REGULARIZED FUNCTIONAL REGRESSION MODELS WITH APPLICATIONS TO BRAIN IMAGING

by

Xuejing Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2013

Doctoral Committee:

Professor Bin Nan, Chair
Assistant Professor Hui Jiang
Associate Professor Timothy D. Johnson
Professor Ji Zhu

To my parents

ACKNOWLEDGEMENTS

First and foremost, I owe my deepest gratitude to my doctoral and academic advisor Dr. Bin Nan, who has supported and supervised me during my Ph.D. study. This dissertation would not have been possible without his immeasurable guidance, encouragement and support. I thank him for introducing me to the wonders of scientific research. I also thank him for bringing me into the realm of brain imaging research. It has been one of the most interesting and challenging research topics I have faced in my statistical life.

I also wish to express my sincerest thanks to Dr. Ji Zhu for sharing his rich knowledge and invaluable experience in statistical learning. I have benefited a great deal from working with him. My great thanks also go to Dr. Timothy D. Johnson and Dr. Hui Jiang for providing me very insightful comments and invaluable advice on my dissertation work.

In addition to my dissertation committee, I would like to thank Dr. Robert A. Koeppe for kindly granting the time to explain the basics of brain imaging data and dedicating his expertise and experience to my dissertation.

I would also like to thank Chris Scheller for his great technical support and Virginia Rogers for her quick responses to my questions regarding the data set used in this dissertation.

I am especially grateful for my parents for their unconditional love, endless patience and being with me through every up, as well as every down. I would not be who I am today if it were not for them.

Last but not least, I would like to thank my dear friends with whom I have shared so much laughter and so many tears throughout my years in Ann Arbor.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

REGULARIZED FUNCTIONAL REGRESSION MODELS WITH APPLICATIONS TO BRAIN IMAGING

by

Xuejing Wang

Chair: Bin Nan

Positron emission tomography (PET) is an imaging technique that provides useful information about brain metabolism to help clinicians in the early diagnosis of Alzheimer's disease (AD). In order to identify the brain areas that show significant signals, many statistical methods have been developed for the analysis of brain imaging data. However, most of them neglect accounting for spatial information in imaging data. One way to address this problem is to treat each image as a realization of a functional predictor. This dissertation includes three research projects concerning regularized functional regression models via Haar wavelets for the analysis of brain imaging data, particularly PET images.

The first project develops a lasso penalized 3D functional linear regression model by viewing PET image as a 3D functional predictor and cognitive impairment as the response variable, aiming to identify the most predictive voxels with the underlying assumption that only a few brain areas are truly predictive. The PET images are obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The second project concerns a lasso penalized 3D functional logistic regression model

for classification of PET images from ADNI database. ADNI participants were classified into three groups during their initial visits: AD, Mild Cognitive Impairment (MCI) and Normal Control (NC). The model is applied to all the pairwise classifications using baseline PET images. The third project develops a regularized 3D multiple functional logistic regression model that can account for the group structure among voxels. Cerebral cortex can be partitioned into multiple regions. Treating each region as a group, within-group and groupwise regularization is imposed into the estimation to identify the most predictive voxels. This model is applied to the prediction of MCI-to-AD conversion using ADNI MCI subjects baseline PET images. All proposed models are evaluated through extensive simulation studies which are based on simulated data and slices extracted from ADNI PET images. Comparisons with existing methods for the prediction performance are also conducted using ADNI data. The results suggest that the proposed models are able to not only identify the predictive voxels, but also achieve higher prediction accuracy than existing methods in general.

CHAPTER I

Introduction

1.1 Motivation

Functional observations, such as curves or images, have become increasingly popular in recent years. Standard multivariate methods are unable to analyze functional observations due to the curse of dimensionality and a failure to account for the correlation between observations. However, statistical analyses of these kinds of observations are required extensively in many fields of research, including economics, biomechanics and medicine. This dissertation focuses mainly on developing some novel methods for analyzing 3D functional data, aiming to overcome the limitations of existing methods.

The primary motivation for this dissertation derives from brain imaging studies, in particular, the application of brain images, such as positron emission tomography (PET) images, in assisting the diagnosis of Alzheimer's disease (AD). It is known that AD has become the most common type of dementia, accounting for 60-80 percent of age-related dementia cases. The disease currently affects about five million people in the US, and the number of victims will significantly increase in the near future barring the development of therapeutics. By recent estimates (Alzheimer's Association, 2010), from 2010 to 2050, the total costs of care for Americans age 65 and older with AD will increase five-fold, from $172 billion to $1.08 trillion per year. Over the past decade, many organizations have been established to avert this tidal wave by funding

research to accelerate the search for cures while improving diagnosis. PET imaging is based on an assumption that brain activity is associated with high radioactivity, and it has been shown to be an important tool to assist with the diagnosis of AD. For example, $^{18}Fluorine$-fluorodeoxiglucose PET (FDG PET) measuring brain glucose metabolism, can show changes of glucose metabolism as the disease progress (Mosconi, 2005). Figure 1.1 shows that glucose metabolism is greatly decreased in some regions for patients with AD. Improving the diagnosis of AD using PET images has attracted more and more attention, and many statistical methods have been developed for this purpose. However, most of the existing methods not only tend to ignore the correlations between voxels, but also fail to identify the truly predictive brain areas. Here we conduct a simple experiment to see how existing methods fail to identify the truly predictive voxels. We use 200 subjects' PET images for this experiment, and only consider one and the same axial slice per subject for simplicity. In the regression framework, the voxels in the selected slice from subject i are treated as X_i with dimension $1 \times p$, where p is the total number of voxels in the slice. We further assume that only two small round regions have nonzero effect on the outcome (see Figure 1.2 (a)), and all other regions have no effect at all. These voxel-level effects can be characterized by the coefficient β. The blue area indicates negative β, while the red area indicates positive β. The binary outcome ($Y_i = 1$ or $Y_i = 0$) is considered for each subject, and thus the logistic regression model is fit as follows:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + X_i^T\beta, i = 1, ..., 200, \tag{1.1}$$

where $\pi_i = P(Y_i = 1|X_i)$. The binary outcomes are simulated following the procedure commonly used in logistic regression, namely that for subject i, Y_i is generated by drawing a random uniform number u on the interval $[0, 1]$, and let $Y_i = 1$ if $\pi_i < u$ and $Y_i = 0$ otherwise. Based on these steps, 200 subjects are divided into two groups:

one with 87 subjects and the other with 113 subjects. The first method considered here is the univariate voxel-wise analysis, and a two-sample t-test at each voxel is carried out. As there are a large number of voxels, multiple comparisons need to be adjusted. We apply the False Discovery Rate (FDR) controlling procedure at the level of 0.005, and the result is presented in Figure 1.2 (b). It can be easily seen that the significant voxels identified by the univariate analysis are much more than the true nonzero voxels, which leads to false positive findings. In addition to the univariate analysis, we also consider the principal component analysis (PCA) by first extracting a small number of PCs and then conducting logistic regression based on the PC scores. The voxels identified in this way are presented in Figure 1.2 (c), showing that the identified voxels are distributed throughout the brain. One reason for this massive association obtained from the existing methods could be the high correlation among voxels.

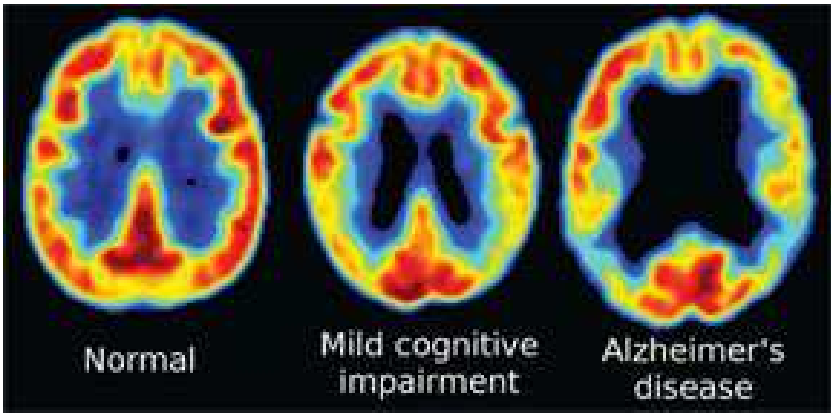

Figure 1.1: FDG PET images show reduced glucose metabolism in temporal and parietal regions in patients with Alzheimer's disease and mild cognitive impairment. *Images courtesy of Suzanne Baker, PhD; William Jagust, MD; and Susan Landau, PhD.*

This dissertation develops some functional regression models by treating PET images as 3D functional observations. The major advantage here is the extra information obtained from viewing all the voxels altogether rather than analyzing them one by one. The main objectives in this dissertation are searching for the predictive voxels

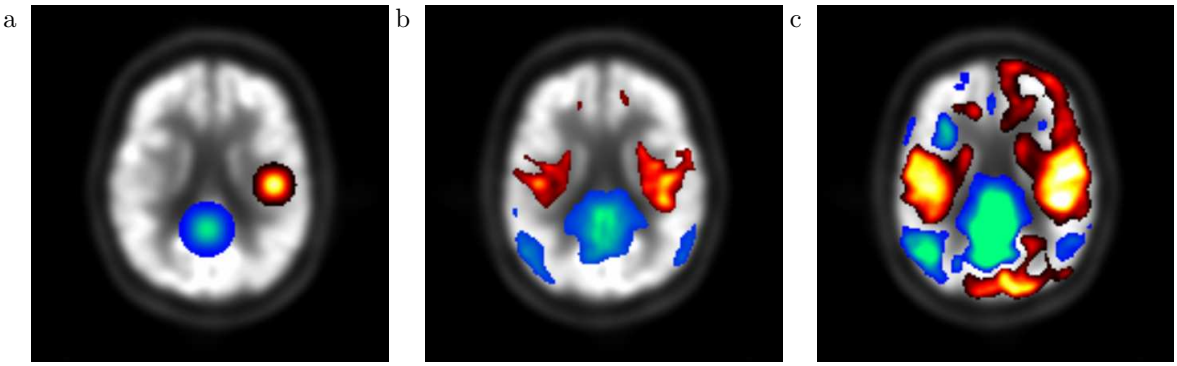

Figure 1.2: (a) The image of true nonzero voxels (β). (b) The image of t statistics showing differences between the two groups. (c) The coefficient (β) image obtained by PCA method.

and improving the performance in predicting the outcomes of interest. The limitations of existing methods can be overcome by considering proper basis functions and regularized estimation. We start with the analysis of 1D functional observations for the illustration of proposed models and then move on to 2D and 3D cases. Comprehensive simulation studies are presented to evaluate the performance for 1D, 2D and 3D functional observations. Finally, we demonstrate how the proposed approaches can be used in practice by applying them to real PET images.

1.2 Outline

The rest of the dissertation is organized as follows. Chapter II develops a lasso penalized 3D functional linear regression model via Haar wavelets by treating each PET image as a realization of a 3D functional predictor and cognition as the response variable. In terms of regularization, the lasso penalty is considered. The main objective is to identify the most predictive voxels. The performance of this approach is examined through a variety of simulation studies and then it is applied to the PET imaging data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The results indicate that the proposed approach is able to provide sparse estimates of the voxel-level effects which would enable researchers to identify

predictive voxels more precisely. Chapter III proposes a regularized 3D functional logistic regression model using the lasso penalty for classification of PET images obtained from the ADNI database. The results show that the proposed approach not only achieves comparable classification accuracy rates to previous studies but also yields sparse estimation. The findings are in general agreement with previous studies. Chapter IV is devoted to the application of a regularized multiple functional regression model that accounts for the group structure among voxels. Cerebral cortex can be partitioned into a number of regions. In order to search for important brain regions, the sparse group lasso penalty which can achieve both with-group and groupwise sparsity is incorporated into the estimation by treating each region as a group. As a result, the predictive groups of voxels and non-predictive groups of voxels can be identified more accurately. Finally, Chapter V presents conclusions and future work.

CHAPTER II

Regularized 3D Functional Linear Regression with Application to PET Images

2.1 Introduction

Alzheimer's disease (AD) has become the most frequent cause of dementia in our increasingly aging societies, representing a significant impact on the US population with 10% prevalence in individuals aged above 70 years old (Plassman et al., 2007). Despite the prevalence, this disease remains quite a mystery so far; there is neither a cure nor a definite treatment to arrest its course, and currently, the only definite way to diagnose AD is to examine the brain tissue after death. According to recent studies (Leifer, 2003), the early diagnosis of AD is of great value since new drug therapies can be used to potentially delay the progression of the disease. To this end, much progress has been made in assisting the diagnosis of AD with the aid of neuroimaging techniques. One such widely used neuroimaging technique is positron emission tomography (PET) imaging, which is one of the most promising tools for the early diagnosis of AD, and it is of great scientific interest to understand the association between PET imaging and cognitive impairment. In particular, the fluorodeoxyglucose (FDG) PET has been used to measure the cerebral glucose metabolic activity for over 20 years. Many studies have shown that reduced metabolic activity in some

regions of the brain such as posterior cingulate, temporal and parietal cortices are highly associated with the progression of memory and cognitive impairments in AD (Foster et al., 1984; Minoshima et al., 1995, 1997b).

FDG PET scans used in this work were obtained from a large multi-center follow-up study on Alzheimer's disease and early dementia, called the Alzheimer's Disease Neuroimaing Initiative (ADNI). The ADNI project was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, to test whether imaging biomarkers could measure the progression of AD. A total of 403 FDG PET scans at approximately 50 different participating sites were acquired for this application, including 102 normal control (NC) subjects, 206 subjects with mild cognitive impairment (MCI), and 95 subjects diagnosed with AD. In this study, we consider the baseline FDG PET scans with a standard $160 \times 160 \times 96$ voxel image grid as the predictor to the cognitive performance as measured by mini-mental state exam (MMSE), which is a questionnaire test that is used to screen for cognitive impairment (Cockrell and Folstein, 1988). The maximum MMSE score is 30, and on average, MMSE scores decline as the disease progresses. The goal of our study is to identify brain subregions that are most closely related to the prediction of MMSE scores.

Many methods have been developed for the analysis of brain imaging data in order to identify disease-related brain subregions. Most of these methods focus on region of interest (ROI) and voxel-based univariate analysis, see for example, Luo and Nichols (2003), Grimmer et al. (2009), Karow et al. (2010) and Habeck et al. (2008), among many others. These methods are intended to provide statistics by doing a separate analysis for each ROI or voxel and then to draw inferences at the region- or voxel-level. As a result of testing millions of hypotheses, appropriate adjustments for multiple comparison have to be considered. In addition to univariate analysis,

a multivariate regression model can be fitted by treating every voxel as a covariate. Since the number of voxels is much larger than the number of scans, some dimension reduction techniques have to be implemented, otherwise the least squares solutions are not reliable. Such analysis, however, may lead to difficulties in interpretations and practical implications. Both the traditional univariate and multivariate approaches have one major limitation in common: they neglect to account for spatial correlations among voxels. These methods are developed without considering the spatial information of the brain, possibly resulting in some loss of information. There is an emerging awareness of the importance of taking such information into account. For example, multivariate analysis can be conducted with a focus on extracting principal components from the images (Friston et al., 1996; Kerrouche et al., 2006). More recently, a variety of Bayesian spatial modeling approaches have been proposed to model the correlation between neighboring voxels, which need to carefully specify the prior distributions, see for example Bowman et al. (2008); Kang et al. (2011). Our way to address this issue is to incorporate voxels' spatial information by treating the 3D image as a whole. We therefore can retain all the information from the original image, and take advantage of the extra information to investigate the relationship between PET images and the diagnosis of disease more efficiently. In this work, we treat PET imaging data as the 3D functional observations, and propose a novel Haar wavelet-based regularized approach to analyze PET imaging data in the framework of functional data analysis.

Functional regression models are known as one of the standard techniques in functional data analysis. It is noted that the models can be defined as functional in one or both of two ways: the response variable is functional; at least one of the covariates is functional. In this work, we focus on the functional linear regression model with a scalar response variable and a single functional predictor. Using the 1D case as an illustration, the functional linear regression model relates a scalar response

variable Y to a functional predictor as follows:

$$Y_i = \beta_0 + \int_0^T X_i(t)\beta(t)\,dt + \epsilon_i, \quad i = 1, \cdots, n, \tag{2.1}$$

where $\beta(t)$ is the regression coefficient function. As this is the 1D case, t refers to the time information, and the spatial information will be incorporated in the 3D model. Regularization methods, such as the roughness penalty approach or using restrictive basis functions (Ramsay and Silverman, 2005) can be implemented to produce an estimator that is meaningful in interpretation and useful in prediction.

For the functional linear regression model (2.1), James et al. (2009) proposed a regularized approach that focuses on producing sparse and highly interpretable estimates of the coefficient function $\beta(t)$. This approach involves first dividing the time period into a fine grid of points, and then using appropriate variable selection methods to determine whether the dth derivative of $\beta(t)$ is zero or not at each of the grid points, i.e. $\beta^{(d)}(t) = 0$ for one or more values of $d \in \{0, 1, 2, \cdots\}$. They proposed the Dantzig selector (Candes and Tao, 2007) and a Lasso-type approach for the estimation of $\beta(t)$ using piecewise constant basis, where the Dantzig selector seems to be more natural. Empirical results show that their methods perform well when p, the number of basis functions, is not too large. When functional data are measured over a very fine grid such as brain imaging data, the Dantzig selector faces the challenge of solving a huge linear programming problem and the Lasso-type algorithm can be extremely slow; note that for the latter the fast shooting algorithm (Fu, 1998; Friedman et al., 2007) does not apply due to the penalty on derivatives. Without imposing sparsity, Reiss and Ogden (2010) considered the functional principal component regression for imaging data.

In this work, we choose the Haar wavelet basis instead of the piecewise constant basis for analyzing 3D imaging data and show that the Haar wavelet-based approach

presents a number of advantages. First, it yields regional sparseness without imposing constraints on derivatives, which is needed in James et al. (2009). In other words, the estimator of the regression coefficient function can be exactly zero over regions where no relationship to the response variable is present by shrinking the corresponding wavelet coefficients of the regression coefficient function to zero. Second, the Haar wavelet transform offers a way to overcome the issue of high multicollinearity caused by high neighboring spatial correlations. Third, our approach is flexible enough to allow the coefficient function to be estimated at different levels of smoothness through choosing different levels of the Haar wavelet decomposition. Fourth, the Haar wavelet transform can be applied as a dimension reduction technique prior to model fitting for high-dimensional imaging data by setting a common set of close to zero wavelet coefficients of PET images to zero, which is an effective way of removing voxels outside brain or in brain ventricles. Zhu et al. (2012) considered a wavelet-based functional mixed model that incorporates functional observations as the fixed-effects and covariates indicating a possible clustering structure as the random effects, and the wavelet-based transformation was conducted. The recent article by Zhao et al. (2012) also considered a wavelet-based approach in functional linear regression. Both papers did not consider sparsity of the coefficient function $\beta(t)$.

The rest of this chapter is organized as follows. In Section 2.2, we review some background on wavelet decomposition and properties of Haar wavelet basis functions using 1D functional linear regression model as an illustration, and then propose the ell_1 regularized shrinkage estimation for general functional data, including both 1D and 3D cases. In Section 2.3, we provide non-asymptotic error bounds for prediction and estimation. To evaluate the numerical performance of our approach, we conduct extensive simulations in Section 2.4. We present the analysis of ADNI 3D FDG PET imaging data in Section 2.5, and present some concluding remarks in Section 2.6.

2.2 Methods

For ease of presentation, we describe the proposed methodology starting with the 1D case given in (2.1), then extend it to the 3D case using a tensor product of three 1D wavelet expansions.

2.2.1 Choice of basis

Basis expansions are commonly used in analyzing functional data. Among a variety of choices of basis expansions, wavelets have the important ability to allow simultaneous time, or space, and frequency localization. Unlike many other commonly used basis systems, wavelet transforms are highly adaptable to different levels of smoothness, and more capable of capturing edges, spikes and other types of discontinuities, especially for wavelet transforms with relatively small support such as the Haar wavelets. Wavelet transforms also provide a powerful tool to compress the data. A compressed approximation of the signal can be achieved by penalizing the wavelet coefficients, which involves discarding least significant coefficients and possibly shrinking the large ones without affecting the main features of the data. Hence it is advantageous to use wavelet transforms to decompose images as well as the regression coefficient function for estimation.

In many applications, it is often the case that the association between $X(t)$ and Y in model (2.1) is sparse, and potentially discontinuous at the boundaries of subregions. In particular, few brain subregions in the aforementioned PET images are believed to be related to cognitive impairment. To better identify such patterns, we choose to use the Haar wavelets. The Haar wavelet transform is easily calculated and affected less by discontinuities. In addition, sparsity of $\beta(t)$ can be recovered by shrinking its wavelet coefficients to zero. The scaling function (also called a father wavelet) ϕ and

the mother wavelet ψ of Haar wavelets defined on $[0,1)$ are given below:

$$\begin{aligned}
\phi(t) &= \begin{cases} 1, & \text{if } 0 \le t < 1; \\ 0, & \text{otherwise;} \end{cases} \\
\psi(t) &= \begin{cases} 1, & \text{if } 0 \le t < 1/2; \\ -1, & \text{if } 1/2 \le t < 1; \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

The Haar wavelet bases are then generated in the form of translations and dilations of the above father and mother wavelet functions as

$$\begin{aligned}
\phi_{j,k}(t) &= \sqrt{2^j}\phi(2^j t - k), \\
\psi_{j,k}(t) &= \sqrt{2^j}\psi(2^j t - k),
\end{aligned}$$

where $j = 0, 1, \cdots$ and $k = 0, 1, \cdots, 2^j - 1$. The index j refers to dilations and k refers to translations and $\sqrt{2}$ is the normalizing factor. It is noted that the basis functions are orthogonal and normalized. Therefore, for a resolution J, the coefficient function $\beta(t)$ in (2.1) defined on $[0,1)$ can be expanded in a Haar wavelet series:

$$\beta(t) = \sum_{k=0}^{2^{j_0}-1} a_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t) + e(t), \tag{2.2}$$

where $a_{j_0,k} = \int_0^1 \beta(t)\phi_{j_0,k}(t)dt$ are the approximation coefficients at the coarsest resolution j_0, $d_{j,k} = \int_0^1 \beta(t)\psi_{j,k}(t)dt$ are the detail coefficients that characterize the finer structures of $\beta(t)$ as j grows, and $e(t)$ is the approximation error that goes to zero as J goes to infinity. The Haar wavelet representation of a signal thus consists of approximations together with details that can provide the desirable frequencies. See e.g. Walker (2008) for details about Haar wavelets.

