# Functional Linear Models for Brain Data

Junier Barbaro Oliva

May 2, 2016

**Abstract**

**Background.** Modern neuroimaging data has provided a much needed window into the intricacies of the human brain. Neuroimaging techniques such as functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), and diffusion tensor imaging (DTI), often contain many thousands of functional observations per subject. While some success has been had using heuristical summary statistics of neuroimaging functional covariates to build predictive models, there is a lack of a flexible and principled framework to build models from neuroimaging techniques.

**Aim.** Our aim is to develop a general, principled framework for supervised learning that scales to many thousands of neuroimaging functional covariates per subject without resorting to heuristical summary statistics. In particular, we look to more accurately regress a subject's age given neuroimaging data through the use of functional covariates than previous approaches that considered only summary statistics.

**Data.** Diffusion magnetic resonance (MR) images were acquired from a total of 90 subjects. The subjects ranged in age from 13 to 60 years old, and included 45 males and 45 females. The subjects had no known history of neurological or mental disorder. We used per-voxel functional covariates of *diffusion orientation distribution functions* (dODFs) in our models and real-valued fractional anisotropy (FA) summary statistic covariates as a baseline. dODFs are functions that represent the amount of water molecules, or spins, undergoing diffusion in different orientations over the $S^2$ sphere. FA values are the peak of the dODF and have previously been used as a summary statistic of dODFs for age regression. In total, 25K voxels were considered in a shared template-space.

**Methods.** We develop and empirically test a linear functional model for neuroimaging data. We use a model that represents neuroimaging functional observations nonparametrically using orthonormal basis projections. The age response is modelled as a sparse additive linear combination of inner products. We show that our model may be optimized using a group-LASSO approach.

**Results.** Our functional sparse linear model was able to predict age with a 29.15% lower mean squared error (MSE) than previous heuristic-based methods that used summary statistics. Our approach yielded an MSE of 58.49. Results were found to be statistically significant with a $p$-value of 0.04 using a paired $t$-test.

**Conclusions.** In this work we have developed and tested a principled framework for building predictive supervised models with functional neuroimaging data. This framework allows one to use functional neuro-data in a statistically principled fashion without resorting to heuristical summary statistics.

***Keywords:*** Neuroimaging, Functional Data, Nonparametric, Sparse Models

# 1 Introduction

Modern data collection has allowed us to collect not just more data, but more complex data. This is especially so for neuroimaging data, where each subject's scan may have hundreds of thousands of voxels associated with it, and each voxel corresponds to a functional object such as a probability density function. It would be beneficial to perform machine learning tasks using these functional objects. However, many existing techniques cannot handle complex, possibly infinite dimensional, objects; hence, one often resorts to the heuristic of representing these complex objects by ad-hoc summary statistics. In this paper, we develop a principled, general framework for regressing a real-valued response when given many functional input covariates. As a case study, we consider the task of predicting a subject's age given diffusion imaging data that contains thousands of functional covariates.

Studies have shown that aging produces several changes in the human brain including synaptic density variations (Huttenlocher and De Courten, 1986), myelin level changes (Benes et al., 1994), and Schwann cell subunit density changes (Kanda et al., 1991). Recently, diffusion tensor imaging (DTI) techniques have made it possible to characterize anatomical tissue in vivo, and numerous studies have linked DTI measurements to brain aging and maturation (Johansen-Berg and Behrens, 2013). In this work, we leverage semi-parametric models to predict and learn the aging process in the human brain. Such results are not only of interest scientifically, but may also serve practical medical applications. For instance, studying abnormal brain aging and maturity can help diagnose cognitive and paediatric, disorders such as ADHD, that alter brain maturation (Shaw et al., 2012).

## 1.1 Problem Being Solved

In this work we regress a subject's age given thousands of voxel diffusion orientation distribution function (dODF) covariates, each of which specify the orientation of water molecule diffusion at a corresponding voxel. There are several unique challenges to performing regression on this data; one's approach must be:

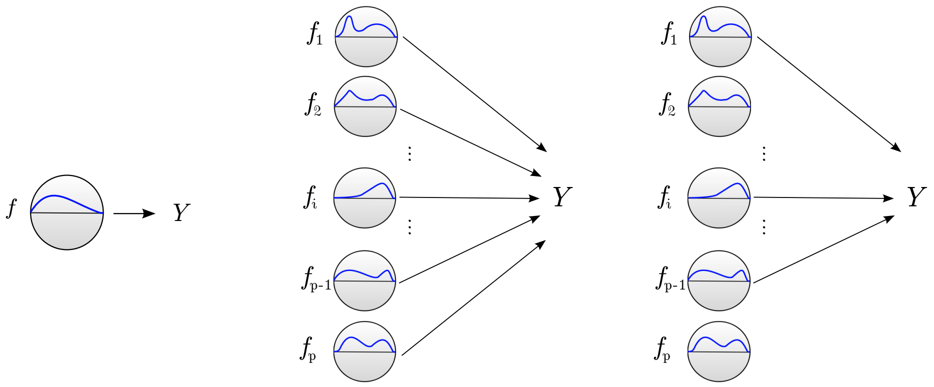
**Interpretable** It is vital that we be able to comprehend some of the underlying mechanics in our model. Being able to understand our model will allow us to not only achieve accurate age predictions but also to uncover details of the aging process of the human brain, which is of scientific and medical interest.

**Flexible** Although some summary statistics of dODFs have proven to be useful previously, we believe there are several reasons to favor a more nonparametric treatment of dODFs:

1. Preprocessing the data into heuristical summary statistics biases the type of scientific discoveries models can uncover; a more nonparametric approach can yield unexpected relations between the response (age) and our functional covariates.
2. Whilst there have been a few effective summary statistics for age prediction, it is unclear if the same statistics will prove effective for other prediction tasks. Nonparametric approaches alleviate the need to develop new statistics for new tasks.
3. The use of ad-hoc summary statistic precludes our ability to theoretically analyze one's model from a general perspective. A nonparametric approach allows us to give sufficient conditions under which our models are effective.
4. By being more flexible, we hypothesize that a nonparametric functional approach will achieve better accuracy.

**Scalable** We will consider thousands of functional covariates, hence we need a framework which can scale to many covariates both computationally and statistically.

**Generalizable** Given the massive number of covariates relative to the number of instances in brain imaging data, it is imperative that our model be able to generalize to held-out data and not over-fit to training instances.



(a) Single Functional Covariate

(b) Multiple Functional Covariates

(c) Sparse Model

Figure 1: (a) Model where mapping takes in a function $f$ and produces a real $Y$. (b) Model where response $Y$ is dependent on multiple input functions $f_1, \ldots, f_p$. (c) Sparse model where response $Y$ is dependent on a sparse subset of input functions $f_1, \ldots, f_p$.

## 1.2 Approach

We take a semi-parametric approach to perform sparse regression with multiple input functional covariates and a real-valued response, the FuSSO: Functional Shrinkage and Selection Operator (Oliva et al., 2014). No parametric assumptions are made on the nature of input functions. We shall model the age response $Y$ as the result of a sparse set of linear combinations of dODF input functions $\{f_j\}$ and other unknown nonparamteric functions $\{g_j\}$ that we optimize for:

$$Y = \sum_j \langle f_j, g_j \rangle. \tag{1}$$

The resulting method is a LASSO-like (Tibshirani, 1996) estimator that effectively zeros out entire functions from consideration in regressing the response. Our approach is particularly adept at solving the unique challenges presented in age regression with dODF covariates since it is:

**Interpretable** A sparse additive functional linear model (Figure 1(c)) is interpretable in several ways. First, sparsity at a voxel level will be indicative of relevant regions in the brain for determining age. Second, the unknown functions $\{g_j\}$ recovered in our model (1) may be analyzed to determine what are important characteristics of dODFs for determining age.

