Using In-Situ Statistics and a Spatially-Aware Kernel for Longitudinal Neuroimaging Analysis

Michael H. Coen[1,2]
mhcoen@biostat.wisc.edu

M. Hidayath Ansari[2]
ansari@cs.wisc.edu

Barbara B. Bendlin[3]
bbb@medicine.wisc.edu

Mark A. Sager[3,4]
masager@wisc.edu

[1]Department of Biostatistics and Medical Informatics
[2]Department of Computer Sciences
[3]Department of Medicine
[4]Wisconsin Alzheimer's Institute

University of Wisconsin-Madison

Abstract

This paper introduces a framework for analyzing longitudinal neuroimaging datasets. We address the problem of detecting subtle, short-term changes in neural structure that are indicative of cognitive decline and correlate with risk factors for Alzheimer's disease. Previous approaches have focused on separating populations with different risk factors based on gross changes, such as decreasing gray matter volume. 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 subject's cognitive test scores from neuroimaging data alone. 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). Previous approaches have focused on separating populations with different risk factors based on gross changes, such as decreasing gray and white matter volume [12, 19, 9] or statistical voxel-based comparisons [11, 7]. 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 subject's cognitive test scores from neuroimaging data alone. This work thereby identifies neural regions that are implicated in cognitive performance.

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 problems, which arise 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 patients 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 [1]. 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

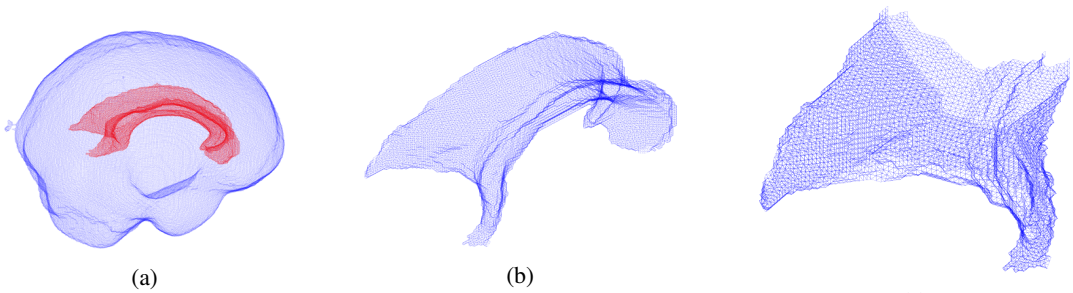
Our approach will begin with a "simple" classification problem. For longitudinal data, one instance of ground truth is the chronological order in which the datasets were collected. Thus, a natural question is: can we determine this order for a given individual? Solving this problem allows us to identify 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 [21]. Data examined here come from the Merit220 and PREDICT cohorts provided by the Wisconsin Registry for Alzheimer's Prevention [17]. 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. In addition to imaging data, each subject provided extensive demographic information. The subjects were also genotyped to determine the presence of the apolipoprotein E allele; the presence of the $\epsilon 4$ allele form places individuals at higher risk for developing AD [2]. 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 Observing and Measuring White Matter

We use measurements of white matter volume and structure obtained through diffusion tensor magnetic resonance imaging (DT-MRI, or DTI). The diffusion properties of water at each voxel are encoded in a *diffusion tensor*, an order 2 tensor describing molecular mobility along 3 directions and correlations among mobilities along those directions.



(a) (b) (c)

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.

2.1.1 Fractional Anisotropy

In addition to the full tensors for each voxel, scalar summary measures may be computed that extract different types of information contained in the tensor matrix. Of interest here is Fractional Anisotropy (FA), a (directionless) value that measures how directional the diffusion of water molecules is in a voxel. FA for a voxel is defined as [10]:

$$\text{FA} = \sqrt{\frac{3}{2}}\sqrt{\frac{(\lambda_1 - \hat{\lambda})^2 + (\lambda_2 - \hat{\lambda})^2 + (\lambda_3 - \hat{\lambda})^2}{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}}$$

where λ_1, λ_2, and λ_3 are the eigenvalues of the diffusion tensor and $\hat{\lambda}$ is the mean of the eigenvalues.

2.2 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.* [6]. The factor score of interest in this paper is the *Speed and Flexibility* factor, a composite measure based on the interference trial from the Stroop Test [20], and Trail Making Test A and B [16].

