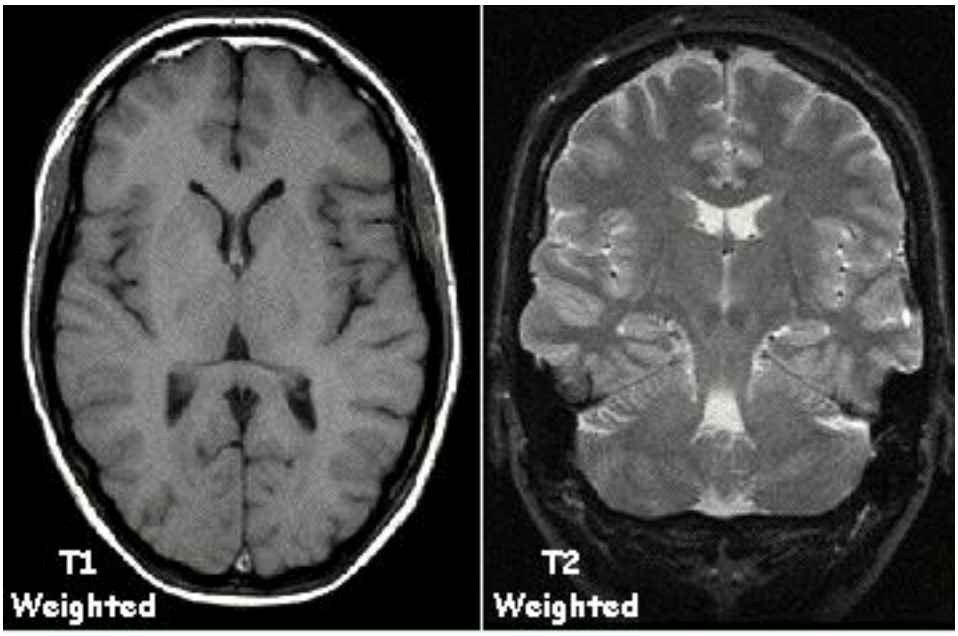



Figure 1: Different Weighted MRI.

Source: Magnetic Resonance Imaging e-tutorials [4]

MRI is a 3D data cube, which is usually composed of many MRI slices. Each MRI slice is a 2D image. The 2D brain image is made up of many pixels, we call each pixel a voxel in MRI/fMRI.

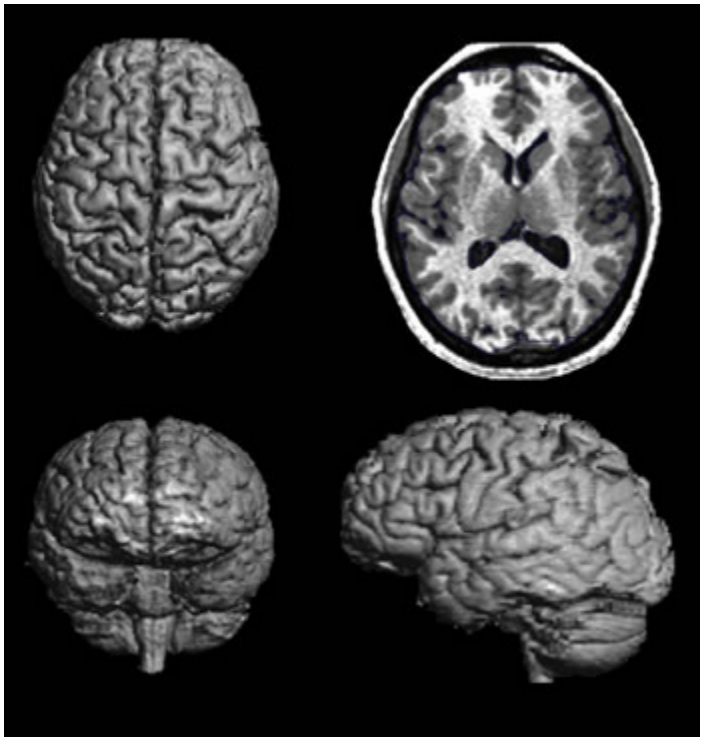


Figure 2: MRI 3D Cube.

Source: University of Missouri [5]

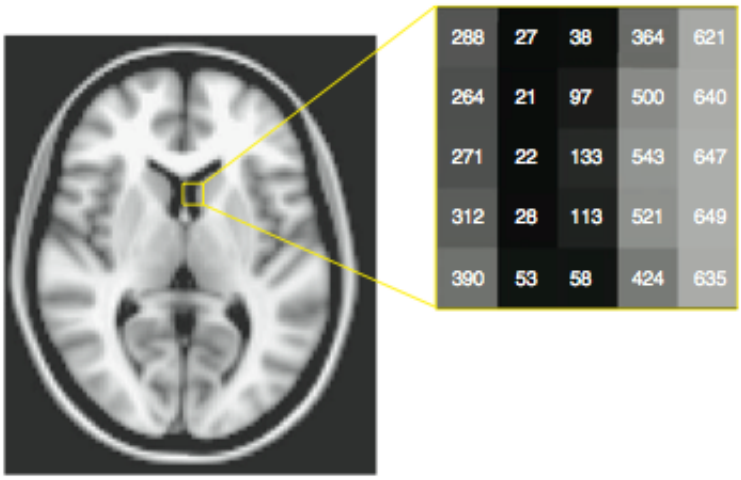



Figure 3: A MRI Slice.

Source: Handbook of Functional MRI Data Analysis [6]

In Figure 3, each voxel is represented by a number. The corresponding numbers for the particular voxels in the closeup section are shown on the right of Figure 3.

### 1.2.2 fMRI

Functional magnetic resonance imaging (fMRI) is a functional neuroimaging procedure using MRI technology that measures brain activity by detecting blood-oxygen-level-dependent (BOLD). The functional magnetic resonance imaging (fMRI) technique is used to indirectly measure brain activity by measuring changes in the local blood oxygen level, which in turn reflects the amount of brain activity. This is based on the fact that when neurons in the local brain area become active, the amount of blood flowing through that corresponding local area is also increased, which leads to a relative surplus in local blood oxygen. The higher blood oxygen level will create higher fMRI signal intensity. This measured signal in fMRI is referred to as the blood oxygenation level dependent (BOLD) [6]. The BOLD signal for an active voxel (blue) and the stimulus time series (red) is shown in Figure 4.

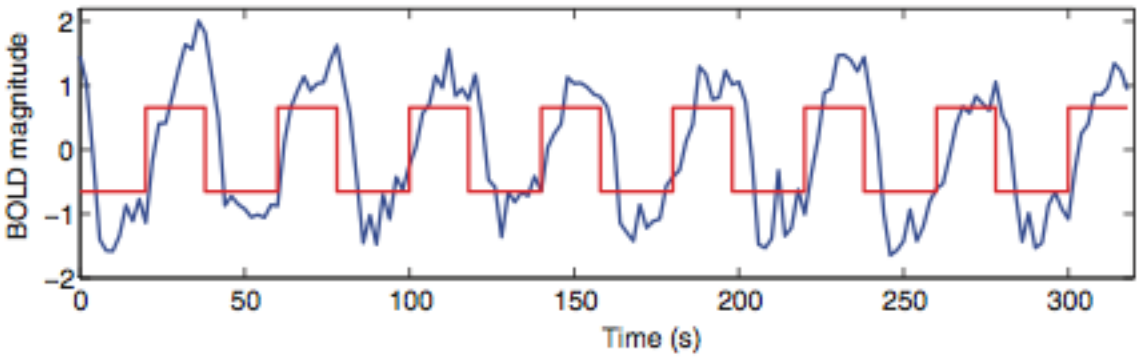



Figure 4: The Blood Oxygenation Level Dependent (BOLD) Signal.

Source: Handbook of Functional MRI Data Analysis [6]

If a voxel becomes active, the BOLD signal will start to gradually increase, and reach its peak after several seconds. Finally it will gradually go back to the normal state. This whole process is called hemodynamic response. "Hemo" means blood, "dynamic" means change. Figure 5 shows the ideal hemodynamic response. This whole process persists up to 20 seconds or more after the stimulus.

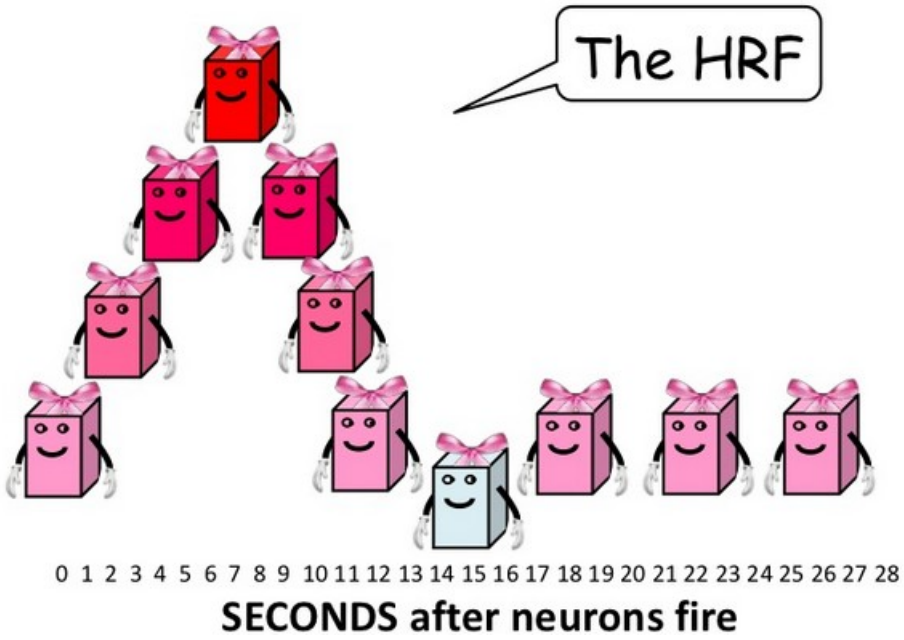



Figure 5: Hemodynamic Response Function.

Source: Intro to fMRI [7]

During an fMRI experiment, a series of 3D brain images are acquired while the subject performs a set of particular cognitive tasks. The fMRI machine will record a number for the magnetism of each voxel at each time point while the external particular task is ongoing, which means if we take a picture every 2 seconds for 5 minutes, we will get 150 numbers for each voxel. As there are 100,000 or more voxels in the whole brain, a huge amount of data will be processed. Hence accurate and

efficient methods are necessary due to the high dimensionality of fMRI data. Figure 6 shows a time series of 3D fMRI images, measure at very Repetition time (TR).

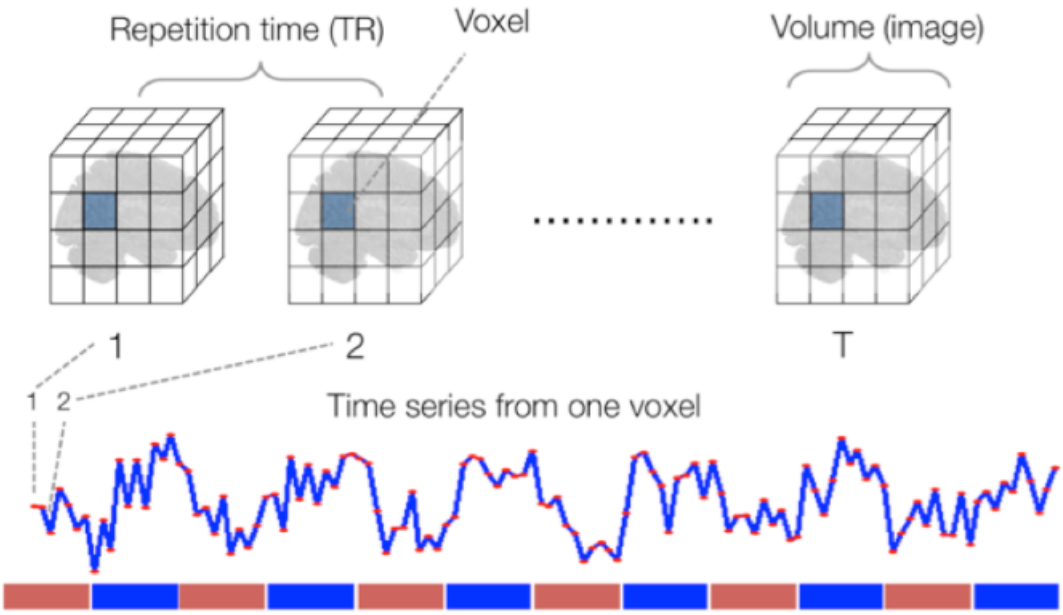



Figure 6: fMRI Time Series Data.

Source: Principles of fMRI. [8]

The general purpose of fMRI study is to analyze each voxel's time series to check whether the BOLD signal changes in response to an external particular cognitive task and hence to infer neuronal activity of the human brain. fMRI is used both in the research world, and in the clinical world. However, unlike MRI, the fMRI technique has not been tested enough for widespread commercial products, so it is mostly used as a research tool.

