A Spatially Sensitive Kernel to Predict Cognitive Performance from Short-Term Changes in Neural Structure

M. Hidayath Ansari[1], **Michael H. Coen**[1,3], **Barbara B. Bendlin**[2], **Mark A. Sager**[2], and **Sterling C. Johnson**[2]

{ansari,mhcoen}@cs.wisc.edu
[1]Dept. of Computer Sciences, [2]Dept. of Medicine
[3]Dept. of Biostatistics and Medical Informatics
University of Wisconsin, Madison, WI 53706, USA

Abstract

This paper introduces a novel framework for performing machine learning on longitudinal neuroimaging datasets. These datasets are characterized by their size, particularly their width (millions of features per data input). Specifically, we address the problem of detecting subtle, short-term changes in neural structure that are indicative of cognitive change and correlate with risk factors for Alzheimer's disease. We introduce a new spatially-sensitive kernel that allows us to reason about individuals, as opposed to populations. In doing so, this paper presents the first evidence demonstrating that very small changes in white matter structure over a two year period can predict change in cognitive function in healthy adults.

1 Introduction

This paper introduces a framework for analyzing *longitudinal* neuroimaging datasets. We address the problem of detecting subtle changes in neural structure that are indicative of cognitive decline and correlate with risk factors for Alzheimer's disease (AD). AD is a form of dementia affecting memory and executive function afflicting over 18 million worldwide. This number is projected to rise manifold to over 70 million by 2050 and cost trillions of dollars in healthcare.

The problem of identifying neural changes that correspond to age and cognitive decline is one that will benefit from a machine learning methodology tailored to its specific needs. Previous approaches to the problem of analyzing neural change have focused on separating populations with different risk factors based on gross changes, such as regional gray matter atrophy (Misra, Fan, and Davatzikos 2009; Smith et al. 2002) or statistical voxel-based comparisons (Le Bihan et al. 2001; Dyrba et al. 2012). In contrast, we introduce a new *spatially-sensitive* kernel that allows us to characterize individuals, as opposed to populations. We use this for both classification and regression, e.g., to predict changes in a participant's cognitive test scores over time using neuroimaging data alone. This is a difficult problem and in solving this, we have been able to identify neural regions that are implicated in cognitive performance and change over time. More generally, our approach introduces a simple paradigm for addressing *wide-data* longitudinal problems. It

is not specific to neuroimage analysis and shares a number of properties that are representative of this class of problem, which arises often in medical and related domains. These properties include:

1. The datasets are *wide* – they have many more features (p) than they do samples (N). For example in an MRI study, we may gather $O(1e6)$ voxels for each of 100 patients. Similarly, a genome-wide association study may have 500,000 single nucleotide polymorphisms (SNPs) measured over a similar number of patients. Because $p \gg N$, linear models are often the tool of choice due to their speed and low variance. However, these models are also often extremely sparse, as described next.

2. Longitudinal studies track changes over time, with the goal of correlating significant features with some outcome or effect. Naturally occurring variations across features can mask these correlations. For example in medical studies based on neuroimaging, most neural variation is non-pathological and unrelated to the study outcome. The desired model is therefore often extremely sparse but identifying significant features may be difficult due to the next issue.

3. We often lack ground truth to validate results. Consider the problem of determining whether healthy participants tracked over time are expected to develop some condition, such as AD. Given the subjects are currently healthy, even if issues (1) and (2) could be ignored, we have few ways to validate any constructed models. Instead, results are often presented as hypothesis tests that distinguish populations, e.g., those with a family history of the disease from control groups. Predictions about specific individuals are therefore elusive, outside of summary statistics for populations of which they are members.

4. It is increasingly common to track longitudinal changes over very short periods of time. In human neuroimaging, this interval has become as short as three months (ADNI). One may ask if there is even a "signal" to find here. How do we know if there is anything meaningful to detect? This is exacerbated when the sampling time frame is much shorter than the onset time of observable phenomena we would like to predict.

1.1 Framework

Our approach will begin with a "simple" classification problem. For longitudinal data, one instance of ground truth is the chronological order in which the datasets were collected. Thus, a natural question is: can we determine this order for a given individual? Solving this problem allows us to identify and rank the most temporally significant (longitudinally) and consistent (cross-sectionally) voxels in our data. We hypothesize that these voxels correlate with other temporally sensitive data, such as cognitive test scores. In confirming this hypothesis using the novel computational methods in Section 3 for the experiments in Section 4, we present the first evidence demonstrating that very small changes in white matter structure over a two year period can predict change in cognitive function in healthy adults.