2.2.2 Model estimation

Rewrite $\beta(t)$ in (2.2) by

$$\beta(t) = B(t)^T \eta + e(t), \tag{2.3}$$

where $B(t)$ denotes the collection of all $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ in the above Haar wavelet expansion, and η is the corresponding wavelet coefficient vector of length p. Plugging (2.3) into (2.1), we obtain

$$Y_i = \beta_0 + \int_0^1 X_i(t) B(t)^T \eta \, dt + \epsilon_i^* = \beta_0 + C_i^T \eta + \epsilon_i^*, \quad i = 1, \cdots, n, \tag{2.4}$$

where $C_i = \int_0^1 X_i(t) B(t) \, dt$. It should be noted that C_i is the wavelet coefficient vector of $X_i(t)$ when we decompose $X_i(t)$ using the same set of Haar wavelet basis functions as those in (2.3). Model (2.4) can then be rewritten as follows:

$$Y = \beta_0 + C\eta + \epsilon^*, \tag{2.5}$$

where $C = [C_1, C_2, \cdots, C_n]^T$, which is an $n \times p$ design matrix in linear model (2.5). Once an estimator $\hat{\eta}$ is obtained from (2.5), $\beta(t)$ can then be estimated by $B(t)^T \hat{\eta}$.

In practice, $X(t)$ is observed on only a finite set of grid points $\{t_1, \cdots, t_p\}$. p is assumed to be a power of 2 for convenience, since Haar wavelet transform performs the operations of averaging and differencing on each pair of values, and the operations are repeated recursively. The wavelet decompositions of $X(t)$ and $\beta(t)$ on those grid points can be performed only at a finite number of levels. Using the usual terminology for Haar wavelets (see e.g. Walker (2008), which is the same as what is used in the Matlab Wavelet Toolbox, 2011b), we define level 1 Haar wavelet decomposition by computing the average and the difference on each consecutive pair of values, and

the maximum level is $\log_2 p$. The level number is directly determined by the integer j_0 in (2.2). For any level of Haar wavelet decomposition, the total number of basis functions $\phi_{j,k}$ and $\psi_{j,k}$ is always p, and the collection of $\phi_{j,k}$ and $\psi_{j,k}$ then forms a set of p-dimensional orthonormal basis functions.

A key advantage of using Haar wavelets is as follows. When $\beta(t) = 0$ in large regions of $t \in [0, 1)$, the coefficient vector η in (2.3)-(2.5) should be sparse, i.e. $\beta(t)$ can be well approximated by an economical wavelet expansion with few nonzero coefficients. To obtain sparse solutions, a variety of variable selection methods can be used, including the lasso (Tibshirani, 1996), the Dantzig selector (Candes and Tao, 2007), the elastic net (Zou and Hastie, 2005), the adaptive lasso (Zou, 2006) and SCAD (Fan and Li, 2001). The latter two methods work for cases with $p < n$ and the method of elastic net is developed for highly correlated covariates. For ultra large values of p, the computational cost of implementing linear programming for solving the Dantzig selector problem could be a huge hurdle. We therefore choose the lasso approach over other procedures in this work, which can be solved by a fast coordinate descent algorithm (Fu, 1998; Daubechies et al., 2004; Friedman et al., 2007; Wu and Lange, 2008).

For a given j_0, which corresponds to a specific level of Haar wavelet expansion, the lasso estimator for η is given by

$$\hat{\eta} = \arg\min_{\eta} \left\{ \frac{1}{n} \|Y - \beta_0 - C\eta\|_2^2 + 2\lambda \|\eta\|_1 \right\}, \tag{2.6}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the ℓ_1 and ℓ_2 norms respectively, and $\lambda \geq 0$ is a tuning parameter. In our estimating procedure, j_0 is another tuning parameter.

It should be noted that in general the Haar wavelet coefficients with larger magnitude are related to salient features. With the Haar wavelet, the magnitude of detail coefficients should be proportional to the differences between every pair of values, i.e.

larger magnitude indicates sharp changes at corresponding locations and zero magnitude of the detail coefficients indicates no change. If both detail and approximation coefficients of the Haar wavelet transform are close to zero, then $\beta(t)$ is close to zero. Thus we are able to obtain a sparse solution of $\beta(t)$ by shrinking its small wavelet coefficients to zero.

2.2.3 Selection of tuning parameters

In addition to the lasso tuning parameter λ in (2.6), we also need to take into account the level of the Haar wavelet decomposition. There should exist an optimal level of decomposition for $\beta(t)$ in terms of certain criteria, such as AIC, BIC, or cross-validation. If the length of observed $X_i(t)$ is p, then the maximum possible level of the discrete Haar wavelet transform is $\log_2 p$, which is relatively small. Moreover, lower levels are usually considered in real applications. Therefore including two tuning parameters does not increase computational burden much.

2.2.4 3D case

A 3D function can be decomposed using the tensor product of three 1D Haar wavelets. In particular, the 3D Haar wavelet transform can be considered as averaging and differencing operations (Muraki, 1992). The averaging operation is constructed by the 3D scaling function below:

$$\phi_{j,\{k,l,m\}}(u,v,w) = \phi_{j,k}(u)\phi_{j,l}(v)\phi_{j,m}(w).$$

The differencing operation is taken in seven directions, which is constructed by the 3D wavelet functions as follows:

$$\psi^1_{j,\{k,l,m\}}(u,v,w) = \phi_{j,k}(u)\phi_{j,l}(v)\psi_{j,m}(w),$$
$$\psi^2_{j,\{k,l,m\}}(u,v,w) = \phi_{j,k}(u)\psi_{j,l}(v)\phi_{j,m}(w),$$
$$\psi^3_{j,\{k,l,m\}}(u,v,w) = \phi_{j,k}(u)\psi_{j,l}(v)\psi_{j,m}(w),$$
$$\psi^4_{j,\{k,l,m\}}(u,v,w) = \psi_{j,k}(u)\phi_{j,l}(v)\phi_{j,m}(w),$$
$$\psi^5_{j,\{k,l,m\}}(u,v,w) = \psi_{j,k}(u)\phi_{j,l}(v)\psi_{j,m}(w),$$
$$\psi^6_{j,\{k,l,m\}}(u,v,w) = \psi_{j,k}(u)\psi_{j,l}(v)\phi_{j,m}(w),$$
$$\psi^7_{j,\{k,l,m\}}(u,v,w) = \psi_{j,k}(u)\psi_{j,l}(v)\psi_{j,m}(w).$$

Let $X_i(u,v,w)$ be a 3D functional predictor and Y_i be a scalar response variable for subject i, $i = 1,\ldots,n$. The 3D functional linear regression model can be written as:

$$Y_i = \beta_0 + \int_0^{T_1}\int_0^{T_2}\int_0^{T_3} X_i(u,v,w)\beta(u,v,w)\,dudvdw + \epsilon_i. \tag{2.7}$$

For a resolution J, the 3D coefficient function $\beta(u,v,w)$ can be approximated by:

$$\sum_{k,l,m=0}^{2^{j_0}-1} a_{j_0,\{k,l,m\}}\phi_{j_0,\{k,l,m\}}(u,v,w) + \sum_{j=j_0}^{J}\sum_{k,l,m=0}^{2^j-1}\sum_{q=1}^{7} d^q_{j,\{k,l,m\}}\psi^q_{j,\{k,l,m\}}(u,v,w). \tag{2.8}$$

Denote the set of all basis functions $\phi_{j,\{k,l,m\}}$ and $\psi^q_{j,\{k,l,m\}}$ in (2.8) by $B(u,v,w)$ and the wavelet coefficients in (2.8) by η, then $\beta(u,v,w)$ can be written as

$$\beta(u,v,w) = B(u,v,w)^T\eta + e(u,v,w). \tag{2.9}$$

Plugging (2.9) into model (2.7), we obtain

$$\begin{aligned} Y_i &= \beta_0 + \int_0^{T_1}\int_0^{T_2}\int_0^{T_3} X_i(u,v,w)B(u,v,w)^T\eta\, dudvdw + \epsilon_i^* \qquad (2.10) \\ &= \beta_0 + C_i^T\eta + \epsilon_i^*, \end{aligned}$$

where $C_i = \int_0^{T_1}\int_0^{T_2}\int_0^{T_3} X_i(u,v,w)B(u,v,w)\, dudvdw$, which is equivalent to the wavelet coefficient vector when we apply the 3D wavelet transform to $X_i(u,v,w)$. Then the methodology proposed in previous subsections for the 1D case applies exactly.

2.3 Theoretical Results

In this section, we provide non-asymptotic theoretical properties of the proposed method following the calculation of Bickel et al. (2009). Proofs are deferred to the Appendix A. Though the results are described using the 1D notation, they hold exactly for the 3D case by viewing variable t as a 3D variable.

Considering the Haar wavelet representation of β with approximation error as in (2.3), if we assume only a few elements of η are nonzero, we can define $A_\eta = \{j : \eta_j \neq 0, j \in \{1,\cdots,p\}\}$. We further denote the cardinality of A_η by $|A_\eta|$, which characterizes the sparsity of η. One common feature for the lasso and the Dantzig selector is that, for the residual vector $\delta = \hat{\eta} - \eta$, the following relation holds with high probability:

$$\|\delta_{A_\eta^c}\|_1 \leq k_0\|\delta_{A_\eta}\|_1,$$

where δ_{A_η} is the subvector of δ consisting of the jth element of δ where $\eta_j \neq 0$, and the constant k_0 is a positive number: $k_0 = 1$ for the Dantzig selector and $k_0 = 3$ for the lasso (see Bickel et al., 2009). To achieve the oracle inequalities for the proposed estimator, we assume the same restricted eigenvalue assumption of Bickel et al. (2009) on the Gram matrix $\frac{1}{n}C^TC$, which is given below.

Assumption . (Restricted Eigenvalue assumption $RE(s, k_0)$) For some integer $1 \leq s \leq p$ and $k_0 > 0$, the following holds:

$$\kappa(s, k_0) = \min_{A_\eta \subseteq \{1,\cdots,p\}: |A_\eta| \leq s} \quad \min_{\delta \neq 0: \|\delta_{A_\eta^c}\|_1 \leq k_0 \|\delta_{A_\eta}\|_1} \frac{\|C\delta\|_2}{\sqrt{n}\|\delta_{A_\eta}\|_2} > 0.$$

We also assume that for a given finite wavelet decomposition with resolution J and corresponding total number of basis functions p, all the diagonal elements of the matrix $\frac{C^T C}{n}$ be equal to 1. Then we have Theorem II.1, Theorem II.2 and Theorem II.3.

Theorem II.1. *Assume that ϵ_i in model (2.1) follows a normal distribution $N(0, \sigma_1^2)$ with $\sigma_1^2 > 0$. Suppose that η is an s-sparse vector, where $1 \leq s \leq p$. Suppose that the assumption $RE(s, k_0)$ is satisfied with $k_0 = 3 + 4/\theta$, for some $\theta > 0$. Let $\hat{\eta}$ be the lasso solution given in (2.6) with $\lambda = a\sigma_1\sqrt{\frac{\log p}{n}}$, $a > 2\sqrt{2}$. Then with probability at least $1 - p^{1-a^2/8}$, we have*

$$\left\| \int_0^T X(t)\hat{\beta}(t)\, dt - \int_0^T X(t)\beta(t)\, dt \right\|_n^2$$
$$\leq (1+\theta) \inf_{\eta \in \mathbb{R}^p:\ |A_\eta| \leq s} \left\{ \left\| \int_0^T X(t)e(t)\, dt \right\|_n^2 + \frac{D(\theta)a^2\sigma_1^2}{\kappa^2} \left(\frac{s \log p}{n} \right) \right\}, \quad (2.11)$$

where $\kappa = \kappa(s, 3 + 4/\theta)$ and $D(\theta)$ is a positive constant depending on θ; and for this given p-dimensional basis, let $\omega = \sup_t |e(t)|$ and suppose that there exists an $M < \infty$ such that $\int |X_i(t)|\ dt \leq M$, then we have

$$\left| \hat{\beta}(t) - \beta(t) \right| \leq \gamma(t)(4 + 4/\theta) \left\{ \frac{4as\sigma_1\sqrt{\frac{\log p}{n}} + 2\kappa\sqrt{s}M\omega}{\kappa^2} \right\} + \omega, \quad (2.12)$$

where

$$\gamma(t) = \sum_{j=1}^{p} \left| \frac{b_j(t)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} C_{ij}^2}} \right|,$$

and $b_j(t)$ *is the* j*-th basis function in* $B(t)$*.*

If we further assume that $\epsilon_i^* \sim N(0, \sigma_2^2)$ in model (2.5), we have the following result.

Theorem II.2. *Assume that* $\epsilon_i^* \sim N(0, \sigma_2^2)$ *with* $\sigma_2^2 > 0$*. The model (2.5) reduces to a linear regression problem. Suppose that* η *is an* s*-sparse vector with* $1 \leq s \leq p$*. Let* $\hat{\eta}$ *be the corresponding lasso solution with* $\lambda = a\sigma_2\sqrt{\frac{\log p}{n}}$*, and* $a > 2\sqrt{2}$*. Also suppose there exists an* $M < \infty$ *such that* $\int |X_i(t)|\ dt \leq M$*. Then under assumption* $RE(s, 3)$*, with probability at least* $1 - p^{1-a^2/8}$*, we have*

$$\left\| \int X(t)\hat{\beta}(t)\, dt - \int X(t)\beta(t)\, dt \right\|_n \leq \frac{4a\sigma_2}{\kappa}\sqrt{\frac{s\log p}{n}} + M\omega, \tag{2.13}$$

$$\left| \hat{\beta}(t) - \beta(t) \right| \leq \gamma(t)\frac{16as\sigma_2}{\kappa^2}\sqrt{\frac{\log p}{n}} + \omega, \tag{2.14}$$

where $\kappa = \kappa(s, 3)$*.*

We would like now to derive the asymptotic rates of convergence for $\hat{\beta}(t)$ for the bound (2.12). Before that, we need to state the general conditions for convergence. These conditions are sufficient to derive Theorem II.3.

C.1 There exists $m > 0$ such that $\omega_p = O(2^{-Jm})$, where J is the finest resolution for a finite Haar wavelet decomposition (see Mallat, 1989).

C.2 There exists a constant $S < \infty$ such that $s \leq S$ for all p.

C.3 For a given t, there exists b_t such that $2^{-Jb_t}\gamma_{n,p}(t)$ is bounded for all n and p.

C.4 For some $\theta > 0$, there exists a J^* and corresponding p^* such that $\kappa_{n,p}(s_{p^*}, 3{+}4/\theta)$ is bounded away from zero for reasonably large enough n.

C.5 For some $\theta > 0$, $\kappa(s_{p_n}, 3 + 4/\theta)$ is bounded away from zero for reasonably large enough n and p_n, and $p_n/n \to 0$.

C.1 states that the approximation error of wavelet decomposition converges to zero at the rate of 2^{-Jm}. C.2 assumes that the maximum number of non-zero wavelet coefficients of $\beta(t)$ is bounded, and C.3 states that $\gamma_{n,p}(t)$ grows no faster than 2^{-Jb_t}. C.4 and C.5 guarantee that assumption $RE(s, k_0)$ is satisfied with $k_0 = 3 + 4/\theta$, for large n or for both large n and p_n.

Theorem II.3. *Assume that the assumptions in Theorem II.1 hold. For a given finite Haar wavelet decomposition, fix $J = J^*$ and then $p = p^*$. Let $\omega_{p^*} = \sup_t |e(t)_{p^*}|$ suppose that η is an s_{p^*}-sparse vector and suppose C.1 through C.4 hold, with $\lambda = a\sigma_1\sqrt{\frac{\log p^*}{n}}$, as $n \to \infty$, we have,*

$$\left|\hat{\beta}_n(t) - \beta(t)\right| \leq O(n^{-\frac{1}{2}}) + O(2^{-J^* m}), \tag{2.15}$$

with probability at least $1 - p^{*1-a^2/8}$.

Suppose that C.4 is replaced with with C.5, and b_t is less than m, if 2^{J_n} grows at the rate of $n^{\frac{1}{2m}}$, we have

$$\left|\hat{\beta}_n(t) - \beta(t)\right| = O(\frac{\sqrt{\log n}}{n^{\frac{m-b_t}{2m}}}), \tag{2.16}$$

2.4 Simulation Studies

To investigate the performance of the proposed Haar wavelet-based approach, we have conducted extensive simulations for both 1D and 3D functional data.

2.4.1 1D simulation

We consider a variety of settings of $X(t)$ and $\beta(t)$. For $X(t) = X^*(t) + \mathcal{E}(t)$ defined on $0 \leq t \leq 1$, where $\mathcal{E}(t) \sim N(0, \sigma_{\mathcal{E}}^2)$ is the noise term over time t, we consider the

following two scenarios:

- Fourier: $X^*(t) = a_0 + a_1 \sin(2\pi t) + a_2 \cos(2\pi t) + a_3 \sin(4\pi t) + a_4 \cos(4\pi t)$,

- B-splines: $X^*(t)$ is a linear combination of cubic B-splines with interior knots at $1/7, \cdots, 6/7$ and coefficients a_i, i.e. $X^*(t) = \sum a_i \phi_i(t)$, where $\phi_i(t)$ are the B-spline basis functions.

In both scenarios, the coefficients $a_i \sim N(0, 1)$. To assess the performance of the proposed approach in identifying continuous and discontinuous signals, we consider two cases of the regression coefficient function $\beta(t)$.

- Case 1: $\beta(t)$ is a smooth function,

$$\beta(t) = \begin{cases} 0.5(\sin(20t - \pi) + 1) & \text{if } \pi/8 \leq t < 9\pi/40, \\ 0, & \text{otherwise} \end{cases}$$

- Case 2: $\beta(t)$ is piecewise constant,

$$\beta(t) = \begin{cases} 1, & \text{if } 0.2 \leq t < 0.3 \\ 0.5, & \text{if } 0.5 \leq t < 0.7 \\ 0, & \text{otherwise.} \end{cases}$$

For each curve $X^*(t)$, we record $p = 128$ equally spaced measurements for convenience. The variance of the noise term $\mathcal{E}(t)$ is set to be $\sigma_{\mathcal{E}}^2 = \frac{1}{p-1} \sum_{j=1}^{p} (X^*(t_j) - \bar{X}^*(t_j))^2$, where $\bar{X}^*(t_j)$ is the mean of $X^*(t_j)$. The error term ϵ in model (2.1) also follows a normal distribution $N(0, \sigma^2)$. The value of σ^2 is determined by the signal-to-noise ratio:

$$\text{SNR} = \frac{\sigma_g^2}{\sigma^2}, \tag{2.17}$$

where σ_g^2 is the sample variance of $g(X_i) = \int X_i(t)\beta(t)\, dt$. The simulation results presented in this work are under SNR = 9. For each of the settings, we use $n = 100$

training observations to fit the model. The optimal tuning parameter is selected by using one of the following methods: (i) validating by a separate validation (SV) data set of the same size; (ii) 5-fold cross-validation (CV); (iii) AIC and (iv) BIC (Zou et al., 2007) given below:

$$\text{AIC} = \frac{\|Y - \hat{g}(X)\|^2}{n\, hatsigma^2} + \frac{2}{n}\hat{df}, \tag{2.18}$$

$$\text{BIC} = \frac{\|Y - \hat{g}(X)\|^2}{n\, hatsigma^2} + \frac{log(n)}{n}\hat{df}, \tag{2.19}$$

where $\hat{df}$ is the number of nonzero elements of $\hat{\eta}$ in model (2.5). We estimate σ^2 by the refitted cross-validation method introduced in Fan et al. (2012). We then generate $n = 10,000$ test observations to calculate the mean squared errors (MSEs) of the corresponding selected models. The procedure is repeated 100 times and the average MSEs and their standard errors (SE) for each of the models are presented in Table 2.1. We also report the percentages of correctly identified zero regions and nonzero regions in Table 2.1. We can see that all four methods perform reasonably well, while SV performs the best but it is not a practical method. CV method seems to have a nice trade-off between the sparsity and the prediction accuracy. Averages of $\hat{\beta}(t)$ estimated using CV method over 100 replications are shown in Figure 2.1.

We also conduct permutation tests to assess the significance of the regularized estimates of $\beta(t)$. For each of the training data set, we generate 200 permutation data sets by randomly shuffling the response values. Using the same model selection technique for each of the 200 permutation data sets, 200 sets of $\hat{\beta}_{\text{perm}}(t)$ are obtained. At each t_j, $j = 1, \cdots, p$, the two-sided critical values are set to be the 2.5th and 97.5th percentiles of $\hat{\beta}_{\text{perm}}(t_j)$ for the significance level of 0.05. Suppose the null hypothesis is $\beta(t_j) = 0$ at each t_j, we will reject the null hypothesis if $\hat{\beta}(t_j)$ is within the critical region. Repeating this permutation process 100 times, we can compute the percentages that we reject null hypothesis at each t_j. The results of the permutation tests

		Average MSE (SE) ($\times 10^{-3}$)		Average percentage (%)	
Type	Method	Case 1	Case 2	Case 1	Case 2
B-spline	SV	0.11 (0.05)	0.19 (0.08)	84.30 (69.20)	96.00 (57.26)
	CV	0.15 (0.11)	0.23 (0.11)	82.95 (69.68)	95.03 (58.90)
	BIC	0.60 (1.96)	1.63 (3.10)	72.70 (96.14)	83.26 (79.36)
	AIC	0.56 (1.96)	1.56 (3.12)	75.80 (93.80)	82.51 (82.27)
Fourier	SV	0.65 (0.30)	1.20 (0.49)	84.00 (70.59)	95.87 (58.93)
	CV	0.92 (0.56)	1.46 (0.63)	82.30 (71.39)	95.56 (55.76)
	BIC	1.12 (0.86)	10.62 (20.69)	72.75 (96.59)	84.03 (67.07)
	AIC	1.05 (1.28)	10.29 (20.82)	75.80 (93.64)	83.85 (69.01)

Table 2.1: Average MSEs with standard errors (SE, in parentheses), and average percentages of correctly identified nonzero and zero elements (in parentheses) over 100 replications for 1D cases.

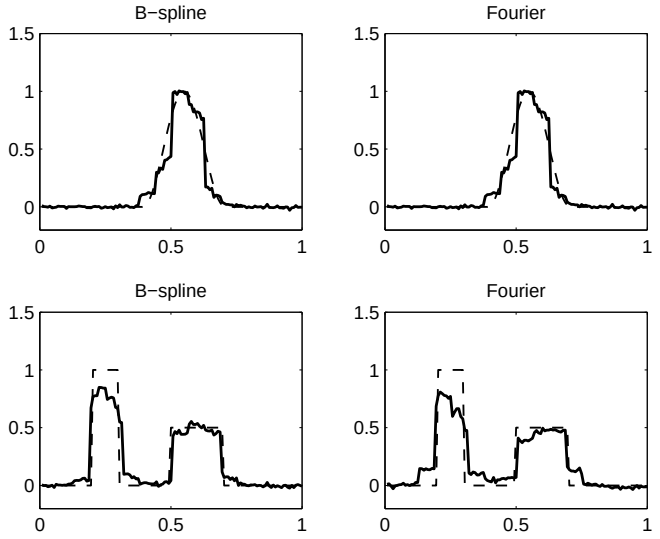

Figure 2.1: Average of $\hat{\beta}(t)$ estimated using 5-fold cross-validation over 100 replications (solid line). The dashed line is true $\beta(t)$. The top panel is for case 1, and the bottom panel is for case 2.

using CV method are presented in Figure 2.2, which shows high rejection frequency in the regions where $\beta(t)$ is nonzero.