## 1.3 Image Preprocessing

The analysis of fMRI data is extremely complex due to its high dimensionality. Moreover, the data is susceptible to other factors such as head movement, variability between individuals, and variability through the time within individuals. Therefore, applying imaging preprocessing to MRI/fMRI data will provide more meaningful interpretation of the analysis results. As neuroimaging techniques matured, many laboratories began to distribute their software packages for MRI/fMRI analysis as open source. The most common packages are: SPM, FSL, AFNI and Brain Voyager shown in Figure 7.


Figure 7: Free Software for Neuroimage Preprocessing.

In our analysis, SPM is used for image preprocessing. Here, we outline the basic preprocessing pipeline for MRI/fMRI analysis as listed in in Statistic Parametric

Mapping (SPM) [9].

- Realignment

  This step will align a time series of images acquired from the same subject. The aim is to remove movement artifacts in fMRI time-series.

- Coregistration

  This step will implement a coregistration between the structural and functional data that maximizes the mutual information.

- Segmentation

  This step will segment the structural image and create grey and white matter images and bias-filed corrected structural image.

- Normalize

  This step normalizes different MRI images into a standard template such as MNI template.

- Smoothing

  This step is used to suppress noise and effects due to the residual differences in functional and structural images.

# CHAPTER 2

# SPATIAL REGULARIZATION FOR MULTITASK LEARNING AND APPLICATION IN FMRI DATA ANALYSIS

## 2.1 Introduction

Functional magnetic resonance imaging is an MRI procedure that measures brain activity by detecting associated changes in blood flow. This is based on the fact that when neurons in the local brain area become active, the amount of blood flowing through that corresponding local area is also increased, which leads to a relative surplus in local blood oxygen [6]. The higher blood oxygen level will create higher fMRI signal intensity. Because the neuronal activity can be indirectly observed via the blood oxygenation level dependent (BOLD) signal contrast, usually the BOLD signal is used to detect the neuronal activity. Functional magnetic resonance imaging(fMRI) has become the most widely used technique to investigate human brain function in the past two decades.

The general purpose of fMRI data studies is to analyze each voxel's time series data to detect whether the BOLD signal changes in response to a particular stimulus and hence to infer neuronal activity of the human brain. However, fMRI data has an extremely complicated structure. The subject's 3D volume brain is divided into a grid of volume boxes, or voxels. The BOLD signal is observed at each voxel at each time point resulting in an enormous amount of data. Hence efficient and accurate models are necessary in detecting accurate neuronal activity.

The analysis of fMRI data is challenging due to artifacts, variability in the data, and high dimensionality. The major components of fMRI analysis include but are

not limited to processing techniques, statistical modeling and inference from the data, and applications in medical diagnosis. The initial development of fMRI was driven by cognitive psychology researchers, who were interested in exploring the brain's active responses to external tasks [6].

One of the most important research areas in fMRI analysis literature has focused on the detection of the activated brain regions associated to human activities or diseases; see e.g. [10, 11, 12, 13, 14, 15] and references therein. This could be modeled from either the voxel level [10, 11] or cluster level [12]. The statistical models for active region detection include the general linear models [10] and autoregressive models [11]. In these models, each voxel is usually treated as a linear regression task. As all the tasks are correlated, considering all voxels together may benefit the modeling and inference. This has driven the use of multitask learning in this area [16, 17, 18, 19, 20].

In general, fMRI analysis methods have centered on the relationship between cognitive variables and individual brain voxels, but it has limits on what can be learned about brain activity by isolated voxel study. The application of the multitask method into fMRI data is motivated by fMRI studies in which functional activity is classified using brain voxels as features. By using multitask, each voxel can be studied as a task. This training process has significant advantages, since the related tasks (voxels) can help one another gain better performance.

Multitask learning (MTL) refers to a machine learning framework that learns multiple related tasks simultaneously to improve generalization. This is especially true when the data is small and the performance of single task learning is not as good. Intuitively it would seem that learning of one task could benefit from the information of closely related tasks. A more formal explanation is that the learning

of related tasks introduces an inductive bias while helping to significantly reduce the variance. MTL has been found successful in the study of many real applications [21, 22, 23, 24, 25, 26, 27]. A variety of techniques and algorithms have been proposed for MTL for different purposes and different problem domains.

The idea of MTL dates back at least to the application of NETtalk to learning both phonemes and their stresses [21, 22], although the concept of MTL was coined much later. In the context of neural network learning, back-propagation was used to learn multiple related tasks that are drawn from the same domain and share the same hidden units [22, 23, 24]. MTL formulation was also proposed for k-nearest neighborhood, kernel regression, and decision tree in [24].

In 1997, Caruana published a paper entitled Multitask Learning [24]. In this paper, Caruana demonstrated that multitask learning could work in many fields. He presents nine kinds of fields often available, and most of the real world problem fall into one of these domains [24]. In addition, he also mentioned in this paper that most problems traditionally used in machine learning have been preprocessed to fit single task learning (STL), thus eliminating the opportunities for multitask Learning before learning was applied.

In 2003, Bart and Tom proposed Bayesian multitask learning in [25]. In this paper, they proposed a hierarchical Bayesian approach to multitask in which some of the model parameters are shared explicitly, and others are soft-shared through a prior distribution. Their method is applied into two real world problems: a school problem and newspaper sales. Finally, they concluded that both problems are modeled better through Bayesian multitask learning.

In [26], Evgeniou and Pontill proposed an approach to multitask learning based on the minimization of regularization functions, which is the first generation of

regularization-based methods from single task to multitask learning. They applied their method into the "school data". The experiments show that their proposed method outperforms other multitask learning methods and largely perform better than single task learning methods. Moreover, their results significantly outperform the Bayesian method of [25].

In recent years, regularization theory was introduced into MTL. Regularized MTL algorithms are usually problem dependent because the penalty term is designed according to prior knowledge of the problem. For instance, by assuming all the tasks share a common component and each task has an additional individual component, the authors in [26] proposed an MTL approach by trading off the size of the common component and the individual components. By adjusting the trade-off parameter, this method allows the data itself to demonstrate how closely the tasks are related and how much improvement can be garnered by learning multiple tasks at the same time.

In some applications, not all tasks share the same components, but there is a cluster structure and only tasks belonging to the same cluster share a common component, while the relationship between tasks from different clusters may be weak. This has motivated the structured regularization for MTL [28, 29]. Temporal priors were introduced in a study of the progression of Alzheimer's disease, where each task is the status of patients at a time point, and the temporal relationship arises naturally [30].

While the multitask learning algorithm is becoming stable and flexible, researchers are paying more attention to apply it to more interesting high dimension medical data: MRI or fMRI data. In high dimensional data analysis such as fMRI data, feature selection is a natural issue and sparse penalty is required. In order to facilitate

sparsity, the adaptive multitask lasso and elastic net were introduced in [18] which utilizes the $l_1/l_q$ mixed matrix norm. Here, $p$-norm is $\|x\|_p := (\sum_{i=1}^{n} |x_i|^p)^{1/p}$, $p \geq 1$. $l_1$ norm is simply the sum of the absolute values of the columns, $\|x\|_1 := \sum_{i=1}^{n} |x_i|$.

One of the most common applications for medical image analysis is the diagnosis of Alzheimer's disease. Papers study this topic include [30], [31], [16], [17]. In [30], Jiayu Zhou and Lei Yuan propose a multitask learning formulation for predicting the disease progression measured by the cognitive scores and selecting markers predictive of the progression. They capture the relatedness among different tasks by a temporal group Lasso regularizer. This method is based on the regularization-based method from [26]. These experimental studies demonstrate the effectiveness of the proposed algorithm for capturing the progression trend, and also show that the markers selected by the proposed algorithm are consistent with the existing findings from other studies. In [17], Biao Jie and Daoqiang Zhang proposed a manifold regularized multitask learning framework to jointly select features from multi-modality data. The experimental results demonstrate the effectiveness of the proposed method.

In [18] and [19], multitask learning method is applied into fMRI data. In [18], the adaptive multitask elastic net method is used to study fMRI data and the results outperform Lasso and Elastic Net. In [19], Nikhil Rao and Christopher Cox proposed a new procedure called sparse overlapping sets Lasso, the experimental results demonstrate this method is better than Lasso and Group Lasso. Although these approaches are useful for fMRI data, they do not directly attempt to include the potential commonalities between voxels. To remedy this, we propose a class of multitask learning methods to extract the spatial information for each voxel that are neighbors by a shared tuning parameter.

In this chapter we propose a spatial regularization approach for MTL and apply it

to fMRI data analysis. In the problem of active region detection using fMRI data, the tasks (brain voxels) are spatially related. It is natural to code the spatial information into the training process to improve the learning performance. Works on this topic include [32, 33]. However, to the best of our knowledge, the idea of coding spatial information in the regularization theory context is new.

The remainder of this chapter is organized as follows: The linear regression model for single task is described in Section 2.2. We develop our spatial regularized multitask learning models and show how to solve the model in Section 2.3. In Section 2.4, the models are tested on both simulated and real fMRI data. We finish with concluding remarks in Section 2.5.

## 2.2 Linear Regression for a Single Task

For fMRI data analysis, our goal is to detect the neuronal activation for each voxel. We can formulate each voxel's time series data by using a standard linear regression model. The most traditional method to solving a linear regression model is the ordinary least square method (OLS, [34]). In linear regression, a scalar response variable $y$ is assumed to be linearly dependent on a set of $p$ predictors. The data is a sample of $n$ observations subject to noise:

$$y_i = x_i\beta + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{1}$$

where $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^p$ is a row vector, and $\beta \in \mathbb{R}^p$ is an unknown column vector. Denote $Y = (y_1, y_2, \ldots, y_n)^\top \in \mathbb{R}^n$ as a column vector of the response values, $X = [x_1; x_2; \ldots; x_n] \in \mathbb{R}^{n\times p}$ the data matrix, and $E = (\epsilon_1, \ldots, \epsilon_n)^\top$ the error vector. We

can rewrite (1) as

$$Y = X\beta + E.$$

The OLS estimator minimizes the sum of squared errors (SSE) made by predicting the true response $y_i$ by $x_i\beta$, that is

$$\hat{\beta} = \arg\min \|Y - X\beta\|_2^2 = \arg\min \sum_{i=1}^{n} (y_i - x_i\beta)^2.$$

Here and in the sequel $\|\cdot\|_q$ denotes the $q$-Euclidean norm for any $1 \leq q \leq \infty$. If $X$ is of full rank, the OLS estimator can be solved by a linear system

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y. \tag{2}$$

The OLS estimator is known as the best linear conditionally unbiased estimator. However, it could be numerically unstable when the matrix $X^\top X$ is singular or has a large conditional number. This is usually the case when $n < p$ or when the predictors are highly correlated. Even when the matrix is well conditioned, it may be beneficial to introduce some bias to facilitate some desired properties (such as sparsity). These considerations have led to the development and application of regularized regression methods such as ridge regression [35], Lasso [36] and Elastic Net [37].

Ridge regression is a method that utilizes Tikhonov regularization of the OLS estimator. It shrinks the coefficients in the estimator by minimizing the penalized SSE where the penalty term $\lambda_2\|\beta\|_2^2$ is determined by a regularization parameter $\lambda_2 > 0$ and the Euclidean 2-norm square of $\beta$, that is

$$\hat{\beta}_{Ridge} = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda_2\|\beta\|_2^2 \right\}. \tag{3}$$

The Ridge regression estimator can also be solved by a linear system, which gives

$$\hat{\beta}_{Ridge} = (X^\top X + \lambda_2 I)^{-1} X^\top Y. \tag{4}$$

Here and in the sequel, $I$ denotes an identity matrix (whose dimension is omitted if it is clear from the context or appears as subscript otherwise).

Although ridge regression is numerically stable, the coefficients are never exactly zero even when the corresponding predictors are irrelevant to the response. To implement variable selection, Tibshirani [36] proposed an alternative regularization approach called least absolute shrinkage and selection operator (Lasso). It minimizes the SSE with an $\ell_1$ norm penalty.

$$\hat{\beta}_{Lasso} = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right\}, \tag{5}$$

where $\lambda_1 > 0$ is the regularization parameter. It is well known that the $\ell_1$ penalty leads to sparse solution. Therefore, Lasso is advantageous for sparse models because of its facilitation of variable selection.

The elastic net (EN, [37]) also combines shrinkage and variable selection, and in addition encourages grouping of variables: groups of highly correlated variables tend to be selected together, whereas the Lasso would only select one variable of the group. To implement the grouping effect, EN comprises both the $\ell_2$ and $\ell_1$ penalty.

$$\hat{\beta}_{EN} = \arg\min_{\beta} \left\{ \|Y - X\beta\| l_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right\}. \tag{6}$$

EN is particularly useful in the "large $p$ small $n$" setting where the number of predictors is much larger than the number of observations. Since the $\ell_1$ norm is not differentiable at 0, the optimization process to solve Lasso and EN is more complicated than ridge regression. The most commonly used solvers include the LARS [38], cyclical coordinate descent [39], etc.