2 Background and Data

The analysis in this paper focuses on the white matter (WM) regions of the brain. Much previous research on Alzheimer's disease has focused on gray matter; white matter has historically been regarded as less relevant to cognition. In recent years, however, the role of white matter in the transfer of information has attracted vigorous interest (Ziegler et al. 2010). Data examined here come from the Merit220 and PREDICT cohorts provided by the Wisconsin Registry for Alzheimer's Prevention (Sager, Hermann, and La Rue 2005). Longitudinal imaging and cognitive testing data were available for 75 subjects, who were healthy and middle-aged (ranging from ages 45 to 70). All tested cognitively normal on neuropsychological assays. A significant percentage (78%) of subjects showed one or more risk factors for AD.

Imaging data consisted of measurements of white matter microstructure obtained through diffusion tensor magnetic resonance imaging (DT-MRI, or DTI). Specifically, we use a summary measure at each voxel called fractional anisotropy (FA). FA is a scalar measure of the directional coherence of water diffusion that reflects tissue microstructure, and is particularly sensitive to white matter organization in the brain (Basser and Pierpaoli 1996).

Each subject additionally provided extensive demographic information. The subjects were also genotyped to determine the presence of the apolipoprotein E (APOE) $\epsilon 4$ allele, which is the strongest genetic risk factor for late onset Alzheimer's disease and is associated with earlier age of onset compared to other forms (Corder et al. 1993). In the experiment detailed in Section 4.2 we examine whether the presence or absence of this allele leads to a difference in the way WM changes over time.

2.1 Neuropsychological Tests

All participants underwent comprehensive neuropsychological testing. Cognitive factor scores were derived from a factor analytic study of the WRAP neuropsychological battery and adapted from work published by Dowling *et al.* (Dowling et al. 2010). Based on prior studies showing a strong relationship between indicators of white matter health and processing speed, the factor score chosen for our experiment was the *Speed and Flexibility* factor, a composite measure based on the interference trial from the Stroop Test (Trenerry et al. 1989), and Trail Making Test A and B (Reitan and Wolfson 2009).

Preprocessing Raw scan data was processed and non-linearly registered to a template image in standard MNI152 space using Tract-Based Spatial Statistics (TBSS) (Smith et al. 2006). Each scan underwent identical preprocessing.

Using the white matter atlas from the Johns' Hopkins University DTI research group (Oishi et al. 2008), we extracted the voxels corresponding to the corpus callosum, superior longitudinal fasciculus, fornix, and cingulum bundle. These regions were chosen based on their vulnerability to Alzheimer's disease (Di Paola et al. 2010; Benitez et al. 2014; Canu et al. 2013). The sizes of these regions range from hundreds to over 20,000 voxels.

Figure 1 illustrates the location and shape of the corpus callosum and details its splenium (back), which is a region known to show significant changes both is healthy aging and AD.

3 A Point Set Approach

In many classification problems, data are often abstracted into a representation, e.g., a vector, that fails to retain their spatial information. This is common with many methods in machine learning. However, given the inherent spatial nature of the voxel data, incorporating the voxel locations into our analysis seemed reasonable. This view has received much scrutiny in clustering (Coen, Ansari, and Fillmore 2010), where set theoretic measures of similarity cannot capture subtle changes in the spatial distribution of data. Rather than serialize the voxels of a brain or region into one vector and lose their locations, we represent them as a point set $B = (V, W)$ where $V \subset \mathbb{R}^3$ is the set of positions of the voxels, $W \subset \mathbb{R}$ and every point v_i in V has a corresponding weight w_i in W. Weights w_i correspond to FA values here.

3.1 Comparison of Point Sets

To work with point set representations, we need a way of measuring the similarity or dissimilarity between different point sets. We compare DTI scans by defining a custom distance between their respective voxel sets. This distance is not a simple point-to-point distance; rather it is between two point sets. For a more detailed analysis of related approaches to this problem, including Pyramid Match Kernel (Grauman and Darrell 2007) and Similarity Distance, see (Coen, Ansari, and Fillmore 2011).

