REGULARIZED 3D FUNCTIONAL REGRESSION FOR BRAIN IMAGE DATA VIA HAAR WAVELETS

Xuejing Wang[†], **Bin Nan**[†,‡], **Ji Zhu**[†,§], and **Robert Koeppe**[†] **for the Alzheimer's Disease Neuroimaging Initiative**[*]
University of Michigan

Abstract

The primary motivation and application in this article come from brain imaging studies on cognitive impairment in elderly subjects with brain disorders. We propose a regularized Haar wavelet-based approach for the analysis of three-dimensional brain image data in the framework of functional data analysis, which automatically takes into account the spatial information among neighboring voxels. We conduct extensive simulation studies to evaluate the prediction performance of the proposed approach and its ability to identify related regions to the outcome of interest, with the underlying assumption that only few relatively small subregions are truly predictive of the outcome of interest. We then apply the proposed approach to searching for brain subregions that are associated with cognition using PET images of patients with Alzheimer's disease, patients with mild cognitive impairment, and normal controls.

Keywords and phrases

Alzheimer's disease; Brain imaging; Functional data analysis; Haar wavelet; Lasso; PET image; Variable selection

1. Introduction

Alzheimer's disease (AD) is 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; 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), 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 early diagnosis of AD with 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 in understanding the association between PET images and cognitive impairment. In particular, the fluorodeoxyglucose (FDG) PET has been used to measure the cerebral glucose metabolic activity for over 20 years.

FDG PET scans used in the preparation of this article were obtained from a large multi-center follow-up study on Alzheimer's disease and early dementia, the Alzheimer's Disease Neuroimaing Initiative (ADNI). A total of 403 FDG PET scans 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 image 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) and Shin et al. (2010), among many others. For AD in particular, several studies have shown that reduced metabolic activity in some regions of the brain such as the posterior cingulate and the temporal and parietal cortices are associated with the progression of cognitive impairment (Foster et al., 1984; Minoshima et al., 1995, 1997). 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 comparisons have to be considered. In the neuroimaging literature, a distinction is often drawn between such univariat analyses and an alternative, multiple covariate regression models that treat every voxel as a covariate. Since the number of voxels is much larger than the number of scans, the ordinary least squares for linear regression cannot be implemented without applying, for example, some dimension reduction techniques. Such analysis, however, may lead to difficulties in interpretations and practical implications. Both the traditional univariate and multiple covariate approaches (if applicable) have one major limitation in common: they 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, multiple covariate 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 of addressing this issue is to treat the entire 3D image as a single functional input, which allows retaining all the information from the original image without modeling the spatial correlation between voxels explicitly. Specifically, in this article, we treat PET image data as the 3D functional observations, and propose a novel Haar wavelet-based regularized approach to analyze PET image 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 article, 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 $X(t)$ as follows:

$$Y_i = \beta_0 + \int_0^T X_i(t)\beta(t)\,dt + \varepsilon_i, \quad i = 1, \cdots, n, \quad (1.1)$$

where $\beta(t)$ is the regression coefficient function, and t refers to time or location. For the 3D case we consider later, t is replaced by the coordinate (u, v, w). Regularization methods, such as the roughness penalty approach or using restricted basis functions (Ramsay and Silverman, 2005) can be implemented to produce an estimator of $\beta(t)$ that is meaningful in interpretation and useful in prediction.

For the functional linear regression model (1.1), James, Wang and Zhu (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 domain 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 image 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 coordinate descent algorithm (Fu, 1998; Daubechies, Defrise and De Mol, 2004; Friedman et al., 2007; Wu and Lange, 2008) does not apply due to the penalty on derivatives. Without imposing sparsity, Reiss and Ogden (2010) considered the functional principal component regression for image data.

In this article, we choose the Haar wavelet basis instead of the piecewise constant basis for analyzing 3D image 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, Wang and Zhu (2009). In other words, by shrinking corresponding wavelet coefficients to zero, the estimator of the regression coefficient function can be exactly zero over regions where no relationship to the response variable is present. 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 image 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 the brain or in the ventricles. It should be noted that a recent article by Zhao, Ogden and Reiss (2012) considered a general wavelet-based Lasso approach in functional linear regression, but only concerned 1D $\beta(t)$. The Haar wavelet transform is a useful tool for image and signal analysis, and has many other applications. For example, Picart, Butenschön and Shutler (2012) and Lovejoy and Schertzer (2012) discussed the use of Haar wavelet transforms in geophysics and climate research.

The rest of this article is organized as follows. In Section 2, we review some background on wavelet decomposition and properties of Haar wavelet basis functions using a 1D functional linear regression model as an illustration, and then propose the ℓ_1 regularized shrinkage estimation for general functional data, including both 1D and 3D cases. To evaluate the numerical performance of our approach, we conduct extensive simulations in Section 3. We present the analysis of ADNI 3D PET image data in Section 4 and make some concluding remarks in Section 5. We also show that the proposed method achieves the desirable non-asymptotic error bounds for prediction and estimation, the so-called oracle inequalities, meaning that the method performs equally well (up to a constant) as if the true subregions with nonzero regression coefficient were given. The theoretical results are provided in the online supplementary material.