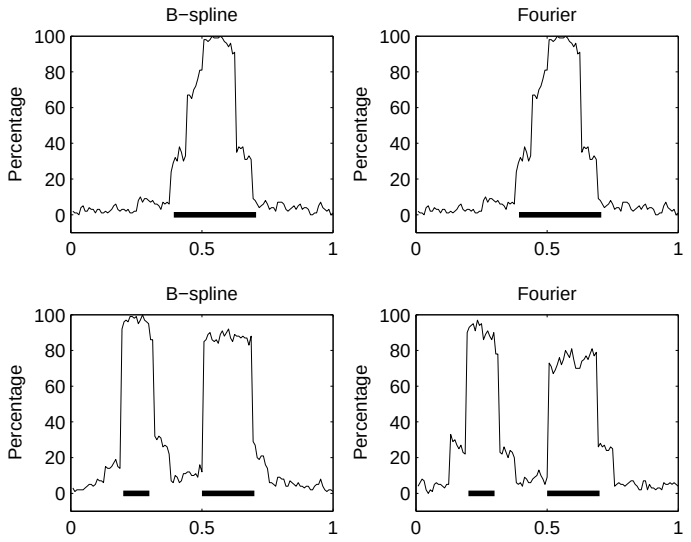

Figure 2.2: Frequency of rejecting the null hypothesis $\beta(t) = 0$ using 5-fold cross-validation based on 100 permutation repetitions. The thick solid horizontal segments indicate the true nonzero regions. The top panel is for case 1, and the bottom panel is for case 2.

2.4.2 3D simulation

For the 3D case, we generate the following type of images $X(u, v, w) = X^*(u, v, w) + \mathcal{E}(u, v, w)$ with:

$$X^*(u, v, w) = a_0 + a_1 \sin(2\pi u) + a_2 \cos(2\pi u) + a_3 \sin(2\pi v) + a_4 \cos(2\pi v) + a_5 \sin(2\pi w) + a_6 \cos(2\pi w),\ 0 \le u, v, w \le 1;$$

where $a_i \sim N(0, 1)$ and $\mathcal{E}(u, v, w) \sim N(0, \sigma_{\mathcal{E}}^2)$ with $\sigma_{\mathcal{E}}^2$ similarly defined as in the 1D case. For simplicity, we record $32 \times 32 \times 32$ equally spaced measurements in the unit cube. We define the coefficient function $\beta(u, v, w)$ as follows,

$$\beta(u, v, w) = \begin{cases} a(\sin(bu + c) + 1)(\sin(bv + c) + 1)(\sin(bw + c) + 1), \text{ if} \\ \quad (u - 7\pi/40)^2 + (v - 7\pi/40)^2 + (w - 7\pi/40)^2 \le (3\pi/40)^2; \\ 0, \text{ otherwise;} \end{cases}$$

where $a = 1/8$, $b = 40/3$ and $c = \pi/6$. Note that $\beta(u, v, w) = 0$ outside a ball that is located in the center of the unit cube. The error term ϵ in model (2.7) also follows a normal distribution $N(0, \sigma^2)$ with SNR $= 9$. We generate 400 training images and apply 3D Haar wavelet transform to decompose each image and obtain the wavelet coefficient matrix. Optimal tuning parameters are selected using the same procedures as for the 1D case. The results are summarized in Table 2.2. Figure 2.3 illustrates the comparison of the true $\beta(u, v, w)$ and the mean estimates of $\beta(u, v, w)$ over 100 replications at five different slices, which shows that our approach can not only detect most of the region where $X(u, v, w)$ is associated with Y, but also identify most of the regions with zero effect.

Method	Average MSE (SE) ($\times 10^{-4}$)	Average percentage (%)
SV	0.97 (0.29)	77.15 (61.97)
CV	1.21 (0.51)	74.25 (57.04)
BIC	4.78 (1.52)	39.48 (99.42)
AIC	4.11 (2.13)	41.86 (98.74)

Table 2.2: Average MSEs with standard errors (SE, in parentheses), and average percentages of correctly identified nonzero and zero elements (in parentheses) over 100 replications for 3D case.

2.5 ADNI FDG PET image analysis

The FDG PET data used in this work were obtained from the ADNI database (adni.loni.ucla.edu). The ADNI was launched in 2003 by NIA, NIBIB, FDA, private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and

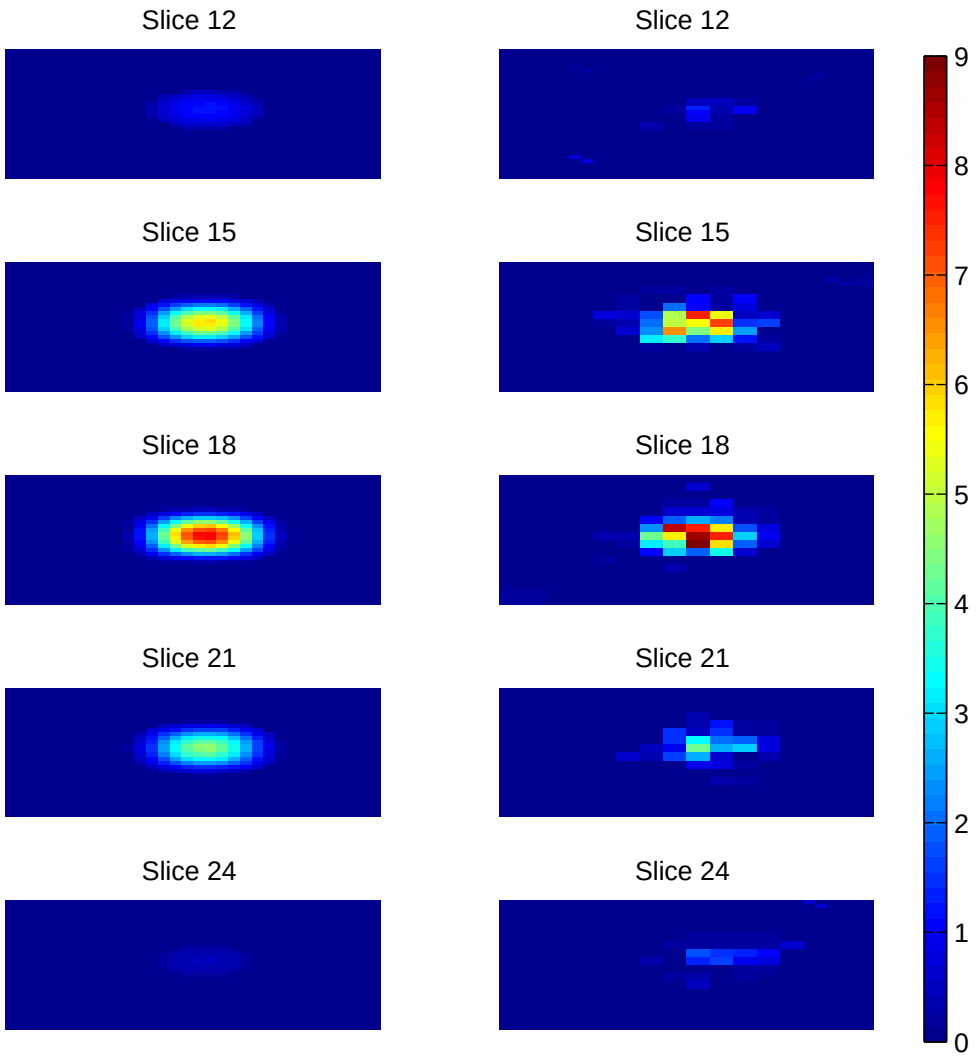

Figure 2.3: The left panel is true $\beta(u, v, w)$ at five selected slices and the right panel is the average of $\hat{\beta}(u, v, w)$ estimated using 5-fold cross-validation over 100 replications at the same five slices.

monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The

initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org.

In the ADNI's FDG PET study, the injected dose of FDG was 5.0 ± 0.5 mCi, and subjects were scanned from 30 to 60 minutes post-injection acquiring 6 five-minute frames. The scans were preprocessed by the following steps: each frame was co-registered to the first frame of the raw image file; six co-registered frames were averaged to create a single 30 minute PET image; each subject's co-registered, averaged PET image from the baseline PET scan was reoriented into a standard $160 \times 160 \times 96$ voxel image grid with 1.5 mm cubic voxels and the anterior-posterior axis of the subject is parallel to a line connecting the anterior and posterior commissures (the AC-PC line). It should be noted that the number of voxels in each image is over 2.4 million, so the approach via linear programming, as in James et al. (2009), can hardly be applied here. The data set consists of 403 scans, including 102 NCs, 206 subjects with MCI, and 95 subjects diagnosed with AD. The demographic characteristics of the 403 subjects are described in Table 2.3. The goal of our analysis is to identify brain subregions that are most closely related to MMSE scores, we therefore choose not to adjust for age and other demographic variables. The summary of MMSE scores among the three groups of participants is given in Figure 2.4. We treat each PET image as a realization of the 3D functional predictor and then fit the 3D functional linear regression model (2.7). The voxel values outside the brain are set to zero prior to implementing the 3D Haar wavelet transform. We further reduce the computational cost by excluding those columns of the wavelet coefficient matrix where all the elements are zero.

In terms of applying 3D Haar wavelet transforms to each subject's PET image data, we consider all the possible levels of the Haar wavelet decompositions. Two

Category	Sex (% male)	Age (SD)	MMSE (SD)
NC ($n = 102$)	60.8%	80.9 (4.7)	28.9 (1.1)
MCI ($n = 206$)	67.0%	79.7 (7.3)	27.2 (1.7)
AD ($n = 95$)	58.9%	80.4 (7.5)	23.4 (2.1)

Table 2.3: Demographics of ADNI participants (n=403)

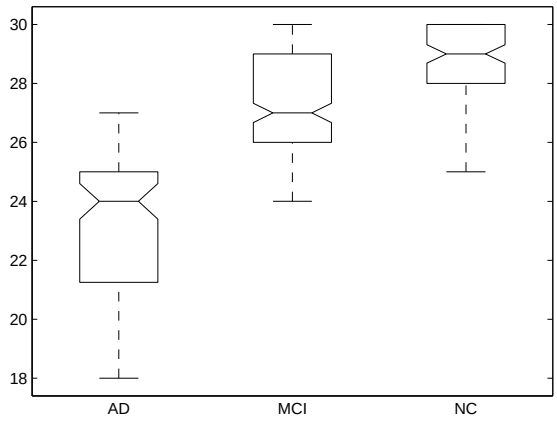

Figure 2.4: Box plots of MMSE scores among AD, MCI, and NC.

tuning parameters are therefore included in the model selection procedure: the level of the 3D Haar wavelet decomposition and the lasso regularization parameter.

Firstly, we evaluate whether the proposed model has a reasonable predictive power for the MMSE score. To ensure that the evaluation is not misleading nor overly optimistic, we employ a technique similar to the leave-one-out cross-validation. Specifically, for each observation, we leave it out as a testing point, use the rest data to fit a model (including the tuning parameter selection) and compute the prediction error on the data point that has been left out. We aggregate these quantities in a way similar to the R-square, i.e. $1 - \sum(y_i - \hat{y}_{i,-i})^2 / \sum(y_i - \bar{y})^2$, where $\hat{y}_{i,-i}$ means the predicted value of y_i is calculated using the estimates obtained from the training data without observation i. The result is 0.26 for the ADNI data set, which suggests that about 26% of the variance among the MMSE scores can be explained by our model.

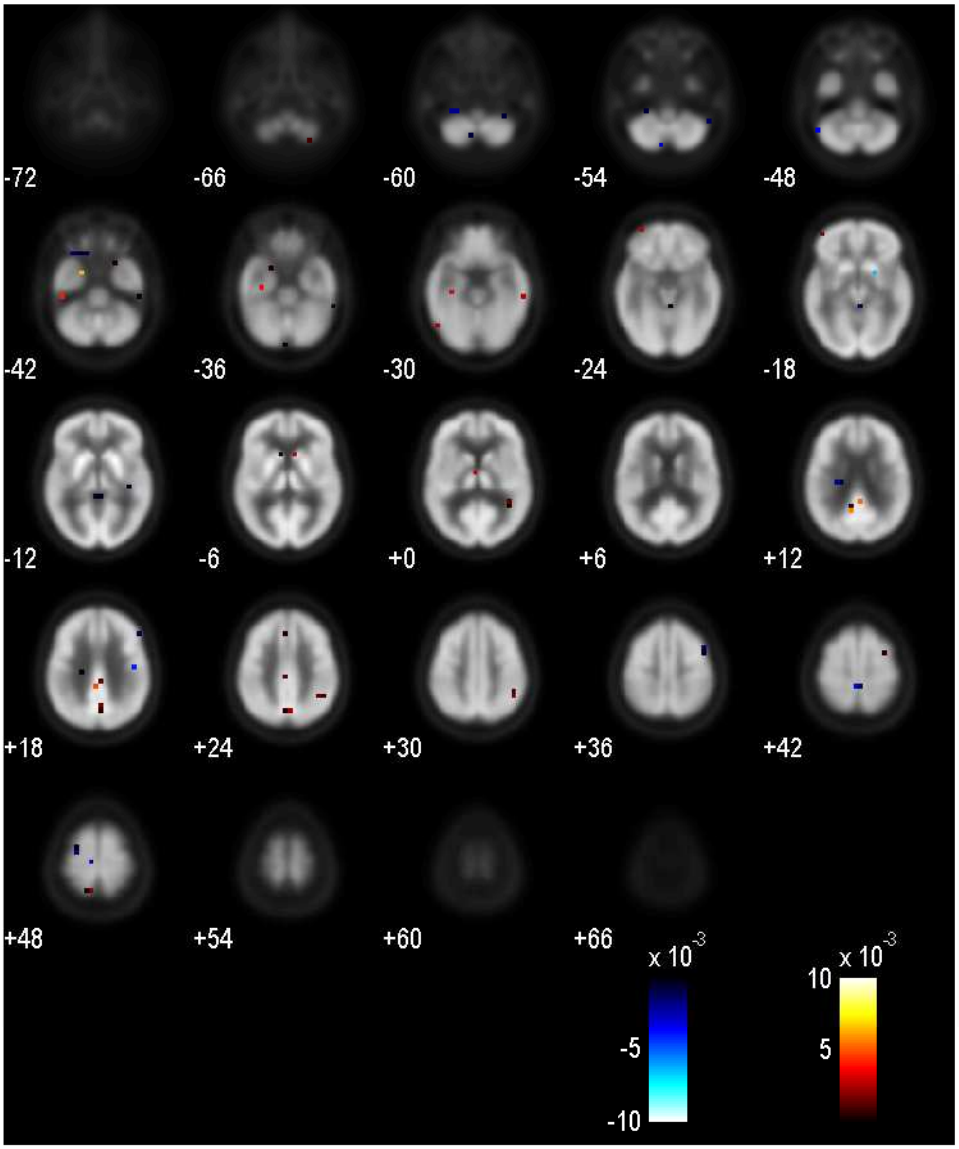

Figure 2.5: Clusters of voxels identified using our approach for the ADNI data.

Secondly, we investigate the voxels that are selected by our method. We use five-fold cross-validation to the full data set to choose the optimal set of tuning parameters. The identified clusters of voxels ($\hat{\beta}(u, v, w) \neq 0$) are shown on selected axial slices in Figure 2.5, which are presented from the bottom of the brain to the

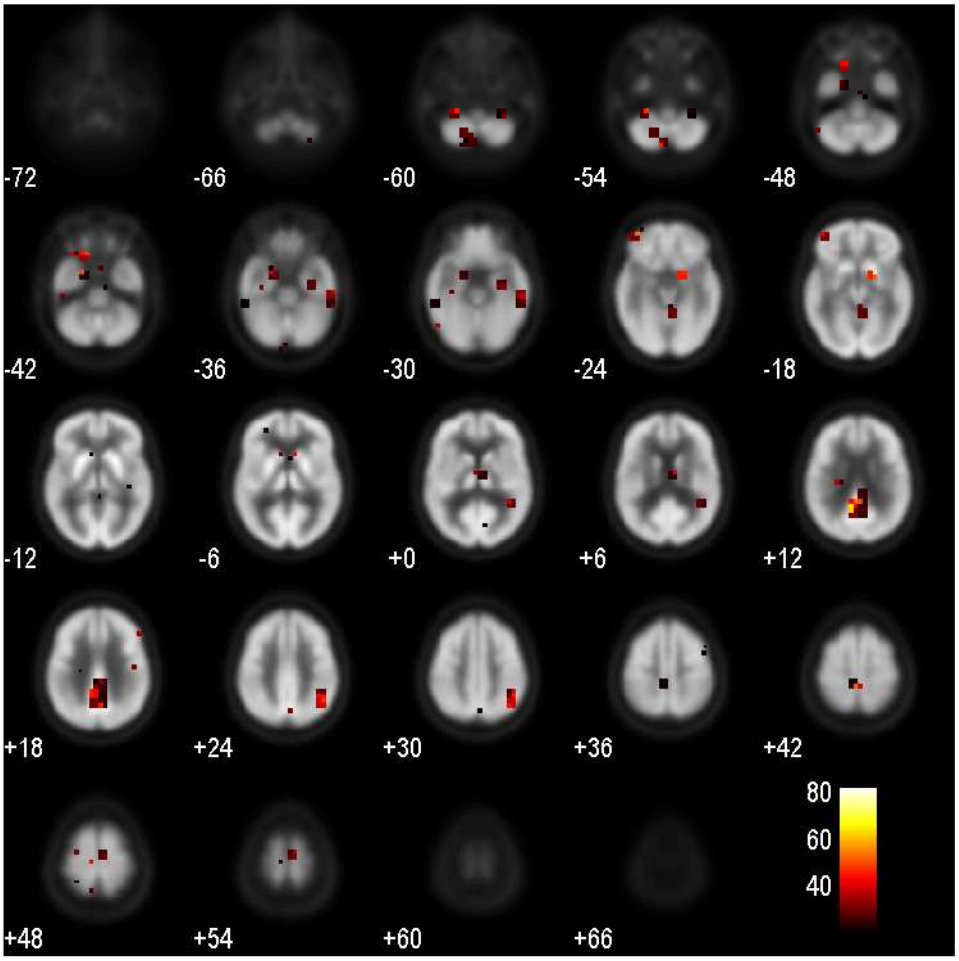

Figure 2.6: Bootstrap inclusion frequencies over 100 bootstrap samples.

top. The clusters of voxels with hot colors show a positive association to prediction of MMSE scores, whereas those with cold colors show a negative association. Each small square represents a small cluster of voxels. To assess the significance of the selected voxels, similar to what we have done in simulation studies, we permute the response variable, i.e., the MMSE score, 200 times. It turns out that 95.3% of the selected voxels are significant at the 5% level. In addition to this pointwise testing, we also consider the global test described by Nichols and Holmes (2001), which provides a

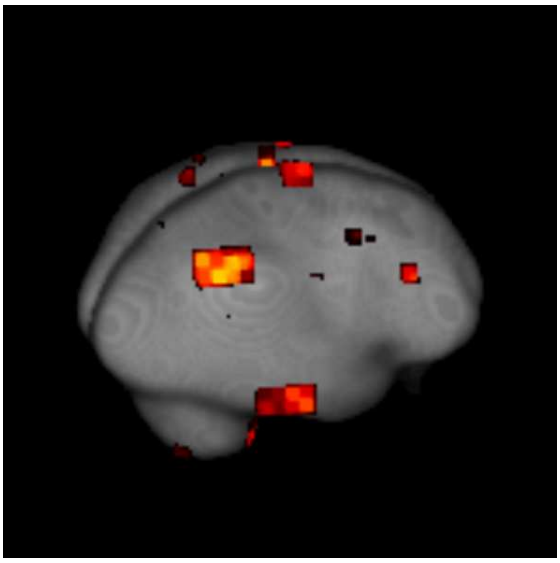

Figure 2.7: Locations of frequently selected voxels in the 3D sagittal view.

way to control the family-wise error rate by comparing $\hat{\beta}(t_j)$ to a "maximal statistic". Due to the computational cost, we also perform 200 permutations, and it turns out that only 15.6% of the selected voxels are significant at the 5% level, which is more conservative than the pointwise testing procedure. To further evaluate the stability of the selection, we generate 100 bootstrap samples and for each bootstrap sample, we apply our method including the tuning parameter selection via five-fold cross-validation. Similar approaches have also been employed by other researchers, such as Sauerbrei and Schumacher (1992), Royston and Sauerbrei (2008) and Meishausen and Bühlmann (2010). To summarize the results, we count the number of times that each voxel is selected over 100 bootstrap samples, and denote it as the bootstrap inclusion frequency (BIF). The voxel BIFs are presented in Figure 2.6. The locations of these more frequently selected voxels are also presented in the 3D sagittal view in Figure 2.7 for the ease of understanding. It can be seen that the highly selected brain regions agree well with the results in Figure 2.5. We note that the clusters of voxels identified in our analysis shown in Figures 2.5 and 2.6 reveal high associations of the expected anatomical regions with the cognitive deficits. For example, the orange ones on slices "+12" and "+18" in Figure 2.5 and the big cluster on the same slices

in Figure 2.6 indicate that the posterior cingulate/precuneus cortex is significantly related to cognitive impairment; the blue ones on slices "-60", "-54" and "-48" in Figure 2.5 and the clusters on the same slices in Figure 2.6 suggest that the medial temporal/hippocampal cortex is also closely involved; the red ones on slices "-42", "-36" and "-30" in Figure 2.5 and the corresponding clusters on the same slices in Figure 2.6 correspond to the lateral temporal cortex. Many studies have demonstrated that the most prominent metabolic abnormalities are found in these regions (see for example Mueller et al. (2005)). In our study, we have particularly found the most predictive voxels of the cognitive impairment in these regions. Other involved regions include the superior lateral parietal cortex and the frontal cortex, which are all known to be related to the progression of Alzheimer's disease.

2.6 Discussion

In this chapter, we propose a highly effective Haar wavelet-based regularization approach that can be easily applied to analyze multi-dimensional functional data. Analysis of the PET imaging data demonstrates that our approach is useful in finding brain subregions that are most responsible for cognitive impairment in elderly people. It has great potential to efficiently assist the diagnosis of disease in neuroimaging studies, yielding easily interpretable results. Our approach is also computationally fast because of the implementation of the coordinate descent algorithm with the MATLAB glmnet package (available at http://www-stat.stanford.edu/ tibs/glmnet-matlab/). We should note that another practical favor of our approach is that the wavelet transform itself can reduce the large volume of brain imaging data. As a result, we can then apply the proposed approach on reduced data sets. In such situations, although the resolution of original PET images is decreased, the results remain largely the same since related subregions are usually not comprised of a single voxel but a cluster of voxels.