**Flexible** By taking a semi-parametric approach we achieve a flexible model that does not assume a specific form to our input functional covariates $\{f_j\}$, nor assumes any specific form to the voxel wide linear functions contributions $\{g_j\}$.

**Scalable** Imposing sparsity allows us to efficiently optimize the FuSSO model using active set and FISTA approaches (Huang et al., 2011).

**Generalizable** The sparsity inducing norms considered will regularize our model and prevent over-fitting to training data.

# 2 Related Work/Background

## 2.1 Age and Neuroimaging

The effects of aging on the human brain have long been studied (Huttenlocher and De Courten, 1986; Kanda et al., 1991; Benes et al., 1994), but it is only recently that we may observe aging in vivo for human subjects through neuroimaging data. Several works have explored associations in neuroimaging to aging and maturity (Gunning-Dixon et al., 2009; Johansen-Berg and Behrens, 2013). More recently several efforts have been put forth to predict age given neuroimging data. For instance, Dosenbach et al. (2010); Franke et al. (2010, 2012) explore age prediction with fMRI data. There has also been limited work performing age prediction with diffusion imaging data. For example, Han et al. (2014) explore predicting the age of a subject given a structural functional network derived from diffusion imaging. Furthermore, Mwangi et al. (2013) research predicting age given summary statistics from voxel dODFs. Most notably, functional anisotropy (FA) values were used as covariates in the age regression task. A higher FA value represents diffusion occurring along one direction but largely restricted in all other directions (Johansen-Berg and Behrens, 2013). Furthermore, higher FA values are often associated with highly myelinated white-matter tracts and vice versa (Kochunov et al., 2012); FA decreases consistently with aging.

## 2.2 Function Covariate Models

There has been a considerable amount of recent work studying linear models that operate over functional covariates. As previously mentioned, we build off of the FuSSO (Oliva et al., 2014), which yields a sparse mapping that takes multiple functional inputs and produces a real-valued output. Other LASSO-like regression estimators that work with functional data include the following. Mingotti et al. (2013) consider a functional output and several real-valued covariates. Here, the estimator finds a sparse set of functions to scale by the real valued covariates to produce the functional response. Also, Zhao et al. (2012); James et al. (2009) study the case when one has one functional covariate $f$ and one real valued response that is linearly dependent on $f$ and some function $g$: $Y = \langle f, g\rangle = \int fg$. Zhao et al. (2012) use an estimator that searches for sparsity across wavelet basis projection coefficients. James et al. (2009) achieve sparsity in the time (input) domain of the $d^{\text{th}}$ derivative of $g$; i.e. $[D^d g](t) = 0$ for many values of $t$ where $D^d$ is the differential operator. Hence, roughly speaking, Zhao et al. (2012); James et al. (2009) look for sparsity across frequency and time domains respectively, for the regressing function $g$. For a general discussion of functional spaces and linear operators on functional covariates see (Ferraty and Vieu, 2006).

# 3 Data

## 3.1 Apparatus and Instrumentation

The diffusion magnetic resonance (MR) images were acquired on a 3 T MRI scanner (Trio, Siemens, Erlangen, Germany) using an 8-channel head coil. The images were acquired by a T2-weighted fast spin echo sequence with parameters TR/TE=5920/102 ms, and slice thickness of 3.0 mm, 256 × 256 acquisition matrix, and 25 × 25 cm field of view rendered in 35 axial slices. See (Yeh and Tseng, 2011) for additional details.

## 3.2 NTU90 Dataset

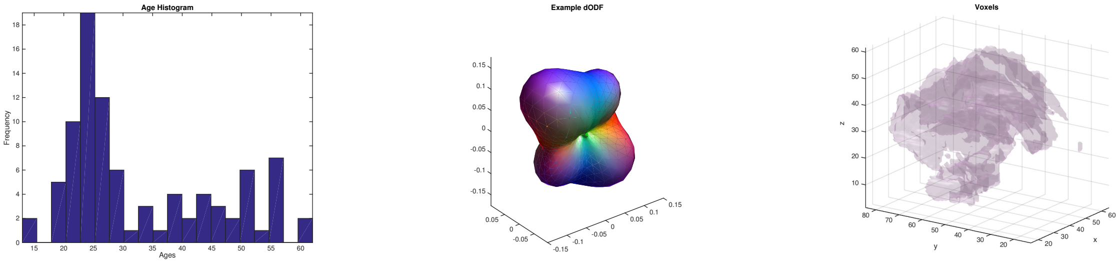
We use the NTU90 dataset (Yeh and Tseng, 2011). Diffusion MR images were acquired from a total of 90 subjects. The subjects ranged in age from 13 to 60 years old, and included 45 males and 45 females. A histogram of ages can be seen in Figure 2(a), the mean age is 32.99 with a standard deviation of 12.68. The subjects had no known history of neurological or mental disorder.

Diffusion orientation distribution functions (dODF) were extracted for over 185,000 white-matter voxels in a shared template space for all subjects. The dODF is a function that represents the amount of water molecules, or spins, undergoing diffusion in different orientations over the $S^2$ sphere (Yeh and Tseng, 2011).

In other words, each dODF is a function with a 3d domain (of spherical coordinates) and a range of reals representing the strength of water diffusion at the given orientations (Figure 2(b)).

We also used the non-functional collection of fractional anisotropy (FA) values for the same white matter voxels as with dODF functions. FA values are the estimated amount of spins that undergo diffusion in the direction of the principle fiber orientation, i.e., the peak of the dODF; FAs have been used as a measure of white matter integrity in the underlying voxel hence making for a descriptive and effective summary statistic of an dODF function for age regression (Mwangi et al., 2013).

For computational considerations, we analyze only the top 25K voxels according to top mean absolute FA value. See Figure 2(c) for the volume of the considered 25K white matter voxels in the shared template space.



(a)

(b)

(c)

Figure 2: (a) Histogram of ages. (b) Example dODF with values plotted for each spherical coordinate. (c) Considered voxels in shared template space.

# 4 Method

Below we expound on our approach for analyzing the NTU90 dataset. First, we go over the FuSSO, our main method for regressing age given dODF functional covariates. After, we describe a particularly useful extension, the Elastic-FuSSO, which can help select correlated voxels. Then, we give details on our representation of dODF functions using a cosine basis set. Later, we expound on a two-step process for building our final model. Lastly, we discuss some selection techniques for selecting relevant voxels.

## 4.1 FuSSO

The FuSSO operates over regression tasks where one regresses a real-valued response given several functional covariates. To better understand the FuSSO's model we draw several analogies to real-valued linear regression and the Group-LASSO (Yuan and Lin, 2006). Note that although for simplicity we illustrate the FuSSO on functions working over a one dimensional domain, it is straightforward to extend the estimator and results to the multidimensional case.