2. Regularized Haar Wavelets Method

For ease of presentation, we describe the proposed methodology starting with the 1D case given in (1.1), then extend it to the 3D case using a tensor product of three 1D wavelet expansions.

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 in this article) 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 (Wand and Ormerod, 2011), which involves shrinking small coefficients to zero 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 (1.1) is sparse, and potentially discontinuous at the boundaries of subregions. In particular, only 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 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:

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

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

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

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

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 $\sqrt{2^j}$ is the normalizing factor. It is noted that the basis functions are orthonormal. Therefore, for a sufficiently fine resolution J, the coefficient function $\beta(t)$ in (1.1) defined on [0, 1) can be expanded in a Haar wavelet series:

$$\beta(t) = \sum_{k=0}^{2^{j_0}-1} a_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{J}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t) + e(t), \quad (2.1)$$

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 more details about Haar wavelets.

2.2. Model estimation

Rewrite $\beta(t)$ in (2.1) by

$$\beta(t) = B(t)^T \eta + e(t), \quad (2.2)$$

where $B(t)$ denotes the collection of all $\varphi_{j,k}(t)$ and $\psi_{j,k}(t)$ in the above Haar wavelet expansion, and η is the corresponding wavelet coefficient vector of length p. Plugging (2.2) into (1.1), we obtain

$$Y_i = \beta_0 + \int_0^1 X_i(t)B(t)^T \eta \, dt + \varepsilon_i^* = \beta_0 + C_i^T \eta + \varepsilon_i^*, \quad i = 1, \cdots, n, \quad (2.3)$$

where $C_i = \int_0^1 X_i(t)B(t)dt$ and $\varepsilon_i^* = \varepsilon_i + \int_0^1 X_i(t)e(t)dt$. 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 (2.2). Model (2.3) can then be rewritten as follows:

$$Y = \beta_0 + C\eta + \varepsilon^*, \quad (2.4)$$

where $C = [C_1, C_2, \cdots, C_n]^T$ is an $n \times p$ design matrix in linear model (2.4). Once an estimator $\hat{\eta}$ is obtained from (2.4), $\beta(t)$ can then be estimated by $B(t)^T \hat{\eta}$.

In practice, $X(t)$ is observed on only a finite set of grid points $\{t_1, \cdots, t_p\}$, which also determines the highest and yet practically meaningful level of decomposition for $\beta(t)$. For the discrete wavelet transform, p is required to be a power of 2. Using the usual terminology for Haar wavelets (see e.g. Walker (2008), and that used in the MATLAB Wavelet Toolbox, 2011b), we define the 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 (2.1). For any level of Haar wavelet decomposition, the total number of basis functions $\varphi_{j,k}$ and $\psi_{j,k}$ is always p, and the collection of $\varphi_{j,k}$ 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 $t \in [0, 1)$ (in the ADNI brain image analysis where t is 3D, this would correspond to that large regions in the brain are not associated with the cognitive performance measured by MMSE), the coefficient vector η in (2.2) should be sparse with $e(t) = 0$ for those regions, i.e. $\beta(t)$ can be well approximated by an economical wavelet expansion with few nonzero coefficients. We consider the Lasso approach (Tibshirani, 1996) and implement the method with the fast coordinate descent algorithm to obtain a desirable sparse solution for the wavelet coefficients.

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

$$\hat{\eta} = \arg\min_{\eta} \left\{ \frac{1}{n} \|Y - \beta_0 - C\eta\|_2^2 + 2\lambda \|\eta\|_1 \right\}, \quad (2.5)$$

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 also a tuning parameter.

It should be noted that in general the Haar wavelet coefficients with large magnitudes are related to salient features. The magnitudes of detail coefficients should be proportional to the differences between every pair of values, i.e. larger magnitudes indicate sharper changes

at corresponding locations and zero magnitudes indicate 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.

2.3. Selection of tuning parameters

In addition to the Lasso tuning parameter λ in (2.5), 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 criterion, 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 by much.

2.4. 3D case

The ADNI's FDG PET brain images are 3D. A 3D function can be decomposed using a 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 constructed by the 3D wavelet functions as follows:

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

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

$$Y_i = \beta_0 + \int_0^{T_1} \int_0^{T_2} \int_0^{T_3} X_i(u,v,w)\beta(u,v,w)\ dudvdw + \varepsilon_i. \quad (2.6)$$

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

$$\sum_{k,l,m=0}^{2^{j_0}-1} a_{j_0,\{k,l,m\}}\phi_{j_0,\{k,l,m\}}(u,v,w)+\sum_{j=j_0}^{J}\sum_{k,l,m=0}^{2^j-1}\sum_{q=1}^{7}d^q_{j,\{k,l,m\}}\psi^q_{j,\{k,l,m\}}(u,v,w). \quad (2.7)$$