But from our study, we found that all these algorithms are only suitable for a single task, which means each time they only study one task individually. In fMRI data, if we assume each voxel is a task, then all these algorithms can only study one

voxel each time without considering brain spatial information. However, in reality, the voxels of the brain normally have some relations with each other, in order to improve this spatial limit we proposed a class of Spatial Multi-task learning algorithms. We present a class of spatial MTL algorithms for fMRI data analysis in section 2.3.

## 2.3 Spatial Multitask Learning

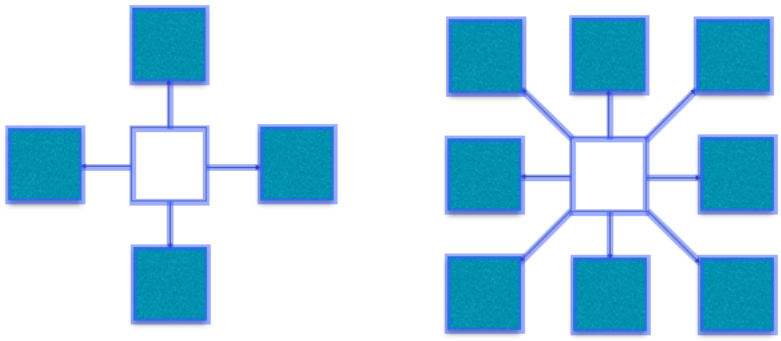
In fMRI studies, one of the important problems is detection of a functional region associated with certain brain activities. For each voxel, this can be done by a linear regression model. As the brain contains thousands of voxels, we need to solve thousands of linear regression problems. Of course one can solve these problems voxel by voxel using the single task learning methods. However, this is suboptimal because each functional region contains multiple voxels that are spatially continuous. As a result, if one voxel is active, then its neighbors are very likely to be active as well. Conversely, if one voxel is inactive, its neighbors are unlikely to be active. We expect such spatial information will benefit the learning performance if it is used in the training process. In some applications, Markov random field is used to incorporate the spatial information: image reconstruction [40] and IMS proteomic data analysis in [41], for instance. In this dissertation, we propose a spatial regularization approach for MTL by using user defined neighborhood structure.

In MTL regression, there are $T \geq 2$ tasks. Assume the $t$-th task has the data matrix $X_t$ and response vector $Y_t$ which are linked by

$$Y_t = X_t\beta_t + E_t.$$

To code the spatial information, we first define a neighborhood system. It is defined by the user and may be quite data dependent. An example of the 4 or 8 nearest

neighborhood systems in two dimensional space is shown in Figure 8.



(a) 4 Neighborhood (b) 8 Neighborhood

Figure 8: Neighborhood Structure for Each Task

Based on the neighborhood system, we define the task similarity coefficient by

$$w_{tk} = \begin{cases} 1, & \text{if task } t \text{ is a neighborhood of task } k; \\ 0, & \text{if task } t \text{ is not a neighborhood of task } k. \end{cases}$$

We assume the neighborhood system is symmetrically defined so that $w_{tk} = w_{kt}$. The penalty term for spatial regularization is defined by

$$\lambda_s \sum_{t,k=1}^{T} w_{tk} \|\beta_t - \beta_k\|_2^2.$$

Here we use a shared tuning parameter, $\lambda_s$, to adjust the spatial information. When $\lambda_s$ becomes large, it forces the neighboring tasks to become very close, while as $\lambda_s$ tends to 0, the tasks are treated as independent.

By applying the spatial penalty to ridge regression, Lasso, and EN, we propose three new MTL algorithms. We discuss their formulation and solution in the next three subsections. In the sequel, we will denote $B = [\beta_1; \beta_2; \ldots; \beta_T] \in \mathbb{R}^{pT}$ as the

column vector composed of all the task coefficients and $W = [w_{tk}]_{t,k=1}^{T}$ as the matrix of the task similarity coefficients, here $\beta_i \in p \times 1, i = 1, 2, ..., T$.

### 2.3.1 Spatial Ridge Regression

When we learn all $T$ ridge regression problems simultaneously and apply the spatial penalty, the resulted MTL learning algorithm, called spatial ridge regression algorithm, takes the form

$$\hat{B}_{SR} = \arg\min_{B} \left\{ \sum_{t=1}^{T} \parallel Y_t - X_t\beta_t \parallel_2^2 + \lambda_2 \sum_{t=1}^{T} \|\beta_t\|_2^2 + \lambda_s \sum_{t,k=1}^{T} \omega_{tk} \|\beta_t - \beta_k\|_2^2 \right\}. \quad (7)$$

It is easy to check that

$$\sum_{t=1}^{T} \parallel Y_t - X_t\beta_t \parallel_2^2 = B^\top SB - 2V^\top B + \sum_{i=1}^{T} \|Y_t\|_2^2,$$

where $S = \text{diag}(X_1^\top X_1, \ldots, X_T^\top X_T)$ and $V = [X_1^\top Y_1; \ldots; X_T^\top Y_T]$. For S, $X_k \in n \times p$, $X_k^\top X_k \in p \times p$, so we have $S$:

$$\begin{pmatrix} X_1^\top X_1 & 0 & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & X_2^\top X_2 & 0 & \ldots & \ldots & \ldots & \ldots \\ \vdots & 0 & \ddots & \ddots & \ldots & \ldots & \ldots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ldots & \ldots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ldots & \ldots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ldots & X_T^\top X_T \end{pmatrix}_{pT \times pT}.$$

Let $d_t = \sum_{k=1}^{T} w_{tk}$, $D_1 = \text{diag}(d_1 I_p, \ldots, d_T I_p)$, and $D_2 = W \otimes I_p$, (where $\otimes$ denotes the kronecker product of two matrices). For $D_1$, we have the following matrix:

$$\begin{pmatrix}
d_1 & 0 & 0 & 0 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & \overset{p}{\ddots} & 0 & 0 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & d_1 & 0 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & 0 & d_2 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & 0 & 0 & \overset{p}{\ddots} & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & 0 & 0 & 0 & d_2 & \dots & \dots & \dots & \dots & \dots & \dots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & 0 & 0 & 0 & d_T & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & 0 & 0 & 0 & 0 & \overset{p}{\ddots} & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & 0 & 0 & 0 & 0 & 0 & d_T
\end{pmatrix}_{pT \times pT}.$$

For $D_2$, we have:

$$\begin{pmatrix}
\omega_{11} & 0 & 0 & \omega_{12} & 0 & 0 & \dots & \dots & \dots & \omega_{1T} & 0 & 0 \\
0 & \overset{p}{\ddots} & 0 & 0 & \overset{p}{\ddots} & 0 & \dots & \dots & \dots & 0 & \overset{p}{\ddots} & 0 \\
0 & 0 & \omega_{11} & 0 & 0 & \omega_{12} & \dots & \dots & \dots & 0 & 0 & \omega_{1T} \\
\omega_{21} & 0 & 0 & \omega_{22} & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & \overset{p}{\ddots} & 0 & 0 & \overset{p}{\ddots} & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & \omega_{21} & 0 & 0 & \omega_{22} & \dots & \dots & \dots & \dots & \dots & \dots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \dots & \dots & \dots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \dots & \dots & \dots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \dots & \dots & \dots \\
\omega_{T1} & 0 & 0 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \omega_{TT} & 0 & 0 \\
0 & \overset{p}{\ddots} & 0 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & 0 & \overset{p}{\ddots} & 0 \\
0 & 0 & \omega_{T1} & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & 0 & 0 & \omega_{TT}
\end{pmatrix}_{pT \times pT}.$$

Define $D = 2D_1 - 2D_2$. Then we have

$$\sum_{t,k=1}^{T} \omega_{tk} \|\beta_t - \beta_k\|_2^2 = B^\top DB.$$

Let $Q = S + \lambda_s D$. The function in (7) that needs to be minimized takes the quadratic form

$$B^\top (Q + \lambda_2 I)B - 2V^\top B + \sum_{i=1}^{T} \|Y_t\|_2^2.$$

It can be solved by a linear system:

$$\hat{B} = (Q + \lambda_2 I)^{-1} V.$$

Noticing that $Q = S + \lambda_S D$, where $S$ is a block diagonal matrix, $D = 2(D_1 - D_2)$ with $D_1$ a diagonal matrix and $D_2$ a sparse matrix, we see that $Q$ is a sparse matrix. Therefore, this linear system can be solved quickly by using the conjugate gradient method.

### 2.3.2 Spatial Lasso

Analogously, the spatial Lasso algorithm takes the form

$$\hat{B}_{SL} = \arg\min_B \left\{ \sum_{t=1}^{T} \| Y_t - X_t \beta_t \|_2^2 + \lambda_1 \sum_{t=1}^{T} \|\beta_t\|_1 + \lambda_s \sum_{t,k=1}^{T} \omega_{tk} \|\beta_t - \beta_k\|_2^2 \right\}. \quad (8)$$

By ignoring the constancy term that does not affect the solution, we need to minimize the $\ell_1$ penalized quadratic function:

$$B^\top QB - 2V^\top B + \lambda_1 \|B\|_1. \quad (9)$$

The parameter for the $l_1$ penalty controls the sparsity. As it increases, the sparsity of the solution increases. Its choice depends on the sparsity of the true model. If the

true model is not sparse, the parameter should be chosen to be small. Similarly, when the true model is sparse, the large parameter should be used. In practice, the true model is unknown and there are many criteria to select it. But usually, we have some priori information to guide us using $l_1$ penalty, which could be model interpretation needs. In fMRI study, since we believe not all voxels are functioning with all activities, thus sparsity is natural.

One of the most popular methods for solving (9) is in the class of iterative shrinkage-thresholding algorithms (ISTA). A Fast Iterative Shrinkage Thresholding Algorithm (FISTA) with the computational simplicity of ISTA but a significantly better global rate of convergence was proposed in [42]. In order to solve (9) by FISTA, we first define a soft thresholding operator on $\mathbb{R}^{pT}$

$$(prox_{\lambda_1\alpha}(z))_i = \begin{cases} z_i - \lambda_1\alpha, & \text{if } z_i > \lambda_1\alpha \\ 0, & \text{if } |z_i| \leq \lambda_1\alpha \\ z_i + \lambda_1\alpha, & \text{if } z_i < -\lambda_1\alpha \end{cases}$$

with some $\alpha \in (0, \frac{1}{\|Q\|})$. Then the spatial Lasso can be solved by using the following iterating steps:

- $B_k = prox_{\lambda_1\alpha}\left(g_k - \alpha(Qg_k - V)\right)$;
- $a_{k+1} = \frac{1+\sqrt{1+4a_k^2}}{2}$;
- $g_{k+1} = B_k + (\frac{a_k - 1}{a_{k+1}})(B_k - B^{k-1})$,

after given suitable initial values of $B, a$, and $g$.

### 2.3.3 Spatial EN

In Spatial Elastic Net, $\lambda_s \sum_{t=1}^{T} \sum_{s=1}^{T} \omega_{ts} \|\beta_t - \beta_s\|_2^2$ is added based on EN Regression. Because $l_1$ penalty is not differential. So we have to use numerical solution to approximate the results.

Spatial EN solves the problem

$$\hat{B}_{SEN} = \arg\min_{B} \left\{ \sum_{t=1}^{T} \|Y_t - X_t\beta_t\|_2^2 + \lambda_1 \sum_{t=1}^{T} \|\beta_t\|_1 + \lambda_2 \sum_{t=1}^{T} \|\beta_t\|_2^2 + \lambda_s \sum_{t,k=1}^{T} \omega_{tk} \|\beta_t - \beta_k\|_2^2 \right\}. \tag{10}$$

By ignoring the constant, we need to minimize

$$B^\top (Q + \lambda_2 I) B - 2V^\top B + \lambda_1 \|B\|_1.$$

The solution to this problem can be obtained by the same procedure as spatial Lasso, except we need to replace $Q$ with $Q + \lambda_2 I$.

## 2.4 Results

We illustrate the power of spatial MTL algorithms by simulation and their application to real fMRI data sets. The performance is compared with the single task learning (STL) method and the regularized MTL algorithm proposed in [26]. All the parameters used in this section are selected by 5 fold cross validation. For the spatial MTL algorithm, there are two or three parameters. An extensive but computationally expensive way to cross validate the parameter is by grid search. To speed up the computation, we adopt a simpler way. We first select the non-spatial parameter (e.g. $\lambda_2$ in spatial ridge or $\lambda_1$ in spatial Lasso) and fix it. Then the spatial parameter $\lambda_s$ is selected. Both steps are done by cross validation.

### 2.4.1 Simulation Data

We first verify the effectiveness of spatial MTL algorithms on simulated data. In this case, since we know the true model, it is easy to compare the performance of different algorithms.

The model and data are generated as follows. We have designed a $10 \times 10$ grid to mimic 100 voxels in a slice of the brain. For each grid, there is an associated input variable and an associated response variable. The array of all input variables mimics the design matrix and the response values mimic the fMRI times series. The response of each grid is computed by the average of input variables associated to the grid itself and its left, right, upper, and lower neighbor grids (if they exist). This gives us 100 tasks in the 100 dimensional input space. For the simulation data, we applied 4-neighborhood structure.