First, consider a model for typical real-valued linear regression with a data-set of input-output pairs $\{(X_i, Y_i)\}_{i=1}^N$:

$$Y_i = \langle X_i, w \rangle + \epsilon_i,$$

where $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^d$, $w \in \mathbb{R}^d$, $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, and $\langle X_i, w \rangle = \sum_{j=1}^d X_{ij} w_j$. If instead one were working with functional data $\{(f^{(i)}, Y_i)\}_{i=1}^N$, where $f^{(i)} : [0, 1] \mapsto \mathbb{R}$ and $f^{(i)} \in L_2[0, 1]$, one may similarly consider a linear model :

$$Y_i = \langle f^{(i)}, g \rangle + \epsilon_i,$$

where, $g \in L_2[0,1]$, and $\langle f^{(i)}, g\rangle = \int_0^1 f^{(i)}(t)g(t)\mathrm{d}t$. If $\Phi = \{\varphi_m\}_{m=1}^{\infty}$ is an orthonormal basis for $L_2[0,1]$ (Tsybakov, 2008) then we have that

$$f^{(i)}(x) = \sum_{m=1}^{\infty} \alpha_m^{(i)} \varphi_m(x), . \tag{2}$$

where, $\alpha_m^{(i)} = \int_0^1 f^{(i)}(t)\varphi_m(t)\mathrm{d}t$. Similarly, $g(x) = \sum_{m=1}^{\infty} \beta_m \varphi_m(x)$. Thus,

$$\begin{aligned} Y_i = \langle f^{(i)}, g\rangle + \epsilon_i &= \langle \sum_{m=1}^{\infty} \alpha_m^{(i)} \varphi_m(x), \sum_{k=1}^{\infty} \beta_k \varphi_k(x)\rangle + \epsilon_i = \sum_{m=1}^{\infty}\sum_{k=1}^{\infty} \alpha_m^{(i)} \beta_k \langle \varphi_m(x), \varphi_k(x)\rangle + \epsilon_i \\ &= \sum_{m=1}^{\infty} \alpha_m^{(i)} \beta_m + \epsilon_i, \end{aligned}$$

where the last step follows from orthonormality of $\Phi$.

Going back to the real-valued covariate case, if instead of having one feature vector per data instance: $X_i \in \mathbb{R}^d$, one had $p$ feature vectors associated to each data instance: $\{X_{ij} \mid 1 \leq j \leq p,\ X_{ij} \in \mathbb{R}^d\}$, an additive linear model may be used for regression:

$$Y_i = \sum_{d=1}^{p} \langle X_{id}, w_d\rangle + \epsilon_i, \text{where } w_1, \ldots, w_d \in \mathbb{R}^d.$$

Similarly, in the functional case one may have $p$ functions associated with data instance $i$: $\{f_j^{(i)} \mid 1 \leq j \leq p,\ f_j^{(i)} \in L_2[0,1]\}$. Then, an additive linear model would be:

$$Y_i = \sum_{j=1}^{p} \langle f_j^{(i)}, g_j\rangle + \epsilon_i = \sum_{j=1}^{p}\sum_{m=1}^{\infty} \alpha_{jm}^{(i)} \beta_{jm} + \epsilon_i, \tag{3}$$

where $g_1, \ldots, g_p \in L_2[0,1]$, and $\alpha_{jm}^{(i)}$ and $\beta_{jm}$ are projection coefficients for $f_j^{(i)}$ and $g_j$ respectively.

Suppose that one has few observations relative to the number of features ($N \ll p$). In the real-valued case, in order to effectively find a solution for $w = (w_1^T, \ldots, w_p^T)^T$ one may search for a group sparse solution where many $w_j = 0$. To do so, one may consider the following Group-LASSO regression:

$$w^{\star} = \underset{w}{\operatorname{argmin}} \frac{1}{2N} \|Y - \sum_{j=1}^{p} X_j w_j\|^2 + \lambda \sum_{j=1}^{p} \|w_j\|, \tag{4}$$

where here $X_j$ is the $N \times d$ matrix $X_j = [X_{1j} \ldots X_{Nj}]^T$, $Y = (Y_1, \ldots, Y_N)^T$, and $\|\cdot\|$ is the Euclidean norm.

If in the functional case (3) one also has that $N \ll p$, one may set up a similar optimization to (4), whose direct analogue is:

$$g^{\star} = \underset{g}{\operatorname{argmin}} \frac{1}{2N} \sum_{i=1}^{N} \left( Y_i - \sum_{j=1}^{p} \langle f_j^{(i)}, g_j\rangle \right)^2 + \lambda \sum_{j=1}^{p} \|g_j\|; \tag{5}$$

equivalently,

$$\beta^{\star} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2N} \sum_{i=1}^{N} \left( Y_i - \sum_{j=1}^{p}\sum_{m=1}^{\infty} \alpha_{jm}^{(i)} \beta_{jm} \right)^2 + \lambda \sum_{j=1}^{p} \sqrt{\sum_{m=1}^{\infty} \beta_{jm}^2}, \tag{6}$$

where $g = \{g_j\}_{j=1}^{p} = \{\sum_{m=1}^{\infty} \beta_{jm}\varphi_m\}_{j=1}^{p}$.

However, it is unrealistic to assume that one is able to directly observe functional inputs $\{f_j^{(i)} \mid 1 \leq i \leq N, 1 \leq j \leq p\}$. Instead, let us suppose that one observes some sort of noisy functional observations $\{\vec{y}_j^{(i)} \mid 1 \leq i \leq N, 1 \leq j \leq p\}$. For instance, we may take these observations to be noisy function evaluations on a fixed grid:

$$\vec{y}_j^{(i)} = \vec{f}_j^{(i)} + \xi_j^{(i)}, \tag{7}$$

$$\vec{f}_j^{(i)} = \left( f_j^{(i)}(1/n),\ f_j^{(i)}(2/n),\ \ldots,\ f_j^{(i)}(1) \right)^T, \tag{8}$$

$$\xi_j^{(i)} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\xi^2 I_n). \tag{9}$$

In this case we observe a grid of $n$ noisy values for each functional input. Then, one may estimate $\alpha_{jm}^{(i)}$ as:

$$\tilde{\alpha}_{jm}^{(i)} = \frac{1}{n} \vec{\varphi}_m^T \vec{y}_j^{(i)} = \frac{1}{n} \vec{\varphi}_m^T (\vec{f}_j^{(i)} + \xi_j^{(i)}) = \bar{\alpha}_{jm}^{(i)} + \eta_{jm}^{(i)} \tag{10}$$

where $\bar{\alpha}_{jm}^{(i)} \equiv \frac{1}{n} \vec{\varphi}_m^T \vec{f}_j^{(i)}$, $\eta_{jm}^{(i)} \equiv \frac{1}{n} \vec{\varphi}_m^T \xi_j^{(i)}$, and $\vec{\varphi}_m = (\varphi_m(1/n),\ \varphi_m(2/n),\ \ldots,\ \varphi_m(1))^T$. Furthermore, we may truncate the number of basis functions used to express $f_j^{(i)}$ to $M_n$, estimating it as:

$$\tilde{f}_j^{(i)}(x) = \sum_{m=1}^{M_n} \tilde{\alpha}_{jm}^{(i)} \varphi_m(x). \tag{11}$$

Using the truncated estimate (11), one has:

$$\langle \tilde{f}_j^{(i)}, g_j \rangle = \sum_{m=1}^{M_n} \tilde{\alpha}_{jm}^{(i)} \beta_{jm}, \text{and } \|\tilde{f}_j^{(i)}\| = \sqrt{\sum_{m=1}^{M_n} (\tilde{\alpha}_{jm}^{(i)})^2}.$$

We note that we can compute functional estimates given other noisy function observations such as: noisy function observations on a random, irregular design; or samples drawn for pdf covariates $f_j^{(i)}$.