CHAPTER III

Classification of PET Images using Regularized 3D Functional Logistic Regression

3.1 Introduction

The clinical diagnosis of Alzheimer's disease (AD), the most common cause of dementia, uses a variety of tests including patient's family history, physical examination, mini-mental state exam and neuroimaging. Recently, functional neuroimaging technologies, such as single photon emission computed tomography (SPECT) and positron emission tomography (PET), are rapidly becoming powerful tools in the diagnosis of AD since these technologies have made it possible to reveal pathophysiological changes before irreversible anatomical changes are present. For example, ^{18}F-Fluorodeoxyglucose (FDG) is a widely used radioactive tracer in PET imaging, and FDG PET provides useful information about the cerebral glucose metabolic rate. Studies have demonstrated reduced glucose metabolism in a small number of brain regions such as the temporal and parietal lobes in AD patients comparing to normal subjects (Hoffman et al., 2000b; Langbaum et al., 2009). As such difference becomes noticeable, researchers are increasingly interest in distinguishing AD patients from normal subjects by utilizing their brain images. As a non-negligible complementary way in the diagnosis of AD, PET imaging has high specificity and sensitivity, even a

long period before full-blown dementia is developed.

A large number of brain imaging studies have been performed in patients with AD and its prodromal stage, mild cognitive impairment (MCI), in an effort to assist in the early diagnosis of AD (Zuendorf et al., 2003; Higdon et al., 2004; Silveira and Marques, 2010; Vemuri et al., 2008; Dehghan et al., 2011; Bonneville et al., 1998; Illán et al., 2011; Stoeckel and Fung, 2007; Hinrichs et al., 2009; Shen et al., 2011; Casanova et al., 2011). Traditional methods to discriminate between patients with AD (or MCI) and normal control subjects are mostly based on voxel-wise analysis. However, each image contains up to millions of voxels, which can be a major cause of practical limitation. To overcome the curse of dimensionality, dimension reduction techniques have been developed prior to classification. One common way is to group the voxels into anatomical regions and average the voxel values within each region of interest (ROI) without taking into account any heterogeneity among the voxels. Prior knowledge of what specific regions may be correlated to the disease is generally desirable, however, this knowledge is not always available in practice. In order to account for spatial correlation between voxels as well as to reduce the dimension of imaging data, principal component analysis (PCA) has been performed, which reduces the feature space to a smaller number of principal components (PCs, called eigenimages) while still preserves the largest portion of variability (Zuendorf et al., 2003; Higdon et al., 2004). The PC scores are then used as predictors in, for example, a logistic regression model. However, each PC is usually comprised of weighted contributions of all voxels within the brain, so PCA is usually less accurate and may blur the true relation between the progression of disease and voxels as what we will show later in simulations.

There has been a growing interest in developing machine learning classification techniques due to the large number of voxels. Support vector machines (SVMs) are one of these techniques used for binary classification (Hastie et al., 2009). They aim to

find the hyperplane that maximizes the distance from the nearest training points while correctly separating two classes. To avoid the curse of dimensionality and improve prediction performance, SVMs are often performed on selected features, including selected voxels (Silveira and Marques, 2010; Vemuri et al., 2008), ROIs (Dehghan et al., 2011), or even PCs (Bonneville et al., 1998; Illán et al., 2011). In an attempt to incorporate the spatial correlation, Stoeckel and Fung (2007) and Hinrichs et al. (2009) presented modified versions of SVMs which are implemented by setting similar weights to neighboring voxels at a very local level. Although SVM-like methods have been shown to achieve high classification accuracy rates, they are not optimized for selecting sensitive and interpretable disease-related brain subregions and fail to provide estimates for the probability that a given subject has the disease or not. To address this issue, logistic regression can be implemented, often with a regularization for variable selection to prevent overfitting. Regularized logistic regression models with the potential of taking into account highly correlated predictors in imaging have been proposed for this purpose. Shen et al. (2011) developed an ROI-based regularized logistic regression model with the elastic net penalty (Zou and Hastie, 2005), a linear combination of lasso and ridge penalties, to classify AD subjects from others. The elastic net penalty enables the selection of groups of highly correlated ROIs. Their method does not consider correlations between voxels within each ROI. Casanova et al. (2011) discussed regularized logistic regression with the elastic net penalty in the context of a large-scale regularization problem in which voxels are used as predictors. For the typical large-p-small-n classification problem, they showed that it can be solved efficiently using the coordinate descent algorithm (Friedman et al., 2010), and it is preferable to SVMs in terms of classification accuracy. However, they did not evaluate the performance in identifying relevant voxels, and in fact they only implemented ridge regression.

We propose a highly effective and computationally efficient regularized functional

logistic regression approach using Haar wavelets, which automatically preserves the spatial information of voxels by viewing each subject's image as a realization of the 3D functional predictor. Functional logistic regression has been used in the classification of functional data. For example, Reiss and Ogden (2010) considered the problem by applying functional PCA to images and demonstrated their method in 2D settings. The general goal of functional data analysis (FDA) is to estimate the coefficient function that describes the association between an outcome and a functional predictor. In this study, we are interested in finding out which voxels are most responsive in the determination of disease status. In particular, we assume that only few brain subregions are predictive of disease status. Properly regularized FDA with Haar wavelet expansion is able to yield a sparse coefficient function estimate (taking value zero at most places in the brain) and also enjoys the advantage of preserving the spatial correlation among voxels. To demonstrate the advantages of the proposed approach, we compare it with other classification methods including regularized voxel-level logistic regression with the elastic net penalty and PCA-based logistic regression.

The data used in this chapter are baseline FDG PET images of 403 subjects from the Alzheimer's Disease Neurological Initiative (ADNI) database, including 95 AD patients, 206 MCI patients and 102 normal controls (NC). The rest of this chapter is organized as follows. We present the proposed approach and also describe two other classification methods in Section 3.2. Numerical results for analyzing the simulated and real data sets are presented in Section 3.3. Final conclusions are provided in Section 3.4.

3.2 Materials and Methods

In this section, we present three regression-based methods for classification of brain images and briefly introduce the data set we use in this work. In particular, we describe how the proposed approach is applied in the functional regression framework

for analyzing brain images, and also explain in detail why we choose to use Haar wavelets.

3.2.1 Haar wavelet based regularized functional logistic regression (HWRFLR)

Logistic regression is commonly used for a binary response variable Y. Functional logistic regression is developed to relate the response variable Y to a functional predictor. Here we treat each subject's 3D brain image as a functional predictor $X_i(u, v, w)$. Suppose Y_i takes values either 0 or 1, indicating the disease status of subject i. We fit the following 3D functional logistic regression model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \iiint X_i(u,v,w)\beta(u,v,w)\,dudvdw, i = 1, ..., n, \tag{3.1}$$

where $\pi_i = P(Y_i = 1|X_i)$ for subject i and $\beta(u, v, w)$ is the 3D regression coefficient function. In this study, we are particularly interested in the assumption that $\beta(u, v, w) = 0$ over large regions, with potential discontinuities of β allowed.

Choosing proper basis functions to represent β in the above regression model is a critical step. Among a variety of basis functions, we choose 3D Haar wavelets to decompose β owing to the following desirable properties. First, the use of Haar wavelets provides a way of overcoming the issue of multicollinearity caused by large spatial correlation among neighboring voxels. Haar wavelets consist of piece-wise constant functions. Our estimation procedure tends to estimate β as a cluster of neighboring voxels instead of a single voxel to be zero or nonzero altogether. Second, as mentioned earlier, we assume that only a few brain subregions are predictive, implying sparsity of the coefficient function. Exact zero regions can be obtained by the sparsity of wavelet coefficients. Third, Haar wavelets can be applied as a signal compression technique. They provide a good approximation of the original function

with only a subset of nonzero wavelet coefficients, which can be achieved by zeroing out the wavelet coefficients that are smaller than a pre-specified threshold value. The dimensionality can thus be reduced if we only consider the nonzero subset.

3D Haar wavelets can be obtained by tensor products of 1D Haar wavelets. For simplicity, we assume that $0 \leq u, v, w \leq 1$. 1D Haar wavelets can be constructed from a mother wavelet function and a scaling function. The mother wavelet function $\psi(t)$ is given by

$$\psi(t) = \begin{cases} 1 & \text{if } 0 \leq t < 1/2, \\ -1 & \text{if } 1/2 \leq t < 1, \\ 0 & \text{otherwise;} \end{cases}$$

and the scaling function $\phi(t)$ is given by

$$\phi(t) = \begin{cases} 1 & \text{if } 0 \leq t < 1, \\ 0 & \text{otherwise.} \end{cases}$$

All 1D Haar wavelets are obtained as translated and dilated versions of the above functions:

$$\begin{aligned} \psi_{j,k}(t) &= \sqrt{2^j}\psi(2^j t - k), \\ \phi_{j,k}(t) &= \sqrt{2^j}\phi(2^j t - k), \end{aligned}$$

where $j = 0, 1, \cdots$ and $k = 0, 1, \cdots, 2^j - 1$. The index j refers to dilations and k refers to translations and $\sqrt{2}$ is the normalizing factor. It can be seen that these basis functions are orthogonal to each other, and the support becomes smaller as j increases. The functions $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ are usually referred to as averaging and differencing operations respectively. Let us now consider tensor products of three elements with each of them being $\phi(t)$ or $\psi(t)$. The total number of different combinations is $2^3 = 8$. The 3D scaling function is the tensor product of three 1D

scaling functions $\phi(t)$. 3D mother wavelet functions are the remaining seven tensor products considering all cross-spatial horizontal, vertical and diagonal directions. 3D Haar wavelets are generated as adapted translations and dilations of these functions. For more details about the construction of 3D Haar wavelets, see Muraki (1992).

We can now decompose $X_i(u, v, w)$ and $\beta(u, v, w)$ by the same set of 3D Haar wavelet basis functions, denoted by $B(u, v, w)$, as follows

$$X_i(u, v, w) = C_i^T B(u, v, w), \quad \beta(u, v, w) = B(u, v, w)^T \eta, \tag{3.2}$$

where C_i is the *known* wavelet coefficient vector of X_i and η is the *unknown* coefficient vector of β. Then by the orthogonality of wavelet basis functions, the 3D functional logistic regression reduces to the following multiple logistic regression by plugging (3.2) into (3.1):

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + C_i^T \eta. \tag{3.3}$$

Once an estimator of η is obtained from (3.3), an estimator of β can be obtained from (3.2).

It should be noted that the wavelet expansion of a given function is determined by the coarsest and finest level of decomposition. In practice, we only observe $X(u, v, w)$ discretely, e.g. at a finite number of uniformly distributed voxels in a cube. Representing observed $X(u, v, w)$ by a set of wavelet coefficients is called the discrete wavelet transform. In this case, the finest level is always given as the operations on adjacent voxels, and thus only the coarsest level needs to be determined, which is refereed to as the level of decomposition in this work.

The estimation of η in (3.3) is accomplished by fitting the model via a penalized

maximum log-likelihood:

$$\max_{\eta} \frac{1}{n} \sum_{i=1}^{n} \{Y_i \log\, \pi_i + (1 - Y_i) \log\, (1 - \pi_i)\} - \lambda \|\eta\|_1, \tag{3.4}$$

where $\|\cdot\|_1$ denote the ℓ_1 norm. Such a penalty is called the lasso penalty (Tibshirani, 1996) that forces many estimated coefficients to be exactly zero. The constant $\lambda \geq 0$ is a tuning parameter that determines how much shrinkage is applied to the vector η. This regularized logistic regression problem can be efficiently solved by the coordinate descent algorithm (Friedman et al., 2010). The estimation of $\beta(u, v, w)$ is then obtained by

$$\hat{\beta}(u, v, w) = B(u, v, w)^T \hat{\eta}.$$

Note that when the wavelet transform is performed under different levels of decomposition, the obtained $\hat{\eta}$ is different, and as a result, the estimator $\hat{\beta}(u, v, w)$ would be different. We set the level of decomposition as another tuning parameter in addition to λ in (3.4), and their optimal values will be determined by certain criterion using a data driven approach.

3.2.2 Elastic net regularized logistic regression (EN-RLR)

The elastic net is considered a generalized version of lasso which encourages a grouping effect by allowing strongly correlated predictors to be in or out of the model together. It also enjoys the computational advantages of lasso. Here we evaluate the performance of this method at the voxel level rather than the ROI level (Shen et al., 2011), which is given by

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \sum_{u,v,w} X_i(u, v, w)\beta(u, v, w), \quad i = 1, \ldots, n, \tag{3.5}$$

where (u, v, w) are integers indicating the location of corresponding voxel. Denote the total number of voxels by p. Since $p \gg n$, regularization is needed to prevent overfitting. The elastic net method maximizes the following regularized log likelihood function

$$\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \{Y_i \log\, \pi_i + (1 - Y_i) \log\, (1 - \pi_i)\} - \lambda P_\alpha(\beta), \tag{3.6}$$

where $P_\alpha(\beta) = \sum_{u,v,w} \{\alpha|\beta(u, v, w)| + (1 - \alpha)\beta(u, v, w)^2\}$. It can be seen that the penalty $\lambda P_\alpha(\beta)$ is a mixture of ℓ_1 and ℓ_2 penalties, and when $\alpha = 1$, (3.6) is simplified to the lasso problem. We set both λ and α as tuning parameters, whereas in Casanova et al. (2011), α is set to be zero to enforce the ℓ_2 penalty, resulting a ridge regression. Note that (3.4) is for the wavelet transformed images whereas (3.6) is for the original images.

3.2.3 Principal component based logistic regression (PC-LR)

PCA is a widely used tool for dimension reduction. It projects the original images onto the eigenspace such that the variance of the projection along each component, the so-called principal component (PC), is maximized (Zuendorf et al., 2003). Each PC is referred to as an eigenimage. As most of the variability of images are captured by a small number of PCs, we retain the first few PCs with greater variances. The associated PC scores are treated as predictors in the logistic regression model. The original coefficient function β can be obtained by the inverse transform of the coefficients of PC scores. The number of PCs used in the regression may affect the classification performance. Thus in this work, we treat the number of PCs as a tuning parameter.

3.2.4 PET imaging data

PET imaging data analyzed in this chapter were obtained from the ADNI database (adni.loni.ucla.edu). The ADNI project was launched in 2003 by the National Insti-

tute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers for disease progression in very early AD is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada.

Detailed information about how FDG-PET images were acquired is available on the ADNI website (http://www.loni.ucla.edu/ADNI/Data/ADNI Data.shtml). The processing steps can be summarized as follows. First, six five-minute frame scans were acquired 30-60 min after injecting FDG to the participants. These frame scans were co-registered to the first frame and then averaged to create a single image. After this step, the co-registered, averaged PET images were reoriented into a standard $160 \times 160 \times 96$ voxel image grid with 1.5 mm cubic voxels and the anterior-posterior axis of the subject is parallel to the anterior commissure-posterior commissure (AC-PC) plane. Finally, smoothing is performed to produce a uniform resolution. The data set used in the present work consists of 403 participants' baseline scans, including 102 NC participants, 206 MCI participants, and 95 AD participants. To reduce the dimensionality, we set the values of voxels outside the brain and in the ventricles to zero and exclude the columns whose elements are all zero in the wavelet coefficient matrix obtained after applying 3D Haar wavelet transform to the images. Typically,

the dimensionality can be reduced to about 700,000 from more than two million, which is a significant decrease.

3.3 Numerical Results

The numerical experiments in this section are performed by Matlab 2011b. The lasso and elastic net problems are solved by the glmnet package for Matlab (available at http://www-stat.stanford.edu/ tibs/glmnet-matlab/). The wavelet transforms are performed by the Matlab Wavelet Toolbox.

3.3.1 Simulations

We conduct simulation studies to evaluate the performance of the three classification methods. The images, covariates in the logistic regression, are obtained from the ADNI data set. For illustrative purposes, we extract the same 160×160 axial slice from each subject and use it as the 2D functional covariate $X_i(u, v)$. For a given coefficient function β, we randomly generate the response variable Y_i from the following 2D model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \iint X_i(u,v)\beta(u,v)\,dudv, i = 1, \ldots, n. \tag{3.7}$$

The regression coefficient function $\beta(u, v)$ is chosen to be nonzero at two small round regions, see Figure 3.2(a). The scale of $\beta(u, v)$, together with the intercept β_0, is adjusted to achieve a Bayes error rate (Fukunaga, 1990) around 0.15. To mimic the sample sizes of AD, MCI and NC in the ADNI data set, we consider two case-control ratios, the ratio of the number of occurrence ($Y_i = 1$) to the number of non-occurrence ($Y_i = 0$), at $r = 1 : 1$ and $r = 1 : 2$. Hence we randomly choose 200 and 300 subjects from the ADNI data set, respectively, for their 2D images. Such designed simulation study keeps the original spatial correlation structure of the ADNI FDG PET images,

whereas the disease status is randomly generated with the original disease status completely ignored.

The selection of tuning parameters is involved in all three methods. We consider a variety of criteria for determining optimal tuning parameters, including cross-validated deviance (CV-DEV), cross-validated misclassification error rates (CV-MER), cross-validated area under the ROC curve (CV-AUC), AIC and BIC, where the use of CV-DEV and CV-MER are discussed in Friedman et al. (2010), and CV-AUC criterion, specially designed for optimizing the classification performance for binary outcomes, is discussed in Jiang et al. (2011). AIC and BIC are criteria that penalize the number of free parameters, which are common for variable selection in high-dimensional models. Note that in the calculation of AIC and BIC, the degrees of freedom (df) need to be determined. An unbiased estimate of df when only ℓ_1 penalty is used is the number of nonzero coefficients in the model (Zou et al., 2007), while an unbiased estimate of df is derived as the trace of the modified hat matrix, when a mixture of penalties is presented, see Zou (2005) for details. In PC-LR, df is estimated as the number of PCs used in the model.

We apply a ten-fold cross validation to evaluate prediction accuracy. Specifically, each simulated data set is randomly partitioned into ten folds. Among them, nine folds are used as the training set to fit model (3.7) by each of the optimal tuning parameter selection criteria; the remaining fold is used as a test set to calculate the predicted probability $\hat{\pi}_i$ for each test observation. The procedure is repeated ten times with each of the ten folds used exactly once as the test set. For predictions based on each cut-off value of $\hat{\pi}_i$, we compute sensitivity and specificity and then construct the empirical ROC curves by changing the cut-off point of $\hat{\pi}_i$. Area under the ROC curve (AUC) is calculated to provide an overall measure of the discriminative ability of each of the three classification models. The procedure is repeated 100 times by generating 100 independent sets of binary response variables. The average ROC curves with

average AUCs in 10 scenarios are presented in Figure 3.1.

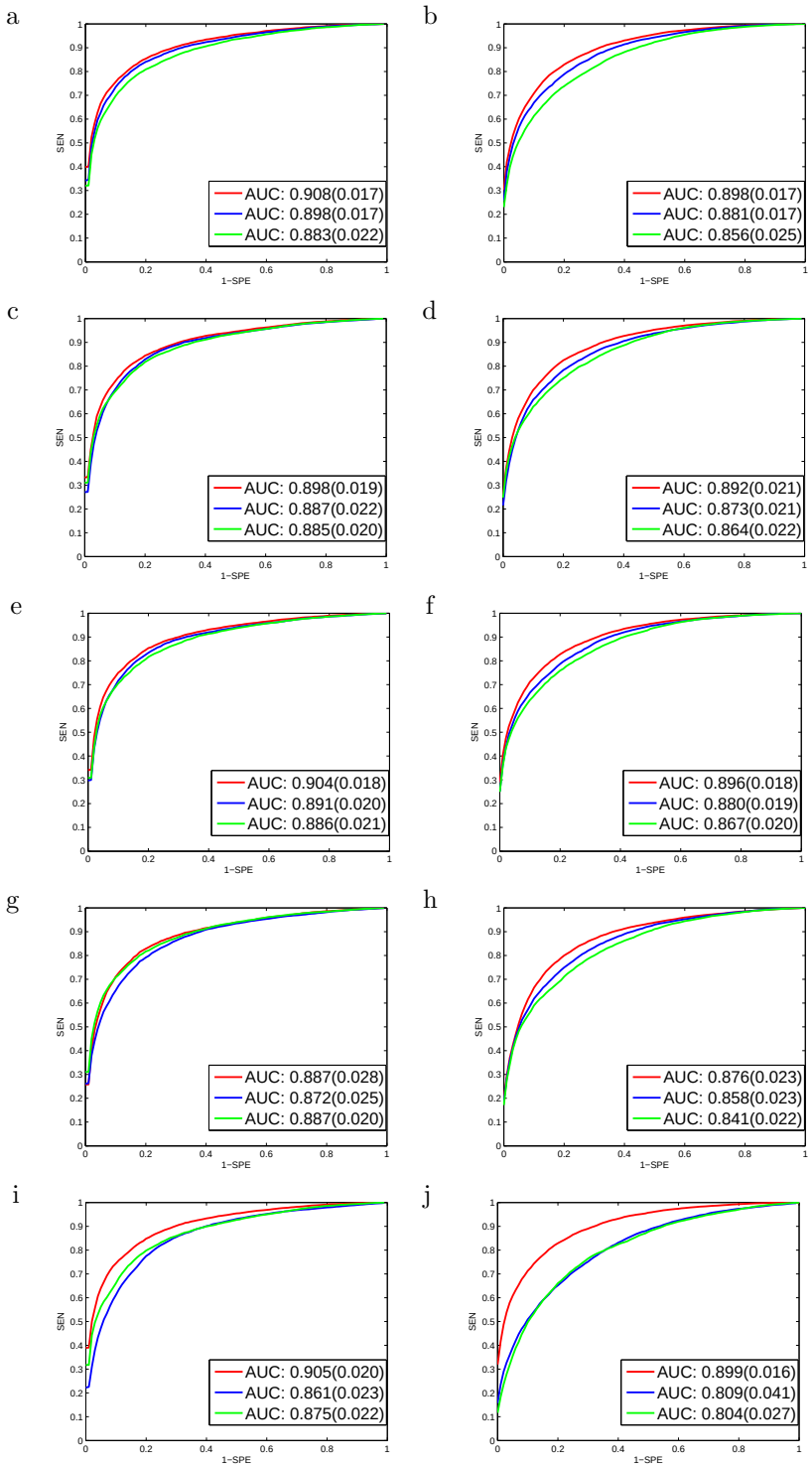

Figure 3.1: Average ROC curves and their average AUCs (standard errors) for three classification methods over 100 simulations (red curve: HW-RFLR; blue curve: EN-RLR; green curve: PC-LR). The left panel is for $r = 1 : 1$, the right panel is for $r = 1 : 2$. The selection criterion used in (a) and (b) is CV-DEV, (c) and (d) is CV-MCR, (e) and (f) is CV-AUC, (g) and (h) is AIC, (i) and (j) is BIC.

In each scenario, two-sample t-tests are performed to test the difference between AUC obtained from the proposed approach HW-RFLR and AUCs from other two methods, respectively. The differences in 9 out of 10 scenarios (except scenario g in Figure 3.1) are statistically significant at the significance level of 0.0025 after the Bonferroni correction, showing that the proposed approach HW-RFLR consistently dominates the other two methods. In general, our simulations indicate that the proposed HW-RFLR approach can achieve higher classification accuracy than EN-RLR and PC-LR for both balanced (i.e. $r = 1 : 1$) and unbalanced (i.e. $r = 1 : 2$) case-control data.