We generate $n = 100$ samples and run spatial MTL algorithms. This process is repeated 20 times and the learning performance is measured by the mean squared error between the estimated model and true model. We compare our algorithms with the STL learning algorithms and the regularized MTL algorithm proposed in [26]. The mean squared error (MSE) and the standard deviation (SD) of these algorithms are reported in Table 2.4.1. It is clear that the MTL is superior to STL. The tasks are related but do not share a common component. The regularized MTL method in [26] is suboptimal. The spatial regularization helps to improve the performance significantly. Since the true model is rather sparse and there is no grouping effect, spatial Lasso performs the best.

Table 1: Mean Squared Error on Simulated Data

| STL Algorithm | MSE (SD) | MTL Algorithm | MSE (SD) |
|---|---|---|---|
| Ridge | 0.1592 (0.0152) | Spatial Ridge | 0.0489 (0.0014) |
| Lasso | 0.1029 (0.0055) | Spatial Lasso | 0.0426 (0.0009) |
| EN | 0.0498 (0.0010) | Spatial EN | 0.0445 (0.0009) |
| | | RMTL in [26] | 0.0742 (0.0010) |

### 2.4.2 Real Data

Neuroscientists have shown that attention to visual motion can increase the activation of certain cortical areas. Decreased or increased activation of specific brain area would lead to the notion that attention is associated with neuronal activity. This study helps us understand the brain functional connectivity. In this dissertation we applied the spatial MTL algorithm to the Attention to Visual motion fMRI data set, which is available on the SPM web site `http://www.fil.ion.ucl.ac.uk/spm/data/attention/`. This dataset was collected by Christian Büchel [43] for a study of finding the brain functional connectivity with visual attention. There are four conditions: F, 'fixation', A, 'attention', N, 'no attention', and S, 'stationary' condition. During the 'no attention' and 'attention' conditions, two hundred and fifty white dots were moving radially from a fixation point towards the border of the screen [43]. During the 'fixation' condition, only the fixation mark was visible. The brain is split into 46 slices, with each slice containing 53×63 voxels. For each voxel, data is collected at 360 time points. Thus, the dimension of the whole fMRI dataset is $53 \times 63 \times 46 \times 360$. We obtained the fMRI data for 2 slices of the brain.

For the real fMRI data analysis, we do not know the true model. We adopt the cross validation error to evaluate the performance of different algorithms. Cross validation error is an unbiased estimator of the mean squared prediction error. Small

cross validation error usually leads to small prediction error and thus is a relatively reliable metric to compare regression algorithms.

In this real data set, there are 4 contiguous blocked image sets: (0016-0105), (0116-0205), (0316-0405), (0416-0505). Each block has 90 time points, so there are 360 data points in the time series. It is natural to use 4 fold cross validation, considering the special property of the fMRI data. Both 4-neighborhood and 8-neighborhood structures are shown in Figure 8. Though no significant difference was noticed between these two structures for the cross validation error result in this real data analysis, it's possible to have a better contrast in other applications. The time complexity increase for 8-neighborhood is minimal because only the sparsity of $D_2$ is slightly increased. All results for real data given here are based on 8-neighborhood structure.

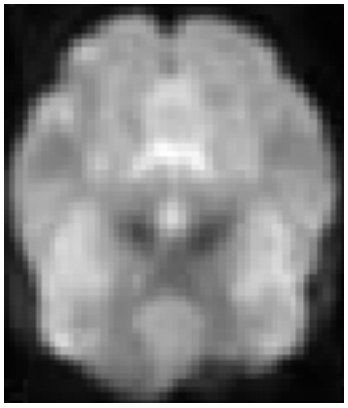
Applying the single task learning and multiple task learning algorithms to the two slices of fMRI data, the cross validation errors are compared in Table 2. On one hand we see spatial MTL slightly improves the result. This indicates that the spatial information does help in the multiple task learning process. On the other hand, we see the improvement is very small. A possible explanation is that, since the design matrix in this study is very simple, the signal is very clear and easy to detect. At the same time, because the noise level is high, the prediction error cannot decrease significantly even if the spatial regularization helps to improve the model estimation.

In this dissertation, we have run 2 slices of the whole brain: slice16, and slice20. Figure 9 (a) and Figure 10 (a) show the functional EPI image for slice 16 and slice 20. Figure 9 (b) - (g) shows the active area of the brain (slice 16) under attention condition by using the estimated $\hat{\beta}$ learned from both STL and MTL algorithms. Figure 10 (b) - (g) shows the active area of the brain (slice 20) under attention condition. The activity of voxels is indicated by the $\hat{\beta}$ values in the regression model – the larger and

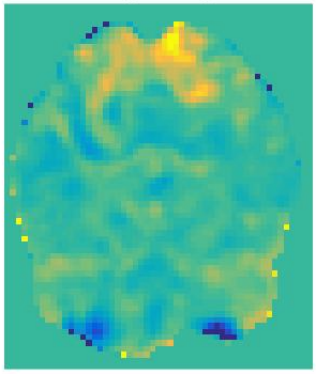
Table 2: The Cross Validation Error of Regression Algorithms on the fMRI Data.

| | Algorithm | Slice 16 | Slice 20 |
|---|---|---|---|
| STL | Ridge | 43.5825 | 33.9385 |
| | Lasso | 43.2791 | 33.7631 |
| | EN | 43.3004 | 33.7600 |
| MTL | Spatial Ridge | 43.1141 | 33.8207 |
| | Spatial Lasso | 43.0957 | 33.7068 |
| | Spatial EN | 43.1211 | 33.7066 |
| | RMTL in [26] | 43.2320 | 33.9384 |

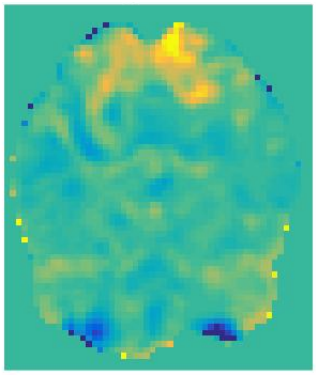
more positive the values, the more active the voxels are. With the naked eye, it is hard to see the difference between the six algorithms. But the numerical values of the $\hat{\beta}$ coefficients do have some small differences. Since the spatial MTL algorithms provide slightly better cross validation error, it is reasonable to assume the active area detection by spatial MTL algorithms is more accurate.



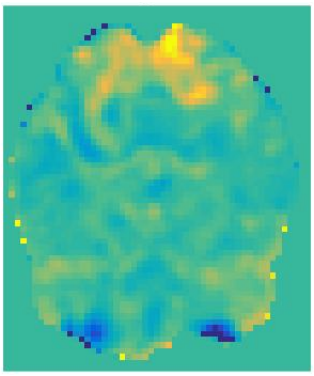
(a) Anatomy Slice 16



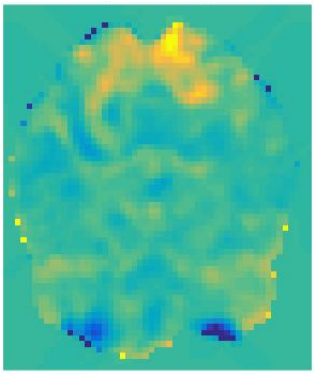
(b) Single Ridge



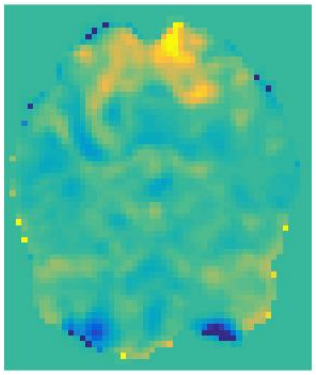
(c) Single Lasso



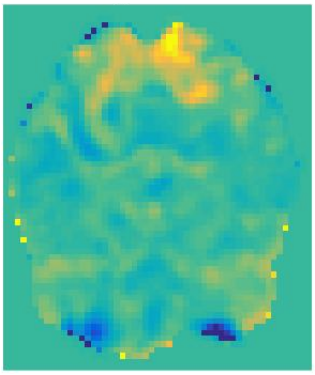
(d) Single EN



(e) Spatial Ridge

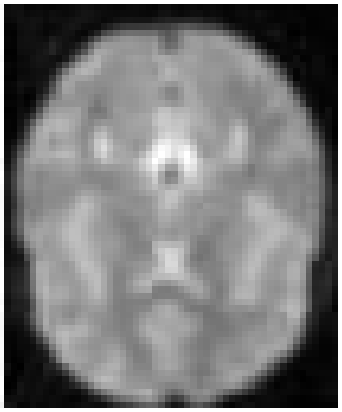


(f) Spatial Lasso

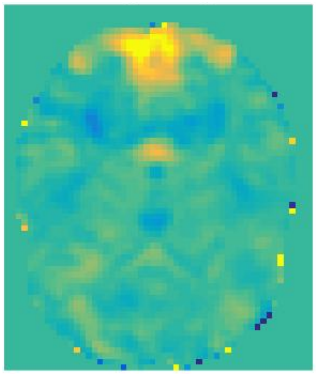


(g) Spatial EN

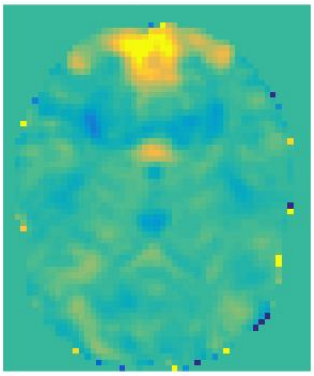
Figure 9: Attention Activation of Slice 16



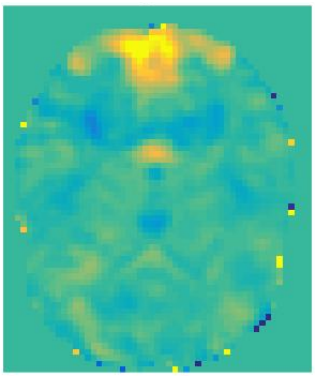
(a) Anatomy Slice 20



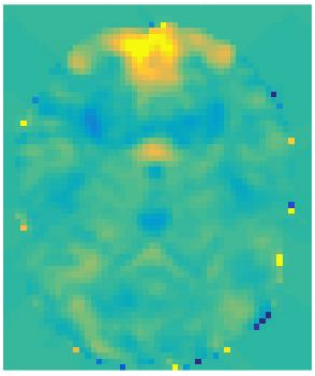
(b) Single Ridge



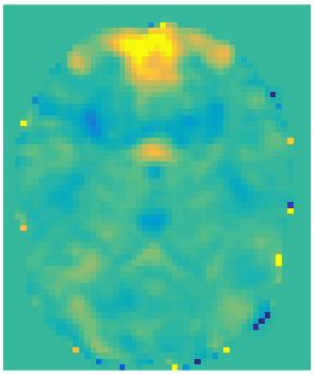
(c) Single Lasso



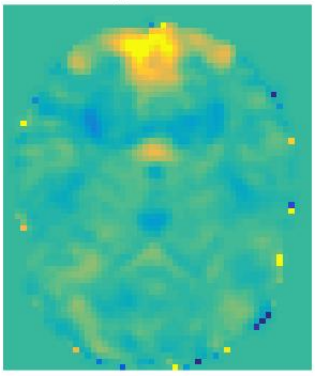
(d) Single EN



(e) Spatial Ridge



(f) Spatial Lasso



(g) Spatial EN

Figure 10: Attention Activation of Slice 20

## 2.5 Conclusion

Motivated by the fMRI data analysis where spatial information is available between voxels, we proposed a class of spatial multiple task learning algorithms for regression. In these methods, we assume the spatially adjacent regression tasks are close. This leads to a natural spatial regularization approach to code the spatial information by using a user-defined neighborhood system. The spatial regularization multiple task learning is shown to be effective in simulated data and real data analysis.

The spatial regularization approach is not necessarily limited to fMRI data analysis. Instead, it may potentially be useful in many fields where spatial information is available, for instance, in environment data from multiple geographical sites. For multiple task learning where no spatial information is available, if soft clustering structure or neighborhood systems could be defined, spatial regularization formulation may also be used, although the regularization term does not code spatial information in this situation. Thus, it would be interesting to further investigate the application domains of spatial regularization in future research. In fMRI data analysis, the real challenges are related to the direction and application of the study. In this chapter, we only developed MTL approaches for analyzing brain activity with visual attention based on 2-dimensional spatial information. MTL scheme(s) using 3-dimensional spacial information and tasks associated with more general regions of interest (ROIs) of the study could be considered.

# CHAPTER 3

# SPATIAL REGULARIZED NEURAL NETWORK AND APPLICATION IN ALZHEIMER'S DISEASE CLASSIFICATION

## 3.1 Introduction

Artificial neural network (ANN) have been around since the 1940s [44]. An artificial neural network is similar to a biological neural network by performing all unites collectively. The term "neural network" usually refers to models in statistics and artificial intelligence.

The idea of a neural network was first introduced by Warren McCulloch and Walter Pitts [44]. However, they did not know how to train the neural network at that time. In 1985, Rumelhart, Hinton, and Williams proposed a training algorithm called backpropagation [45]. However, backpropagation is computationally slow. The hardware power between the 1980s and early 1990s could not effectively train neural networks. In fact, the benefits of neural networks have not been recognized until recent advancements in computer computation power.