Using the approximations (11), (6) becomes:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \frac{1}{2N} \sum_{i=1}^{N} \left( Y_i - \sum_{j=1}^{p} \sum_{m=1}^{M_n} \tilde{\alpha}_{jm}^{(i)} \beta_{jm} \right)^2 + \lambda \sum_{j=1}^{p} \sqrt{\sum_{m=1}^{M_n} \beta_{jm}^2} \tag{12}$$

$$= \underset{\beta}{\text{argmin}} \frac{1}{2N} \|Y - \sum_{j=1}^{p} \tilde{A}_j \beta_j\|^2 + \lambda \sum_{j=1}^{p} \|\beta_j\|, \tag{13}$$

where $\tilde{A}_j$ is the $N \times M_n$ matrix with values $\tilde{A}_j(i, m) = \tilde{\alpha}_{jm}^{(i)}$ and $\beta_j = (\beta_{j1}, \ldots, \beta_{jM_n})^T$. Note that one need not consider projection coefficients $\beta_{jm}$ for $m > M_n$ since such projection coefficients will not decrease the MSE term in (12) (because $\tilde{\alpha}_{jm}^{(i)} = 0$ for $m > M_n$), and $\beta_{jm} \neq 0$ for $m > M_n$ increases the norm penalty term in (12). Hence we see that our sparse functional estimates are a Group-LASSO problem on the projection coefficients.

**Theory** It is worth briefly mentioning that the FuSSO is able to recover the sparsity pattern asymptotically; i.e., that the FuSSO estimate is sparsistent. Under technical assumptions (see (Oliva et al., 2014) for details) we have that:

**Theorem 1**: $\mathbb{P}\left(\hat{S}_N = S\right) \to 1$, where $\hat{S}_N$ is the selected support (i.e. nonzero $g_j$) with $N$ instances and $S$ is the true support.

## 4.2 The Elastic-FuSSO

Although there are some technical assumptions under which the FuSSO is asymptotically sparsistent, there are a few drawbacks to using the FuSSO in real world datasets, most of which stem from limitations of the group-LASSO. For instance, Liu and Zhang (2009) show that a solution to (13) that has at most $N$ non-zero groups always exists. Since we wish to optimize (13) using active-set approaches that consider a growing set of groups, this means that we will likely reach a solution that has saturated at $N$ voxels. This is unfortunate since we suspect that more than 90 voxels will be relevant in age regression. Furthermore, it has been empirically observed that the LASSO and similar estimators have difficulty selecting correlated sets of covariates (Zou and Hastie, 2005). Given the amount of correlations present in neuroimaging data, the inability to select correlated voxels also presents a challenge for us.

There are a few ways to mitigate these shortcomings without directly extending the FuSSO. For instance, below we shall describe variable selection techniques that work through repeated trials. However, such approaches prove to be computationally intensive given the need to solve many optimization problems. Instead, we also consider a simple extension to the FuSSO (13) that will allow us to easily select a non-saturated, correlated support.

The extension we consider, which we coin the "Elastic-FuSSO," is inspired by the elastic-net (Zou and Hastie, 2005). As with the elastic-net, we add a squared 2-norm penalty to our optimization problem:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2N} \|Y - \sum_{j=1}^{p} \tilde{A}_j \beta_j\|^2 + \sum_{j=1}^{p} \lambda \|\beta_j\| + \lambda_e \|\beta_j\|^2, \tag{14}$$

This simple extension will allow us to bypass the aforementioned shortcomings of the FuSSO. For instance, quickly analyzing stationarity conditions for optimality yields that for non-zero groups $\beta_j$, $\beta_k$:

$$\lambda_e(\beta_j - \beta_k) = \frac{1}{N}(\tilde{A}_j - \tilde{A}_k)^T \left( \sum_{l=1}^{p} \tilde{A}_l \beta_l - Y \right) + \lambda \left\| \frac{\beta_j}{\|\beta_j\|} - \frac{\beta_k}{\|\beta_k\|} \right\|,$$

and

$$\begin{aligned} \|\beta_j - \beta_k\| &\leq \frac{1}{\lambda_e} \left\| \tilde{A}_j - \tilde{A}_k \right\| \|r(\lambda, \lambda_e)\| + \frac{\lambda}{\lambda_e} \left\| \frac{\beta_j}{\|\beta_j\|} - \frac{\beta_k}{\|\beta_k\|} \right\| \\ &\leq \frac{1}{\lambda_e} \left\| \tilde{A}_j - \tilde{A}_k \right\| \|r(\lambda, \lambda_e)\| + \frac{\sqrt{2}\lambda}{\lambda_e} \sqrt{1 - \cos(\theta_{jk})}, \end{aligned}$$

where $r(\lambda, \lambda_e)$ is the residual under penalties $\lambda$ and $\lambda_e$, $r(\lambda, \lambda_e) = \frac{1}{N}(\sum_{l=1}^{p} \tilde{A}_l \beta_l - Y)$, and $\theta_{jk}$ is the angle between $\beta_j$, $\beta_k$. Hence, we see that similar designs for the $j^{\text{th}}$, $k^{\text{th}}$ covariates yield similar selected functions $g_j$ and $g_k$.

We shall see empirically that the Elastic-FuSSO (14) is able to select large correlated supports like repeated trial based methods, which we describe below.

## 4.3 Voxel Selection Through Repeated Trials

Although any single optimization of (13) will likely yield only at most $N$ selected voxels, we may select a support larger than $N$ by running multiple optimizations (trials) with slightly different sets of input instances. We briefly describe such an approach below.

### 4.3.1 Stability Selection

An approach to selecting one's support with repeated trials is stability selection (Meinshausen and Bühlmann, 2010), where one runs multiple optimizations on a subset of instances. In particular, we consider the complementary pairs stability selection variant given $B \in \mathbb{N}$, $\lambda > 0$, $\tau > 0$ (Shah and Samworth, 2013):

1. Let $\{(\mathcal{I}_{2j-1}, \mathcal{I}_{2j}) \mid j = 1, \ldots, B\}$ be independent randomly chosen pairs of subsets of $\{1, \ldots, N\}$ such that $\mathcal{I}_{2j-1} \cap \mathcal{I}_{2k} = \emptyset$, and $|\mathcal{I}_{2j-1}| = |\mathcal{I}_{2j}| = \lfloor \frac{N}{2} \rfloor$.
2. For each $\mathcal{I}_k$ solve for the optimal $\beta$ with the instances in $\mathcal{I}_k$ under regularization penalty $\lambda$ (13).
3. For each $j = 1, \ldots, p$ select the $j^{\text{th}}$ covariate if $\frac{1}{2B} \sum_{k=1}^{2B} \mathbb{I}_{\mathcal{I}_k}\{\|\hat{\beta}_j\| > 0\} \geq \tau$, where $\mathbb{I}_{\mathcal{I}_k}\{\|\hat{\beta}_j\| > 0\}$ indicates if the $j^{\text{th}}$ covariate was selected when optimizing over the instances in $\mathcal{I}_k$.

## 4.4 Two-Stage Estimator

Another drawback to naively applying the FuSSO estimator in real world datasets is that one is unable to decouple the shrinkage and selection effects of the penalty parameters. As discussed by Meinshausen (2007), the $\lambda$ penalty parameter in LASSO-like problems will both select and shrink covariates. That is, the same $\lambda$ parameter will not only make unselected covariates 0, but will also shrink selected covariates towards 0. As such, it may not be optimal to perform both shrinkage and selection in this coupled manner. Instead, we take a two-stage approach inspired by Meinshausen (2007): first, we solve an Elastic-FuSSO optimization problem to determine the model's support; second, using only the covariates found in the support from the first step we solve a ridge regression problem.