In addition to the classification performance, we also assess the performance in identifying nonzero regions of $\beta(u, v)$ on the 160×160 grid. To this end, we fit model (3.7) with the optimal tuning parameters selected via one of the five criteria mentioned above. Figure 3.2 shows the median estimates of $\beta(u, v)$ for the 100 repetitions under the selection criterion CV-AUC which is robust to outliers. The results of $r = 1 : 1$ and $r = 1 : 2$ are similar, so we only present the results of $r = 1 : 1$. From Figure 3.2 we see that HW-RFLR and EN-LR methods not only yield sparse estimates of $\beta(u, v)$, but also correctly pick up the two nonzero regions of $\beta(u, v)$, whereas PC-LR method completely fails to recognize the true nonzero regions.

3.3.2 ADNI FDG PET image analysis

In this subsection, we apply all three methods described in Section 3.2 to the ADNI FDG PET imaging data for discriminating AD from NC, AD from MCI, and MCI from NC. We treat each PET image as a realization of the 3D functional predictor and then fit 3D functional logistic regression model (3.1) for each pairwise classification. In each comparison, $Y = 1$ indicates a more severe state. Similar to simulation studies, we assess the classification performance using the ten-fold cross validation. To examine the overall discriminative power, we plot the cross-validated ROC curves

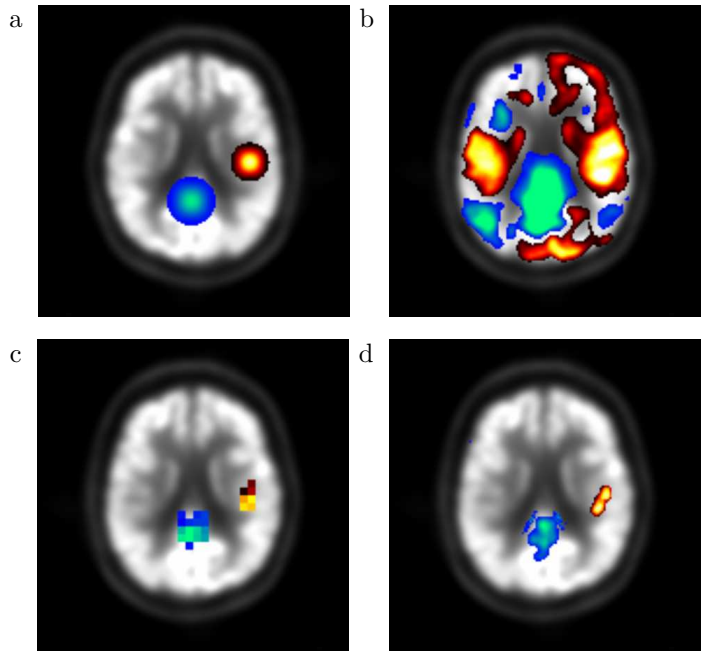

Figure 3.2: Comparison of median of $\hat{\beta}(u, v)$ using three methods over 100 replications for the simulated data sets under the case of $r = 1 : 1$. (a) true $\beta(u, v)$, (b) median of $\hat{\beta}_{PC-LR}(u, v)$, (c) median of $\hat{\beta}_{HW-RFLR}(u, v)$, (d) median of $\hat{\beta}_{EN-RLR}(u, v)$.

in Figure 3.3. Corresponding AUCs are also provided. As the asymptotic behaviors of the lasso estimator still remains an open problem in very high-dimensional settings, the confidence intervals of the reported AUCs are not provided here. We use CV-AUC and CV-DEV as the criteria to select the tuning parameters, as CV-AUC is the most intuitive approach for binary classification and CV-DEV provides highest AUCs in simulation studies. It can be seen that all three methods work well in discriminating AD from NC. The proposed HW-RFLR performs best for the more difficult classification tasks of MCI vs NC and AD vs MCI. Moreover, the proposed HW-RFLR is the most computationally efficient approach. The estimated regression coefficient function for each classification by the proposed HW-RFLR approach with CV-AUC as the tuning parameter selection criterion are given in Figure 3.4 (the results from CV-DEV are almost identical, thus omitted here). The voxels with cold colors are negatively associated with the more severe disease state, whereas the voxels with warm colors indicate a positive relationship. We can see from Figure 3.4 that

some subregions are found to be predictive in all three classifications, such as posterior cingulate and precuneus, which have been found to be the most discriminative subregions in several ROI-based analyses (Langbaum et al., 2009; Rabinovici et al., 2010; Minoshima et al., 1997a).

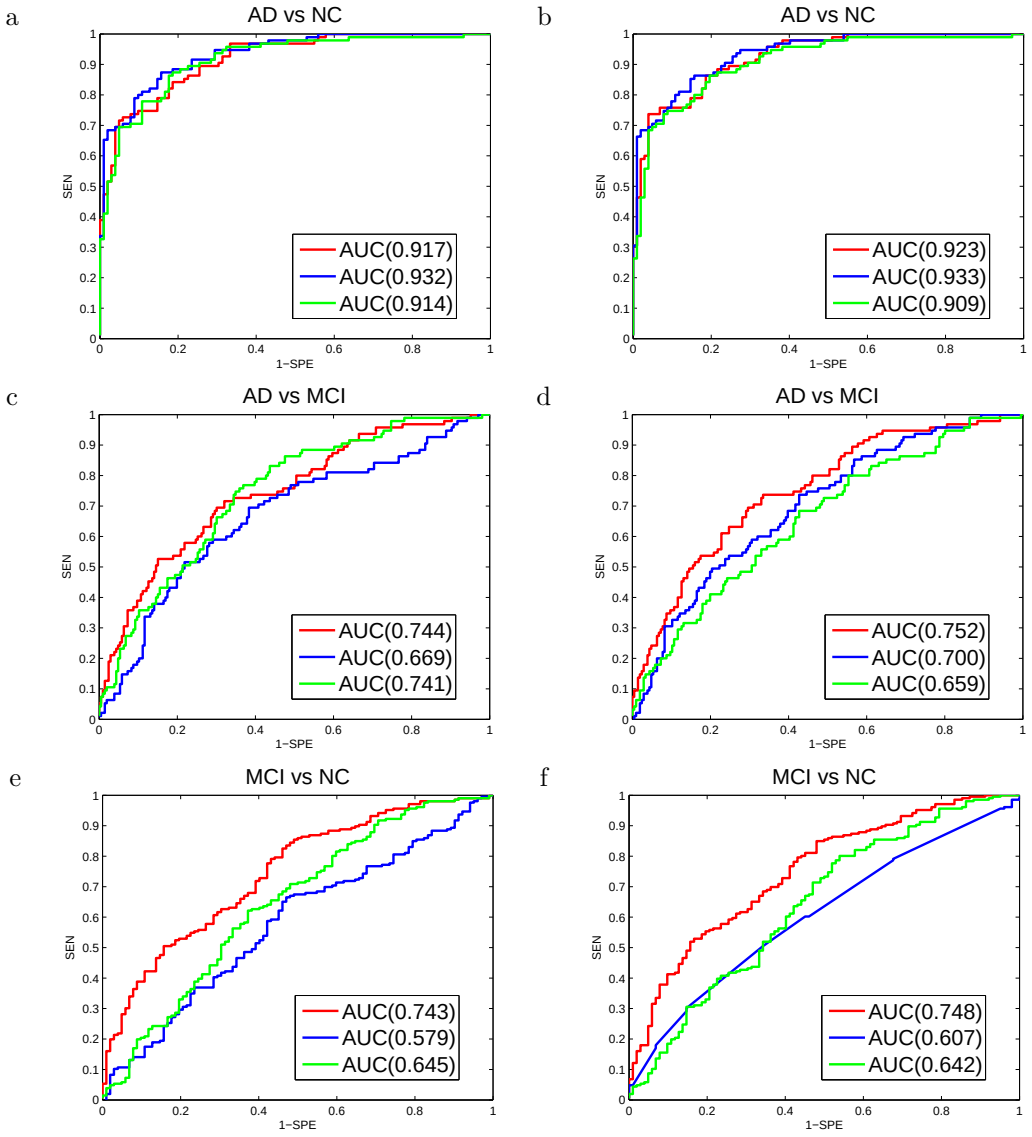

Figure 3.3: ROC curves and their AUC for three classification methods under CV-AUC and CV-DEV for pairwise classification (red curve: HW-RFLR; blue curve: EN-RLR; green curve: PC-LR). The left panel is under CV-AUC, and the right panel is under CV-DEV.

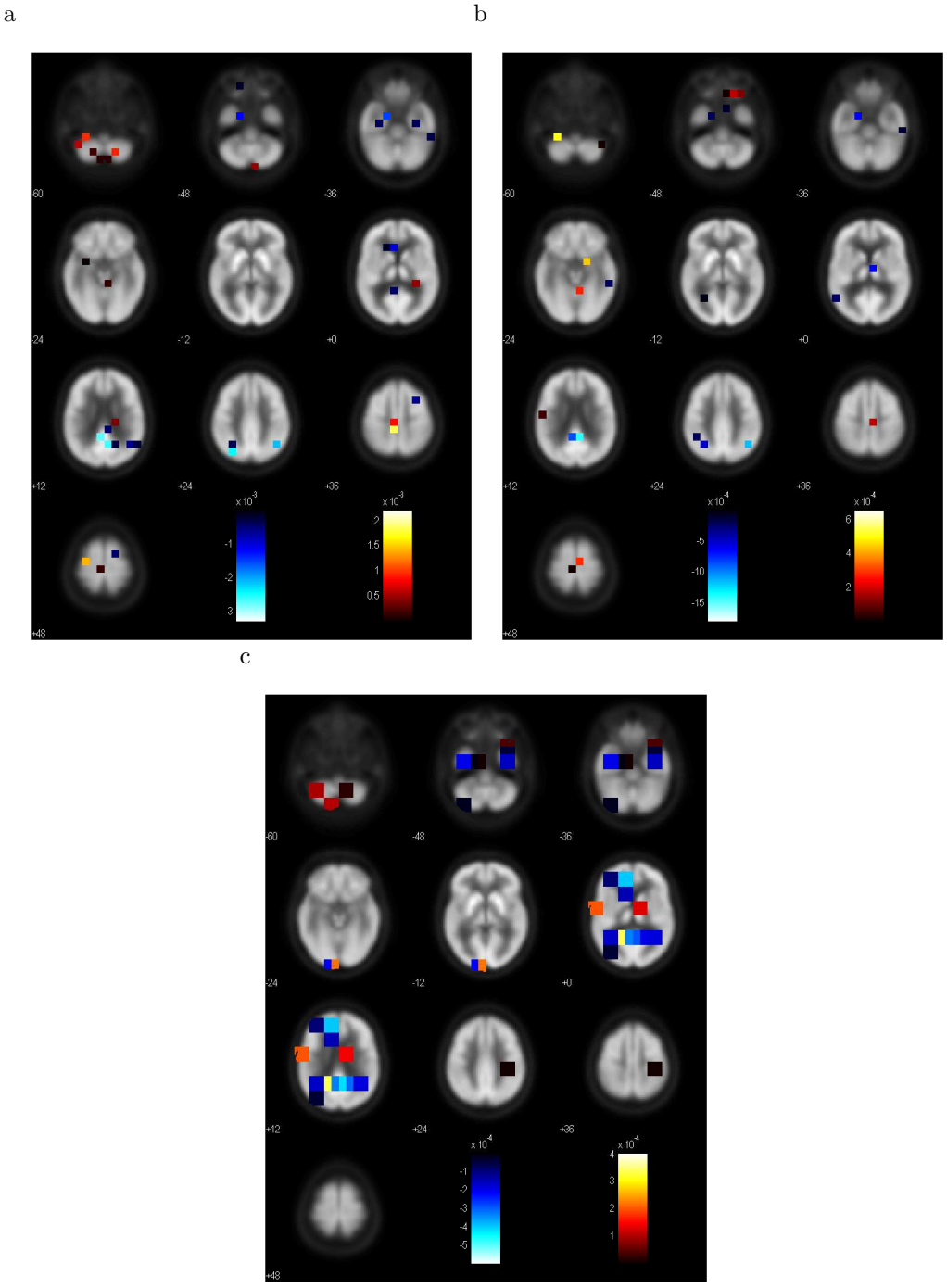

Figure 3.4: Voxels identified using the proposed HW-RFLR under CV-AUC criterion for pairwise classification. (a): AD vs NC; (b) AD vs MCI; (c) MCI vs NC.

3.4 Discussion

In this chapter, we describe a Haar wavelet-based approach for classifying brain images in the framework of 3D functional data analysis. This approach is demonstrated to not only achieve high classification accuracy, but also be more likely to identify the most responsive clusters of voxels. The proposed HW-RFLR method does not impose smoothness requirement on the regression coefficient function, thus has the potential to identify the boundaries of truly predictive subregions. Our numerical results demonstrate that the proposed HW-RFLR can achieve higher classification accuracy than other methods. It should be noted that although many previous studies reported classification accuracy rates using FDG PET imaging data, most of them did not perform the selection of voxels. We emphasize that the proposed HW-RFLR method integrates voxel selection into the estimation procedure, which is useful when only few brain subregions are related to the disease status. The proportional odds model can also be considered to address the classification problems in this work by treating each disease status as a class, but the limitation of this model should be noted. The proposed HW-RFLR method is also computationally efficient partly due to the fact that Haar wavelets can further compress the data by thresholding the absolute value of wavelet coefficients without losing the ability of preserving spatial correlations among voxels.

CHAPTER IV

Classification of PET Images using Regularized 3D Multiple Functional Logistic Regression

4.1 Introduction

With the rapid development of neuroimaging technologies, large-scale brain imaging data are acquired more frequently nowadays for research and clinical applications. For example, positron emission tomography (PET) imaging, measuring the metabolic activity in the brain, is being used increasingly to better understand the progression of neurodegenerative diseases, such as Alzheimer's disease (AD) (Hoffman et al., 2000a; Silverman et al., 2001). Although the high resolution scanners can provide more detailed information of the brain pathology, one challenge for analyzing large-scale brain imaging data is the large number of voxels, which can easily go beyond a million, whereas the sample sizes are usually in the order of a few hundred. Traditionally, the analysis of large-scale brain imaging data has relied on massive univariate voxel-wise analyses, where the issue of multiplicity must be addressed, and modeling the spatial correlation is a difficult problem. To overcome these challenges, sparse modeling techniques have been developed in recent years, aiming to find a small set of voxels that are well suited for interpretation and can be used to predict the outcome more accurately.

Sparse modeling techniques are often implemented by introducing sparsity into the model estimation, leading to a parsimonious model by removing irrelevant features. In regression settings, among a variety of sparse modeling techniques, the lasso regularization (Tibshirani, 1996), which minimizes the ℓ_1 norm of regression coefficients, has long been known as a practical approach. Moreover, techniques that take advantage of problem specific information can often lead to higher accuracy. In brain imaging data, it is known that voxels are spatially correlated (Frackowiak et al., 2004). The elastic net (Zou and Hastie, 2005), as a variant of lasso, is a method to account for the correlation among voxels. Casanova et al. (2011) and Janousova et al. (2012) implemented the elastic net penalty in penalized logistic regression models. Incorporating this penalty into the estimation encourages strongly correlated variables to be either in or out of the model together. In addition to the spatial correlation structure, it is also known that voxels can be partitioned into different groups. For example, the cerebral cortex can be divided into a number of regions according to the structure or cytoarchitectonics (Garey, 2006). Regularization via the elastic net is not able to reveal the underlying group structure in its solution and, thus, is not an optimal approach in the situation where identifying the predictive groups is of concern. Liu et al. (2012) applied the tree-guided group lasso (Kim and Xing, 2009) which can account for the hierarchical spatial relationships of the voxels, in order to identify grouped voxels for brain disease classification. Although this method is shown to improve the interpretation of identified voxels, it assumes that the group structure information is unknown. Usually when this information is available, it would be beneficial to incorporate group selection into variable selection. As this information is available in the present work, we propose to account for it using groupwise regularization, with the underlying assumption that only a few brain areas (also referred to as the "groups") are truly predictive of the outcome and within each selected group, maybe only part of it are truly predictive. Given this group-wise and within group sparsity

assumption, the sparse group lasso penalty can be implemented into the estimation (Friedman et al., 2010; Simon et al., 2012). Zhou et al. (2012) considered a similar assumption and developed the fused sparse group lasso method for longitudinal voxel selection, with the sparse group lasso as a special case at one single time point.

Unlike the above-mentioned sparse modeling techniques, we investigate the use of sparse group lasso penalized estimation in the context of functional regression models by viewing each 3D brain image as a 3D functional predictor, where the sparsity is imposed on the coefficients from expanding the regression coefficient function in terms of some basis functions. The reason why we view the image as a 3D functional predictor and impose sparsity of the coefficients is that we intend to preserve the spatial information of the images, which is not accounted for by the above-mentioned sparse modeling techniques while achieving desirable sparse estimation. In functional regression models it is important to choose proper basis functions to represent the functional predictor and the regression coefficient function. Considering both sparsity and spatial correlation among voxels, we choose to apply the 3D Haar wavelet transform on the images due to its attractive properties. These properties are twofold: (i) the discontinuity of Haar wavelet basis functions not only allows spatial localization of the predictive voxels, but also enables sparse estimation of the voxel-level effects; (ii) the averaging and differencing operations from the Haar wavelet transform can alleviate the problem caused by ignoring the correlation among voxels. In presence of the group structure among voxels, we consider a multiple functional regression model which allows us to perform the wavelet transform on each group respectively rather than the whole image in order to define groups easily for the coefficients obtained from the wavelet transform. In this manner, the group-level and within group effects can be estimated in a more flexible way. As sparsity is achieved by incorporating regularization into the model estimation, we call our approach regularized multiple functional regression to distinguish from other existing sparse modeling techniques.

We evaluate the performance of the proposed approach on classification tasks of PET imaging data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. One of the primary goals of ADNI has been to test whether serial PET images can be used to measure the progression of mild cognitive impairment (MCI) and early AD and finally aid the researchers and clinicians to treat the disease in a more effective way. The participants of ADNI were classified into three groups at their initial visits: AD, MCI and Normal Control (NC). Many existing studies of ADNI have been focused on the pairwise classifications among three groups (AD vs NC, AD vs MCI, and MCI vs NC), however, there has been a growing interest in studying the clinical change of MCI patients to assist in the early diagnosis of AD, in particular, to predict the conversion from MCI to AD using the brain imaging data (Misra et al., 2009; Davatzikos et al., 2011; Zhang and Shen, 2012; Eskildsen et al., 2013). During the follow-up (from 6 months up to 84 months), some of the MCI patients have converted to AD, while others have not. They are referred to as MCI converters (MCI-C) and MCI non-converters (MCI-NC), respectively. In the present work, our main objective is to predict the conversion at future time points using the baseline PET imaging data of MCI subjects.

The rest of this chapter is organized as follows. In Section 4.2, we propose the regularized multiple functional regression model. In Section 4.3, we illustrate the performance of the proposed approach through simulation studies and compare it to other methods. We then validate the approach by applying it to the MCI conversion data obtained from the ADNI database. Finally, we conclude with a discussion in Section 4.4.

4.2 Materials and Methods

In this section, we mainly present in detail the proposed regularized multiple functional regression approach that accounts for the group structure of the voxels.

Moreover, we consider an ad hoc approach, namely that of applying the regularized logistic regression on the voxels directly without treating the images as functional observations.

4.2.1 Haar-wavelet-based regularized multiple functional logistic regression (HW-RMFLR)

To preserve the spatial information of the imaging data, we treat each subject's image as a realization of 3D functional predictor $X_i(u, v, w)$, and assume that Y_i takes values either 0 or 1, indicating the disease status of subject i. In Chapter III, we proposed the 3D functional logistic regression model as follows:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \int\int\int X_i(u, v, w)\beta(u, v, w)\, dudvdw \tag{4.1}$$

where $\pi_i = P(Y_i = 1|X_i)$ for subject i and $\beta(u, v, w)$ is the 3D regression coefficient function. For simplicity, we assume that $0 \leq u, v, w \leq 1$.

In order to achieve sparse estimation and identify the correlated voxels together, we consider 3D Haar wavelets as the basis functions to decompose $X_i(u, v, w)$ and $\beta(u, v, w)$. 3D Haar wavelets are tensor products of 1D Haar wavelet basis functions. 1D Haar wavelets are given by a pair of so-called *father* wavelet function $\phi(t)$ and *mother* wavelet function $\psi(t)$, where $\phi(t)$ is given by

$$\phi(t) = \begin{cases} 1 & \text{if } 0 \leq t < 1, \\ 0 & \text{otherwise,} \end{cases}$$

and $\psi(t)$ is given by

$$\psi(t) = \begin{cases} 1 & \text{if } 0 \leq t < 1/2, \\ -1 & \text{if } 1/2 \leq t < 1, \\ 0 & \text{otherwise.} \end{cases}$$

All 1D Haar wavelets are obtained by applying translations and dilations on the above functions:

$$\psi_{j,k}(t) = \sqrt{2^j}\psi(2^j t - k),\ \phi_{j,k}(t) = \sqrt{2^j}\phi(2^j t - k),$$

where $j = 0, 1, \cdots$ and $k = 0, 1, \cdots, 2^j - 1$. Index j refers to dilations whereas index k refers to translations, and $\sqrt{2}$ is the normalizing factor. The functions $\phi_{j,k}(t)$ and $\psi_{j,k}(t)$ are usually considered as averaging and differencing operations respectively. The tensor products are constructed by three elements, each of which can be $\phi(t)$ or $\psi(t)$. The 3D version of the *father* wavelet function $\phi(u, v, w)$ is given by the tensor product of three 1D functions $\phi(u)\phi(v)\phi(w)$. 3D *mother* wavelet functions are designed as the other seven combinations of tensor products accounting for all cross-spatial horizontal, vertical and diagonal directions. The operations such as translations and dilations can be adapted to the 3D functions accordingly. For more details about the construction of 3D Haar wavelets, see Muraki (1992).

Since it is known that the cerebral cortex can be partitioned into a number of predefined regions, a group structure exists among voxels. It is impossible to account for the group structure among voxels if the entire image $X_i(u, v, w)$ is decomposed by one set of 3D Haar wavelets. Hence, we propose the multiple functional regression model which allows us to account for the group structure by applying a wavelet transform to each group. Assuming that the cerebral cortex is divided into G regions, we rewrite model (4.1) into the multiple functional regression model as follows:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{g=1}^{G} \int\int\int X_{i,g}(u, v, w)\beta_g(u, v, w)\, dudvdw. \tag{4.2}$$

In model (4.2), we decompose $X_{i,g}(u, v, w)$ and $\beta_g(u, v, w)$ in region g by the same

set of 3D Haar wavelet basis functions denoted by $B_g(u,v,w)$:

$$X_{i,g}(u,v,w) = C_{i,g}^T B_g(u,v,w),\ \beta_g(u,v,w) = B_g(u,v,w)^T \eta_g. \tag{4.3}$$

Plugging (4.3) into (4.2), the model can be re-expressed as:

$$\begin{aligned} \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \sum_{g=1}^{G} \int\int\int C_{i,g}^T B_g(u,v,w) B_g(u,v,w)^T \eta_g \, dudvdw \\ &= \beta_0 + \sum_{g=1}^{G} C_{i,g}^T \eta_g, \end{aligned} \tag{4.4}$$

where the integral cancels because wavelet basis functions are orthogonal, and $\eta_g,\ g = 1,...,G$ is the wavelet coefficient vector to be estimated in region g.