Denote the set of all basis functions $\varphi_{j,\{k,l,m\}}$ and $\psi^q_{j,\{k,l,m\}}$ in (2.7) by $B(u,v,w)$ and the wavelet coefficients in (2.7) by η, then $\beta(u,v,w)$ can be written as

$$\beta(u,v,w)=B(u,v,w)^T\eta+e(u,v,w). \quad (2.8)$$

Plugging (2.8) into model (2.6), we obtain

$$\begin{aligned} Y_i&=\beta_0+\int_0^{T_1}\int_0^{T_2}\int_0^{T_3}X_i(u,v,w)B(u,v,w)^T\eta\,dudvdw+\varepsilon_i^*\\ &=\beta_0+C_i^T\eta+\varepsilon_i^*, \end{aligned} \quad (2.9)$$

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 $X_i(u,v,w)$, and

$\varepsilon_i^*=\varepsilon_i+\int_0^{T_1}\int_0^{T_2}\int_0^{T_3}X_i(u,v,w)e(u,v,w)\,dudvdw$. Then the methodology proposed in the previous subsections for the 1D case can be applied directly.

Following the calculations of Bickel, Ritov and Tsybakov (2009), we can show that our proposed method also enjoys the non-asymptotic oracle inequalities similar to the linear model with high-dimensional covariates. All the theoretical results are provided in the online supplementary material. It should be noted that the results are derived using the 1D notation for the estimator $\hat{\beta}(t)$ for simplicity, but they hold exactly for the 3D case of $\hat{\beta}(u,v,w)$.

3. 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. The results for 1D cases can be easily visualized, whereas the 3D case mimics the brain images more naturally.

3.1. 1D simulation

We consider a variety of settings of $X(t)$ and $\beta(t)$. For $X(t)=X^*(t)+\varepsilon(t)$ defined on $0 \leq t \leq 1$, where $\varepsilon(t)\sim N(0,\sigma^2_\varepsilon)$ is the noise term independent of 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. $X^*(t)=\Sigma a_i\varphi_i(t)$, where $\varphi_i(t)$ 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 $\varepsilon(t)$ is set to be $\sigma_\varepsilon^2 = \frac{1}{p-1}\sum_{j=1}^p (X^*(t_j) - \overline{X}^*(t_j))^2$, where $X^*(t_j)$ is the mean of $X^*(t_j)$. The error term ε in model (1.1) also follows a normal distribution $N(0, \sigma^2)$. The value of σ^2 is determined by the signal-to-noise ratio:

$$\text{SNR} = \frac{\sigma_g^2}{\sigma^2}, \quad (3.1)$$

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 article are under SNR = 9, which is also considered in Example 4 of Tibshirani (1996). For the Lasso method, Zou (2006) observed that smaller SNR usually yields smaller relative prediction error. 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, Hastie and Tibshirani, 2007) given below:

$$\text{AIC} = \frac{\|Y - \hat{g}(X)\|^2}{n\hat{\sigma}^2} + \frac{2}{n}\hat{df}, \quad (3.2)$$

$$\text{BIC} = \frac{\|Y - \hat{g}(X)\|^2}{n\hat{\sigma}^2} + \frac{log(n)}{n}\hat{df}, \quad (3.3)$$

where $\hat{df}$ is the number of nonzero elements of $\hat{\eta}$ in model (2.4). We estimate σ^2 by the refitted cross-validation method introduced in Fan, Guo and Hao (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 1. We also report the percentages of correctly identified zero regions and nonzero regions in Table 1. We can see that all four methods perform reasonably well, while the non-practical SV method performs the best. The CV method seems to have a nice trade-off between sparsity and prediction accuracy. Averages of the estimates of $\beta(t)$ using the CV method over 100 replications are shown in Figure 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 the null hypothesis at each t_j. The results of the permutation tests using the CV method are presented in Figure 2, which shows high rejection frequency in the regions where $\beta(t)$ is nonzero.

3.2. 3D simulation

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

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

where $a_i \sim N(0, 1)$ and $\varepsilon(u, v, w) \sim N(0, \sigma_\varepsilon^2)$ with σ_ε^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,

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

where $a = 1/8$, $b = 40/3$ and $c = \pi/6$. Note that $\beta(u, v, w)$ is zero outside a ball that is located in the center of the unit cube. The error term ε in model (2.6) also follows a normal distribution $N(0, \sigma^2)$ with SNR = 9. We generate 400 training images and apply the 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. Figure 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 performs reasonably well in detecting signals based on visual inspection and on the high percentage of correctly identified non-zeros and zeros reported in Table 2.

4. ADNI PET analysis

The FDG PET data used in the preparation of this article 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 in developing new treatments, monitoring treatment effectiveness, and lessening 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, Wang and Zhu (2009), is too computationally expensive for this application. 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 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 4. We treat each PET image as a realization of the 3D functional predictor and then fit the 3D functional linear regression model (2.6). 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 the 3D Haar wavelet transforms to each subject's PET image data, we consider all the possible levels of the Haar wavelet decompositions. Two 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 employ a 10-fold cross-validation to evaluate the predictive power of the proposed method. Specifically, for each set of 10% observations, we leave them out as a test set, use the remaining data as the training data to fit a model (including selecting the tuning parameters via five-fold cross-validation) and compute the prediction error on the data points that have been left out. We aggregate these quantities by using the predictive R-square given by $1 - \Sigma(y_i - \hat{y}_{i,-i})^2/\Sigma(y_i - \bar{y})^2$, where $\hat{y}_{i,-i}$ denotes the predicted value of y_i calculated by using the estimator obtained from the training data generated from the cross-validation. The result is 0.26 for the ADNI data set, whereas the standard R-square is 0.51, suggesting a moderate predictive power of the model.