Suppose that in addition to our training data as in (14) we also have a holdout validation set of $N^{(t)}$ instances with a corresponding vector of responses $Y^{(t)} \in \mathbb{R}^{N^{(t)}}$ and matrices of projection coefficients $\tilde{A}_j^{(t)} \in \mathbb{R}^{N^{(t)} \times M_n}$ for $j = 1, \ldots, p$. We shall select the following three parameters $\lambda \in \{\lambda^{(1)}, \ldots, \lambda^{(m_1)}\}$, $\lambda_e \in \{\lambda_e^{(1)}, \ldots, \lambda_e^{(m_2)}\}$, $\lambda_r \in \{\lambda_r^{(1)}, \ldots, \lambda_r^{(m_3)}\}$ using a grid search and the following two-stage procedure:

1. Solve for
$$\hat{\beta}^{(\lambda,\lambda_e)} = \operatorname*{argmin}_{\beta} \frac{1}{2}\|Y - \sum_{j=1}^{p} \tilde{A}_j \beta_j\|^2 + \sum_{j=1}^{p} \lambda\|\beta_j\| + \lambda_e\|\beta_j\|^2.$$
2. Solve a ridge regression problem based on the found support $S^{(\lambda,\lambda_e)} = \{j \mid \|\hat{\beta}_j^{(\lambda,\lambda_e)}\| > 0\}$:
$$\hat{\beta}^{(\lambda,\lambda_e,\lambda_r)} = \operatorname*{argmin}_{\beta} \frac{1}{2}\|Y - \sum_{j \in S^{(\lambda,\lambda_e)}} \tilde{A}_j \beta_j\|^2 + \lambda_r \sum_{j \in S^{(\lambda,\lambda_e)}} \|\beta_j\|^2.$$
.

We then choose the model with the lowest held-out MSE: $\text{MSE}(\lambda, \lambda_e, \lambda_r) = \frac{1}{N^{(t)}}\|Y^{(t)} - \sum_{j=1}^{p} \tilde{A}_j^{(t)} \beta_j\|^2$. We note that given the support $S^{(\lambda,\lambda_e)}$ one can efficiently solve $\hat{\beta}^{(\lambda,\lambda_e,\lambda_r)}$ for a range of $\lambda_r$ using the Singular Value Decomposition and the Woodbury Identity. Furthermore, if $\lambda^{(1)} > \lambda^{(2)} > \ldots > \lambda^{(m_1)}$, then we can efficiently solve for $\hat{\beta}^{(\lambda^{(j)},\lambda_e)}$ by warm starting at $\hat{\beta}^{(\lambda^{(j-1)},\lambda_e)}$. Thus, although we are validating through a grid search on 3 parameters, we may do so relatively efficiently.

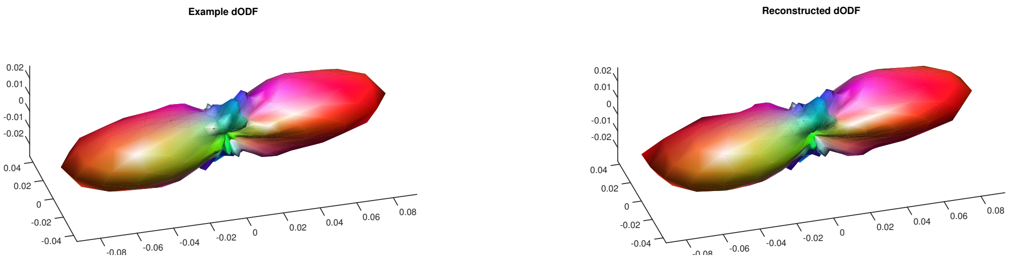
## 4.5 dODF Basis Projection Coefficients

Lastly, we briefly discuss our procedure for representing our input dODF functions (Figure 3(a)) using projection coefficients. As previously discussed one must select an orthonormal basis for use of the FuSSO model. For our analysis we consider using the cosine basis. For 1d domains the cosine basis is:

$$\varphi_0(x) = 1, \ \varphi_m(x) = \sqrt{2}\cos(m\pi x) \text{ for } m = 1, 2, \ldots.$$

We take dODFs as having a two dimensional domain of azimuth and elevation, hence we use the outer product of the cosine basis above. That is, we project dODFs into the following set of basis function:

$$\{\varphi_\gamma(x) = \varphi_{\gamma_1}(x)\varphi_{\gamma_2}(x) \ : \ \gamma \in \mathbb{N}^2, \ \|\gamma\| \leq M\}.$$

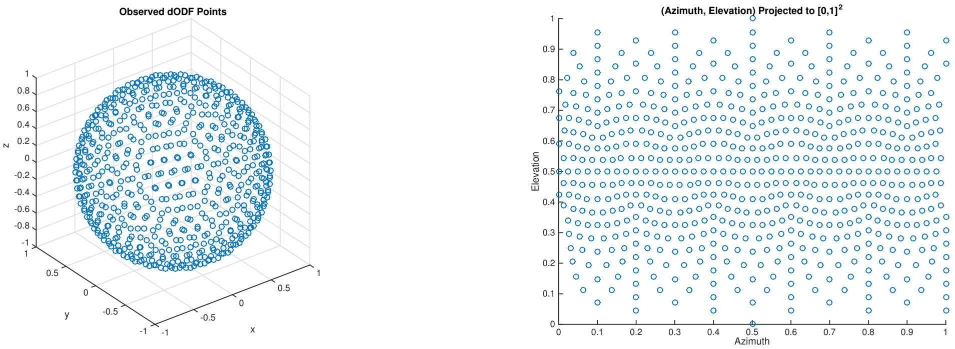


(a)

(b)

Figure 3: (a) Example dODF observation. (b) Reconstructed dODF corresponding to estimated projection coefficients found on basis set.

$M$ was cross-validated and chosen to be $M = \sqrt{208}$. Recall that each input dODF function $f$ was observed with $n = 642$ points $\{(z_k, y_k)\}_{k=1}^n$, where $z_k \in S^2$ are input points on the unit sphere (Figure 4(a)) and $y_k$ are the corresponding dODF function values. Note that we rescale the (azimuth, elevation) values of $\{z_k\}_{k=1}^n$ to the unit cube $[0,1]^2$ (see Figure 4(b)); denote the rescaled inputs as $\{x_k\}_{k=1}^n$.



(a)

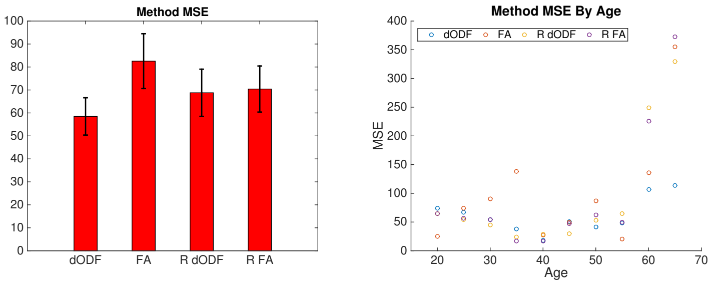
(b)

Figure 4: (a) $\{z_k\}_{k=1}^n$, the 3d points on unit sphere on which dODF function values are observed at. (b) $\{x_k\}_{k=1}^n$, the (azimuth, elevation) of 3d points projected onto the unit square.

Given that the rescaled azimuth and elevation observation points are not in a uniform grid we estimate the projection coefficient of the dODF $f$, $\tilde{\alpha} \in \mathbb{R}^{M_n}$, with a least squares solution. Let $\vec{y} = (y_1, \ldots, y_n)$ be the vector of dODf values corresponding to $f(x_k)$, $\{\gamma_1, \ldots, \gamma_{M_n}\} = \{\gamma \in \mathbb{N}^2,\ \|\gamma\| \leq M\}$, and $\Phi \in R^{n \times M_n}$ be the matrix such that $\Phi_{k,l} = \varphi_{\gamma_l}(x_k)$; Then, our vector of projection coefficients $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_{M_n})$ corresponding to basis functions $\varphi_{\gamma_1}, \ldots, \varphi_{\gamma_{M_n}}$ respectively is:

$$\tilde{\alpha} = (\Phi^T \Phi)^{-1} \Phi^T \vec{y}.$$

As can be seen in Figure 3, the estimated set of projection coefficients do a good job of representing the input functions.