To account for both group-wise and within-group sparsity of η, $\eta = [\eta_1,...,\eta_G]$, we incorporate the sparse group lasso penalty into the following objective function:

$$\max_{\eta} \frac{1}{n} \sum_{i=1}^{n} \{Y_i \log \pi_i + (1-Y_i) \log(1-\pi_i)\} + \lambda_1 \sum_{g=1}^{G} \sqrt{\omega_g \sum_{j=1}^{p_g} \eta_{gj}^2} + \lambda_2 \sum_{g=1}^{G} \sum_{j=1}^{p_g} |\eta_{gj}|, \tag{4.5}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are tuning parameters and ω_g is a weight coefficient indicating the group size. It can be seen that the penalty is a mixture of ℓ_1 and ℓ_2 penalties, where ℓ_2 penalty identifies the important groups and ℓ_1 penalty eliminates unimportant variables within identified important groups. We implement the state-of-the-art software SLEP 4.1 (Liu et al., 2009) to solve (4.5). After obtaining $\hat{\eta}$, we then derive $\hat{\beta}_g(u,v,w)$ of region g by $B_g(u,v,w)^T \hat{\eta}_g$.

In practice, the 3D functional predictor $X(u,v,w)$ is observed over a fine 3D grid. We apply the 3D Haar wavelet transform to $X_g(u,v,w),\ g = 1,...,G$, respectively. One of the practical considerations is that each brain region has an irregular shape. For the ease of application, we consider the smallest cuboid that contains the region g and set the voxels outside of the region but in the cuboid to zero, then apply

the wavelet transform on the cuboid. After obtaining $\hat{\beta}_g(u,v,w)$, we only record it at the locations (u,v,w) where $X_g(u,v,w)$ is nonzero, and combine the recorded $\hat{\beta}_g(u,v,w)$ into the whole coefficient function $\hat{\beta}(u,v,w)$. In addition to λ_1 and λ_2, we also include the level of decomposition of the 3D Haar wavelet transform as a third tuning parameter. Ideally, the optimal level of decomposition could be different across regions. But in consideration of the computational feasibility in practice, we assume it to be the same.

4.2.2 Sparse group lasso regularized logistic regression (SGL-RLR)

An alternative approach is to estimate β directly from the following logistic regression model:

$$\begin{aligned} \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \sum_{u,v,w} X_i(u,v,w)^T \beta(u,v,w) \\ &= \beta_0 + \sum_{g=1}^{G} \sum_{u,v,w \in R_g} X_{i,g}(u,v,w)^T \beta_g(u,v,w), \end{aligned}$$

where (u,v,w) are integers indicating the location of the voxel in region g, denoted by R_g. Each $\beta_g(u,v,w)$ is treated as a regression parameter and the sparse group lasso penalty is implemented directly on these parameters. It should be noted that it is different from the proposed approach in Section 4.2.1 where we implement the sparse group lasso penalty on the coefficient η. Regression parameters β can be obtained by maximizing the regularized log likelihood function,

$$\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \{Y_i \log\, \pi_i + (1-Y_i) \log\, (1-\pi_i)\} - P_{\lambda_1,\lambda_2}(\beta), \tag{4.6}$$

where $P_{\lambda_1,\lambda_2}(\beta) = \lambda_1 \sum_{g=1}^{G} \sqrt{\omega_g \sum_{u,v,w \in R_g} \beta_g(u,v,w)^2} + \lambda_2 \sum_{g=1}^{G} \sum_{u,v,w \in R_g} |\beta_g(u,v,w)|$ is the sparse group lasso penalty with tuning parameters λ_1 and λ_2, and a weight coefficient ω_g indicating the group size.

A similar approach is considered in Zhou et al. (2012), where the fused sparse group lasso is developed for the longitudinal voxel selection and the sparse group lasso is included for variable selection as a special case. Here we evaluate the performance of the sparse group lasso penalty in the regularized logistic regression and compare it to the proposed approach HW-RMFLR.

4.2.3 PET imaging data

PET imaging data used in this work were obtained from the ADNI database (adni.loni.ucla.edu). The ADNI project was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers for disease progression in very early AD is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada.

Participants in the ADNI study were classified into three groups during their initial visits: AD, MCI and NC based on the baseline diagnosis. The data set used in the present work consists of 203 MCI subjects' baseline FDG PET scans. Details about how PET imaging is performed on each participant can be found on the ADNI website (http://www.loni.ucla.edu/ADNI/Data/ADNI Data.shtml). Briefly, multiple

PET frame scans were acquired to show the brain activity 30-60 minutes after the injection of FDG. These scans were co-registered to the first scan and averaged to a single averaged scan which was then reconstructed on a standard $160 \times 160 \times 96$ voxel grid with 1.5 mm cubic voxels along the anterior commissure-posterior commissure (AC-PC) plane. The final images were obtained after smoothing each of the above-mentioned image to produce a uniform isotropic resolution. PET images used in this work are segmented by Brodmann areas (Garey, 2006), as a result, the voxels in the brain are grouped into 106 Brodmann areas, which constitutes the group structure of the voxels. The voxels not indexed by Brodmann areas are not used in the analyses in this work. Brodmann areas have been widely used in the study of pathological process of AD, and many Brodmann areas are demonstrated to be related to AD (Galton et al., 1999; Querbes et al., 2009; Fouquet et al., 2009).

4.3 Numerical Results

All the numerical experiments in this section are carried out by Matlab 2011b. The wavelet transforms are performed by the Matlab Wavelet Toolbox.

4.3.1 Simulations

In this subsection, we perform simulation studies to validate the proposed approach, and compare it to several other sparse modeling methods. Particularly, we compare the performance in identifying predictive voxels among HW-RMFLR, SGL-RLR, and Haar-wavelet-based regularized functional logistic regression (HW-RFLR) introduced in Chapter III. 2D simulations are performed based on one slice for convenience. For each subject i, we select the same axial slice (the dimension is 160×160) and treat it as $X_i(u, v)$, and fit the following 2D multiple functional logistic regression

model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{g=1}^{G} \int\int X_{i,g}(u,v)\beta_g(u,v)\,dudv. \tag{4.7}$$

We apply the 2D Haar wavelet transform to each region on the slice. Based on the Brodmann areas, voxels on the selected slices can be grouped into 18 regions, i.e. $G = 18$, see Figure 4.1 (a) for the region segmentation on the selected slice. We consider the case of $\beta(u,v)$ where part of the voxels in two regions are assumed to have nonzero effect (see Figure 4.1 (b). For subject i, the disease status Y_i is generated by drawing a random uniform number u on the interval $[0,1]$, and denote $Y_i = 1$ if $\pi_i < u$ and $Y_i = 0$ otherwise. As it is a classification problem, the minimum error achievable is referred to as Bayes error rate (Fukunaga, 1990). We set it about 0.28 in this work, as the prediction of conversion to AD from MCI is known to be a difficult task. In the simulations, we keep the case-to-control ratio around 0.5. Note that the scale of $\beta_g(u,v)$ and the intercept β_0 can be adjusted to satisfy these requirements. We apply the five-fold cross validation to select the optimal set of tuning parameters. We consider two criteria: cross-validated deviance (CV-DEV), and cross-validated area under the ROC curve (CV-AUC). In order to evaluate the performance in identifying zeros and nonzeros of $\beta(u,v)$, we perform the five-fold cross validation on the data set to select the tuning parameters, and then obtain $\hat{\beta}(u,v)$. We generate the binary outcome Y_i 10 times to obtain 10 data sets due to the high computational cost. We repeat the whole procedure for each data set and then obtain the median estimate of $\hat{\beta}(u,v)$. The median estimates derived from all the methods under two criteria are presented in Figure 4.2. We also present the average percentages of correctly identified zeros and nonzeros over 10 simulations in Table 4.1, showing that in general, the proposed approach HW-RMFLR achieves the highest rates in correctly identified nonzeros among three methods, with no sacrifice in identifying zeros.

The average ten-fold cross-validated AUCs of the three methods, however, are

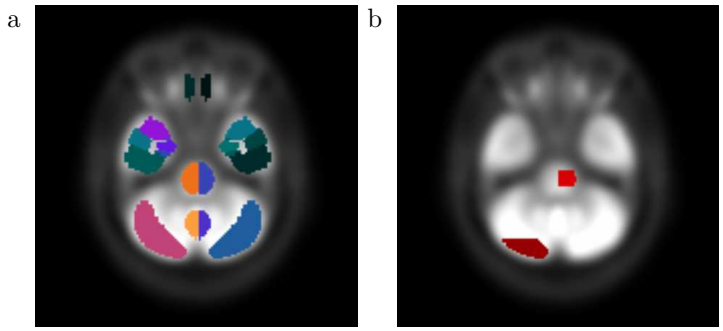

Figure 4.1: (a) Region segmentation on the selected slice; (b) true $\beta(u, v)$.

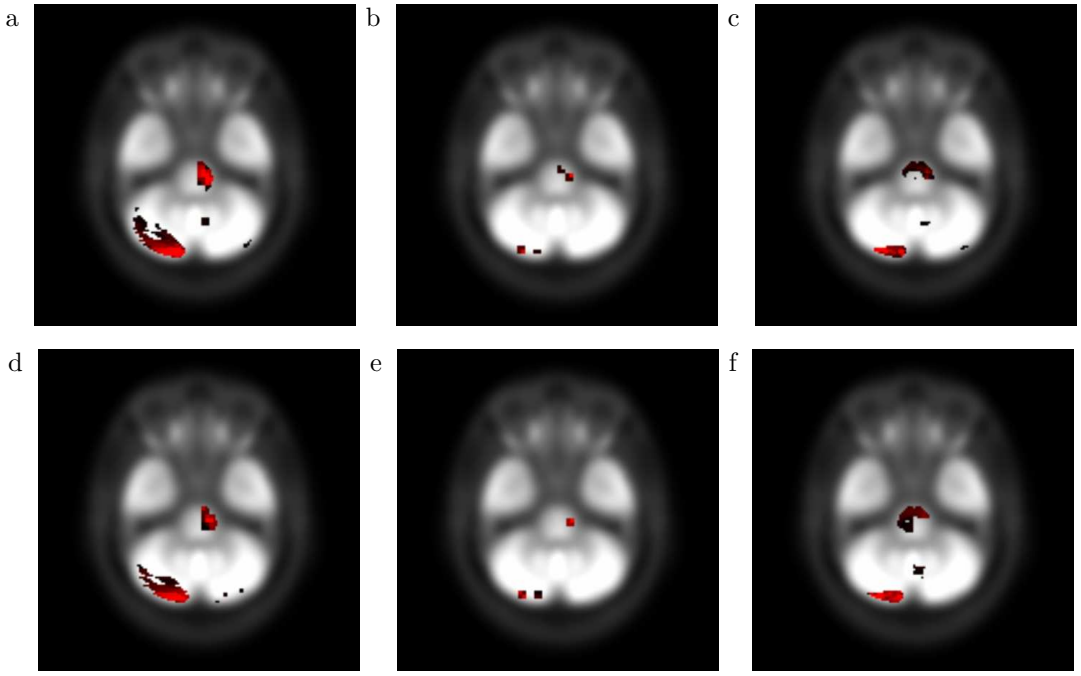

Figure 4.2: Median estimates of $\hat{\beta}(u, v)$ over 10 simulations. (a) HW-RMFLR with CV-DEV; (b) HW-RFLR with CV-DEV; (c) SGL-RLR with CV-DEV; (d) HW-RMFLR with CV-AUC; (e) HW-RFLR with CV-AUC; (f) SGL-RLR with CV-AUC.

quite similar (all of them are about 0.77), showing that they have similar classification performance under this simulation setting.

4.3.2 Predicting MCI-to-AD conversion using ADNI FDG PET images

In this subsection, the three methods considered in Section 4.3.1 are applied to the classification of two types of MCI subjects from the ADNI data set. We also include the principal component based logistic regression (PC-LR) proposed in Chapter III.

Approach	Criterion	Zeros	Nonzeros
HW-RMFLR	CV-DEV	97.9%	74.3%
	CV-AUC	98.5%	67.9%
HW-RFLR	CV-DEV	99.6%	21.2%
	CV-AUC	99.7%	18.9%
SGL-RLR	CV-DEV	97.9%	43.0%
	CV-AUC	98.1%	40.9%

Table 4.1: Average percentages of correctly identified zeros and nonzeros over 10 simulations.

To reduce the computational cost, we reduce the dimension of the images from $160 \times 160 \times 96$ to $80 \times 80 \times 48$ by combining every two neighboring voxels. The reduced dimensional images are also segmented by Brodmann areas correspondingly.

In the ADNI procedure, PET scans and clinical diagnosis were performed at baseline, 6 months, 12 months, 18 months, 24 months, 36 months for MCI subjects. With additional funding, in the form of a Grand Opportunities grant, the ADNI study moved into the ADNI GO study in 2010 for an additional 2-year period. Moreover, while the ADNI GO project continues, ADNI began its third phase in 2011, which is known as ADNI 2, to further identify who may be at risk of developing AD. Therefore, the MCI subjects continued to receive follow-up at 48 months, 60 months, 72 months and 84 months when our data were acquired in March 2013. In this work, we focus on the prediction of MCI-to-AD conversion using baseline scans. The status of AD conversion was observed at each of the follow-up time points. Predicting the conversion has been studied in recent years. Misra et al. (2009) applied a high-dimensional pattern classification method based on regional volumetric features to predict the MCI-to-AD conversion using MRI scans in the ADNI database. As their study was finished in 2008, the average follow-up time of the subjects included in the study was only 15 months, but they reported high predictive performance (AUC=0.77) using the leave-one-out cross validation. Although this result is promising, they only observed

27 MCI-C, which may cause difficulty in comparing with other studies that have been published more recently with longer term follow-ups. Davatzikos et al. (2011) considered MRI imaging data together with many other biomarkers and clinical variables. Although they reported an AUC of 0.734 using the five-fold cross validation, their method is not able to select important regions as a derived measure from the whole brain was used. Zhang and Shen (2012) investigated the prediction of conversion at different time points using both baseline and longitudinal multi-modality data including MRI, PET and cognitive scores. They focused on the longitudinal changes of the brain regions in addition to prediction of the conversion and the subjects included in their work must have all imaging data at five different time points (baseline, 6 months, 12 months, 18 months and 24 months), which reduced the sample size (only 88 subjects were included) as there are many dropouts in the ADNI study. They reported an AUC of 0.768 from the leave-one-out cross validation using both baseline and longitudinal multi-modality data and an AUC of 0.676 using PET imaging data only. Eskildsen et al. (2013) considered the conversion at a number of time points prior to the diagnosis: 6 months, 12 months, 24 months and 36 months, using MRI scans, and further examined the classification between MCI-C and MCI-NC at these time points separately. In their study, subjects who did not convert to AD over the course of the ADNI study were considered as MCI-NC and baseline scans were used for MCI-NC. For those converters, scans at time point T prior to the conversion were used in the analysis at T. A potential drawback of this study is that the images from converters and non-converters were obtained at different time points. For example, the scans used in the analysis at 6 months prior to the conversion consist of baseline scans of MCI-NC and the scans obtained at 6 months prior to conversion for MCI-C. In other words, the images represent two extreme classes, which largely improves the classification accuracy and yields a high AUC of 0.81. This method is not applicable in practice to predict future conversions. To overcome this potential drawback, we

consider the prediction of conversion at multiple time points in a different way and only use MCI subjects' baseline PET imaging data. For example, if we consider the conversion at the time point T, subjects dropping out without conversion before T are excluded from the study; subjects having not converted to AD at T are classified as MCI-NC; and subjects who converted to AD by T are considered as MCI-C. Since there are more dropouts after 48 months, we only consider four time points in this work: 12 months, 24 months, 36 months and 48 months. The number of subjects who exhibited conversion or not at these time points are summarized in Table 4.2. It can be seen that the total number of eligible subjects decreases at later follow-ups. The frequencies of all MCI subjects' follow-up time are presented in Figure 4.3. The maximum follow-up time is 84 months, and the mean follow-up time is 42.4 months. It is also worth noting that a collection of recent findings in prediction of MCI-to-AD conversion are reviewed in Eskildsen et al. (2013), where Misra et al. (2009) and Davatzikos et al. (2011) are among the best results in terms of AUC.

Months	Number of MCI	Number of MCI-C
12 months	186	24
24 months	172	58
36 months	160	77
48 months	137	86

Table 4.2: Summary of conversion at different time points.

We use the ten-fold cross validation to evaluate the performance for the classification tasks at the four time points. Cross-validated ROC curves with AUCs are presented in Figure 4.4. It indicates that the classification performance is slightly increased at later follow-ups. This makes sense since MCI-NC defined in the proposed way are more likely to be real non-converters (i.e. stable MCI) and more different from MCI-C, if they have not converted after a long period. Moreover, in most cases, the proposed approach HW-RMFLR achieves the best classification performance among

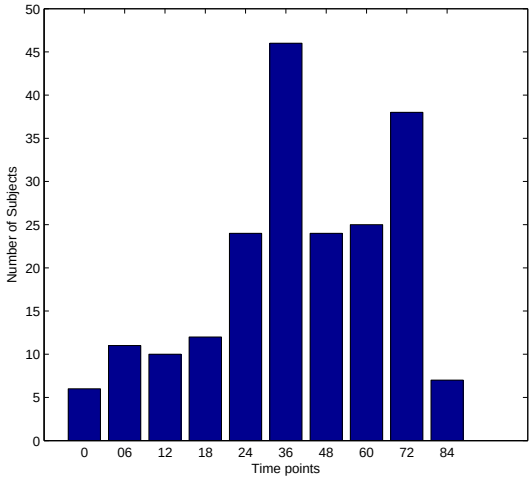

Figure 4.3: Histogram of subjects' last observed time points.

the four methods. We achieve an AUC of 0.749 under CV-DEV for the classification between MCI-C and MCI-NC at 48-month time point, which is comparable to previous studies using multi-modality data, given that we only use PET imaging data. In the ADNI data set, the baseline clinical variables including apolipoprotein E4 (APOE4) allele frequencies and mini mental state exam (MMSE) scores, and demographic variables, such as age, education (in years) and gender, are also available. We reanalyze the data by including these additional variables into the functional logistic regression model at 48-month time point using the CV-DEV as the criterion to select tuning parameters. We also conduct an analysis using only these additional variables in the logistic regression model. It should be noted that the model can be fit directly without tuning parameter selection as there are only five covariates. The classification performance is evaluated using the ten-fold cross-validation procedure. The results are presented in Figure 4.5, showing the ROC curves from the analyses using PET images and other variables, PET images only, and other variables only, respectively. The corresponding AUC values are 0.764, 0.749 and 0.676, showing that the AUC can be slightly improved by including other variables, and imaging analysis does provide added value in the prediction of AD conversion.

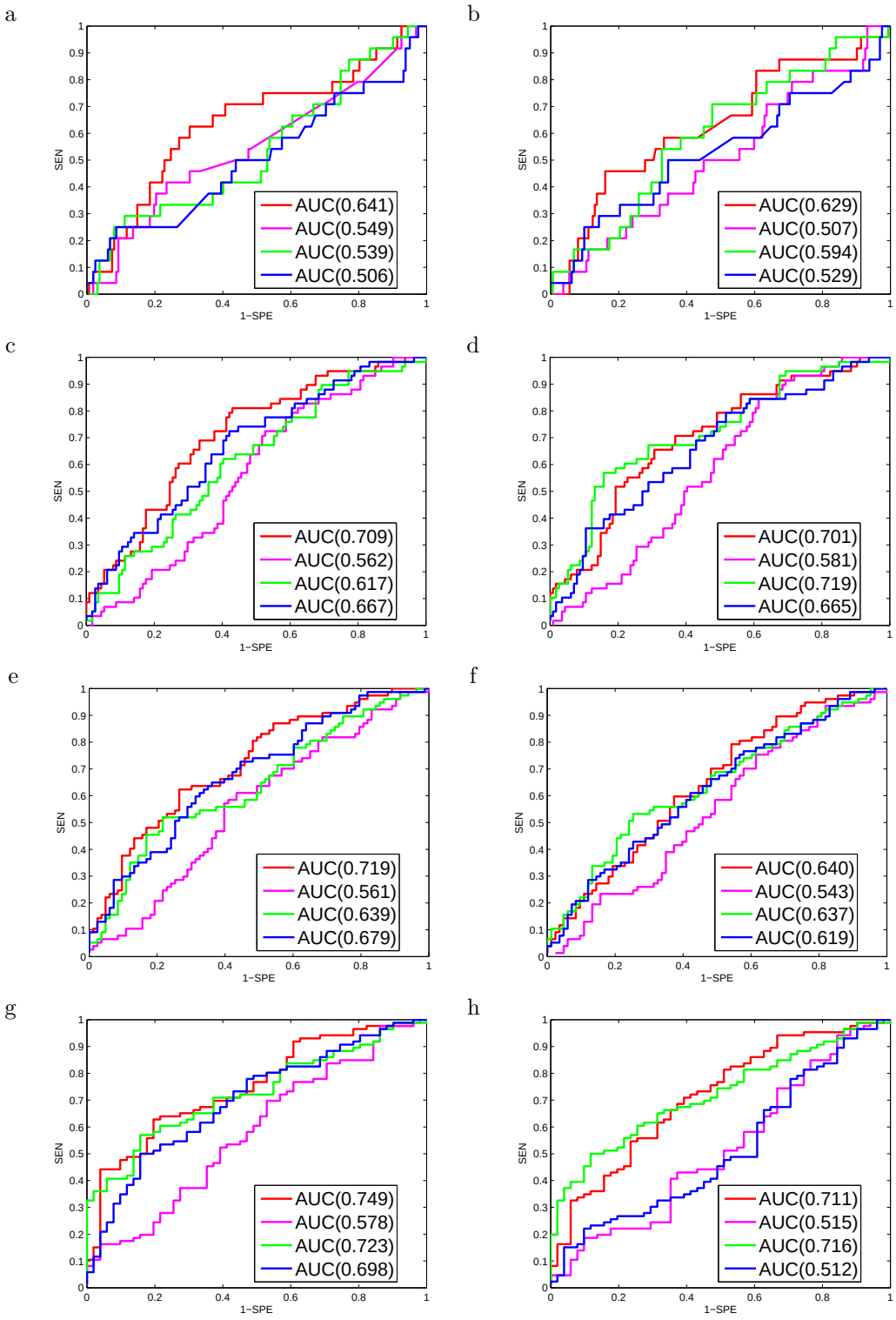

Figure 4.4: ROC curves and their AUCs for four classification methods (red curve: HW-RMFLR; magenta curve: HW-RFLR; blue curve: SGL-RLR; green curve: PC-LR) at four time points. Left panel: CV-DEV; right panel: CV-AUC. (a) and (b) 12 months, (c) and (d) 24 months, (e) and (f) 36 months, (g) and (h) 48 months.