The neural network model is particularly useful in applications where the complexity of the data is high. Therefore, it is a powerful tool in many practical applications, such as aerospace, electronics, robotics and so on [46]. A neural network is not only good at fitting non-linear functions, but also efficient in recognizing patterns. Hence neural networks have been widely applied in the fields of speech and image recognition [47]. In [48], the neural network model is used to perform word recognition. In [49], it is applied in a speech synthesizer. The research for face detection applications can be found in [50, 51, 52, 53]. In addition, neural networks have also been widely

applied in disease diagnosis. In [54], an artificial neural network is used to classify brain cancer to specific diagnostic categories based on their gene expression signatures. The diagnosis of Hepatic Fibrosis is also evaluated using a neural network in [55] by considering the regions of interests (ROI) from the liver MRI images as input features. Neural networks have also been used for early detection of Alzheimer's disease in [56]. The EGG data has been used as inputs to NN model, the single unit output indicates whether the subject is AD or normal. Recently, deep learning has emerged as a relatively new advanced technique in neural network studies. In 2012, Alex Krizhevsky proposed a deep learning algorithm: deep convolutional neural networks in [57]. They introduce convolution to image recognition networks, and achieved better, more accurate results than previous state-of-the-art results. Since then, convolution neural networks have been widely applied to fields with large scale data such as video classification and human action classification [58, 59, 60].

Alzheimer's disease (AD) is a progressive, irreversible, degenerative brain disorder, causing impaired memory, thinking and behavior. The symptoms of AD usually develop slowly and get worse and worse over time. In the absence of AD, the human brain often can live up to the age of 100 and beyond [61]. The research to uncover the mysteries of human brain will help us to find the new ways to treat, prevent and cure AD. In past decades, MRI has been widely used as an image aid tool for clinical disease diagnosis. In consequently, many researchers have studied Alzheimer's disease by using MRI brain image data. A lot of research has been done by using structural MRI brain images for classification between AD and healthy controls. According to the features being extracted from the structural MRI, the existing classification methods can be roughly divided into three categories [62] : voxel level, cortical thickness, and hippocampus.

For voxel levels, some researchers use whole brain gray matter as the input features directly, then Support Vector Machine (SVM) is used for the classification. SVM successfully separate patients with AD from healthy aging subjects in [61]. Some other researchers prefer to reduce the dimensionality of the input features first, for example in [63], only 90 regions of interested (ROI) extracted from the brain are set as input features. In addition to SVM, the unsupervised method such as PCA, ICA have also been applied to AD classification[64].

In addition to voxel levels, there has also been a considerable research focus on cortical thickness. The cortical thickness represents a direct index of atrophy, which is properly known to be affected in Alzheimer's disease. For the cortical thickness method, the cortical surface with a lot of vertices should be generated from the MRI image first. Then the cortical structures segmented from the surface will be set as input features. The work includes [65, 66].

Unlike the voxel level considering the whole brain, the cortical thickness considers the whole cortical surface. The third category method only considers the hippocampus of the human brain, research includes [67, 68, 69]. The hippocampus is located in the medial temporal lobe of the brain, which is underneath the cortical surface. In Alzheimer's disease, the hippocampus is believed to be one of the first regions of the brain to become damaged. This leads to memory loss and disorientation associated with the condition.

By considering the brain spatial information, we use voxel level method in this paper. Related to the problem of AD detection using MRI data, the brain voxels are spatially related. It is natural to code the 3-dimensional spatial information in the training process to improve the learning performance. The idea of coding spatial information in the regularization theory contexts is proposed in our previous chapter, and

has been successfully applied in fMRI data [70]. Since computer hardware power has been improved, neural networks have been widely applied in different fields. However, to the best of our knowledge, the idea of coding spatial information for AD classification in the regularization neural network context is still new. Consequently, this current condition has led us to propose a spatial regularization approach for neural network and apply it to structural MRI Alzheimer's disease classification. We tested our proposed model on two types of data. For the first type, the whole brain grey matter voxels are considered as input features. For the second type, we extracted 5 regions of interest (ROI) from the whole brain by using SPM MNI template, and these 5 ROI are considered as input features. The 5 regions we selected are: Hippocampus, Amygdala, Temporal Lobe, Frontal Lobe, and Parietal Lobe. These regions are chosen based on the current publications for AD study. The AD subjects have shown atrophy patterns in these 5 regions [71, 72, 73, 74, 68, 75, 76, 77, 78]. Real application results show satisfactory performance from spatial regularization neural network for both whole-brain data and ROI data.

The remainder of this chapter is organized as follows. In section 3.2.1, the neural network model is described and we also show how to solve the model. The regularized neural network model is explained in section 3.2.2. We develop our spatial regularized neural network model in Section 3.2.3. In Section 3.3, the three models are tested on real MRI AD data. We finish with concluding remarks in Section 3.4.

## 3.2 Methods

### 3.2.1 Neural Network

A typical neural network is shown in Figure 11. It has at least 3 layers: an input layer, a hidden layer, and an output layer. The neural network accepts an input and returns an output. The hidden layer helps the neural network understand the input data, and form the output.

A primary question here is how many hidden layers should be used. In [79], Hornik and Kurt have proved that a single hidden layer can function as a universal approximator. In other words, a single hidden layer neural network should be able to approximate any output from any input as long as it has enough neurons in a single layer [44]. However, the neural network with a larger number of hidden layers could facilitate a more complex model at the cost of expensive computation.

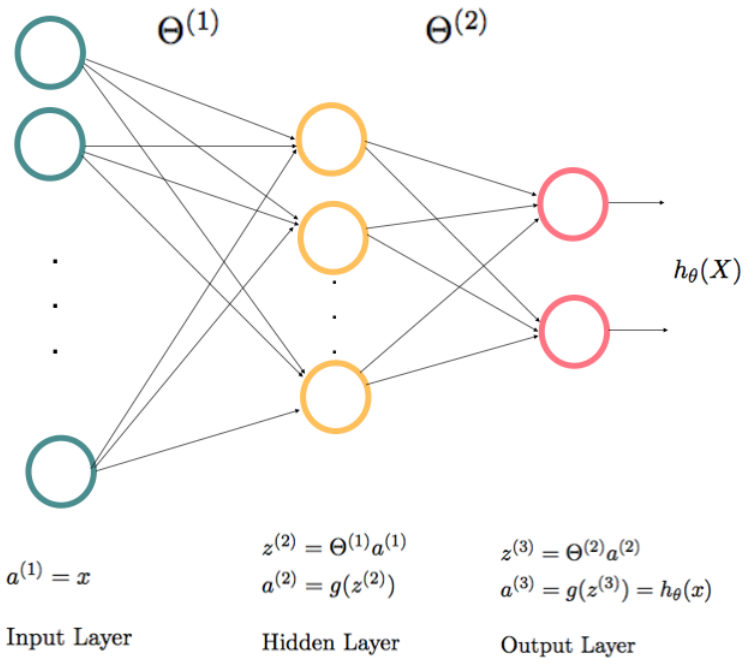



Figure 11: Neural Network Model

In a neural network, activation functions are needed to establish bounds for the output [44]. There are many activation functions available for a neural network. The two most common activation functions are step function and sigmoid function. In this dissertation, we use sigmoid function: $g(z) = sigmoid(z) = \frac{1}{1+e^{-z}}$.

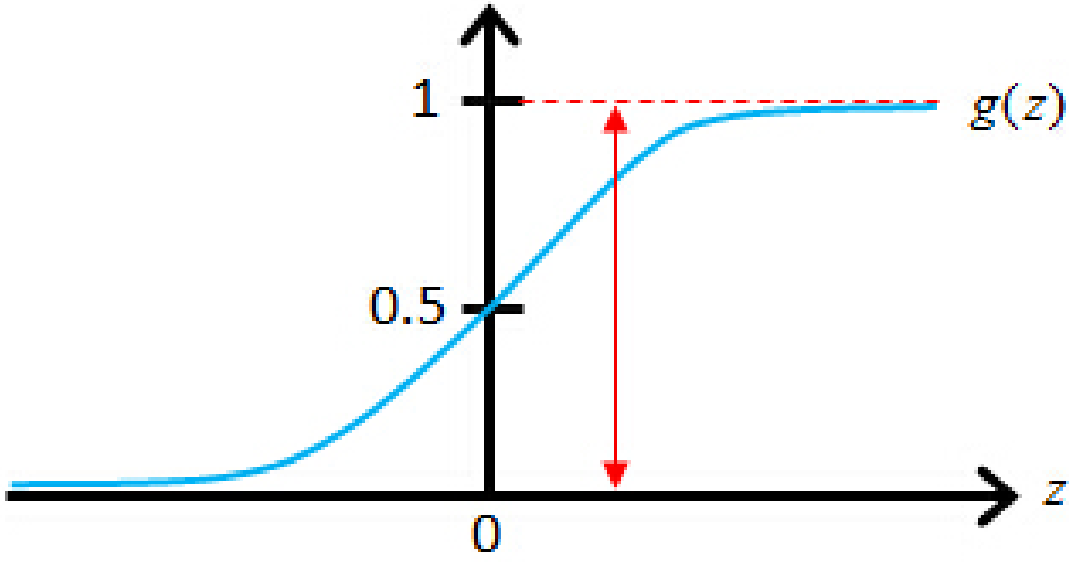



Figure 12: Sigmoid Function

For a classification problem, we could use a neural network to classify the input into one or more classes. When a neural network has to choose between two options like true or false, this is a binary classification. A binary classification needs a single output neural network to do classification. In other words, we could classify the input into two categories by using the single output. If we want to predict more than two categories, we need more than two outputs. This is called Multi-Class classification. In this dissertation, we use binary classification, because we only have two categories: Alzheimer disease(AD) and Control subject(CS). Typically, we use a log loss function to evaluate a neural network.

The objective function for the binary classification is:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\Bigg[ - y^{(i)}log(h_\theta(x^{(i)})) - (1-y^{(i)})log(1-h_\theta(x^{(i)}))\Bigg]. \tag{11}$$

where $X = \Bigg[(x^{(1)}), (x^{(2)}), ..., (x^{(m)})\Bigg] \in \mathbb{R}^{n\times m}$ , $y = [1, 1, ..., 0]^\top \in \mathbb{R}^m$ . X is the input data with m training data, $x^{(i)} \in \mathbb{R}^n$ is a column vector, $y^{(i)} \in \mathbb{R}$, $(i = 1, ..., m)$. The column vector $y$ is the known labels for X. In this paper, AD is labeled as '1', CS is labeled as '0'. $h_\theta(x^{(i)})$ $(i = 1, ..., m)$ is neural network's classification prediction for $i_{th}$ subject calculated in Figure 11.

$$\begin{cases} y^{(i)} = 1, & \text{if } h_\theta(x^{(i)}) \geq 0.5. \\ y^{(i)} = 0, & \text{if } h_\theta(x^{(i)}) < 0. \end{cases}$$

To train the neural network, we implement the classic back propagation, which is the most common method to train a neural network.

Back propagation was introduced by Rumelhart, Hinton, and Williams in [45]. The principle of back propagation is based on gradient descent. Gradient descent refers to calculating an individual gradient for each node in the neural network for each sample. The error function calculates the difference between the expected output and actual output of the neural network [44]. Based on these derivatives, the training algorithm will decide whether the weights of the node should be increased or decreased. In other words, back propagation optimized individual weights with derivatives. In turn, this optimization will decrease the total error of the neural network. The specific steps are described as following:

**1.** Set the network architecture.

- Set the the number of hidden layers, number of neurons in each layer, and activation function.
- Input layer size: $n$
- Hidden layer size:$h_1$
- Num of labels: $h_2$

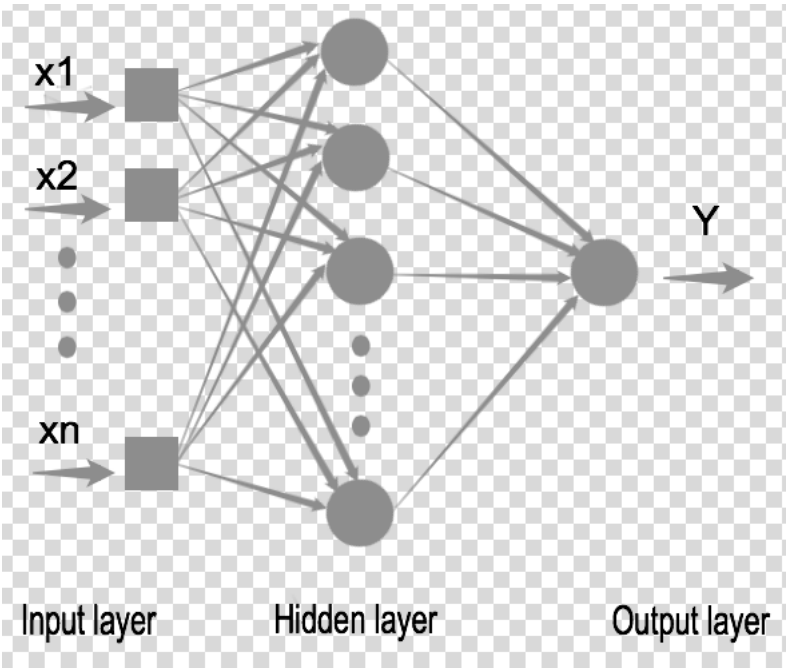



Figure 13: Neural Network Model

**2.** Initialize weights randomly.

- Initialize each $\Theta_{ij}^{(l)}$ to a random value in $[-\epsilon, \epsilon]$
- $\Theta^{(1)} = \text{rand}(h_1,\, n) \times (\, 2\epsilon\text{\_init} - \epsilon\text{\_init}\,)$
- $\Theta^{(2)} = \text{rand}(h_2,\, h_1) \times (\, 2\epsilon\text{\_init} - \epsilon\text{\_init}\,)$

**3.** Forward propagation.

Use forward propagation to determine the output node given the training data (X,y):

- Input layer: $a^{(1)} = X$
- Hidden layer: $z^{(2)} = \Theta^{(1)}a^{(1)}$, $a^{(2)} = g(z^{(2)})$
- Output layer: $z^{(3)} = \Theta^{(2)}a^{(2)}$, $a^{(3)} = g(z^{(3)}) = h_\theta(X)$

**4.** Calculate gradient.

We need to calculate:

$$\frac{\partial J(\Theta)}{\partial \Theta_{ij}^{(l)}} = \frac{\partial J(\Theta)}{\partial z^{(l+1)}} \frac{\partial z^{(l+1)}}{\partial \Theta_{ij}^{(l)}}. \tag{12}$$

We need to compute the error $\delta_j^{(l)}$ for each node $j$ in layer $l$, this error term measures the difference between the actual output of this node and the expected output. In other words, how much error this node is responsible for in this output. To calculate the error, we need to start with the output layer and work backwards through the neural network. Because we need to propagate the errors backwards through the neural network. We don't need to calculate the error term for input layer, because the gradient calculation does not need this.