(a)

(b)

Figure 5: (a) MSE using LOOCV of Elastic-FuSSO with functional dODF covariates (dODF), elastic-net with real-valued FA covariates (FA), ridge regression with dODF (R dODF), and ridge regression with FA (R FA). Standard-error bars are shown in black. (b) MSE of respective methods by age.

# 5 Results

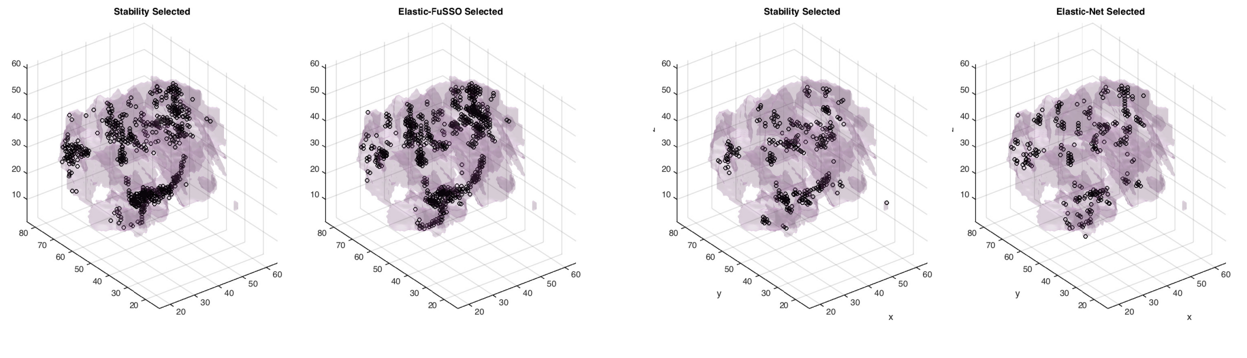
## 5.1 Age Prediction Error

We cross-validated the mean squared error (MSE) of age prediction under leave-one-out cross-validation (LOOCV) using a two-stage procedure as described above for Elastic-FuSSO with functional dODF covariates and elastic-net with real-valued FA covariates. We report the respective MSEs below in Figure 5(a). As can be seen, the Elastic-FuSSO using functional dODF covariates with an MSE of 58.49 is able to achieve a 29.15% lower MSE than the conventional approach of performing elastic-net regression of revalued FA summary statistics with an MSE of 82.55. A paired $t$-test found this difference to be statistically significant with a $p$-value of 0.044. We note that the variance of the age response is 160.69, and that the respective $R^2$ values for the dODF and FA models were 0.64 and 0.49 respectively. We also ran non-sparse ridge regression for both dODF and FA covariates and observed a MSE of 68.78 and 70.940 respectively. It is interesting to note that the dODF performance is not worsened by sparsity, in fact it is improved; this is perhaps because we are able to recover more relevant features to age prediction and can hence hone in on relevant areas. We also plotted the MSE by age groups (per each 5 years) in Figure 5(b). It can be seen that the dODF Elastic-FuSSO model can leverage functional covariates and sparsity to achieve accurate predictions even for older subjects.

## 5.2 Voxel Selection

We report selected voxels using both dODF and FA covariates below. We ran stability selection as described in Section 4.3.1 using 2096 trials with FuSSO (13) for dODF covariates and the LASSO for FA covariates with thresholds of $\tau = 0.01$. We show results of example selected supports in Figure 6. It is interesting to note that the Elastic-FuSSO extension we proposed is able to select a similar set of voxels to the stability selection support. We see that the average nearest-neighbor distance of voxels in the Elastic-FuSSO support to the stability selected support is 0.45 voxels, where the average nearest-neighbor distance of voxels selected in a random support to the stability selected support is 3.84 voxels. Hence, even though the Elastic-FuSSO only runs a single trial it can select a robust and correlated support akin to that selected using multiple trials and more computational resources.

We also show the brain maps of the selected supports below in Figure 7. The identified selected regions and their definitions are as follows: *cerebellar pathways* - classically a motor area, but it is linked to temporal processing and language; *red nucleus* - a main output nucleus of the cerebellum, it plays critical roles in motor control and coordination; *medial lemniscus* - the main communication pathway for somatosensory information to the cortex; *internal capsule* - the main descending pathway that links the motor cortex to the spine; *corona radiata* - a major communication cation system that links the neocortex to subcortical



(a) dODF Stability (b) dODF Elastic (c) FA Stability (d) FA Elastic

Figure 6: dODF selected voxels with (a) stability selection, and (b) Elastic-FuSSO. FA selected voxels with (c) stability selection, and (d) Elastic-net. We see that although the elastic varients run only a single trial, they recover a similar support to stability selection.

areas and is critical for a whole host of cognitive functions (along with motor and sensory processing); *corpus callosum* - major pathway that allows the two cortical hemispheres to talk to each other; *inferior longitudinal fasciculus* - a critical part of the visual pathway that does object processing.

# 6 Analysis

There are several indications that suggest that the dODF functional covariate model is superior to the real valued FA model. First and foremost, we are able to achieve significantly better predictions of age using the Elastic-FuSSO with dODF covariates. Secondly, the model derived using dODFs is more flexible and comprehensive than one that only relies on FA values. We explore this last point below.

Recall that the FuSSO model operates by a sparse linear combination of inner products between dODFs input covariates $\{f_j\}_{j=1}^p$ and unknown functions $\{g_j\}_{j=1}^p$ (1), which we regress through their projection coefficient $\{\beta_j\}_{j=1}^p$ (13). That is, for each selected voxel, we have a $g_j$ function in our model used in an inner product, and studying it yields insight into how our model operates.