2.3 Preprocessing

Raw scan data was processed and non-linearly registered to a template image in standard MNI152 space using Tract-Based Spatial Statistics (TBSS) [18]. Regions of interest such as the corpus callosum and cingulum bundle were extracted using the white matter atlas from Johns Hopkins University [13]. Figure 1 illustrates the location and shape of the corpus callosum and details its splenium (back), which is a region of much interest.

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 [4], 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 $V \subset \mathbb{R}^3, W \subset \mathbb{R}$ 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 similarity between their respective voxel sets. This similarity is not a simple point-to-point similarity; rather it is between two point sets. We review below one such measure of similarity; for a more detailed analysis of related approaches to this problem, see [5].

Random Fourier Features These were introduced to transform data into a form where linear operations can approximately simulate kernel evaluations [14]. In this work, a map $\tilde{\Phi}$ (the "lifting" function) is applied to each data point in $\mathbb{R}^d$, transforming it into an element of $\mathbb{R}^{2D}$, a $2D$-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 ($2D$) of the lifted space. For the kernel $K(\boldsymbol{x}, \boldsymbol{y}) = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2}}$, the approximate lifting map $\hat{\Phi}_D : \mathbb{R}^d \rightarrow \mathbb{R}^{2D}$ is defined as follows:

$$\hat{\Phi}(\boldsymbol{x}) = [cos(\omega_1 \boldsymbol{x}), \ldots, cos(\omega_D \boldsymbol{x}), sin(\omega_1 \boldsymbol{x}), \ldots, sin(\omega_D \boldsymbol{x})]$$

for $\boldsymbol{x} \in \mathbb{R}^d$ where elements of ω_i's are random and normally distributed, 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.* [15] 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 $\hat{\Phi}(X) = \sum_{\boldsymbol{v_i} \in V_X} w_i \hat{\Phi}(\boldsymbol{v_i})$.

3.2 Identifying subsets of informative voxels

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:

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

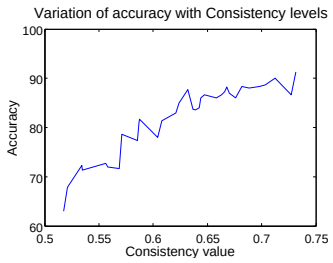


Figure 2: This figure plots variation of classification accuracy with CONS. See text for more detail.

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 the subject population.

We also define an additional quantity called CONSISTENCY (or CONS) for a voxel as follows, measuring the percentage of subjects who show the same sign change in that voxel from time 1 to time 2:

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

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

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:

$$\widehat{V}_Q(\tau) = \{v_i | v_i \in V, Q(v_i) > \tau\} \text{ and } \Delta\widehat{W}_Q(\tau) = \{w_i | v_i \in \widehat{V}_Q(\tau)\} \tag{4}$$

$$\widehat{V}_{\text{CONS}}(\tau) = \{v_i | v_i \in V, \text{CONS}(v_i) > \tau\} \text{ and } \Delta\widehat{W}_{\text{CONS}}(\tau) = \{w_i | v_i \in \widehat{V}_{\text{CONS}}(\tau)\} \tag{5}$$

For an analysis of CONS and its utility in the experiments below, see Figure 2. At each CONS level (x-axis) we randomly chose 50 voxels, and computed accuracy (y-axis) for the before-after prediction using only those voxels. The random subset was varied with every fold while cross-validating.

4 Experiments & Analysis

We present three experiments conducted on the data set in Section 2.3. 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? 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 are unable to solve this problem for healthy patients better than chance.

4.1.1 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.

4.1.2 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.

Region	$\lvert\Delta\widehat{V}_{\text{CONS}}(\tau = 0.7)\rvert$	Accuracy
Corpus Callosum (whole)	3429 voxels	96%
Corpus Callosum (splenium)	463 voxels	97.3%
Corpus Callosum (genu)	364 voxels	90.7%
Cingulum bundle (R & L)	776 voxels	97.3%

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 $\lvert\Delta\widehat{V}_{\text{CONS}}\rvert$ across the different folds in each experiment.