Random Fourier Features These were introduced to transform data into a form where linear operations can approximately simulate kernel evaluations (Rahimi and Recht 2007). In this work, a map $\tilde{\Phi}$ (the "lifting" function) is applied to each d-dimensional data point in $\mathbb{R}^d$, transforming it into an element of $\mathbb{R}^D$, a D-dimensional approximation of a reproducing kernel Hilbert space (RKHS). This mapping is randomized and similarity-preserving; a shift-invariant kernel in the original space is approximately equal to the inner product in the new space, where the approximation can be made as precise as possible by varying the dimensionality
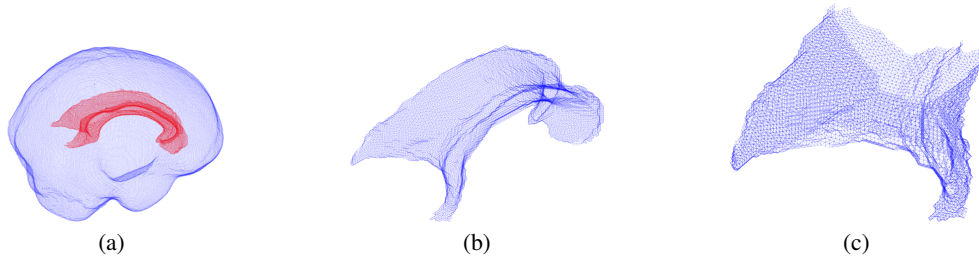
Figure 1: (a) The blue outer mesh is a 3-D view of a representation of the surface of the human brain. The red inner mesh outlines the corpus callosum. (b) A view of the corpus callosum in isolation. The corpus callosum is a thick band of nerve fibers that connects the left and right hemispheres of the brain. (c) A view of the splenium of the corpus callosum, which contains over 12,000 voxels. The splenium of the corpus callosum carries fibers that connect the bilateral temporal, parietal and occipital lobes.

(D) of the lifted space. For the kernel $K(\boldsymbol{x}, \boldsymbol{y}) = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2}}$, the approximate lifting map $\hat{\Phi}_D : \mathbb{R}^d \to \mathbb{R}^D$ is defined as follows: $\hat{\Phi}(\boldsymbol{x}) =$

$$[cos(\omega_1 \boldsymbol{x}), \ldots, cos(\omega_{D/2} \boldsymbol{x}), sin(\omega_1 \boldsymbol{x}), \ldots, sin(\omega_{D/2} \boldsymbol{x})]$$

for $\boldsymbol{x} \in \mathbb{R}^d$ where elements of ω_i's are drawn i.i.d from a standard normal distribution and

$$\langle \hat{\Phi}(\boldsymbol{x}), \hat{\Phi}(\boldsymbol{y}) \rangle \simeq K(\boldsymbol{x}, \boldsymbol{y}) = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2}} \text{ for any } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$$

Raman *et al.* (Raman, Phillips, and Venkatasubramanian 2011) applied this approximate lifting map in representing point sets as elements of an RKHS. The map is applied to each point in a point set, and the whole set is then represented as a single vector by summing the lifted representations of the constituent points. The summed vector is normalized to unit length to eliminate differences caused by differing set cardinalities. The similarity between two point sets X and Y is defined as the dot product between the vectors representing them. We extend this formulation of point set similarity to incorporate weights for each point, so that the final expression for similarity between two point sets $X = (V_X, W_X)$ and $Y = (V_Y, W_Y)$ becomes $\langle \frac{\hat{\Phi}(X)}{\|\hat{\Phi}(X)\|}, \frac{\hat{\Phi}(Y)}{\|\hat{\Phi}(Y)\|} \rangle$, where $\hat{\Phi}(X) = \sum_{\boldsymbol{v_i} \in V_X} w_i \hat{\Phi}(\boldsymbol{v_i})$. For a visual demonstration of lifting in a toy example, see Figure 2.

Given the large number of available voxels in our neuroimaging data, we combined longitudinal and cross-sectional data to identify those that had comparatively *large*, *consistent*, and *similar* values in all difference images corresponding to a class. Our hypothesis is that the voxels that change similarly in all subjects (cross-sectionally) across time (longitudinally) are the ones most sensitive to temporal ordering. Towards this, we define a "Q-value" for each voxel as follows:

$$Q(v_i) = \frac{\text{mean}(\text{FA}_i^1 - \text{FA}_i^2)}{\text{var}(\text{FA}_i^1 - \text{FA}_i^2)} \tag{1}$$

where FA_i^1 is the FA value at voxel i at time 1, FA_i^2 the value at time 2, and mean and variance are computed cross-sectionally over all subjects.
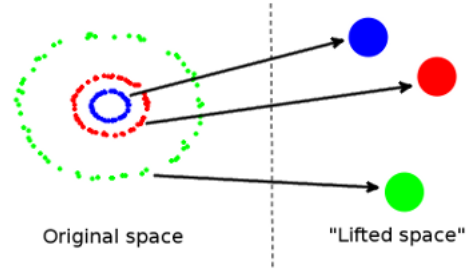

Figure 2: This figure provides an illustration of the "lift" operation described in Section 3.1. The three clusters of points in the original two-dimensional space (colored blue, red, and green respectively) are transformed into singular points in a much higher D-dimensional space. An approximation of their relative positions in a two-dimensional projection is shown on the right hand side of the figure. Notice that the "distance" relationships between the clusters on the left are preserved in the new space, in that blue is closer to red than green which is furthest from the others.