First let's define:

$$\delta^{(l)} = \frac{\partial J(\Theta)}{\partial z^{(l)}}. \tag{13}$$

For output layer 3, we need to calculate error term $\delta^{(3)}$:

$$\delta^{(3)} = \frac{\partial J(\Theta)}{\partial z^{(3)}} = (a^{(3)})^T - y. \tag{14}$$

For hidden layer 2, we need to calculate error term $\delta^{(2)}$:

$$\begin{aligned}\delta^{(2)} &= \frac{\partial J(\Theta)}{\partial z^{(2)}} = \frac{\partial J(\Theta)}{\partial z^{(3)}}\frac{\partial z^{(3)}}{\partial z^{(2)}} \\ &= {\Theta^{(2)}}^T \delta^{(3)} .* g\prime(z^{(2)}).\end{aligned} \tag{15}$$

After replace equation 13 into 12, we obtain the gradient for the objective function:

$$\frac{\partial J(\Theta)}{\partial \Theta^{(l)}} = \frac{1}{m}\delta^{(l+1)}(a^{(l)})^T. \tag{16}$$

**5.** Update the weights to minimize cost function:

$$\Theta_{ij}^{(l)} = \Theta_{ij}^{(l)} + \Delta\Theta_{ij}^{(l)} = \Theta_{ij}^{(l)} - \alpha\frac{\partial J(\Theta)}{\partial \Theta_{ij}^{(l)}}.$$

Since we obtain the gradient, now we could train the neural network by minimizing the objective function $J(\Theta)$.

### 3.2.2 Regularized Neural Network

Although a neural network is a very powerful technique for the classification, it is more prone to overfitting when there are a huge number of features. For example, for the data used in this dissertation, the dimension of the input feature is $79 \times 95 \times 79 = 592895$, because each subject brain has $79 \times 95 \times 79 = 592895$ voxels. In order to avoid overfitting, we could implement the regularization approach. There are two common regularization techniques to reduce overfitting: $l_1$ and $l_2$ regularization [80]. Both regularization techniques will add a weight penalty to the neural network, but they calculate this penalty differently. $l_1$ regularization is used to create sparsity in the neural network. The $l_2$ regularization is utilized to create lower weight values in

the neural network. The lower weight values typically lead to less overfitting [44]. In this dissertation, we implement $l_2$ regularization to reduce overfitting.

The objective function of $l_2$ regularization neural network for binary classification is:

$$J_\lambda(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^{(i)}log(h_\theta(x^{(i)})) - (1-y^{(i)})log(1-h_\theta(x^{(i)}))\right] + \frac{\lambda_2}{2m}\left[\sum_{j=1}^{h_1}\sum_{k=1}^{n}(\theta_{j,k}^{(1)})^2 + \sum_{j=1}^{h_2}\sum_{k=1}^{h_1}(\theta_{j,k}^{(2)})^2\right], \tag{17}$$

where $n$ is the total number of input layer features, $h_1$ is the total number of the hidden layer nodes, and $h_2$ is the total number of the output layer nodes. $\lambda_2$ is the parameter to control the importance the $l_2$ penalty. $\lambda_2 = 0$ means $l_2$ regularization is not considered at all.

Since a $l_2$ penalty is added in regularization neural network, derivatives of the regularization term for the gradient should be added:

$$\frac{\partial J_\lambda(\Theta)}{\partial \Theta^{(l)}} = \frac{1}{m}\delta^{(l+1)}(a^{(l)})^T + \frac{\lambda}{m}\Theta^{(l)}. \tag{18}$$

In this dissertation, we add bias for the input layer and hidden layer weight. The bias has a constant value of 1, it is not connected to the previous layer. We don't add regularization for the bias. Since we have obtained the gradient, now we could train the regularization neural network by minimizing the objective function $J_\lambda(\Theta)$.

### 3.2.3 Spatial Regularized Neural Network

In the problem of Alzheimer's disease classification using structural MRI data, the brain voxels are spatially related. In other words, if one voxel is abnormal, its neighbors are more likely to be abnormal, and vice versa. So it is natural to code the

spatial information into the neural network training process to improve the learning performance.

The idea of coding spatial information in the regularization theory contexts is proposed in our paper [70], and has been successfully applied in fMRI data analysis. However, we only consider the 2-dimension spatial information in [70], because we only studied one slice of the brain which is 2-dimensional. In this chapter, our AD data is the whole 3-dimensional brain, so we need to consider the 3-dimension spatial information.

To code the spatial information, we need to define a neighborhood system first. We define the voxel similarity coefficient by :

$$w_{kt} = \begin{cases} 1, & \text{if voxel } k \text{ is a neighborhood of voxel } t. \\ 0, & \text{if voxel } k \text{ is not a neighborhood of voxel } t. \end{cases}$$

Figure 14 shows three different 3-dimensional neighborhood structures: (a) 6 neighborhood, (b) 18 neighborhood, (c) 26 neighborhood. In this dissertation, we consider 26 neighbors for each voxel. If the voxel has no such neighbors, the coefficient is set to 0.

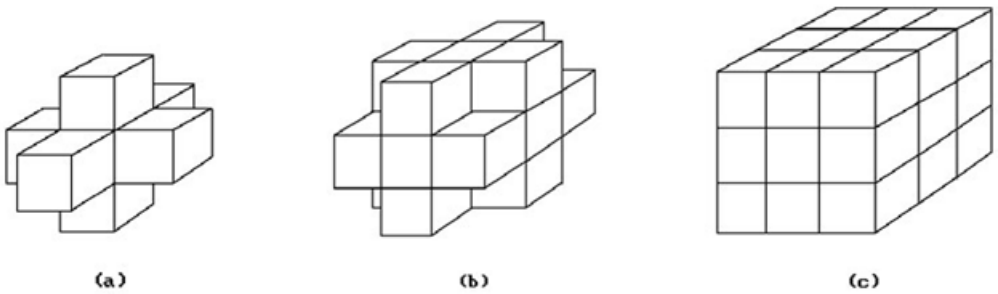



Figure 14: Neighborhood Structure for Each Voxel

The objective function of spatial regularization neural network for the binary

classification is:

$$
\begin{aligned}
J_{S\lambda}(\theta) = \ & \frac{1}{m}\sum_{i=1}^{m}\Bigg[-y^{(i)}log(h_\theta(x^{(i)})) - (1-y^{(i)})log(1-h_\theta(x^{(i)}))\Bigg] \\
& + \frac{\lambda_2}{2m}\Bigg[\sum_{j=1}^{h_1}\sum_{k=1}^{n}(\theta_{j,k}^{(1)})^2 + \sum_{j=1}^{h_2}\sum_{k=1}^{h_1}(\theta_{j,k}^{(2)})^2\Bigg] \\
& + \frac{\lambda_s}{2m}\sum_{j=1}^{h_1}\sum_{k=1}^{n}\sum_{t=1}^{n}\omega_{kt}(\theta_{j,k}^{(1)} - \theta_{j,t}^{(1)})^2,
\end{aligned}
\tag{19}
$$

where $\lambda_s$ is the spatial parameter to control the spatial penalty. $\lambda_s = 0$ means the spatial penalty is not considered at all and each voxel is treated independently. When $\lambda_s$ is increased, neighbors are more close to each other.

In the previous section, we added regularization for both input layer weights $\Theta^{(1)}$ and hidden layer weights $\Theta^{(2)}$. However, we only added spatial regularization for the input layer $\Theta^{(1)}$ in this section, because we know that the input layer has the 3-dimensional information of the whole brain. The spatial information from the input layer will help the neural network to improve the learning performance. However, we don't know the information for the hidden layer, because the hidden layer is a black box.

Because a spatial penalty is added in a spatial regularization neural network, we should add the derivatives of the spatial regularization term for the gradient:

$$
\begin{aligned}
\frac{\partial J_{S\lambda}(\Theta)}{\partial \Theta^{(2)}} &= \frac{1}{m}\delta^{(3)}(a^{(2)})^T + \frac{\lambda_2}{m}\Theta^{(2)} \\
\frac{\partial J_{S\lambda}(\Theta)}{\partial \Theta^{(1)}} &= \frac{1}{m}\delta^{(2)}(a^{(1)})^T + \frac{\lambda_2}{m}\Theta^{(1)} + \frac{\lambda_s}{m}(B_1 - B_2),
\end{aligned}
\tag{20}
$$

where

$$B_1 = [\sum_{t=1}^{n} \omega_{1t}\theta_1^{(1)}, \sum_{t=1}^{n} \omega_{2t}\theta_2^{(1)}, ..., \sum_{t=1}^{n} \omega_{nt}\theta_n^{(1)}],$$
$$B_2 = [\sum_{t=1}^{n} \omega_{1t}\theta_t^{(1)}, \sum_{t=1}^{n} \omega_{2t}\theta_t^{(1)}, ..., \sum_{t=1}^{n} \omega_{nt}\theta_t^{(1)}],$$
$$\Theta^{(1)} = [\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_n^{(1)}].$$

$\theta_i^{(1)}$ is a column vector with dimension of $h_1$, $i = 1, 2, ..., n$. Both $B_1$ and $B_2$ are matrix with dimension of $h_1 \times n$. We don't add regularization for the bias. Since we have obtained the gradient, now we can train the spatial regularization neural network by minimizing the objective function $J_{S\lambda}(\Theta)$.

## 3.3 Results

We have evaluated the performance of spatial regularization neural network with the application on real MRI data. The performance is compared with neural network (NN), regularized neural network (RNN), and spatial regularization neural network (SRNN) proposed in section 3.2.3. For the real MRI data, we obtained 80 subjects in total: 42 Alzheimer's disease (AD) and 38 Control subject (CS). All data can be downloaded from the Alzheimer's disease Neuroimaging Initiative (ADNI) website `http://adni.loni.usc.edu`. All data is normalized by using SPM12 (`http://www.fil.ion.ucl.ac.uk/spm/software/spm12/`).

We tested all three models in section 3.2 on two types of data. The first type data is based on the whole brain gray matter shown in Figure 16. The whole brain gray matter is considered as input features.