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 5, which are presented from the bottom of the brain to the 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 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 way to control the family-wise error rate by comparing $\hat{\beta}(t_j)$ to a "maximal statistic". 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 also have 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 6. The locations of these more frequently selected voxels are also presented in the 3D sagittal view in Figure 7 for ease of understanding. It can be seen that the highly selected brain regions agree well with the results in Figure 5. We note that the clusters of voxels identified in our analysis shown in Figures 5 and 6 reveal high associations of the expected anatomical regions with cognitive deficits. For example, the orange ones on slices "+12" and "+18" in Figure 5 and the big cluster on the same slices in Figure 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 5 and the clusters on the same slices in Figure 6 suggest that the medial temporal/hippocampal cortex is also closely involved; the red ones on slices "−42", "−36" and "−30" in Figure 5 and the corresponding clusters on the same slices in Figure 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, Foster et al. (1984); Minoshima et al. (1995, 1997); 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.

5. Discussion

In this article, we propose a highly effective Haar wavelet-based regularization approach that can be easily applied to analyzing multidimensional functional data. Analysis of the PET image 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 coordinate descent algorithm with the MATLAB glmnet package. For example, the real data analysis of 403 subjects' PET image data can be finished in less than two hours on a 64-bit Intel Xeon 3.33GHz server with about 35GB of RAM, including the selection of tuning parameters. We should note that another practical advantage of our approach is that the wavelet transform itself can reduce the dimensionality of the large volume of brain image data. As a result, we can then apply the proposed approach on reduced data sets. In such situations, although the resolution of the original PET images is decreased, the results remain largely the same since the related subregions are usually not comprised of a single voxel but of a cluster of voxels.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Bickel P, Ritov Y, Tsybakov A. Simultaneous analysis of Lasso and Dantzig Selector. Annals of Statistics. 2009; 4:1705–1732.

Bowman FD, Caffo B, Bassett SS, Kilts C. A Bayesian hierarchical framework for spatial modeling of fMRI data. NeuroImage. 2008; 39:146–156. [PubMed: 17936016]

Candes E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. Annals of Statistics. 2007; 35:2313–2351.

Cockrell JR, Folstein MF. Mini-Mental State Examination (MMSE). Psychopharmacology Bulletin. 1988; 24:689–692. [PubMed: 3249771]

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

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

Foster NL, Chase TN, Mansi L, Brooks R, Fedio P, Patronas NJ, Chiro GD. Cortical abnormalities in Alzheimer's disease. Annals of Neurology. 1984; 16:649–654. [PubMed: 6335378]

Friedman J, Hastie T, Höfling H, Tibshirani R. Pathwise coordinate optimization. Annals of Applied Statistics. 2007; 1:302–332.

Friston KJ, Poline JB, Holmes AP, Frith C, Frackowiak RSJ. A multivariate analysis of PET activation studies. Human Brain Mapping. 1996; 4:140–151. [PubMed: 20408193]

Fu WJ. Penalized regressions: the bridge versus the lasso. Journal of Computational and Graphical Statistics. 1998; 7:397–416.

Grimmer T, Henriksen G, Wester HJ, Förstl H, Klunk WE, Mathis CA, Kurz A, Drzezga A. Clinical severity of Alzheimer's disease is associated with PIB uptake in PET. Neurobiology of Aging. 2009; 30:1902–1909. [PubMed: 18346821]

James GM, Wang J, Zhu J. Functional linear regression that's interpretable. Annals of Statistics. 2009; 37:2083–2108.

Kang J, Johnson TD, Nichols TE, Wager TD. Meta analysis of functional neuroimaging data via Bayesian spatial point processes. Journal of the American Statistical Association. 2011; 106:124–134. [PubMed: 21706069]

Kerrouche N, Herholz K, Mielke R, Holthoff V, Baron J-C. 18F-FDG PET in vascular dementia: differentiation from Alzheimer's disease using voxel-based multivariate analysis. Journal of Cerebral Blood Flow and Metabolism. 2006; 26:1213–1221. [PubMed: 16525414]

Leifer BP. Early diagnosis of Alzheimer's disease: clinical and economic benefits. Journal of the American Geriatrics Society. 2003; 51:S281–S288. [PubMed: 12801384]

Lovejoy S, Schertzer D. Haar wavelets, fluctuations and structure functions: convenient choices for geophysics. Nonlinear Processes in Geophysics. 2012; 19:513–527.

Luo WL, Nichols TE. Diagnosis and exploration of massively univariate neuroimaging models. NeuroImage. 2003; 19:1014–1032. [PubMed: 12880829]

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

Minoshima S, Frey KA, Koeppe RA, Foster NL, Kuhl DE. A diagnostic approach in Alzheimer's disease using three-dimensional stereotactic surface projections of fluORine-18-FDG PET. Journal of Nuclear Medicine. 1995; 36:1238–1248. [PubMed: 7790950]

Minoshima S, Giordani B, Berent S, Frey KA, Foster NL, Kuhl DE. Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. Annals of Neurology. 1997; 42:85–94. [PubMed: 9225689]

Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jaqust W, Trojanowski JQ, Toga AW, Beckett L. 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. 2005; 1:55–66.