4.1.3 Classification & Accuracy

We trained a support vector machine (SVM) with the kernel derived from Random Fourier Features (Section 3.1) to classify "positive" and "negative" difference images. Accuracy was determined with 10-fold cross validation. CONS and voxel selection were re-calculated per fold in order to prevent any information leakage from the test set during training. 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 400 well-chosen voxels are sufficient to achieve a classification accuracy of 96%. Figure 2 show a further analysis of the variation of classification accuracy with random subsets of 50 voxels chosen at different levels of Q and CONS.

4.1.4 Identification of Regions of Consistent Cross-Sectional Change

The hypothesis of this experiment was that there exist voxels that undergo consistent and similar changes 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 3 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.

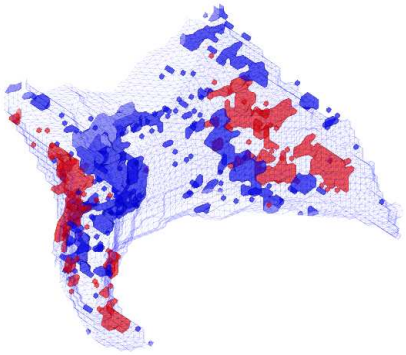


Figure 3: This figure illustrates the portions of the splenium of the corpus callosum that contain voxels with high CONS value. Red voxels indicate a consistent increase in FA value across subjects, while blue represents a consistent decrease.

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 genotypes? Prior studies have established that subjects with the $\epsilon 4$ allele are at higher risk for developing AD [2]. We attempt to answer this question by predicting the ApoE $\epsilon 4$ status (i.e. the presence or absence 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: within

Method	Parameters	Accuracy
Lasso logistic regression [8]	λ = .011	70%
SVM, Lifted kernel	$2D = 500, \sigma = 1, C = 1$	58%
SVM, Gaussian kernel	$\sigma = 1, C = 1$	57%
Baseline Random Guessing		54%

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

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 used 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 ϵ4 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. The data themselves are difficult to work with, as 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 (§ 2.2) 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 [8]. 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. Doing so 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.

Initially, the chance of finding a successful solution to this problem seemed implausible. Our output variable is the sign of a small difference that appears to fluctuate around zero at random. However, 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. No significant improvement was seen for other parameters on competing approaches.

These results are quite surprising. Although achieving 70% accuracy seems a modest achievment, 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
Ridge logistic regression [8]	$\lambda = .013$	75%
SVM, Lifted kernel	$2D = 500, \sigma = 1, C = 1$	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

4.3.1 Clustering

In general, we prefer as few explanatory variables in a model as we can get away with. 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. The regions corresponding to clustered voxels are shown in Figure 4.

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. Because $p = 30$ here, which is the number of regions, we are no longer dealing with wide data, alleviating many of the concerns that they raise. 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

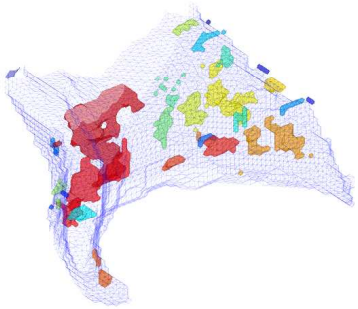


Figure 4: A view of voxels clustered by Q values. Colors correspond to different clusters.

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 kernel can be used to reliably classify longitudinal neuroimages based on small changes in their white matter structure. This task cannot be solved by human experts. We then used the voxels that enabled this classification to predict changes in the significant cognitive factor of Speed and Flexibility. 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, certain portions of the splenium show a decrease in FA from the first time point to the second time point approximately 2 years later. Given what is known about aging in general, this was expected. More unexpected were the portions of white matter tracts that showed an increase in FA from time 1 to time 2. (The red regions of Figure 3.) 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 lessor extent parietal cortices do change with age. Studies on white matter in the frontal cortex of rhesus macaques indicate that age is associated with loss of nerve fibers, but that this degenerative process may be accompanied by continued myelination [3]. It is possible that changes occurring over time include both loss

of fibers and regenerative myelination.

It is also important to keep in mind that white matter development over the lifespan is nonlinear. Certain areas of the brain continue to myelinate well into adulthood. It is possible that we are capturing patterns of white matter change that are not necessarily related to a degenerative effect of time passing, 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.

References

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

[2] J. Ashford. APOE Genotype effects on Alzheimers disease Onset and Epidemiology. *Journal of Molecular Neuroscience*, 23(3):157–165, 2004.