We also define an additional quantity called CONSISTENCY (CONS) for a voxel as follows:

$$\text{POS}_i = \frac{1}{\#subjects} \sum_{\text{subjects}} \{\text{FA}_i^1 - \text{FA}_i^2 > 0\} \tag{2}$$

$$\text{CONS}_i = \max(\text{POS}_i, 1 - \text{POS}_i) \tag{3}$$

Note that POS is defined as a sum of indicator functions. CONSISTENCY in a voxel measures the percentage of subjects who show the same sign change in that voxel from time 1 to time 2.

For a point set $R = (V, W)$ (such as those corresponding to a WM region), we define $\Delta R = (V, \Delta W)$ where ΔW is the change in FA from time 1 to time 2. We set thresholds on Q and CONS to identify subsets of "informative" voxels $\Delta \widehat{R}_Q(\tau) = (\widehat{V}_Q(\tau), \Delta \widehat{W}_Q(\tau))$ and $\Delta \widehat{R}_{\text{CONS}}(\tau) = (\widehat{V}_{\text{CONS}}(\tau), \Delta \widehat{W}_{\text{CONS}}(\tau))$ where:
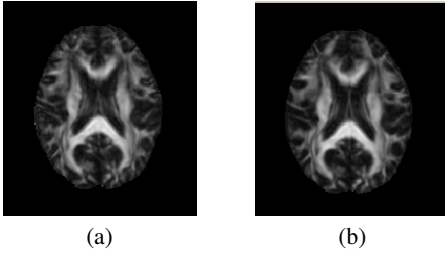
| (a) | (b) |

Figure 3: Two axial slices from DTI scans of the same participant taken approximately two years apart. In our first task, we treated the order of the scans as unknown and proceeded to use data from 75 subjects to predict the order. The images shown here are slices from the full three-dimensional scan. The analysis is performed on the full scans.

$$\widehat{V}_Q(\tau_Q) = \{v_i | v_i \in V, Q(v_i) > \tau_Q\} \text{ and} \quad (4)$$

$$\Delta\widehat{W}_Q(\tau_Q) = \{w_i \ s.t. \ v_i \in \widehat{V}_Q(\tau_Q)\} \quad (5)$$

$$\widehat{V}_{\text{CONS}}(\tau_C) = \{v_i | v_i \in V, \text{CONS}(v_i) > \tau_C\} \text{ and} \quad (6)$$

$$\Delta\widehat{W}_{\text{CONS}}(\tau_C) = \{w_i \ s.t. \ v_i \in \widehat{V}_{\text{CONS}}(\tau_C)\} \quad (7)$$

4 Experiments & Analysis

We present three experiments conducted on the data set in Section 2.1. These demonstrate application of our framework to detecting minute, short-term changes in WM structure and relating them to changes in cognitive test scores and genetic biomarkers.

4.1 Before vs. After

Our goal is to determine the temporal ordering in pairs of scans for an individual. Given two scans, which was taken earlier? (see Figure 3 for an example) Our approach is to exploit voxels that undergo changes that are consistent and similar *across* subjects. This problem is challenging for several reasons: 1) The time period between scans is extremely short (1.5-2 years) and the subtle changes in the scans are believed to be largely unrelated to cognition; 2) All subjects are healthy and middle-aged and do not exhibit any pathology; 3) Domain experts in neuroscience and radiology we have tested are unable to solve this problem for healthy patients better than chance.

Experimental Setup For each of the 75 subjects, we construct two "difference" images. The first subtracts the latter image from the earlier one (the "positive difference image"), and the second by reverses the order of subtraction (the "negative difference image"). This is done so that when given two new images from a single subject with no ordering information, we perform the subtraction in an arbitrary manner and compute to which set of difference images this new difference image is more "similar," using the kernel in Section 3.1.