Muraki S. Approximation and rendering of volume data using wavelet transforms. In. Proceedings of Visualization 1992. 1992:21–28.

Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Human Brain Mapping. 2001; 15:1–25. [PubMed: 11747097]

Picart SS, Butenschön M, Shutler JD. Wavelet-based spatial comparison technique for analysing and evaluating two-dimensional geophysical model fields. Geoscientific Model Development. 2012; 5:223–230.

Plassman BL, Langa KM, Fisher GG, Heeringa SG, Weir DR, Ofstedal MB, Burke JR, Hurd MD, Potter GG, Rodgers WL, Steffens DC, Willis RJ, Wallace RB. Prevalence of dementia in the United States: the aging, demographics, and memory study. Neuroepidemiology. 2007; 29:125–132. [PubMed: 17975326]

Ramsay, J.; Silverman, B. Functional data analysis. 2. Springer; 2005.

Reiss PT, Ogden RT. Functional generalized linear models with images as predictors. Biometrics. 2010; 66:61–69. [PubMed: 19432766]

Royston, P.; Sauerbrei, W. Multivariable Model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. 1. Wiley; 2008.

Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. Statistics in Medicine. 1992; 11:2093–2109. [PubMed: 1293671]

Shin J, Lee S, Kim S, Kim S, Cho S, Kim Y. Voxel-based analysis of Alzheimer's disease PET imaging using a triplet of radiotracers: PIB, FDDNP, and FDG. NeuroImage. 2010; 52:488–496. [PubMed: 20385246]

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

Walker, JS. A primer on wavelets and their scientific applications. 2. Chanpman & Hall/CRC; 2008.

Wand MP, Ormerod JT. Penalized wavelets: embedding wavelets into semiparametric regression. Electronic Journal of Statistics. 2011; 5:1654–1717.

Wu TT, Lange K. Coordinate descent algorithms for lasso penalized regression. Annals of Applied Statistics. 2008; 2:224–244.

Zhao Y, Ogden RT, Reiss PT. Wavelet-based LASSO in functional linear regression. Journal of Computational and Graphical Statistics. 2012; 21:600–617. [PubMed: 23794794]

Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101:1418–1429.