[3] Michael P Bowley, Howard Cabral, Douglas L Rosene, and Alan Peters. 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, 2010.

[4] Michael H. Coen, M. Hidayath Ansari, and Nathanael Fillmore. Comparing clusterings in space. In *ICML 2010: Proceedings of the 27th International Conference on Machine Learning*, 2010.

[5] Michael H. Coen, M. Hidayath Ansari, and Nathanael Fillmore. Learning from spatial overlap. In *AAAI '11: Proceedings of the 25th National Conference on Artificial Intelligence*, pages 177–182. AAAI Press, 2011.

[6] N Maritza Dowling, Bruce Hermann, Asenath La Rue, and Mark A Sager. Latent structure and factorial invariance of a neuropsychological test battery for the study of preclinical alzheimers disease. *Neuropsychology*, 24(6):742, 2010.

[7] Martin Dyrba, Michael Ewers, Martin Wegrzyn, Ingo Kilimann, Claudia Plant, Annahita Oswald, Thomas Kirste, and Stefan Teipel. 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, Oct 2012. Springer Berlin / Heidelberg.

[8] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[9] Håkon Grydeland, Lars T Westlye, Kristine B Walhovd, and Anders M Fjell. Improved prediction of Alzheimer's disease with longitudinal white matter/gray matter contrast changes. *Human Brain Mapping*, 2012.

[10] P. Hagmann, L. Jonasson, P. Maeder, J. Thiran, V. Wedeen, and R. Meuli. Understanding Diffusion MR Imaging Techniques: From Scalar Diffusion-weighted Imaging to Diffusion Tensor Imaging and Beyond. 2006.

[11] Denis Le Bihan, Jean-Franois Mangin, Cyril Poupon, Chris A. Clark, Sabina Pappata, Nicolas Molko, and Hughes Chabriat. Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging*, 13(4):534–546, 2001.

[12] Chandan Misra, Yong Fan, and Christos Davatzikos. 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, 2009.

[13] Kenichi Oishi, Karl Zilles, Katrin Amunts, Andreia Faria, Hangyi Jiang, Xin Li, Kazi Akhter, Kegang Hua, Roger Woods, Arthur W. Toga, G. Bruce Pike, Pedro Rosa-Neto, Alan Evans, Jiangyang Zhang, Hao Huang, Michael I. Miller, Peter C.M. van Zijl, John Mazziotta, and Susumu Mori. Human brain white matter atlas: Identification and assignment of common anatomical structures in superficial white matter. *NeuroImage*, 43(3):447 – 457, 2008.

[14] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20:1177–1184, 2007.

[15] Parasaran Raman, Jeff M. Phillips, and Suresh Venkatasubramanian. Spatially-aware comparison and consensus for clusterings. In *Proceedings of SIAM International Conference on Data Mining (SDM)*, 2011.

[16] Ralph M Reitan and Deborah Wolfson. The halstead–reitan neuropsychological test battery for adultstheoretical, methodological, and validational bases. *Neuropsychological Assessment of Neuropsychiatric and Neuromedical Disorders*, page 1, 2009.

[17] Mark A Sager, Bruce Hermann, and Asenath La Rue. 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, 2005.

[18] Stephen M. Smith, Mark Jenkinson, Heidi Johansen-Berg, Daniel Rueckert, Thomas E. Nichols, Clare E. Mackay, Kate E. Watkins, Olga Ciccarelli, M. Zaheer Cader, Paul M. Matthews, and Timothy E.J. Behrens. Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *NeuroImage*, 31(4):1487 – 1505, 2006.

[19] Stephen M Smith, Yongyue Zhang, Mark Jenkinson, Jacqueline Chen, PM Matthews, Antonio Federico, Nicola De Stefano, et al. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *NeuroImage*, 17(1):479–489, 2002.

[20] Max R Trenerry, B Crosson, J DeBoe, and WR Leber. *Stroop Neuropsychological Screening Test Manual*. Psychological Assessment Resources, 1989.

[21] David A. Ziegler, Olivier Piguet, David H. Salat, Keyma Prince, Emily Connally, and Suzanne Corkin. *Neurobiology of Aging*, volume 31, chapter Cognition in healthy aging is related to regional white matter integrity, but not cortical thickness, pages 1912–1926. Elsevier, Nov 2010.