Region	$|\Delta\widehat{V}_{\text{CONS}}|$	Accuracy	
		RFF	SD
C. Callosum (whole)	3429 voxels	96%	86.7%
C. Callosum (splenium)	463 voxels	97.3%	90.7%
C. Callosum (genu)	364 voxels	90.7%	86.7%
Cingulum bundle	776 voxels	97.3%	89.3%

Table 1: Classification results for predicting the before image from the later image using four different WM regions. τ was fixed at 0.7 for all experiments, and the number of voxels reported is the mean cardinality of the set $|\Delta\widehat{V}_{\text{CONS}}|$ across the different folds in each experiment. Accuracies are reported for two different point set comparison techniques: Random Fourier features (RFF) and Similarity Distance (SD) (see Section 3.1.)

Baseline Since there are an equal number of positive and negative difference images, the baseline accuracy for this experiment is 50%. We applied two classification methods for comparisons with our method.

Region-wide means. A standard approach for characterizing images is to compare mean FA values over a whole WM region across one time point. The classification rule "the image with the higher mean is the earlier image" achieves an accuracy rate of 57% on the splenium of the corpus callosum - little better than random chance. The reason for this is that not all voxels show a decrease in FA value over time; in fact some voxels show an increase. Change in one direction offsets change in the other direction, leading to a low accuracy. This insight leads us to the next baseline method.

Sign-weighted voxel means. The sign of $\mathcal{Q}$ indicates whether the voxel saw an overall increase or decrease in its value over all subjects. As in the earlier method, we compute the mean FA value within a region, but this time weighted by the sign of $\mathcal{Q}$ for that voxel. Applying the same classification rule yields an accuracy of 82% for the same region. All 12,729 voxels in the region are required to achieve this accuracy.

Classification & Accuracy We trained a support vector machine (SVM) with kernels derived from Random Fourier Features (Section 3.1) and Similarity Distance (Coen, Ansari, and Fillmore 2011) to classify "positive" and "negative" difference images. Accuracy was determined with ten-fold cross validation. CONS and voxel selection were recalculated per fold in order to prevent any information leakage from the test set during training. τ was chosen via a beam search. The effect on accuracy was negligible within a range between 0.65 and 0.75 for CONS. The 10-fold cross-validation accuracy in predicting "before" scans from "after" scans (i.e. "positive" difference images from "negative" difference images) is shown for different WM regions in Table 1. As the table shows, approximately 450 well-chosen voxels in the splenium are sufficient to achieve a classification accuracy of 96%.

Identification of Regions of Consistent Cross-Sectional Change The hypothesis of this experiment was that there exist voxels that undergo consistent and similar changes
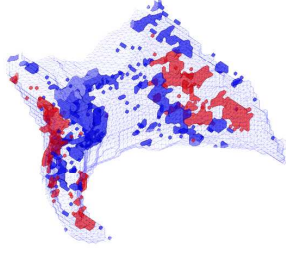
Figure 4: This figure illustrates the portions of the splenium of the corpus callosum that contain voxels with high CONS value. Red voxels indicate a consistent increase in FA value across subjects, while blue represents a consistent decrease.

across subjects, and identification of these voxels would help in characterizing cross-sectional FA change. The experimental results in the previous section show that this hypothesis holds. We now pinpoint those voxels and visualize them in the context of the WM region they belong to. Voxels can be distinguished based on whether they show an upward trend in FA value or a downward trend. Figure 4 shows that voxels tend to be spatially proximal to other voxels of the same type. We note that this naturally-occurring "clustering" of nearby voxels with similar trends is readily apparent even when no smoothing is applied to the data. Further study of these regions and the trends within them will be useful in understanding patterns of age-related change in FA. Of particular interest are the correlations between FA changes, demyelination, and cognitive impairment, as discussed in Section 5.

4.2 APOE Status Classification

We now wish to apply the framework developed above to a different problem, one with higher clinical relevance: is there a difference in the way that WM changes in subjects with different APOE (see Section 2) genotypes? Prior studies have established that subjects with the $\epsilon 4$ allele are at higher risk for developing AD (Corder et al. 1993). We attempt to answer this question by predicting the APOE $\epsilon 4$ status (i.e. the presence (APOE +ve) or absence (APOE -ve) of this allele) based on the changes in FA values. This experiment is similar to the previous one. Rather than have two sets of positive and negative difference images, we take just one (positive difference images) and group them by the APOE $\epsilon 4$ status of the subjects they correspond to. We transform these images into point sets and apply a slightly different voxel selection scheme than before (because the task is different): within each group we identify the voxels that exhibit increases and decreases most consistently, and take the union across both groups:

$$\Delta \widehat{R}_{\text{CONS}}(\tau) = \underset{\text{ApoE -ve}}{\Delta} \widehat{R}_{\text{CONS}}(\tau) \cup \underset{\text{ApoE +ve}}{\Delta} \widehat{R}_{\text{CONS}}(\tau)$$