Zou H, Hastie T, Tibshirani R. On the degrees of freedom of the lasso. Annals of Statistics. 2007; 35:2173–2192.
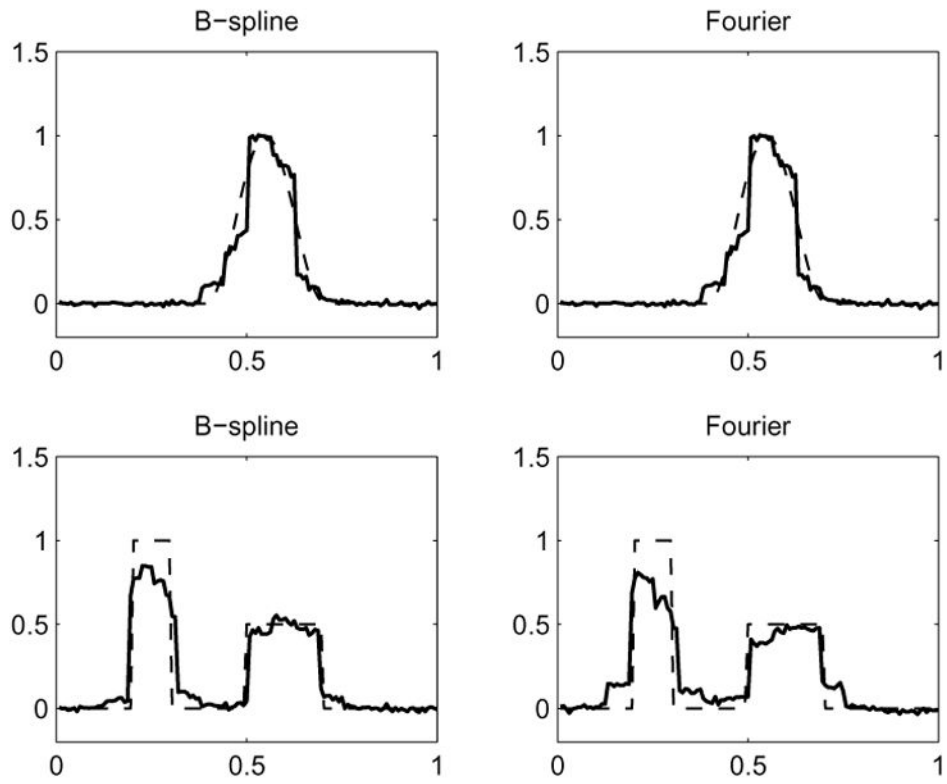
Fig 1.
Average of $\hat{\beta}(t)$ estimated using 5-fold cross-validation with 100 replications (solid line). The dashed line is the true $\beta(t)$. The top panel is for case 1, and the bottom panel is for case 2.
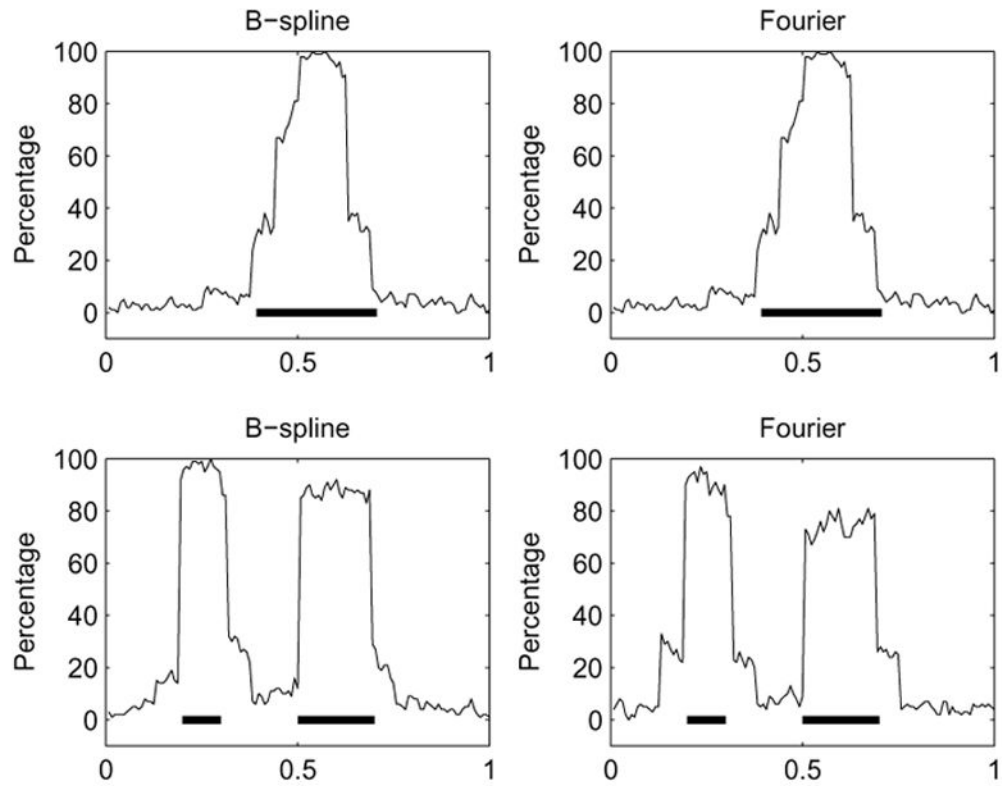
Fig 2.
Frequency of rejecting the null hypothesis $\beta(t) = 0$ using 5-fold cross-validation based on 100 permutation repetitions. The thick solid horizontal segments indicate the true nonzero regions. The top panel is for case 1, and the bottom panel is for case 2.
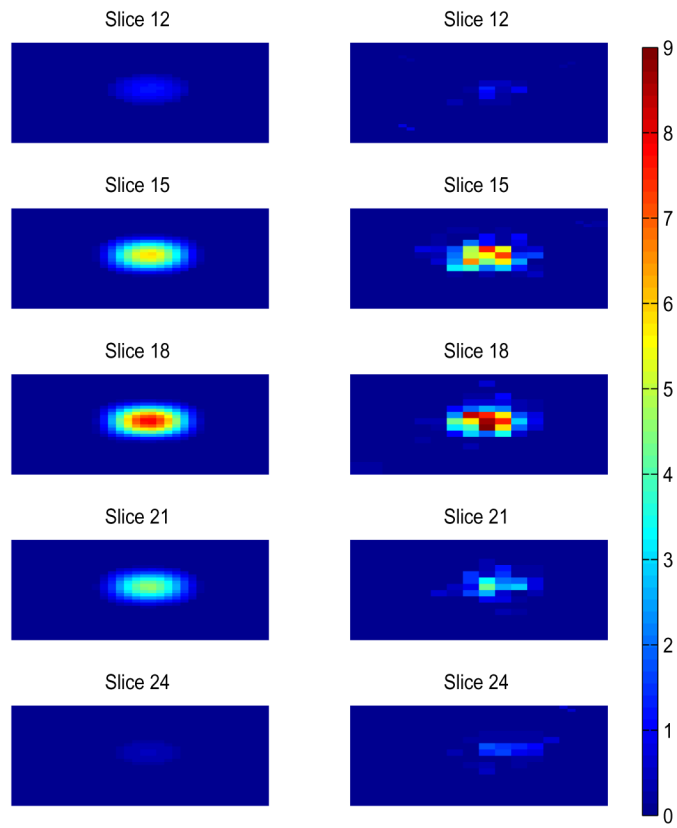
Fig 3.
The left panel is true $\beta(u, v, w)$ at five selected slices and the right panel is the average of $\hat{\beta}(u, v, w)$ estimated using 5-fold cross-validation over 100 replications at the same five slices.
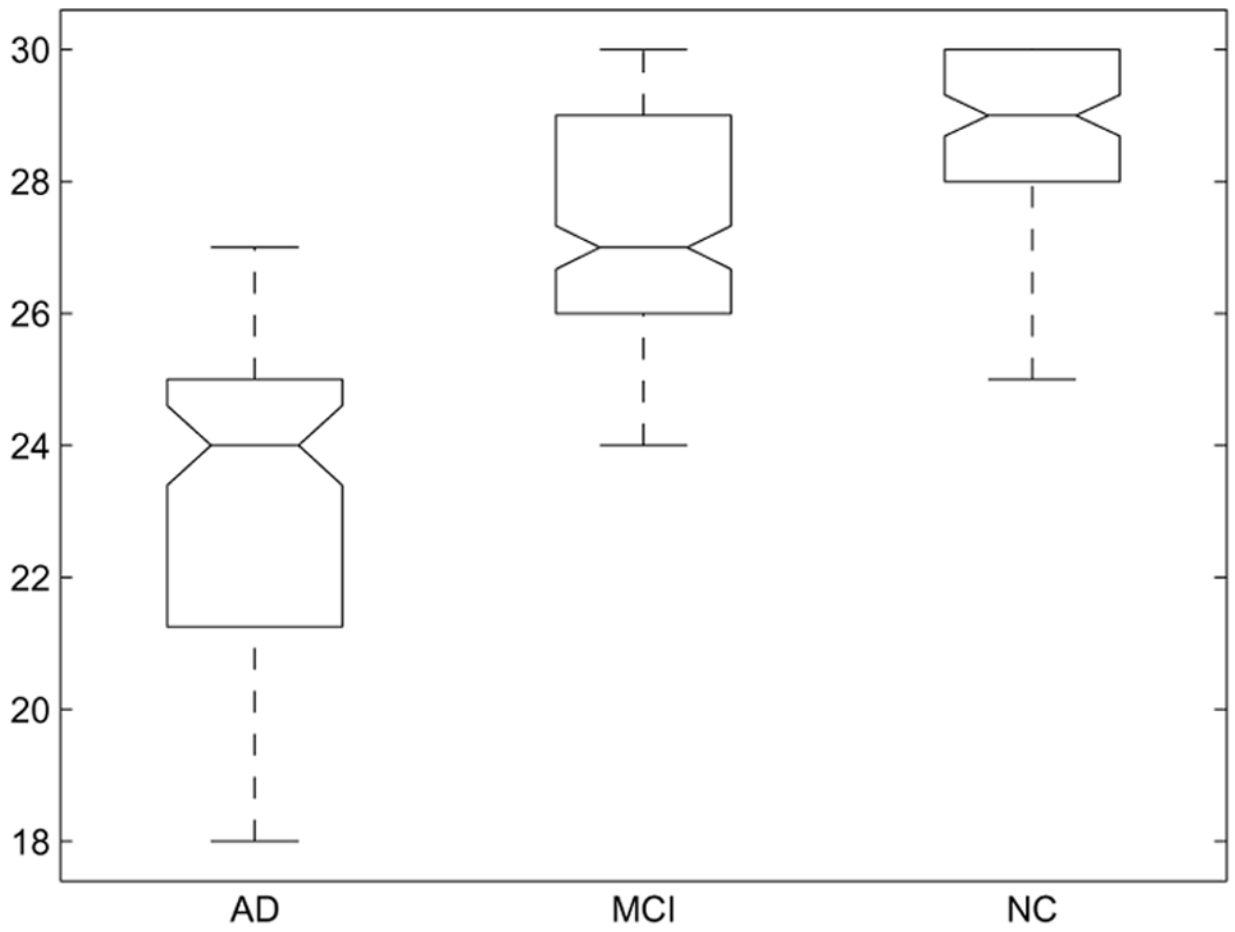
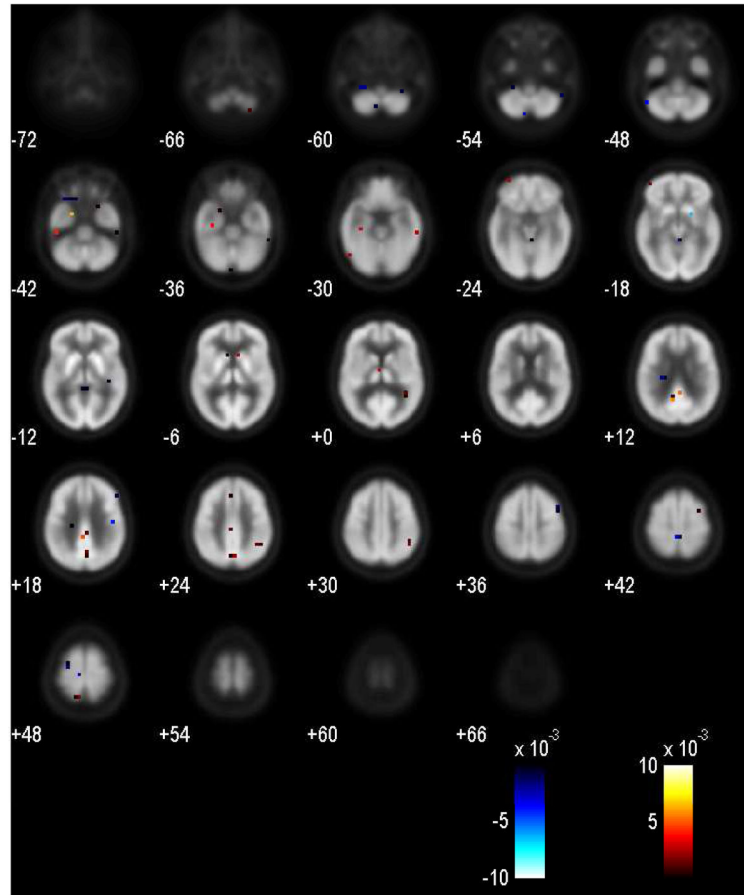**Fig 4.**
Box plots of MMSE scores among AD, MCI, and NC.