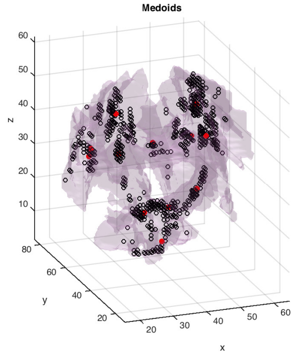
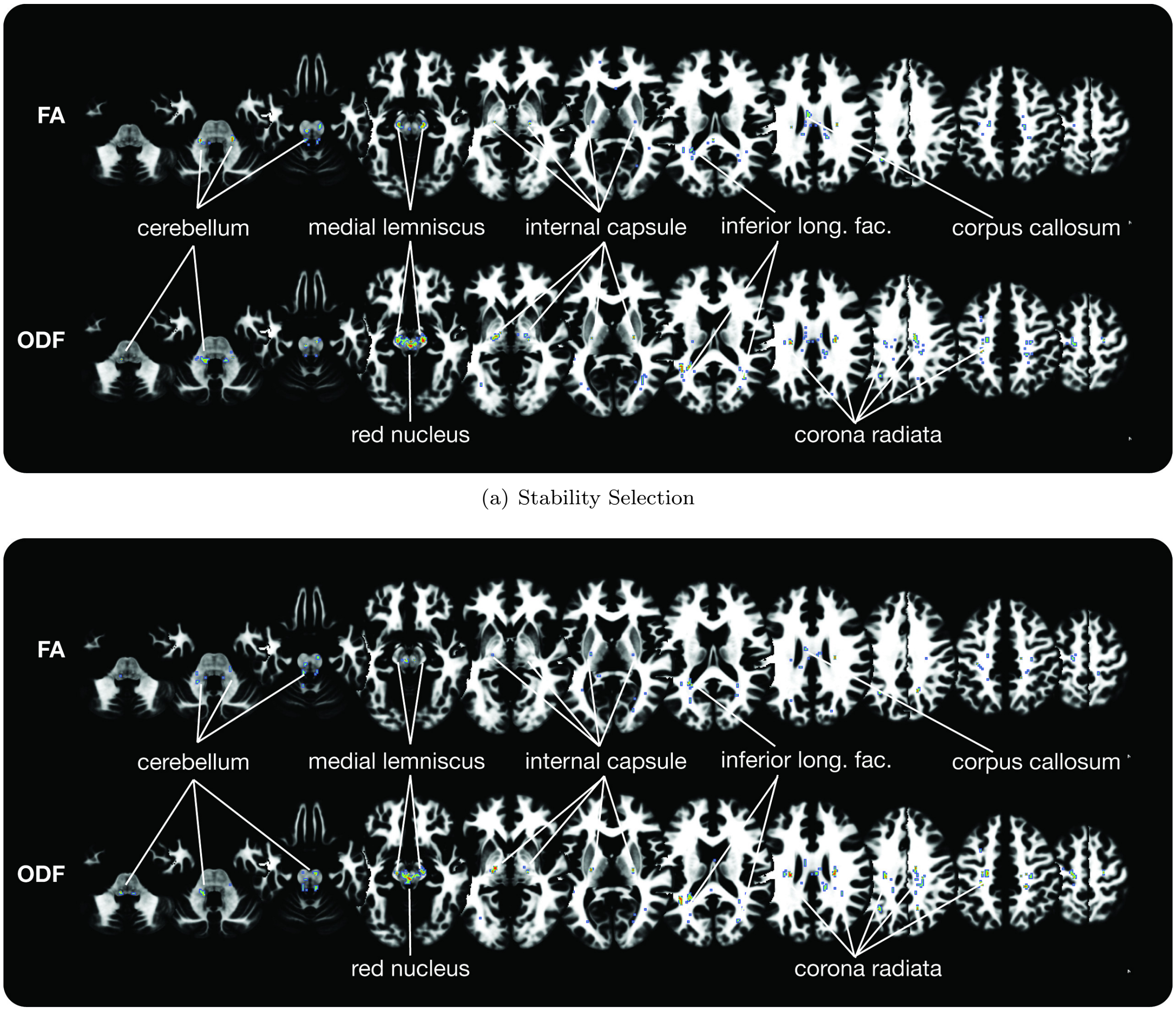


Figure 8: 12 chosen medoids (red, chosen with k-medoids) on selected voxels (black) using the Elastic-FuSSO.

We study our model's $\{g_j\}_{j=1}^p$ functions as follows. First, we cluster the voxels in the support of our Elastic-FuSSO model trained on all instances (Figure 7(b)) using k-medoids with $K = 12$ (see Figure 8 for the medoids found). Let $\{j_1, \ldots, j_{12} \leq p\}$ be the voxel indices for the medoids. Also, let $I_{<25} = \{i \mid Y_i < 25\}$ be the indices of instances corresponding to subjects younger than 25, and $I_{\geq 50} = \{i \mid Y_i \geq 50\}$ be the instance indices for subjects 50 or older. At each of the medoids we plot the average dODFs for young subjects $\bar{f}_{<25}^{(j)}(x) = \frac{1}{|I_{<25}|}\sum_{i \in I_{<25}} f_i^{(j)}(x)$ for each medoid $j = j_1, \ldots, j_{12}$ on row a of Figure 9. Similarly, we plot the mean dODFs for older subjects in row d of Figure 9. One way to visualize the differences in dODFs for younger and older subjects is to plot the negative and positive parts for their differences, $\bar{f}_{-}^{(j)}(x) = (\bar{f}_{\geq 50}^{(j)}(x) - \bar{f}_{<25}^{(j)}(x))_-$ and $\bar{f}_{+}^{(j)}(x) = (\bar{f}_{\geq 50}^{(j)}(x) - \bar{f}_{<25}^{(j)}(x))_+$, shown in rows b and e of Figure 9 respectively. Note that $\bar{f}_{-}^{(j)}$ will show the orientations at which dODFs are stronger for younger instances and $\bar{f}_{+}^{(j)}$ will show the orientations at which dODFs are stronger for older instances. We also plot the negative parts of the found model functions $\{(g_{j_1})_-, \ldots, (g_{j_{12}})_-\}$ and the positive



(a) Stability Selection

(b) Elastic Selection

Figure 7: Brain maps for dODF selected voxels with (a) stability selection, and (b) Elastic-FuSSO. Brain maps for FA selected voxels with (c) stability selection, and (d) Elastic-net. Most areas found, like the cerebellar pathways, are linked to motor skills. However, we also found areas that have been linked to higher order cognition, such as the corona radiata.

parts $\{(g_{j_1})_+, \ldots, (g_{j_{12}})_+\}$ shown in rows c and f of Figure 9 respectively.

It can be clearly seen that the $g_j$ functions are heavily influenced by the differences in dODFs of young and old subjects. Moreover, there are several voxels where the most discrepant orientations do not coincide with the orientations that have the largest values (those large value orientations are the ones used for FA values). When this is the case, it appears that the discrepant orientations are correlated with the underlying $g_j$ functions. Thus, it seems that our Elastic-FuSSO model can uncover other relevant directions to aging. Lastly, it is also interesting to note that although older subject do seem to have smaller magnitude dODFs in general, there are a few voxels at which older subject's dODFs have orientations for which they are of higher value; again, it seems as though the model's $g_j$ are influenced by this.