As the classification performance under the criterion CV-DEV is in general better than that under CV-AUC, we present the voxels identified under the criterion CV-DEV at different time points in Figure 4.6, and the 3D views of the identified voxels in Figure 4.7. It can be seen that many regions are identified at more than one time points, suggesting a strong association with the prediction of conversion to AD from MCI. These regions include posterior cingulate cortex, superior temporal gyrus, inferior temporal gyrus, primary visual cortex and cerebellum. The present findings are also largely consistent with previous research studies mentioned above.

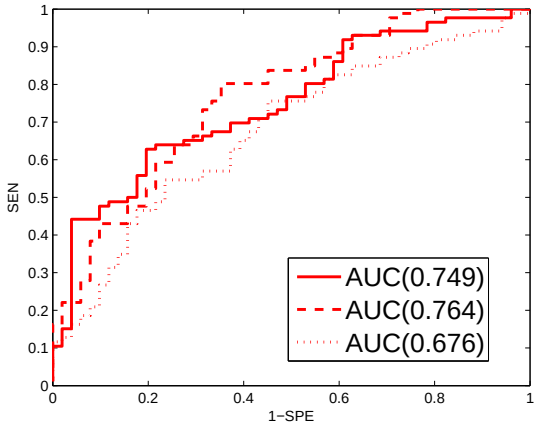

Figure 4.5: ROC curves and their AUCs at 48 months (solid line: PET images only; dashed line: PET images and other variables; dotted line: other variables only).

4.4 Discussion

In this chapter, we propose a regularized multiple functional regression approach via Haar wavelets to identify the predictive voxels of the classification tasks of interest. Simulation results show that this approach can improve the accuracy in identifying the truly predictive brain areas. We then apply this approach to the prediction of conversion to AD among MCI subjects. In the real application, the proposed approach is shown to achieve better classification performance and offer an effective way to identify the predictive brain areas to the conversion.

It should be noted the this conversion exhibits a complex mechanism and our analysis has limitations. Firstly, subjects dropping out without conversion to AD before our selected time points are excluded, as it is not clear if these subjects are MCI-C or MCI-NC. Secondly, we observe that 10 out of 203 MCI subjects have exhibited reversion to normal cognition during the follow-up, but in this work we treat them as MCI-NC. Thirdly, subjects have different follow-up times, and the conversion occurs at different time points. We should note that the use of survival analysis by treating the conversion time as the time-to-event variable may overcome some of the limitations of recent studies.

We demonstrate the performance of the proposed approach on the classification tasks between MCI-C and MCI-NC using baseline PET images from ADNI. The model can be naturally applied to the classification using MRI images, or even multi-modality imaging data where higher predictive performance is expected.

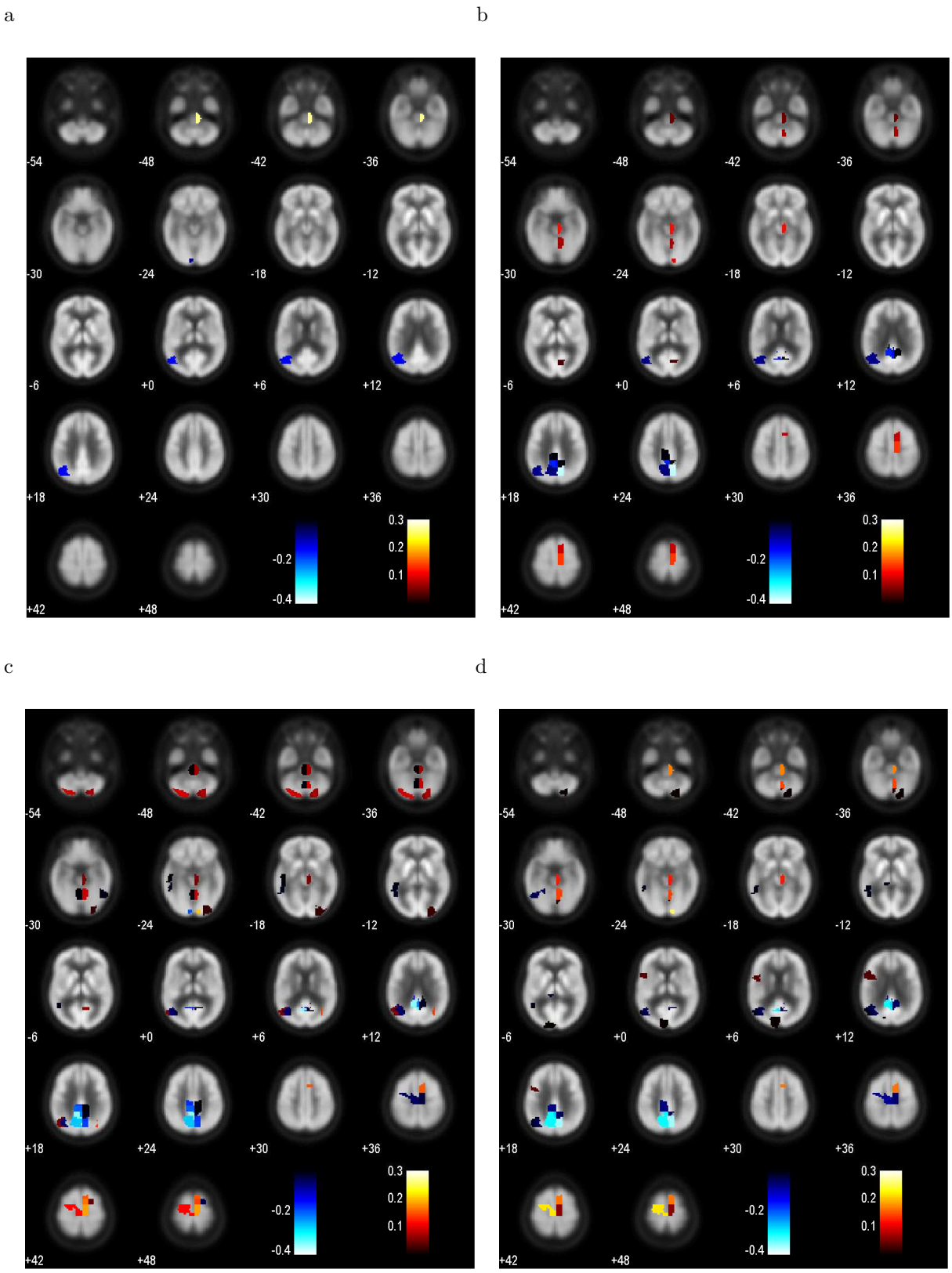

Figure 4.6: Voxels identified using the proposed approach HW-RMFLR under CV-DEV criterion at four time points. (a) 12 months, (b) 24 months, (c) 36 months, (d) 48 months.

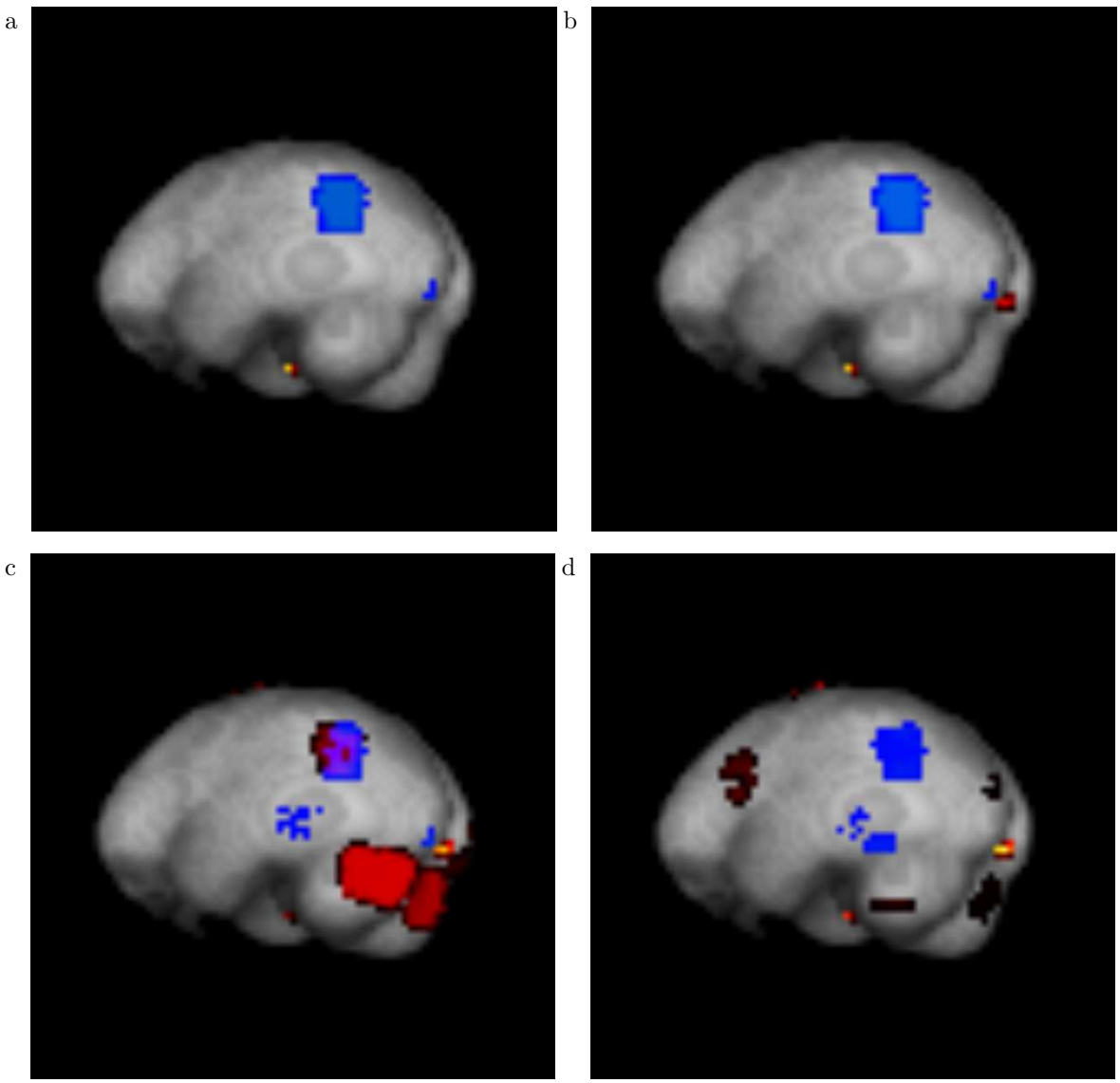

Figure 4.7: 3D sagittal views of the voxels identified using the proposed approach HW-RMFLR under CV-DEV criterion at four time points. (a) 12 months, (b) 24 months, (c) 36 months, (d) 48 months.

CHAPTER V

Conclusions and Future work

This chapter summarizes the contributions of this dissertation and discusses future directions.

5.1 Conclusions

The goal of the dissertation is to develop new approaches for analyzing large-scale brain imaging data to assist the diagnosis of Alzheimer's disease (AD). The dissertation demonstrates that the proposed approaches can overcome the limitations of existing methods, and significantly improve performance in prediction of the disease status and identification of the predictive brain areas.

This dissertation concerns some Haar wavelet-based regularized functional regression models for the analysis of 3D brain imaging data by treating each 3D image as a realization of a 3D functional predictor. All the models are developed in the context of functional data analysis. It is important to choose proper basis functions in functional data analysis. We use the Haar wavelet transform because of the attractive characteristics of Haar wavelets which can account for the spatial correlations among voxels and achieve sparse estimation. It should be noted that the underlying assumption of the proposed models is that only a few brain areas are truly predictive of the outcome of interest, while most existing methods fail to provide sparse estimates of

the voxel-level effects.

Chapter II focuses on the regularized functional linear regression via Haar wavelets. Analysis of the PET imaging data obtained from the ADNI database shows that the proposed approach is useful in identifying the most relevant voxels for cognitive impairment in elderly people. The response variable used here is the MMSE score which is obtained from questionnaires. The most predictive voxels are located in the brain regions that are found to be related to the progression of AD in many previous studies. Chapter III concentrates on the classification problem and develops regularized functional logistic regression via Haar wavelets. As there are three groups of subjects in the ADNI database: AD, MCI and NC. The proposed approach is applied to all the three pairwise classifications. The results demonstrate that the approach not only achieves higher classification accuracy rates than other classification methods, but also identifies the most predictive voxels of the classification. Chapter IV also focuses on the classification problem, in particular, the classification of two types of MCI subjects: one exhibits conversion to AD during follow-up and the other does not. This classification problem is the most difficult one among the classification problems considered in the dissertation, which has also attracted a lot of attention in recent years. To improve the performance, regularized multiple functional regression that can account for the group structure among voxels is developed. It is different from previous chapters in that the Haar wavelet transform is applied within each brain region instead of the whole brain. The results indicate that the proposed approach yields an AUC which is comparable to AUCs reported in previous studies using a variety of biomarkers, clinical measures as well as imaging data, given that the proposed approach has only used the PET imaging data as the predictors.

The major contributions of the dissertation are threefold. First, the proposed approaches in this dissertation can preserve the spatial information among voxels by treating each image as a 3D functional observation. Second, the proposed approaches

offer an effective way in identifying the most predictive voxels of the outcome by using Haar wavelets as basis functions in functional regression models. Third, the underlying assumption is that only a few brain areas are truly predictive of the outcome, while most of the brain areas are not predictive of the outcome. By imposing the sparsity assumption into the estimation, these areas can be identified. The proposed approaches are different from traditional methods for analyzing brain imaging data. For example, the massive association discovered by the traditional univariate analysis or principal component analysis could probably be caused by the high correlation among voxels. Because the human brain has a very complex structure, and every small part of it can have different functions, the proposed approaches can play a very influential role in brain imaging studies.

5.2 Future work

We want to re-emphasize the practical application of the prediction of MCI to AD conversion in Chapter IV, since the scientific discoveries in this kind of research have great potential for early diagnosis of AD. As discussed in Section 4.4, the conversion exhibits a very complex mechanism due to various conversion time points and the inevitable loss to follow up. In this dissertation, the subjects who drop out of the study without conversion are excluded from the analysis which could lead to bias in the estimation. This limitation can be overcome by using survival analysis for censored times. However, there is still an issue that some of the subjects may not be at risk of conversion at all, and in this case, the cure model may be a useful tool (Sy and Taylor, 2000).

The methods developed in this dissertation also point to other directions for future work, including the extension to analysis of imaging data obtained from other modalities. Large-scale neuroimaging data are very common nowadays in brain imaging studies, such as magnetic resonance imaging (MRI), functional MRI (fMRI) and elec-

troencephalography (EEG). In fact, the models developed in this dissertation can be directly extended to structural MRI data. Both fMRI and EEG data consist of a time series at each voxel. Because each time series can also be treated as 1D functional data, we can still apply functional regression models to estimate the parameters of interest. Overall, the proposed regularized functional regression models provide an effective way for analyzing large-scale imaging data.

APPENDICES

APPENDIX A

Proofs of Theorems II.1, II.2 and II.3

Proof of Theorem II.1

Let $\hat{\eta}$ be the lasso solution given in (2.6) with $\lambda = a\sigma_1\sqrt{\frac{\log p}{n}}$, $a > 2\sqrt{2}$. In order to prove Theorem II.1, we need the following Lemma 1, which is modified from Lemma B.1 of Bickel et al. (2009). The proof of Lemma 1 follows similarly and thus is omitted.

Lemma 1. Assume the same assumptions as in Theorem II.1, then with probability at least $1 - p^{1-a^2/8}$ we have

$$\begin{aligned}
&\left\| \int_0^T X(t)\hat{\beta}(t)\, dt - \int_0^T X(t)\beta(t)\, dt \right\|_n^2 + \lambda\|\hat{\eta} - \eta\|_1 \\
&\qquad \leq \left\| \int_0^T X(t)e(t)\, dt \right\|_n^2 + 4\lambda \sum_{j\in A_\eta} |\hat{\eta}_j - \eta_j| \\
&\qquad \leq \left\| \int_0^T X(t)e(t)\, dt \right\|_n^2 + 4\lambda\sqrt{s}\sqrt{\sum_{j\in A_\eta} |\hat{\eta}_j - \eta_j|^2}
\end{aligned}$$

where $e(t)$ is the approximation error in (2.2) and $\|\cdot\|_n$ is the empirical norm, which is defined as $\|g\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^n g_i^2}$

The result in (2.11) is obtained immediately by applying Theorem 5.1 in Bickel et al. (2009). Now we only need to prove (2.12).

By assumption $RE(s, 3+4/\theta)$, we have

$$\begin{aligned}
\kappa^2\|\delta_{A_\eta}\|_2^2 &\leq \frac{1}{n}\|C\delta\|_2^2 \\
&= \frac{1}{n}(\hat{\eta}-\eta)^T C^T C(\hat{\eta}-\eta) \\
&= \|C\hat{\eta}-C\eta\|_n^2 \\
&\leq \left(\left\|C\hat{\eta}-\int_0^T X(t)\beta(t)\,dt\right\|_n + \left\|\int_0^T X(t)e(t)\,dt\right\|_n\right)^2. \qquad \text{(A.1)}
\end{aligned}$$

Lemma 1 implies,

$$\begin{aligned}
\left\|\int_0^T X(t)\hat{\beta}(t)\,dt - \int_0^T X(t)\beta(t)\,dt\right\|_n^2 &= \left\|C\hat{\eta}-\int_0^T X(t)\beta(t)\,dt\right\|_n^2 \\
&\leq \left\|\int_0^T X(t)e(t)\,dt\right\|_n^2 + 4\lambda\sqrt{s}\|\delta_{A_\eta}\|_2. \qquad \text{(A.2)}
\end{aligned}$$

Combining (A.1) with (A.2) , we find

$$\begin{aligned}
\kappa\|\delta_{A_\eta}\|_2 &\leq \left\|C\hat{\eta}-\int_0^T X(t)\beta(t)\,dt\right\|_n + \left\|\int_0^T X(t)e(t)\,dt\right\|_n \\
&\leq \sqrt{\left\|\int_0^T X(t)e(t)\,dt\right\|_n^2 + 4\lambda\sqrt{s}\|\delta_{A_\eta}\|_2} + \left\|\int_0^T X(t)e(t)\,dt\right\|_n. \qquad \text{(A.3)}
\end{aligned}$$

Subtracting the second term of (A.3) from both sides, and then squaring both bides,

we have

$$\left(\kappa\|\delta_{A_\eta}\|_2 - \left\|\int_0^T X(t)e(t)\,dt\right\|_n\right)^2 \leq \left\|\int_0^T X(t)e(t)\,dt\right\|_n^2 + 4\lambda\sqrt{s}\|\delta_{A_\eta}\|_2. \tag{A.4}$$

To solve the quadratic inequality for $\|\delta_{A_\eta}\|_2$, we first need to expand the left side of (A.4), which yields

$$\kappa^2\|\delta_{A_\eta}\|_2^2 - 2\kappa\|\delta_{A_\eta}\|_2\left\|\int_0^T X(t)e(t)\,dt\right\|_n \leq 4\lambda\sqrt{s}\|\delta_{A_\eta}\|_2.$$

This implies

$$\begin{aligned}\|\delta_{A_\eta}\|_2 &\leq \frac{4\lambda\sqrt{s} + 2\kappa\left\|\int_0^T X(t)e(t)\,dt\right\|_n}{\kappa^2} \\ &\leq \frac{4\lambda\sqrt{s} + 2\kappa M\omega}{\kappa^2}.\end{aligned}$$

We also observed that the following relations hold with $k_0 = 3 + 4/\theta$:

$$\|\delta\|_1 = \|\delta_{A_\eta}\|_1 + \|\delta_{A_\eta^c}\|_1 \leq (1+k_0)\|\delta_{A_\eta}\|_1 \leq (1+k_0)\sqrt{s}\|\delta_{A_\eta}\|_2.$$

Then we have,

$$\|\delta\|_1 \leq (4 + 4/\theta)\sqrt{s}\left\{\frac{4\lambda\sqrt{s} + 2\kappa M\omega}{\kappa^2}\right\}.$$

Now let D be a diagonal matrix with $\sqrt{\sum_{i=1}^n C_{ij}^2}$ as the jth diagonal element. We can then rewrite model (2.5) as

$$Y = \beta_0 + C\eta + \epsilon^* = \beta_0 + C\left(\sqrt{n}D^{-1}\right)\left(\frac{1}{\sqrt{n}}D\right)\eta + \epsilon^* = \beta_0 + \tilde{C}\tilde{\eta} + \epsilon^*, \tag{A.5}$$

where $\tilde{C} = C\left(\sqrt{n}D^{-1}\right)$ and $\tilde{\eta} = \left(\frac{1}{\sqrt{n}}D\right)\eta$. Then the diagonal elements of the matrix $\frac{\tilde{C}^T\tilde{C}}{n}$ are equal to 1. Therefore, we obtain

$$\begin{aligned}
\left|\hat{\beta}(t) - \beta(t)\right| &\leq \left|\hat{\beta}(t) - B(t)^T\eta\right| + |e(t)| \\
&= \left|B(t)^T\hat{\eta} - B(t)^T\eta\right| + |e(t)| \\
&\leq \sqrt{n}\|B(t)^TD^{-1}\|_1\|\hat{\tilde{\eta}} - \tilde{\eta}\|_1 + \omega \\
&= \gamma(t)\|\hat{\tilde{\eta}} - \tilde{\eta}\|_1 + \omega \\
&\leq \gamma(t)(4 + 4/\theta)\left\{\frac{4as\sigma_1\sqrt{\frac{\log p}{n}} + 2\kappa\sqrt{s}M\omega}{\kappa^2}\right\} + \omega,
\end{aligned}$$

where $\hat{\tilde{\eta}} = (\frac{1}{\sqrt{n}}D)\hat{\eta}$ and $\gamma(t) = \sqrt{n}\|B(t)^TD^{-1}\|_1 = \sum_{j=1}^{p}\left|\frac{b_j(t)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}C_{ij}^2}}\right|$.

Proof of Theorem II.2

If we assume that $\epsilon_i^* \sim N(0, \sigma_2^2)$ in model (2.5), then this model is the special case in Section 6 of Bickel et al. (2009). Let $\hat{\eta}$ be the corresponding lasso solution with $\lambda = a\sigma_2\sqrt{\frac{\log p}{n}}$, $a > 2\sqrt{2}$, then Lemma 2 here is obtained directly from their Theorem 6.2.

Lemma 2. Assume the same assumptions as in Theorem II.2. Then under assumption $RE(s, 3)$, with the probability at least $1 - p^{1-a^2/8}$ we have

$$\|\hat{\eta} - \eta\|_1 \leq \frac{16as\sigma_2}{\kappa^2}\sqrt{\frac{\log p}{n}}, \tag{A.6}$$

$$\|C(\hat{\eta} - \eta)\|_2^2 \leq \frac{16a^2s\sigma_2^2}{\kappa^2}\log p, \tag{A.7}$$

where $\kappa = \kappa(s, 3)$.