Fig 5.
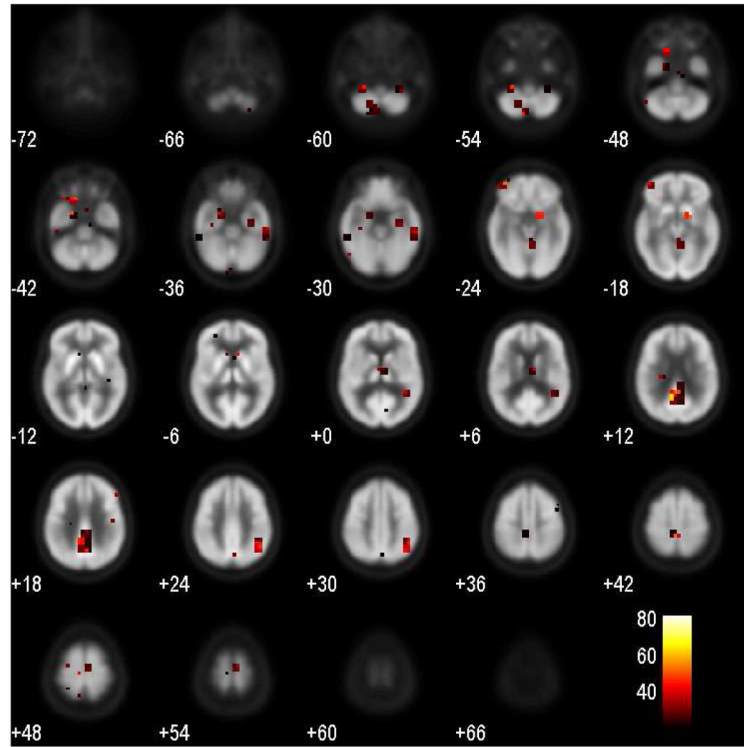Clusters of voxels identified using our approach for the ADNI data.