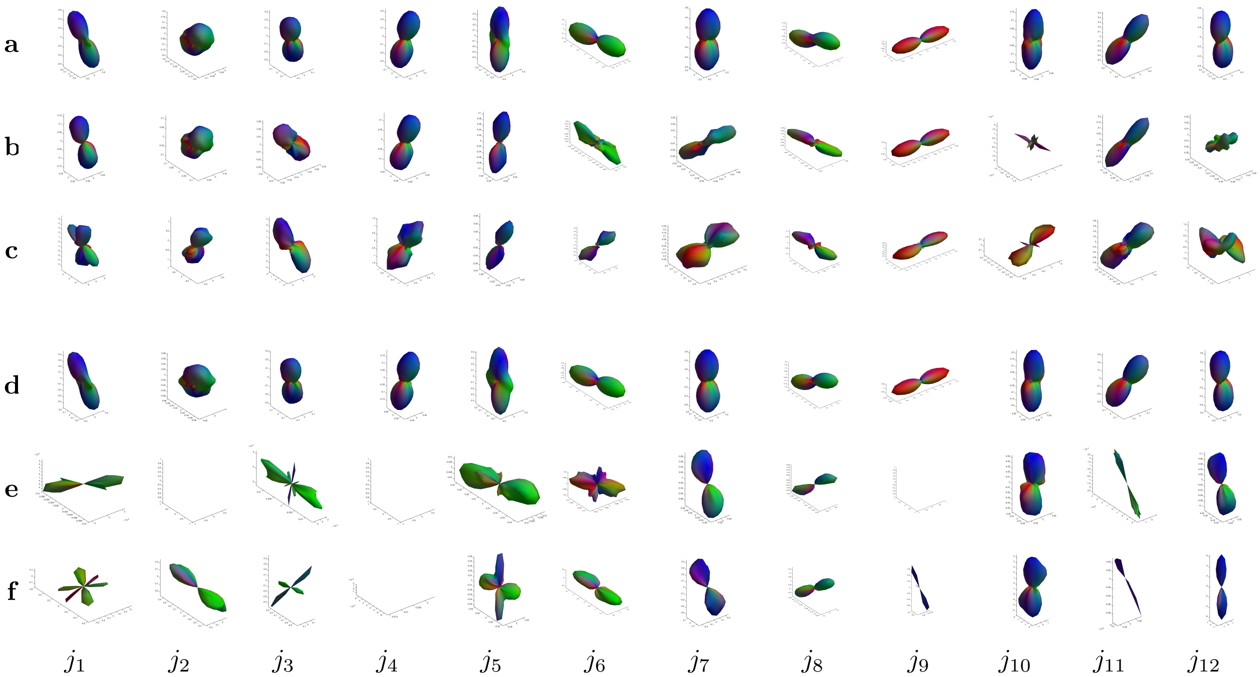


Figure 9: Columns $j_1, \ldots, j_{12}$ correspond to medoids shown in Figure 8. Row a: mean dODF for subjects younger than 25, $\bar{f}^{(j)}_{<25}$. Row b: negative part of difference between dODF of older and younger subjects, $\bar{f}^{(j)}_{-}$. Row c: negative part of $g_j$ model function. Row d: mean dODF for subjects older than 50, $\bar{f}^{(j)}_{\geq 50}$. Row e: positive part of difference between dODF of older and younger subjects, $\bar{f}^{(j)}_{+}$. Row f: positive part of $g_j$ model function. One may see that model functions are heavily influenced by the differences in dODFs of young and old subjects, and tend to focus in on discrepant orientations for age prediction.

# 7 Conclusion

In this work we have developed and tested a principled framework for building predictive supervised models with functional neuroimaging data. This framework allows one to use functional neuro-data in a statistically principled fashion without resorting to heuristical summary statistics.

Our approach is particularly adept at dealing with the unique challenges of performing regression with functional DTI data. We have shown that the Elastic-FuSSO provides a model that is interpretable. By inspecting the support found using our method we were able to find several areas of the human brain related to aging; these areas like the cerebellar pathways are mostly linked to motor skills, however, we also found that areas in the corona radiata, which have been linked to higher order cognition, were also of use. Furthermore, our approach was flexible. Due to the semi-parametric nature of the Elastic-FuSSO we were able to recover contributions of dODFs from non-principle directions, a task not possible before using the standard approach of FA-value summary statistic. Moreover, our framework is scalable; we were able to learn using thousands of functional covariates using FISTA and active-set optimization techniques. Lastly, we saw that our approach was generalizable; the regularization used was able to avoid over-fitting and achieve a low test MSE.

We show a significant improvement for age prediction over conventional summary statistics based methods. Furthermore, we show that our model is able to discover more intricate relations in the diffusion-imaging data than was previously afforded with summary statistics.

### 7.1 Limitations and Future Work

One clear limitation of the the Elastic-FuSSO model is that the underlying functional additive effects are linear through an inner-product (1). Although such a model is interpretable, it is unable to find non-linear relationships. One way to overcome this while still uncovering a sparse model is to consider a sparse additive model of functionals in some reproducing kernel Hilbert space (RKHS): $Y = \sum_{j=1}^{p} g_j(f_j)$ where here $g_j$ are elements in a RKHS over functions in $L_2$. Another drawback to the Elastic-FuSSO, is the need to a priori choose a basis to operate over. Perhaps a data-driven basis such as one found with functional-PCA may perform better in practice. Lastly, it may be worth considering a model that operates over both functional and real-valued covariates jointly.

### 7.2 Acknowledgements

This work is supported in part by NSF grants IIS1247658 and IIS1250350. Furthermore, I wish to thank my DAP committee and all my collaborators for their useful suggestions.

## References

F. M. Benes, M. Turtle, Y. Khan, and P. Farol. Myelination of a key relay zone in the hippocampal formation occurs in the human brain during childhood, adolescence, and adulthood. *Archives of general psychiatry*, 51(6):477–484, 1994.

N. U. Dosenbach, B. Nardos, A. L. Cohen, D. A. Fair, J. D. Power, J. A. Church, S. M. Nelson, G. S. Wig, A. C. Vogel, C. N. Lessov-Schlaggar, et al. Prediction of individual brain maturity using fmri. *Science*, 329(5997):1358–1361, 2010.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer, 2006.

K. Franke, G. Ziegler, S. Klöppel, C. Gaser, A. D. N. Initiative, et al. Estimating the age of healthy subjects from t 1-weighted mri scans using kernel methods: Exploring the influence of various parameters. *Neuroimage*, 50(3):883–892, 2010.

K. Franke, E. Luders, A. May, M. Wilke, and C. Gaser. Brain maturation: predicting individual brainage in children and adolescents using structural mri. *NeuroImage*, 63(3):1305–1312, 2012.

F. M. Gunning-Dixon, A. M. Brickman, J. C. Cheng, and G. S. Alexopoulos. Aging of cerebral white matter: a review of mri findings. *International journal of geriatric psychiatry*, 24(2):109–117, 2009.

C. E. Han, L. R. Peraza, J.-P. Taylor, and M. Kaiser. Predicting age across human lifespan based on structural connectivity from diffusion tensor imaging. In *Biomedical Circuits and Systems Conference (BioCAS), 2014 IEEE*, pages 137–140. IEEE, 2014.

J. Huang, S. Zhang, H. Li, and D. Metaxas. Composite splitting algorithms for convex optimization. *Computer Vision and Image Understanding*, 115(12):1610–1622, 2011.

P. R. Huttenlocher and C. De Courten. The development of synapses in striate cortex of man. *Human neurobiology*, 6(1):1–9, 1986.

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

H. Johansen-Berg and T. E. Behrens. *Diffusion MRI: from quantitative measurement to in vivo neuroanatomy*. Academic Press, 2013.

T. Kanda, H. TSUKAGOSHI, M. ODA, K. MIYAMOTO, and H. TANABE. Morphological changes in unmyelinated nerve fibres in the sural nerve with age. *Brain*, 114(1):585–599, 1991.

P. Kochunov, D. Williamson, J. Lancaster, P. Fox, J. Cornell, J. Blangero, and D. Glahn. Fractional anisotropy of water diffusion in cerebral white matter across the lifespan. *Neurobiology of aging*, 33(1): 9–20, 2012.

H. Liu and J. Zhang. Estimation consistency of the group lasso and its applications. In *International Conference on Artificial Intelligence and Statistics*, pages 376–383, 2009.

N. Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

N. Mingotti, R. E. Lillo, and J. Romo. Lasso variable selection in functional regression. 2013.

B. Mwangi, K. M. Hasan, and J. C. Soares. Prediction of individual subject's age across the human lifespan using diffusion tensor imaging: A machine learning approach. *Neuroimage*, 75:58–67, 2013.

J. B. Oliva, B. Póczos, T. Verstynen, A. Singh, J. Schneider, F.-C. Yeh, and W.-Y. Tseng. Fusso: Functional shrinkage and selection operator. *Artificial Intelligence and Statistics (AISTATS)*, 33, 2014.

R. D. Shah and R. J. Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80, 2013.

P. Shaw, M. Malek, B. Watson, W. Sharp, A. Evans, and D. Greenstein. Development of cortical surface area and gyrification in attention-deficit/hyperactivity disorder. *Biological psychiatry*, 72(3):191–197, 2012.

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

A. B. Tsybakov. *Introduction to nonparametric estimation.* Springer, 2008.

F.-C. Yeh and W.-Y. I. Tseng. Ntu-90: a high angular resolution brain atlas constructed by q-space diffeomorphic reconstruction. *NeuroImage*, 58(1):91–99, 2011.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Y. Zhao, R. T. Ogden, and P. T. Reiss. Wavelet-based lasso in functional linear regression. *Journal of Computational and Graphical Statistics*, 21(3):600–617, 2012.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.