It follows from (A.7) that

$$\|C(\hat{\eta}-\eta)\|_n = \frac{1}{\sqrt{n}}\|C(\hat{\eta}-\eta)\|_2 \le \frac{4a\sigma_2}{\kappa}\sqrt{\frac{s\log p}{n}}.$$

We then have

$$\begin{aligned}
\left\|\int_0^T X(t)\hat{\beta}(t)\,dt - \int_0^T X(t)\beta(t)\,dt\right\|_n &\le \|C(\hat{\eta}-\eta)\|_n + \left\|\int_0^T X(t)e(t)\,dt\right\|_n \\
&= \|\tilde{C}(\hat{\tilde{\eta}}-\tilde{\eta})\|_n + \left\|\int_0^T X(t)e(t)\,dt\right\|_n \\
&\le \|\tilde{C}(\hat{\tilde{\eta}}-\tilde{\eta})\|_n + M\omega \\
&\le \frac{4a\sigma_2}{\kappa}\sqrt{\frac{s\log p}{n}} + M\omega.
\end{aligned}$$

By Lemma 2, we have

$$\|\hat{\tilde{\eta}}-\tilde{\eta}\|_1 \le \frac{16as\sigma_2}{\kappa^2}\sqrt{\frac{\log p}{n}}.$$

Therefore, we obtain

$$\begin{aligned}
\left|\hat{\beta}(t)-\beta(t)\right| &\le \left|\hat{\beta}(t)-B(t)^T\eta\right| + |e(t)| \\
&= \left|B(t)^T\hat{\eta}-B(t)^T\eta\right| + |e(t)| \\
&\le \sqrt{n}\|B(t)^TD^{-1}\|_1\|\hat{\tilde{\eta}}-\tilde{\eta}\|_1 + \omega \\
&= \gamma(t)\|\hat{\tilde{\eta}}-\tilde{\eta}\|_1 + \omega \\
&\le \gamma(t)\frac{16as\sigma_2}{\kappa^2}\sqrt{\frac{\log p}{n}} + \omega.
\end{aligned}$$

Proof of Theorem II.3

Proof. From (2.12), we have

$$\begin{aligned}\left|\hat{\beta}(t)-\beta(t)\right| &\leq \frac{4}{\kappa^2_{n,p^*}}\gamma_{n,p^*}(t)(4+4/\theta)s_{p^*}^{\frac{3}{2}}a\sigma_1\sqrt{\frac{\log p^*}{n}}+\{\frac{2}{\kappa_{n,p^*}}\gamma_{n,p^*}(t)(8+8/\theta)s_{p^*}M+1\}\omega_{p^*}\\ &\leq O(n^{-\frac{1}{2}})+O(2^{-J^*m})\end{aligned}$$

Note that $\frac{\gamma_{n,p^*}(t)}{\kappa^2_{n,p^*}}$ and $\frac{\gamma_{n,p^*}(t)}{\kappa_{n,p^*}}$ are bounded for large n by C.3 and C.4. s_{p^*} is bounded by C.2.

If $n\to\infty$ and $2^{J_n}\to\infty$ ($p_n\to\infty$),

$$\begin{aligned}\left|\hat{\beta}(t)-\beta(t)\right| &\leq \frac{4}{\kappa^2_{n,p_n}}\gamma_{n,,p_n}(t)(4+4/\theta)s_{p_n}^{\frac{3}{2}}a\sigma_1\sqrt{\frac{\log p_n}{n}}+\{\frac{2}{\kappa_{n,p_n}}\gamma_{n,p_n}(t)(8+8/\theta)s_{p^*}M+1\}\omega_{p_n}\\ &\leq \frac{\sqrt{\log n}2^{J_nb_t}}{\sqrt{n}}\{\frac{4}{\kappa^2_{n,p_n}}2^{-J_nb_t}\gamma_{n,,p_n}(t)(4+4/\theta)S^{\frac{3}{2}}a\sigma_1\frac{\sqrt{\log p_n}}{\sqrt{\log n}}\}\\ &\quad+\frac{\sqrt{\log n}\sqrt{n}}{\sqrt{n}2^{J_n(m-b_t)}}\{\frac{2}{\kappa_{n,p_n}}2^{-J_nb_t}\gamma_{n,p_n}(t)(8+8/\theta)SM+2^{-J_nb_t}\}\frac{\omega_{p_n}2^{J_nm}}{\sqrt{\log n}}\\ &= \frac{\sqrt{\log n}2^{J_nb_t}}{\sqrt{n}}K_1+\frac{\sqrt{\log n}\sqrt{n}}{\sqrt{n}2^{J_n(m-b_t)}}K_2\\ &= \frac{\sqrt{\log n}n^{\frac{b_t}{2m}}}{\sqrt{n}}\{(\frac{2^{J_n}}{n^{\frac{1}{2m}}})^{b_t}K_1+\frac{n^{\frac{m-b_t}{2m}}}{2^{J_n(m-b_t)}}K_2\}\\ &= O(\frac{\sqrt{\log n}}{n^{\frac{m-b_t}{2m}}})\end{aligned}$$

By C.1, C.2, C.3 and C.5, K_1 and K_2 are bounded, and by $2^{J_n}=O(n^{\frac{1}{2m}})$, $(\frac{2^{J_n}}{n^{\frac{1}{2m}}})^{b_t}K_1$ and $\frac{n^{\frac{m-b_t}{2m}}}{2^{J_n(m-b_t)}}K_2$ are bounded. □

APPENDIX B

Detailed Explanation for Choosing Haar Wavelets

The functional linear regression model is written as:

$$Y = \beta_0 + \int_0^T X(t)\beta(t)\,dt + \epsilon. \tag{B.1}$$

In many cases, the response variable Y is determined only by a small region of the predictor $X(t)$. The approaches developed in this dissertation can provide us with a sparse solution of $\beta(t)$, as if $\beta(t)$ is zero in many regions, the Haar wavelet coefficients of $\beta(t)$ are usually sparse too. Therefore, we apply the variable selection methods, such as the lasso, to select the nonzero wavelet coefficients in order to achieve sparse estimation.

Now we want to illustrate this by a simple example. Suppose that $\beta(t)$ is a 1D signal and collected at 8 time points. $\beta(t) = [\beta(t_1), ..., \beta(t_8)] = [0, 0, 4, 6, 5, 5, 0, 0]$. Applying level-1 Haar wavelet transform to $\beta(t)$, we will obtain the wavelet coefficient vector c_1. The Haar transformation matrix for this case is shown below:

$$H_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

$$\beta H_1^T = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 0 & 4 & 6 & 5 & 5 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix} = c_1$$

in which $\frac{1}{\sqrt{2}}$ is the scaling factor and $c_1 = [0, 5\sqrt{2}, 5\sqrt{2}, 0, 0, -\sqrt{2}, 0, 0]$ is the wavelet coefficient vector. Conversely, we have this relationship: $c_1 H_1 = \beta$. The first four coefficients in c_1 corresponding to the first four rows of H_1 are called approximation coefficients (a_1), and the last four coefficients in c_1 corresponding to the last four rows of H_1 are called detail coefficients (d_1). The first, third and last elements of d_1 are zero, meaning that there are no changes between the first two, third two and

last two elements of $\beta(t)$. The first and last elements of a_1 are zero, indicating that the averages of the first and the last two elements of $\beta(t)$ are zero. This example shows that zeros in the detail coefficients do not always indicate zeros in the original signal, but no change between the two neighboring elements (for level-1 Haar wavelet decomposition). If the corresponding approximation coefficients are also zero, then some neighboring points are exactly zeros in the original signal. In this case, the first and the last elements of both a_1 and d_1 are zero, showing that the first and the last two elements of the original signal are zero.

Similarly, the applications of Haar wavelet transform at other levels to $\beta(t)$ are shown below. Let H_2 denote the level-2 Haar wavelet transform matrix, and H_3 denote the level-3 Haar wavelet transform matrix.

$$H_2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

$$\beta H_2^T = \begin{bmatrix} 0 & 0 & 4 & 6 & 5 & 5 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & -\frac{1}{2} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{2} & 0 & -\frac{1}{2} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{2} & 0 & -\frac{1}{2} & 0 & 0 & 0 & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{2} & 0 & -\frac{1}{2} & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} \end{bmatrix} = c_2$$

where $c_2 = [5, 5, -5, 5, 0, -\sqrt{2}, 0, 0]$.

$$H_3 = \begin{bmatrix} \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$$

$$\beta H_3^T = \begin{bmatrix} 0 & 0 & 4 & 6 & 5 & 5 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2} & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & \frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} & -\frac{1}{2} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & \frac{1}{2} & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & -\frac{1}{2} & 0 & 0 & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 & -\frac{1}{2} & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} \end{bmatrix} = c_3$$

where $c_3 = [5, 0, -5, 5, 0, -\sqrt{2}, 0, 0]$.

By observing the structures of different levels Haar wavelet transform, we find that the numbers of nonzero elements in the wavelet coefficient vectors are different. It is not surprising to see that $\hat{\beta}(t)$ is often piece-wise constant after applying the lasso method for the selection of important elements of wavelet coefficients because small detail coefficients are set to zero. Moreover, higher-level Haar wavelet transform tends to give us piecewise constant solutions on a coarser scale. Considering all of the above, it would be natural to assume that there exists an optimal wavelet decomposition level for a given signal in terms of measures of prediction accuracy in the regression setting, therefore, we set it as one of the tuning parameters in this dissertation.

BIBLIOGRAPHY

BIBLIOGRAPHY

Alzheimer's Association (2010). Changing the trajectory of Alzheimer's disease: a national imperative. Accessed April 2013, URL `http://www.alz.org/alzheimers_disease_trajectory.asp`.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, **4** 1705–1732.

Bonneville, M., Meunier, J., Bengio, Y. and Soucy, J.-P. (1998). Support Vector Machines for improving the classification of brain PET images. *SPIE Medical Imaging*, **3338** 264–273.

Bowman, F. D., Caffo, B., Bassett, S. S. and Kilts, C. (2008). A Bayesian hierarchical framework for spatial modeling of fMRI data. *NeuroImage*, **39** 146–156.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, **35** 2313–2351.

Casanova, R., Maldjian, J. A. and Espeland, M. A. (2011). Evaluating the impact of different factors on voxel-based classification methods of ADNI structural MRI brain images. *International Journal of Biomedical Data Mining, DOI:10.4303/ijbdm/B110102*, **1**.

Cockrell, J. and Folstein, M. (1988). Mini-mental state examination (MMSE). *Psychopharmacol Bulletin*, **24** 689–692.

Daubechies, I., Defrise, M. and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, **57** 1413–1457.

Davatzikos, C., Bhatt, P., Shaw, L. M., Batmanghelich, K. N. and Trojanowski, J. Q. (2011). Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of Aging*, **32** e19–27.

Dehghan, H., Pouyan, A. A. and Hassanpour, H. (2011). SVM-based diagnosis of the Alzheimer's disease using 18F-FDG PET with Fisher discriminate rate. *18th Iranian Conference on Biomedical Engineering* 37–42.

Eskildsen, S. F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J. C. and Collins, D. L. (2013). Prediction of Alzheimer's disease in subjects

with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *NeuroImage*, **65** 511–521.

Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series* B, **74** 37–65.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96** 1348–1360.

Foster, N. L., Chase, T. N., Mansi, L., R.Brooks, Fedio, P., Patronas, N. J. and Chiro, G. D. (1984). Cortical abnormalities in Alzheimer's disease. *Annals of Neurology*, **16** 649–654.

Fouquet, M., Desgranges, B., Landeau, B., Duchesnay, E., Mézenge, F., Sayette, V., Viader, F., Baron, J.-C., Eustache, F. and Chételat, G. (2009). Longitudinal brain metabolic changes from amnestic mild cognitive impairment to Alzheimer's disease. *Brain*, **132** 2058–2067.

Frackowiak, R., Ashburner, J., Penny, W., Zeki, S. and Friston, K. (2004). *Human Brain Functions*. 2nd ed. Academic Press.

Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, **1** 302–332.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33** 1–22.

Friston, K., Poline, J., Holmes, A., Frith, C. and Frackowiak, R. (1996). A multivariate analysis of PET activation studies. *Human Brain Mapping*, **4** 140–151.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, **7** 397–416.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. 2nd ed. Academic Press.

Galton, C., Patterson, J., K.and Xuereb and Hodges, J. (1999). Atypical and typical presentations of Alzheimer's disease: a clinical, neuropsychological, neuroimaging and pathological study of 13 cases. *Brain*, **123** 484–498.

Garey, L. (2006). *Brodmann's Locallisation in the Cerebral Cortex*. Springer.

Grimmer, T., Henriksen, G., Wester, H. J., Förstl, H., Klunk, W. E., Mathis, C. A., Kurz, A. and Drzezga, A. (2009). Clinical severity of Alzheimer's disease is associated with PIB uptake in PET. *Neurobiology of Aging*, **30** 1902–1909.

Habeck, C., Foster, N., Perneczky, R., Kurz, A., Alexopoulos, P., Koeppe, R., Drzezga, A. and Stern, Y. (2008). Multivariate and univariate neuroimaging biomarkers of Alzheimer's disease. *NeuroImage*, **40** 1503–1515.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer.

Higdon, R., Foster, N. L., Koeppe, R. A., DeCarli, C. S., Jagust, W. J., Clark, C. M., Barbas, N. R., Arnold, S. E., Turner, R. S., Heidebrink, J. L. and Minoshima, S. (2004). A comparison of classification methods for differentiating fronto-temporal dementia from Alzheimer's disease using FDG-PET imaging. *Statistics in Medicine*, **23** 315–326.

Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M. K. and Johnson, S. C. (2009). Spatially augmented LP boosting for AD classification with evaluations on the ADNI dataset. *NeuroImage*, **48** 138–149.

Hoffman, J., Welsh-Bohmer, K., Hanson, M., Crain, B., Hulette, C., Earl, N. and Coleman, R. (2000a). FDG PET imaging in patients with pathologically verified dementia. *Journal of Nuclear Medicine*, **41** 1920–1928.

Hoffman, J. M., Welsh-Bohmer, K. A., Hanson, M., Crain, B., Hulette, C., Earl, N. and Coleman, R. E. (2000b). FDG PET imaging in patients with pathologically verified dementia. *Journal of Nuclear Medicine*, **41** 1920–1928.

Illán, I., Górriz, J., Ramíreza, J., Salas-Gonzalez, D., López, M., Segovi, F., Chaves, R., Gómez-Rio, M. and Puntonet, C. (2011). 18F-FDG PET imaging analysis for computer aided Alzheimer's diagnosis. *Information Sciences*, **181** 903–916.

James, G. M., Wang, J. and Zhu, J. (2009). Functional linear regression that's interpretable. *Annals of Statistics*, **37** 2083–2108.

Janousova, E., Vounou, M., Wolz, R., Gray, K. R., Rueckert, D. and Montana, G. (2012). Biomarker discovery for sparse classification of images. *Annals of the British Machine Vision Association*, **2012** 1–11.

Jiang, D., Huang, J. and Zhang, Y. (2011). The cross-validated AUC for MCP-logistic regression with high-dimensional data. *Statistical Methods in Medical Research, Available Online: 28 Nov 2011*.

Kang, J., Johnson, T. D., Nichols, T. E. and Wager, T. D. (2011). Meta analysis of functional neuroimaging data via Bayesian spatial point processes. *Journal of the American Statistical Association*, **106** 124–134.

Karow, D., McEvoy, L., Fennema-Notestine, C., Hagler, D. J., Jennings, R., Brewer, J., Hoh, C. and Dale, A. (2010). Relative capability of MR imaging and FDG PET to depict changes associated with prodromal and early Alzheimer disease. *Radiology*, **256** 932–942.

Kerrouche, N., Herholz, K., Mielke, R., Holthoff, V. and Baron, J.-C. (2006). 18f-FDG PET in vascular dementia: differentiation from Alzheimer's disease using voxel-based multivariate analysis. *Journal of Cerebral Blood Flow and Metabolism*, **26** 1213–1221.

Kim, S. and Xing, E. P. (2009). Tree-guided group lasso for multi-task regression with structured sparsity. *Arxiv preprint arXiv:0909.1373*.

Langbaum, J. B., Chen, K., Lee, W., Reschke, C., Bandy, D., Fleisher, A., Alexander, G., Foster, N., Weiner, M., Koeppe, R., Jagust, W. and Reiman, E. (2009). Categorical and correlational analyses of baseline fluorodeoxyglucose positron emission tomography images from the Alzheimer's Disease Neuroimaging Initiative (ADNI). *NeuroImage*, **45** 1107–1116.

Leifer, B. (2003). Early diagnosis of Alzheimer's disease: clinical and economic benefits. *Journal of the American Geriatrics Society*, **51** S281–S288.

Liu, J., Ji, S. and Ye, J. (2009). SLEP:sparse learning with efficient projections. *Arizona State University*.

Liu, M., Zhang, D., Yao, P. and Shen, D. (2012). Tree-guided sparse coding for brain disease classification. *2012 International Conference on Medical Image Computer-Assisted Intervention*, **15** 239–247.

Luo, W. and Nichols, T. E. (2003). Diagnosis and exploration of massively univariate neuroimaging models. *NeuroImage*, **19** 1014–1032.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11** 674–693.

Meishausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series* B, **72** 417–473.

Minoshima, S., Frey, K. A., Koeppe, R. A., Foster, N. L. and Kuhl, D. E. (1995). A diagnostic approach in Alzhimer's disease using three-dimensiaonal stereotactic surface projections of fluorine-18-FDG PET. *Journal of Nuclear Medicine*, **36** 1238–1248.

Minoshima, S., Giordani, B., Berent, S., Frey, K., Foster, N. and Kuhl, D. (1997a). Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. *Annals of Neurology*, **42** 85–94.

Minoshima, S., Giordani, B., Berent, S., Frey, K. A., Foster, N. L. and Kuhl, D. E. (1997b). Metabolic reduction in the posterior cingulate cortex in very early Alzhimer's disease. *Annals of Neurology*, **42** 85–94.

Misra, C., Fan, Y. and Davatzikos, C. (2009). Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI. *NeuroImage*, **44** 1415–1422.

Mosconi, L. (2005). Brain glucose metabolism in the early and specific diagnosis of Alzheimer's disease. FDG-PET studies in MCI and AD. *European Journal of Nuclear Medicine and Molecular Imaging*, **32** 466–510.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jaqust, W., Trojanowski, J. Q., Toga, A. W. and Beckett, L. (2005). Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's and Dementia: The Journal of the Alzheimer's Association*, **1** 55–66.

Muraki, S. (1992). Approximation and rendering of volume data using wavelet transforms. In *Proceedings of Visualization 1992*. Boston, 21–28.

Nichols, T. E. and Holmes, A. P. (2001). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, **15** 1–25.

Plassman, B. L., Langa, K. M., Fisher, G. G., Heeringa, S. G., Weir, D. R., Ofstedal, M. B., Burke, J. R., Hurd, M. D., Potter, G. G., Rodgers, W. L., Steffens, D. C., Willis, R. J. and Wallace, R. B. (2007). Prevalence of dementia in the United States: the aging, demographics, and memory study. *Neuroepidemiology*, **29** 125–132.

Querbes, O., Aubry, F., Pariente, J., Lotterir, J.-A., J.-F., D., Duret, V., Puel, M., Berry, I., Fort, J.-C. and Celsis, P. (2009). Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain*, **132** 2036–2047.

Rabinovici, G. D., Furst, A. J., Alkalay, A., Racine, C. A., O'Neil, J. P., Janabi, M., Baker, S. L., Agarwal, N., Bonasera, S. J., Mormino, E. C., Weiner, M. W., Gorno-Tempini, M. L., Rosen, H. J., Miller, B. L. and Jagust, W. J. (2010). Increased metabolic vulnerability in early-onset Alzheimers disease is not related to amyloid burden. *Brain*, **133** 512–528.

Ramsay, J. and Silverman, B. (2005). *Functional data analysis*. 2nd ed. Springer.

Reiss, P. T. and Ogden, R. T. (2010). Functional generalized linear models with images as predictors. *Biometrics*, **66** 61–69.

Royston, P. and Sauerbrei, W. (2008). *Multivariable Model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. Wiley.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine*, **11** 2093–2109.

Shen, L., Kim, S., Qi, Y., Inlow, M., Swaminathan, S., Nho, K., Wan, J., Risacher, S. L., Shaw, L. M., Trojanowski, J. Q., Weiner, M. W. and Saykin, A. J. (2011). Identifying neuroimaging and proteomic biomarkers for MCI and AD via the elastic net. In *Multimodal Brain Image Analysis*, vol. 7012 of *Lecture Notes in Computer Science*. Springer, 27–34.

Silveira, M. and Marques, J. (2010). Boosting Alzheimer disease diagnosis using PET images. *2010 20th International Conference on Pattern Recognition* 2556–2559.

Silverman, D., Small, G., Chang, C., Lu, C., Kung De Aburto, M., Chen, W., Czernin, J., Rapoport, S., Pietrini, P., Alexander, G., Schapiro, M., Jagust, W., Hoffman, J., Welsh-Bohmer, K., Alavi, A., Clark, C., Salmon, E., de Leon, M., Mielke, R., Cummings, J., Kowell, A., Gambhir, S., Hoh, C. and Phelps, M. (2001). Positron emission tomography in evaluation of dementia: regional brain metabolism and long-term outcome. *Journal of the American Medical Association*, **286** 2120–2127.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2012). A sparse-group lasso. *Journal of Computational and Graphical Statistics, DOI:10.1080/10618600.2012.681250, Available Online: 15 May 2012*.

Stoeckel, J. and Fung, G. (2007). SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information. *Journal of Knowledge and Information Systems*, **23** 243–258.

Sy, J. and Taylor, J. M. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, **56** 227–236.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series* B, **58** 267–288.

Vemuri, P., Gunter, J. L., Senjem, M. L., Whitwell, J. L., Kantarci, K., Knopman, D. S., Boeve, B. F., Petersen, R. C. and Jack, C. R. (2008). Alzheimers Disease diagnosis in individual subjects using structural MR images: Validation studies. *NeuroImage*, **39** 1186–1197.

Walker, J. S. (2008). *A primer on wavelets and their scientific applications*. 2nd ed. Chapman & Hall/CRC.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, **2** 224–244.

Zhang, D. and Shen, D. (2012). Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *Plos One*, **7**.

Zhao, Y., Ogden, R. T. and Reiss, P. T. (2012). Wavelet-based LASSO in functional linear regression. *Journal of Computational and Graphical Statistics*, **21** 600–617.

Zhou, J., Liu, J., Nrayan, V. A. and Ye, J. (2012). Modeling disease progression via fused sparse group lasso. In *KDD'12 Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining*. 1095–1103.

Zhu, H., Brown, P. and Morris, J. (2012). Robust classification of functional and quantitative image data using functional mixed models. *Biometrics*, **68** 1260–1268.

Zou, H. (2005). *Some perspectives of sparse statistical modeling*. Ph.D. thesis, Department of Statistics, Stanford University.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101** 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series* B, **67** 301–320.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the degrees of freedom of the lasso. *Annals of Statistics*, **35** 2173–2192.

Zuendorf, G., Kerrouche, N., Herholz, K. and Baron, J. C. (2003). Efficient principal component analysis for multivariate 3D voxel-based mapping of brain functional imaging data sets as applied to FDG-PET and normal aging. *Human Brain Mapping*, **18** 13–21.