We use the kernel defined in Section 3.1 in an SVM to differentiate between these two classes of point sets. The baseline accuracy for this experiment is 62.7%, since 47 out of 75 subjects are APOE $\epsilon 4$ negative. The best cross-validated

accuracy of 76% was obtained using the whole body of the corpus callosum, with $\tau = 0.63$ (corresponding to approximately 600 voxels).

4.3 Regression

We would like to model changes in subjects' neuropsychological test scores using FA differences observed over time. Even employing the Q score defined above to prune the space of voxels, it remains the case that $p > N$. Fitting multivariate linear models in this case cannot be done without constraints. Common approaches that limit model exploration including stepwise, best-subset, lasso, and ridge regression. The latter two are often combined via elastic net regularization. There are many ways to validate these models including: using adjusted R^2 values, cross-validation, hold out sets, and checking the distributions of the residuals. However, with a limited number of samples N, evaluating the assessments themselves is difficult. None of the differences between earlier and later test scores is statistically significant according to paired t-tests adjusted for inequality of variances. Scatterplots of earlier vs. later test scores fit lines of slope 1 with relatively high R^2. In these cases, even null models perform well.

While most of the study's cognitive tests had negative adjusted R^2 values when fit to linear models using the high Q voxels from Section 4.1, the Speed and Flexibility score (§ 2.1) yielded an adjusted R^2 of almost .4. ANOVA analysis revealed wide levels of variability within the model, suggesting that while Q is useful for "screening" informative voxels, it may not be sufficient for model feature selection.

To better manage the need for constrained variable selection with wide data, we used the coordinate descent approach for lasso and ridge in (Friedman, Hastie, and Tibshirani 2010). To make the results easier to interpret, we modified our approach to perform logistic regression on the *signs* of the test score changes, viewed as binomial distributions. This normalizes the error penalty and allows us to pose a well-defined problem: Can changes in neuroimaging data predict whether a subject's score for some neuropsychological test has increased or decreased? One might suppose that cognitive abilities uniformly deteriorate monotonically with age. However, evidence does not bear this out, as discussed in Section 5.

Lasso logistic regression via coordinate descent run 100 times with 10-fold cross validation achieved a classification accuracy of 70% with shrinkage parameter $\lambda = .011$, which corresponds to the λ within one standard error of the minimum. Results for this and other methods are shown in Table 2.

These results are quite surprising. Although achieving 70% accuracy seems a modest achievement, consider that this prediction is made using voxel-based neuroimaging data selected because they were able to accurately answer our initial "Which image came first?" question. Within their own representation, the outcome data do not appear separable. But when viewing them from the neuroimaging perspective, we can classify them.

Method	Parameters	Accuracy
Lasso logistic regression (Friedman et al. 2010)	$\lambda = .011$	70%
SVM, Lifted kernel	$2D = 500, C = 1$	58%
SVM, Gaussian kernel	$\sigma = 1, C = 1$	57%
Baseline Random Guessing		54%

Table 2: Classification results for predicting Speed and Flexibility from voxels

Method	Parameters	Accuracy
Ridge logistic regression (Friedman et al. 2010)	$\lambda = .013$	75%
SVM, Lifted kernel	$2D = 500, C = 1$	55.7%
SVM, Gaussian kernel	$\sigma = 1, C = 1$	58.5%
Baseline Random Guessing		54%

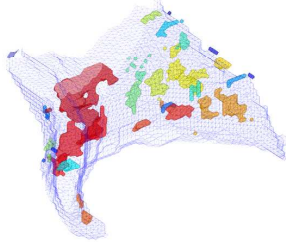Table 3: Classification results for predicting Speed and Flexibility from 30 clusters of voxels

Figure 5: A view of voxels clustered by Q values. Colors correspond to different clusters.

4.4 Clustering

In general, we prefer as few explanatory variables in a model as possible. Wide linear models always raise the specter of overfitting and are notoriously difficult to interpret, particularly when constructed with lasso. E.g., one cannot determine the significance of variables by the magnitude of their coefficients. Following on the spatial point set approach in Section 3, we cluster the voxels based on spatial proximity and their Q values. Simple linkage-based clustering connects voxels with their neighbors if their Q values are within ρ percent of each other. We typically take $\rho = 15$ and specify the maximum number of desired clusters as 30. Emerging from the clustering was the observation that spatially adjacent voxels are likely to have similar Q values. Figure 5 shows regions corresponding to clustered voxels.