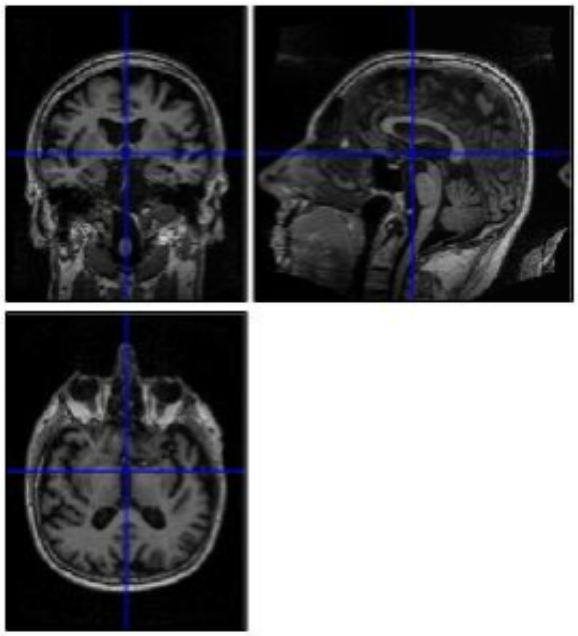


Figure 15: Whole Brain

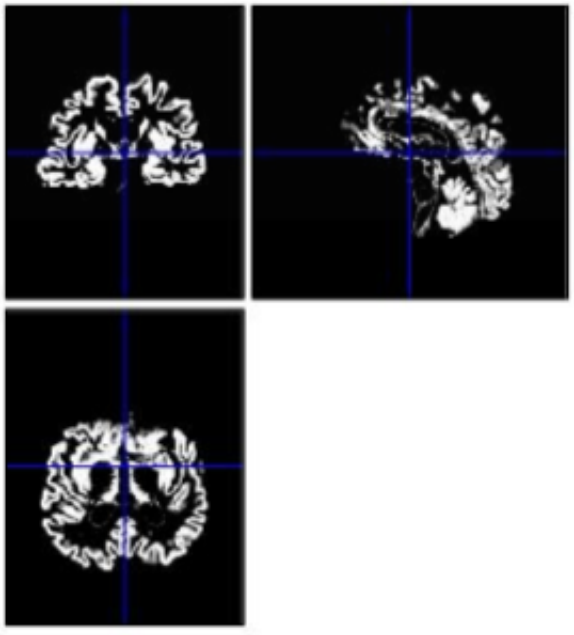


Figure 16: Whole Brain Gray Matter

The second type of data is based on ROI data shown in Figure 17. For the second type, we extracted 5 regions of interest (ROI) from the whole brain by using SPM MNI template, and these 5 ROI are considered as input features. The 5 regions we selected are: Hippocampus, Amygdala, Temporal Lobe, Frontal Lobe and Parietal Lobe. The AD subjects have shown atrophy patterns in these 5 regions [71, 72, 73, 74, 68, 75, 76, 77, 78].

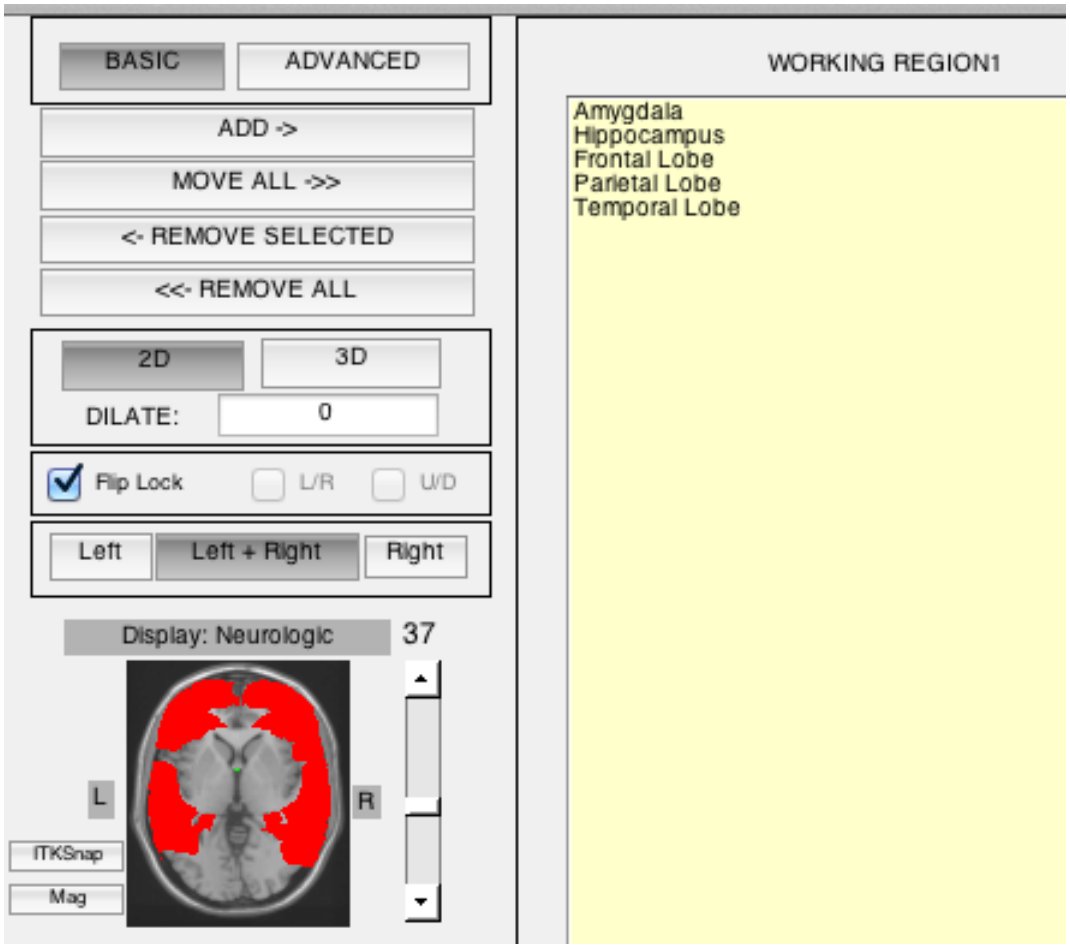



Figure 17: Regions of Interested

The classification results for whole brain grey matter is in Table 3 and the classification results for ROI is in Table 4. All the parameters for the algorithms are selected by 5 fold cross validation. For SRNN, we first select the non-spatial parameter $\lambda_2$ for

regularization neural network and fix it. Then the spatial parameter $\lambda_s$ is selected. Both steps are done by cross validation. The learning performance is measured by cross validation training classification accuracy, testing classification accuracy, testing sensitivity, and testing specificity.

Table 3: Classification Accuracy for Whole Brain Grey Matter

| Methods | Training Accuracy | Testing Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|
| NN | 78.58 % | 62.62 % | 81.67 % | 37.22 % |
| RNN | 92.42 % | 72.38 % | 85.83 % | 54.44 % |
| SRNN | 99.67 % | 73.81 % | 90.00 % | 49.44 % |

Table 4: Classification Accuracy for Regions of Interested (ROI)

| Methods | Training Accuracy | Testing Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|
| ROI_NN | 89.58 % | 65.50 % | 65.50 % | 65.50 % |
| ROI_RNN | 100 % | 93.12 % | 87.25 % | 99.00 % |
| ROI_SRNN | 100 % | 94.63 % | 91.00 % | 98.44 % |

From Table 3 and Table 4, it is clear that RNN increases the classification accuracy significantly compared to NN, because the regularization technique reduces the overfitting problem. Moreover, SRNN is superior to RNN, which indicates that spatial information further improves the learning performance. In addition, we can see that SRNN has the best testing sensitivity among all three methods. Although the testing specificity for SRNN is not as good as RNN, it is still better than NN. For the disease diagnosis, sensitivity is usually more important than specificity. These results demonstrate that our proposed spatial penalty not only helps increase the classification accuracy in the neural network, but also keeps the tradeoff between sensitivity and specificity.

## 3.4 Conclusion

As computer computation power has improved in recent years, neural networks have been widely applied in Alzheimer's disease diagnosis. However, the first layer of this method is based on an individual brain voxel, which means a neural network learns each voxel individually without considering the brain spatial information. Motivated by MRI data analysis where spatial information is available between voxels, we proposed a spatial regularization neural network. In this method, we assumed the spatially adjacent voxels are close. This leads to a natural spatial regularization approach to code the spatial information by using a user defined 3-dimensional neighborhood system. Real application results show that spatial regularization neural network increases the classification accuracy.

However, in this dissertation, we only used a single hidden layer for the neural network by considering the computational cost. Although researchers have proved that a single hidden layer can function as a universal approximator, brain data has more complex patterns than we think. The neural network with a larger number of hidden layers could be considered to facilitate a more complex model.

# CHAPTER 4

# SUMMARY AND FUTURE WORK

## 4.1 Summary

In this dissertation, we focused on the study of machine learning techniques for high-dimensional neuroimaging data. Two kinds of neuroimaging data have been studied: functional MRI data and MRI data. fMRI data is used to gain insight into brain activities of vision motion. MRI data is used to study the Alzheimer's disease classification.

In Chapter 2, we used a General Linear Model to formulate the fMRI data, where each voxel is assumed as a task, and proposed a spatial regularization approach for multitask learning which incorporates spatial information provided by each task's neighborhood. In the problem of active region detection using fMRI data, the tasks (brain voxels) are spatially related. It is natural to code the spatial information into the training process to improve the learning performance. The contribution for our proposed algorithm is that we define a new spatial penalty term for spatial regularization provided by each task's neighborhood using multitask learning (MTL). A class of spatial multitask learning models: MTLRidge, MTLLasso, MTLEN was proposed. Simulation and real application results show satisfactory performance from spatial multitask learning algorithms.

Since computer processing power has improved in recent years, neural networks have been widely applied in different fields. However, to the best of our knowledge, the idea of coding spatial information for AD classification in the regularization neural network context is still new. In Chapter 3, we proposed a spatial regularization

approach for a neural network and applied it to structural MRI Alzheimer's disease classification. Instead of considering 2-dimensional spatial information, we considered 3-dimensional spatial information for SRNN. We tested our proposed model on both whole-brain gray matter data and ROI data. Real application results show that our proposed spatial regularization neural network increased the classification accuracy.

## 4.2 Future Work

In Chapter 2, based on the fMRI data analysis where spatial information is available between voxels, we proposed a class of spatial multiple task learning algorithms for regression. In this method, we used a universal tuning parameter, $\lambda_s$, to control the spatial information for all voxels. The more optimal way is to select a specific tuning parameter for each individual voxel. Although this is very challenging work due to the high dimensionality of fMRI data, this should further improve the current learning performance.

Relative to the problem of AD detection using MRI data, the brain voxels are spatially related. It is natural to code the 3-dimensional spatial information in the training process to improve the learning performance. In consequently, this discovery has led us to propose a spatial regularization approach for neural network and apply it to structural MRI Alzheimer's disease classification in chapter 3. However, we only used a single hidden layer for the neural network by considering the computational cost. Although researchers have proved that a single hidden layer can function as a universal approximator, the brain data is more complex than original anticipated. The neural network with a larger number of hidden layers could facilitate a more complex model at the cost of expensive computation.

Autism spectrum disorder (ASD) is a neurally based psychiatric disorder [81],

which is characterized by the impaired development of social interaction and communication skills [82]. Although strong genetic factors are suspected [81], ASD continues to be diagnosed using symptom-based clinical criteria [82] and its etiology remains unestablished. One recent topic garnering significant attention is the study of functional connectivity of autism and controls [83, 84]. Further study in this area could provide helpful information in gaining a better understanding of the neuronal pathology of autism in children.

# REFERENCES

[1] S. A. Bunge and I. Kahn, "Cognition: An overview of neuroimaging techniques," *Encyclopedia of Neuroscience*, vol. 2, pp. 1063–1067, 2009.

[2] J. P. Hornak, "The basics of MRI." `http://www.cis.rit.edu/htbooks/mri/index.html`, 2008. [Online].

[3] S. Chen, *New Statistical Techniques for High-dimensional Neuroimaging Data*. PhD thesis, Emory University, 2012.

[4] C. Weegenaar, "Magnetic resonance imaging e-tutorail." `http://www.slideshare.net/rnja8c/fmri-study-design`, 2008. [Online].

[5] K. Aldridge, "3D MRI reconstructions of an adult human brain." `http://web.missouri.edu/~aldridgek/human-brains.shtml`, 2010. [Online].

[6] R. A. Poldrack, J. A. Mumford, and T. E. Nichols, *Handbook of Functional MRI Data Analysis*. Cambridge University Press, 2011.

[7] R. James, "Intro to fMRI." `http://www.slideshare.net/rnja8c/fmri-study-design`, 2011. [Online].

[8] T. D. Wager and M. A. Lindquist, "Principles of fMRI." `https://leanpub.com/principlesoffmri/read`. [Online].

[9] J. Ashburner, G. Barnes, and so on, "SPM12 manual," *Functional Imaging Laboratory Wellcome Trust Centre for Neuroimaging*, pp. 1–449, 2014.

[10] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. Frackowiak, "Statistical parametric maps in functional imaging: a general linear approach," *Human Brain Mapping*, vol. 2, no. 4, pp. 189–210, 1994.

[11] K. J. Worsley, C. Liao, J. Aston, V. Petre, G. Duncan, F. Morales, and A. Evans, "A general statistical analysis for fMRI data," *Neuroimage*, vol. 15, no. 1, pp. 1–15, 2002.

[12] A. Dove, S. Pollmann, T. Schubert, C. J. Wiggins, and D. Y. von Cramon, "Prefrontal cortex activation in task switching: an event-related fMRI study," *Cognitive Brain Research*, vol. 9, no. 1, pp. 103–109, 2000.

[13] K. J. Friston, A. Holmes, J.-B. Poline, C. J. Price, and C. Frith, "Detecting activations in PET and fMRI: levels of inference and power," *Neuroimage*, vol. 4, no. 3, pp. 223–235, 1996.

[14] D. Y. Kimberg, G. K. Aguirre, and M. D'Esposito, "Modulation of task-related neural activity in task-switching: an fMRI study," *Cognitive Brain Research*, vol. 10, no. 1, pp. 189–196, 2000.