Fig 6.
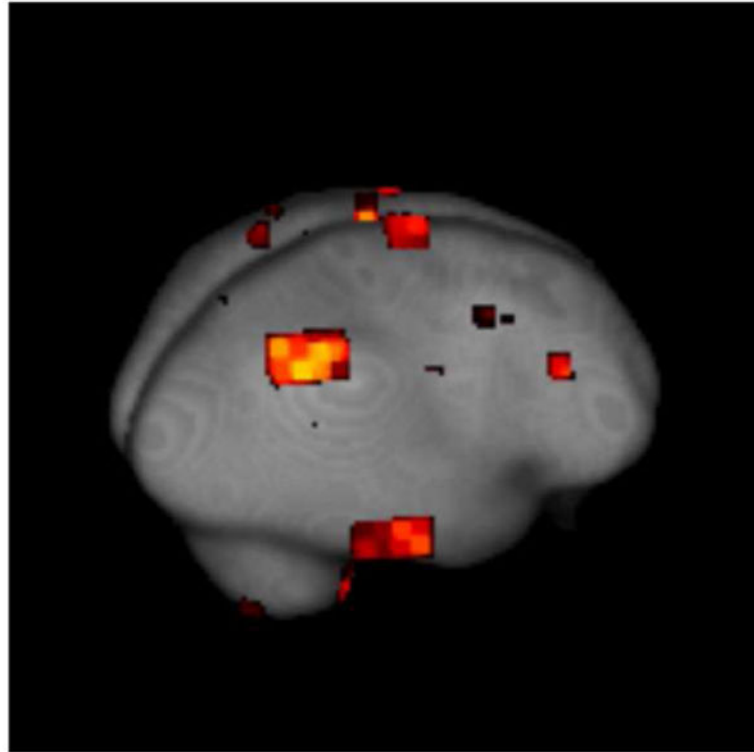Bootstrap inclusion frequencies of the voxels over 100 bootstrap samples.

Fig 7.
Locations of the frequently selected voxels in the 3D sagittal view.

Table 1

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

Type	Method	Average MSE (SE) (×10⁻³)		Average percentage (%)			
		Case 1	Case 2	Case 1		Case 2	
				Nonzero	Zero	Nonzero	Zero
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

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

| Method | Average MSE (SE) (×10⁻⁴) | Average percentage (%) | |
		Nonzero	Zero
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 3

Demographics of ADNI participants (n=403)

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)