Because the clusters are internally consistent with respect to Q values, we used their mean FA values in a ridge logistic regression analysis to predict the sign of the change in the Speed and Flexibility score. We are no longer dealing with wide data since the number of regions $p = 30$ here. While one might imagine the clustering process is lossy, the clusters are better predictors than the voxels used in the previous model. Ridge logistic regression via coordinate descent run 100 times with 10-fold cross validation achieved a classification accuracy of 75% for shrinking parameter $\lambda = 0.13$, as chosen above. Results for this and other methods are shown in Table 3. No significant improvement was seen for other parameters on competing approaches.

5 Discussion

This paper presents a new approach for longitudinal analysis of neuroimaging data. From a computational perspective, our approach relies on the spatial nature of the data both for defining a new kernel and for clustering voxels based on their perceived quality or Q value. We demonstrated this ker-

nel can be used to reliably classify longitudinal neuroimages based on small changes in their white matter structure. We then used the voxels that enabled this classification to predict changes in the significant cognitive factor of Speed and Flexibility, a cognitive function known to be tightly associated with white matter health. While a relationship between speed based cognitive tests and white matter microstructure has been *qualitatively* examined in cross-sectional studies, this is the first work to determine that change in FA over two years can predict change in cognitive function in healthy adults.

From a neuroscience perspective, this work found that over time, portions of the splenium show a decrease in FA between time points separated by 2 years. While this was expected due to aging, more unexpected were the portions of white matter tracts that showed FA increase (the red regions of Figure 4.) The splenium of the corpus callosum carries fibers that connect the bilateral temporal, parietal and occipital lobes. While occipital brain regions do not show high levels of change with age, the temporal and to a lesser extent parietal cortices do change with age. Studies on frontal cortex WM in rhesus macaques indicate that age is associated with loss of nerve fibers but that this degenerative process may be accompanied by continued myelination (Bowley et al. 2010). It is possible that changes occurring over time include both loss of fibers and regenerative myelination. However, in the absence of post-mortem pathological findings, this interpretation is speculative. Another possibility is that our findings reflect the continued myelination that occurs in aging. Visuo-motor skill training in young adults has been shown to increase FA over time (Scholz et al. 2009), and Lövdén et al have shown that experience-dependent changes in FA occur even in older age (Lövdén et al. 2010).

It is possible that we are capturing patterns of white matter change that are not necessarily related to a degenerative effect of time, rather, the effect may reflect continued plasticity in the brain. This is underscored by the tight relationship found with the Speed and Flexibility factor score. Because speed of neural conduction relies on intact myelin, it is not surprising that cognitive speed of processing is linked with white matter health.

6 Acknowledgements

References

Alzheimers Disease Neuroimaging Initiative (ADNI): http://www.adni-info.org.

Basser, P., and Pierpaoli, C. 1996. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor mri. *Journal of Magnetic Resonance, Series B* 111(3):209–219.

Benitez, A.; Fieremans, E.; Jensen, J. H.; Falangola, M. F.; Tabesh, A.; Ferris, S. H.; and Helpern, J. A. 2014. White matter tract integrity metrics reflect the vulnerability of late-myelinating tracts in Alzheimer's disease. *NeuroImage: Clinical* 4:64–71.

Bowley, M. P.; Cabral, H.; Rosene, D. L.; and Peters, A. 2010. Age changes in myelinated nerve fibers of the cingulate bundle and corpus callosum in the rhesus monkey. *Journal of Comparative Neurology* 518(15):3046–3064.

Canu, E.; Agosta, F.; Spinelli, E. G.; Magnani, G.; Marcone, A.; Scola, E.; Falautano, M.; Comi, G.; Falini, A.; and Filippi, M. 2013. White matter microstructural damage in Alzheimer's disease at different ages of onset. *Neurobiology of aging* 34(10):2331–2340.

Coen, M. H.; Ansari, M. H.; and Fillmore, N. 2010. Comparing clusterings in space. In *ICML 2010: Proceedings of the 27th International Conference on Machine Learning*.

Coen, M. H.; Ansari, M. H.; and Fillmore, N. 2011. Learning from spatial overlap. In *AAAI '11: Proceedings of the 25th National Conference on Artificial intelligence*, 177–182. AAAI Press.