[15] H. L. Gallagher, F. Happé, N. Brunswick, P. C. Fletcher, U. Frith, and C. D. Frith, "Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks," *Neuropsychologia*, vol. 38, no. 1, pp. 11–21, 2000.

[16] J. Wan, Z. Zhang, J. Yan, T. Li, B. D. Rao, S. Fang, S. Kim, S. L. Risacher, A. J. Saykin, and L. Shen, "Sparse bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in alzheimer's disease," in

*Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 940–947, IEEE, 2012.

[17] B. Jie, D. Zhang, B. Cheng, and D. Shen, "Manifold regularized multi-task feature selection for multi-modality classification in alzheimer's disease," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, pp. 275–283, Springer, 2013.

[18] X. Chen, J. He, R. Lawrence, and J. G. Carbonell, "Adaptive multi-task sparse learning with an application to fMRI study," in *Proceedings of SIAM International Conference on Data Mining (SDM)*, pp. 212–223, SIAM, 2012.

[19] N. Rao, C. Cox, R. Nowak, and T. T. Rogers, "Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis," in *Advances in Neural Information Processing Systems*, pp. 2202–2210, 2013.

[20] K.-J. Lee, G. L. Jones, B. S. Caffo, and S. S. Bassett, "Spatial bayesian variable selection models on functional magnetic resonance imaging time-series data," *Bayesian Analysis (Online)*, vol. 9, no. 3, pp. 699–732, 2014.

[21] T. J. Sejnowski and C. R. Rosenberg, "Nettalk: A parallel network that learns to read aloud," *Neurocomputing: Foundations of Research*, pp. 661–672, 1988.

[22] T. G. Dietterich, H. Hild, and G. Bakiri, "A comparative study of ID3 and backpropagation for english text-to-speech mapping.," in *Machine Learning*, vol. 18, pp. 51–80, Kluwer Academic Publishers, 1995.

[23] R. A. Caruana, "Multitask connectionist learning," in *Proceedings of the 1993 Connectionist Models Summer School*, pp. 372–379, 1993.

[24] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[25] B. Bakker and T. Heskes, "Task clustering and gating for bayesian multitask learning," *The Journal of Machine Learning Research*, vol. 4, pp. 83–99, 2003.

[26] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 109–117, ACM, 2004.

[27] J. Bi, T. Xiong, S. Yu, M. Dundar, and R. B. Rao, "An improved multi-task learning approach with applications in medical diagnosis," in *Machine Learning and Knowledge Discovery in Databases*, pp. 117–132, Springer, 2008.

[28] A. Agarwal, S. Gerber, and H. Daume, "Learning multiple tasks using manifold regularization," in *Advances in Neural Information Processing Systems*, pp. 46–54, 2010.

[29] J. Zhou, J. Chen, and J. Ye, "Clustered multi-task learning via alternating structure optimization," in *Advances in Neural Information Processing Systems*, pp. 702–710, 2011.

[30] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 814–822, ACM, 2011.

[31] D. Zhang, D. Shen, A. D. N. Initiative, *et al.*, "Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease," *Neuroimage*, vol. 59, no. 2, pp. 895–907, 2012.

[32] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby, "Beyond mind-reading: multi-voxel pattern analysis of fMRI data," *Trends in Cognitive Sciences*, vol. 10, no. 9, pp. 424–430, 2006.

[33] M. Smith and L. Fahrmeir, "Spatial bayesian variable selection with application to functional magnetic resonance imaging," *Journal of the American Statistical Association*, vol. 102, no. 478, pp. 417–431, 2007.

[34] P. Whittle, *Prediction and Regulation by Linear Least-Square Methods*. University of Minnesota Press, 1983.

[35] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[36] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[37] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[38] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, *et al.*, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[39] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[40] R. G. Aykroyd and S. Zimeras, "Inhomogeneous prior models for image reconstruction," *Journal of the American Statistical Association*, vol. 94, no. 447, pp. 934–946, 1999.

[41] L. Xiong and D. Hong, "Incorporating spatial information in IMS data analysis to optimize classification accuracy using Markov Random Field and MCMC method," in *Statistical Analysis of Spectrometry Based Proteomics and Metabolomics Data,* Frontiers in Probability and Statistics Series, pp. xx–xx, Springer, New York, (to appear).

[42] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[43] C. Büchel and K. Friston, "Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI," *Cerebral Cortex*, vol. 7, no. 8, pp. 768–778, 1997.

[44] J. Heaton, *Deep Learning and Neural Networks*. Heaton Research, Inc., 2015.

[45] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," tech. rep., DTIC Document, 1985.

[46] M. T. Hagan, H. B. Demuth, M. H. Beale, and O. De Jesús, *Neural Network Design*, vol. 20. PWS publishing company Boston, 1996.

[47] R. P. Lippmann, "An introduction to computing with neural nets," *ASSP Magazine, IEEE*, vol. 4, no. 2, pp. 4–22, 1987.

[48] K. J. Lang, A. H. Waibel, and G. E. Hinton, "A time-delay neural network architecture for isolated word recognition," *Neural Networks*, vol. 3, no. 1, pp. 23–43, 1990.

[49] S. S. Fels and G. E. Hinton, "Glove-talk: A neural network interface between a data-glove and a speech synthesizer," *Neural Networks, IEEE Transactions on*, vol. 4, no. 1, pp. 2–8, 1993.

[50] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 1, pp. 23–38, 1998.

[51] H. Rowley, S. Baluja, T. Kanade, *et al.*, "Rotation invariant neural network-based face detection," in *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, pp. 38–44, IEEE, 1998.

[52] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *Neural Networks, IEEE Transactions on*, vol. 8, no. 1, pp. 98–113, 1997.

[53] H. Rowley, S. Baluja, T. Kanade, *et al.*, "Neural network-based face detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 1, pp. 23–38, 1998.

[54] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673–679, 2001.

[55] H. Kato, M. Kanematsu, X. Zhang, M. Saio, H. Kondo, S. Goshima, and H. Fujita, "Computer-aided diagnosis of hepatic fibrosis: preliminary evaluation of MRI texture analysis using the finite difference method and an artificial neural network," *American Journal of Roentgenology*, vol. 189, no. 1, pp. 117–122, 2007.

[56] A. Petrosian, D. Prokhorov, W. Lajara-Nanson, and R. Schiffer, "Recurrent neural network-based approach for early recognition of alzheimer's disease in EEG," *Clinical Neurophysiology*, vol. 112, no. 8, pp. 1378–1387, 2001.

[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[58] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 1717–1724, IEEE, 2014.

[59] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 1725–1732, IEEE, 2014.

[60] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 221–231, 2013.

[61] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, N. C. Fox, C. R. Jack, J. Ashburner, and R. S. Frackowiak, "Automatic classifi-

cation of MRI scans in alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, 2008.

[62] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, O. Colliot, A. D. N. Initiative, *et al.*, "Automatic classification of patients with alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database," *Neuroimage*, vol. 56, no. 2, pp. 766–781, 2011.

[63] B. Magnin, L. Mesrob, S. Kinkingnéhun, M. Pélégrini-Issac, O. Colliot, M. Sarazin, B. Dubois, S. Lehéricy, and H. Benali, "Support vector machine-based classification of alzheimer's disease from whole-brain anatomical MRI," *Neuroradiology*, vol. 51, no. 2, pp. 73–83, 2009.

[64] W. Yang, R. L. Lui, J.-H. Gao, T. F. Chan, S.-T. Yau, R. A. Sperling, and X. Huang, "Independent component analysis-based classification of alzheimer's MRI data," *Journal of Alzheimer's Disease: JAD*, vol. 24, no. 4, pp. 775–783, 2011.

[65] Y. Cho, J.-K. Seong, Y. Jeong, S. Y. Shin, A. D. N. Initiative, *et al.*, "Individual subject classification for alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data," *Neuroimage*, vol. 59, no. 3, pp. 2217–2230, 2012.

[66] J. P. Lerch, J. Pruessner, A. P. Zijdenbos, D. L. Collins, S. J. Teipel, H. Hampel, and A. C. Evans, "Automated cortical thickness measurements from MRI can accurately separate alzheimer's patients from normal elderly controls," *Neurobiology of Aging*, vol. 29, no. 1, pp. 23–30, 2008.

[67] E. Gerardin, G. Chételat, M. Chupin, R. Cuingnet, B. Desgranges, H.-S. Kim, M. Niethammer, B. Dubois, S. Lehéricy, and L. Garnero, "Multidimensional classification of hippocampal shape features discriminates alzheimer's disease and mild cognitive impairment from normal aging," *Neuroimage*, vol. 47, no. 4, pp. 1476–1486, 2009.

[68] M. J. West, P. D. Coleman, D. G. Flood, and J. C. Troncoso, "Differences in the pattern of hippocampal neuronal loss in normal ageing and alzheimer's disease," *The Lancet*, vol. 344, no. 8925, pp. 769–772, 1994.

[69] M. J. West, C. H. Kawas, W. F. Stewart, G. L. Rudow, and J. C. Troncoso, "Hippocampal neurons in pre-clinical alzheimer's disease," *Neurobiology of Aging*, vol. 25, no. 9, pp. 1205–1212, 2004.

[70] X. Yang, Q. Wu, J. Zou, and D. Hong, "Spatial regularization for multitask learning and application in fMRI data analysis," *British Journal of Mathematics & Computer Science*, vol. 14, no. 4, pp. 1–13, 2016.

[71] M. Marnane, S. Mortazavi, J. Li, A. Kim, C. Keller, I. R. Mackenzie, and G.-Y. Hsiung, "High prevalence of co-morbidity in pathologically confirmed alzheimer disease in a canadian regional brain biobank (p6. 221)," *Neurology*, vol. 86, no. 16 Supplement, pp. P6–221, 2016.

[72] D. L. Giraldo, J. D. García-Arteaga, and E. Romero, "Finding regional models of the alzheimer disease by fusing information from neuropsychological tests and structural mri images," in *SPIE Medical Imaging*, pp. 97852H–97852H, International Society for Optics and Photonics, 2016.

[73] M. D. Greicius, G. Srivastava, A. L. Reiss, and V. Menon, "Default-mode network activity distinguishes alzheimer's disease from healthy aging: evidence from functional mri," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 13, pp. 4637–4642, 2004.

[74] D. Chan, N. C. Fox, R. I. Scahill, W. R. Crum, J. L. Whitwell, G. Leschziner, A. M. Rossor, J. M. Stevens, L. Cipolotti, and M. N. Rossor, "Patterns of temporal lobe atrophy in semantic dementia and alzheimer's disease," *Annals of neurology*, vol. 49, no. 4, pp. 433–442, 2001.

[75] C.-A. Cuénod, A. Denys, J.-L. Michot, P. Jehenson, F. Forette, D. Kaplan, A. Syrota, and F. Boller, "Amygdala atrophy in alzheimer's disease: an in vivo magnetic resonance imaging study," *Archives of neurology*, vol. 50, no. 9, pp. 941–945, 1993.

[76] A. Smith, K. Jobst, M. Szatmari, A. Jaskowski, E. King, A. Smith, A. Molyneux, M. Esiri, B. McDonald, and N. Wald, "Detection in life of confirmed alzheimer's disease using a simple measurement of medial temporal lobe atrophy by computed tomography," *The Lancet*, vol. 340, no. 8829, pp. 1179–1183, 1992.

[77] A. Brun, "Frontal lobe degeneration of non-alzheimer type. i. neuropathology," *Archives of gerontology and geriatrics*, vol. 6, no. 3, pp. 193–208, 1987.

[78] H. A. Crystal, D. S. Horoupian, R. Katzman, and S. Jotkowitz, "Biopsy-proved alzheimer disease presenting as a right parietal lobe syndrome," *Annals of neurology*, vol. 12, no. 2, pp. 186–188, 1982.

[79] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991.

[80] A. Y. Ng, "Feature selection, $l_1$ vs. $l_2$ regularization, and rotational invariance," in *Proceedings of The Twenty-first International Conference on Machine Learning*, p. 78, ACM, 2004.

[81] R.-A. Müller, N. Kleinhans, N. Kemmotsu, K. Pierce, and E. Courchesne, "Abnormal variability and distribution of functional maps in autism: an fMRI study of visuomotor learning," *American Journal of Psychiatry*, 2014.

[82] T. Iidaka, "Resting state functional magnetic resonance imaging and neural network classified autism and control," *Cortex*, vol. 63, pp. 55–67, 2015.

[83] H. Koshino, P. A. Carpenter, N. J. Minshew, V. L. Cherkassky, T. A. Keller, and M. A. Just, "Functional connectivity in an fMRI working memory task in high-functioning autism," *Neuroimage*, vol. 24, no. 3, pp. 810–821, 2005.

[84] M. A. Just, V. L. Cherkassky, T. A. Keller, R. K. Kana, and N. J. Minshew, "Functional and anatomical cortical underconnectivity in autism: evidence from an fMRI study of an executive function task and corpus callosum morphometry," *Cerebral Cortex*, vol. 17, no. 4, pp. 951–961, 2007.