Corder, E.; Saunders, A.; Strittmatter, W.; Schmechel, D.; Gaskell, P.; Small, G.; Roses, A.; Haines, J.; and Pericak-Vance, M. A. 1993. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261(5123):921–923.

Di Paola, M.; Di Iulio, F.; Cherubini, A.; Blundo, C.; Casini, A.; Sancesario, G.; Passafiume, D.; Caltagirone, C.; and Spalletta, G. 2010. When, where, and how the corpus callosum changes in MCI and AD. a multimodal MRI study. *Neurology* 74(14):1136–1142.

Dowling, N. M.; Hermann, B.; La Rue, A.; and Sager, M. A. 2010. Latent structure and factorial invariance of a neuropsychological test battery for the study of preclinical Alzheimers disease. *Neuropsychology* 24(6):742.

Dyrba, M.; Ewers, M.; Wegrzyn, M.; Kilimann, I.; Plant, C.; Oswald, A.; Kirste, T.; and et al., S. T. 2012. Combining DTI and MRI for the automated detection of Alzheimer's disease using a large European multicenter dataset. In *Multimodal Brain Image Analysis*, volume 7509 of *Lecture Notes in Computer Science*. Nice, France: Springer Berlin / Heidelberg. in press.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* 33(1):1.

Grauman, K., and Darrell, T. 2007. The pyramid match kernel: Efficient learning with sets of features. *The Journal of Machine Learning Research* 8:725–760.

Le Bihan, D.; Mangin, J.-F.; Poupon, C.; Clark, C. A.; Pappata, S.; Molko, N.; and Chabriat, H. 2001. Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging* 13(4):534–546.

Lövdén, M.; Bodammer, N. C.; Kühn, S.; Kaufmann, J.; Schütze, H.; Tempelmann, C.; Heinze, H.-J.; Düzel, E.; Schmiedek, F.; and Lindenberger, U. 2010. Experience-dependent plasticity of white-matter microstructure extends into old age. *Neuropsychologia* 48(13):3878–3883.

Misra, C.; Fan, Y.; and Davatzikos, C. 2009. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. *Neuroimage* 44(4):1415.

Oishi, K.; Zilles, K.; Amunts, K.; and et al., S. M. 2008. Human brain white matter atlas: Identification and assignment of common anatomical structures in superficial white matter. *NeuroImage* 43(3):447 – 457.

Rahimi, A., and Recht, B. 2007. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems* 20:1177–1184.

Raman, P.; Phillips, J. M.; and Venkatasubramanian, S. 2011. Spatially-aware comparison and consensus for clusterings. In *Proceedings of SIAM International Conference on Data Mining (SDM)*.

Reitan, R. M., and Wolfson, D. 2009. The Halstead–Reitan Neuropsychological Test Battery for AdultsTheoretical, Methodological, and Validational Bases. *Neuropsychological Assessment of Neuropsychiatric and Neuromedical Disorders* 1.

Sager, M. A.; Hermann, B.; and La Rue, A. 2005. Middle-aged children of persons with Alzheimers disease: APOE genotypes and cognitive function in the Wisconsin Registry for Alzheimers Prevention. *Journal of geriatric psychiatry and neurology* 18(4):245–249.

Scholz, J.; Klein, M. C.; Behrens, T. E.; and Johansen-Berg, H. 2009. Training induces changes in white-matter architecture. *Nature neuroscience* 12(11):1370–1371.

Smith, S. M.; Zhang, Y.; Jenkinson, M.; Chen, J.; Matthews, P.; Federico, A.; De Stefano, N.; et al. 2002. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *Neuroimage* 17(1):479–489.

Smith, S. M.; Jenkinson, M.; Johansen-Berg, H.; Rueckert, D.; Nichols, T. E.; Mackay, C. E.; Watkins, K. E.; Ciccarelli, O.; Cader, M. Z.; Matthews, P. M.; and Behrens, T. E. 2006. Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *NeuroImage* 31(4):1487 – 1505.

Trenerry, M. R.; Crosson, B.; DeBoe, J.; and Leber, W. 1989. *Stroop Neuropsychological Screening Test Manual*. Psychological Assessment Resources.

Ziegler, D. A.; Piguet, O.; Salat, D. H.; Prince, K.; Connally, E.; and Corkin, S. 2010. *Neurobiology of aging*, volume 31. Elsevier. chapter Cognition in healthy aging is related to regional white matter integrity, but not cortical thickness, 1912–1926.