Spring 5-12-2017

Prediction of Disease Status Based on MRI Brain Scans Using Sparse Principal Component Analysis

Tejal Vashi

ABSTRACT

PREDICTION OF DISEASE STATUS BASED ON MRI BRAIN SCANS USING SPARSE PRINCIPAL COMPONENT ANALYSIS

By

TEJAL PANKAJ VASHI

APRIL 24TH, 2017

INTRODUCTION: Alzheimer's Disease is a neurodegenerative disorder that affects millions of individuals worldwide and the association of brain regions to diagnosis is not presently known. Current methods for diagnosis are not sufficient, with the only true method for knowing if an individual has Alzheimer's Disease being a post mortem analysis of brain tissue. Due to the high dimension of data, a classic principal component analysis to determine which variables to include in a model would not suffice. Sparse Principal Component Analysis deals with the limitations of Classic PCA and can produce which variables are highly correlated to include.

AIM: Compare the results of logistic regression, classic principal component analysis, and sparse principal component analysis to determine the variables to include in a model to differentiate between Mild Cognitive Impairment and Alzheimer's Diagnosis.

METHODS: We analyzed brain scans from the Alzheimer's Disease Neuroimaging Initiative. Variables were predefined by the dataset by individual. We used these variables to run a regular logistic regression on all the variables, ran classic PCA on every stepwise increase in components included in the model, and finally ran the Sparse PCA model, comparing error rate to differentiate between the models and select the variables to include.

RESULTS: We identified the error rate for every model, with SPCA with 8 components and a tuning parameters of 6 having the lowest, and then the variables included in that model were selected as the variables for prediction.

DISCUSSION: By applying this method to high dimensional brain scan data, we identified 59 variables to include in the model. Majority of these 59 variables agreed with the current literature for association with Alzheimer's Disease.

PREDICTION OF DISEASE STATUS BASED ON MRI BRAIN SCANS USING SPARSE PRINCIPAL COMPONENT ANALYSIS

by

TEJAL PANKAJ VASHI

B.S., UNIVERSITY OF GEORGIA

A Thesis Submitted to the Graduate Faculty
of Georgia State University in Partial Fulfillment
of the
Requirements for the Degree

MASTER OF PUBLIC HEALTH

ATLANTA, GEORGIA
30303

APPROVAL PAGE

PREDICTION OF DISEASE STATUS BASED ON MRI BRAIN SCANS USING SPARSE PRINCIPAL COMPONENT ANALYSIS

by

TEJAL PANKAJ VASHI

Approved:

___DR. RUIYAN LUO_____
Committee Chair

__DR. IKE OKOSUN_____
Committee Member

__April 24th, 2017_____
Date

Acknowledgments

I would like to express my deepest regards for all the people in my life who have made this possible, from my family and their love and support, to the professors who have taught me for their guidance and instruction in my education. I would not have made it here without you.

Similarly, my deepest appreciation to my committee chair, Dr. Ruiyan Luo, for her eternal patience, amazing poise, and unfailing guidance through the process of completing a thesis as well as her amazing teaching in lectures. Without her guidance, teaching, and persistent help, this would not have been possible, so thank you for all you have knowingly (and unknowingly) taught me.

A most heartfelt thanks to Dr. Ike Okosun, for sitting on my committee and for the amazing and wonderful direction he has given me through my master degree. Thank you for always having a spare minute (hour) to sit and talk with me, and for your priceless advice. It is greatly appreciated.

And finally, to God, for You have walked with me this far, and for that I am eternally grateful.

Author's Statement Page

In presenting this thesis as a partial fulfillment of the requirements for an advanced degree from Georgia State University, I agree that the Library of the University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote from, to copy from, or to publish this thesis may be granted by the author or, in his/her absence, by the professor under whose direction it was written, or in his/her absence, by the Associate Dean, School of Public Health. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without written permission of the author.

___TEJAL P. VASHI_____
Signature of Author

TABLE OF CONTENTS

ACKNOWLEDGMENTS 4

LIST OF TABLES.....7

LIST OF FIGURES.....8

INTRODUCTION.....9

METHODS AND MATERIALS.....12

RESULTS.....15

DISCUSSION AND CONCLUSION.....18

REFERENCES.....25

APPENDIX.....28

List of Tables

Table 1 Correlation Matrix For the Principal Component Analysis (Restricted to the first 10 components for clarity)

Table 2 Percent of Total Variance Explained by Principal Component

Table 3 Error Rate by Total Number of Components Included In Logistic Regression Model

Table 4 Error Rate by Total Number of Components Included and Tuning Parameter Value

List of Figures

Figure 1 Screeplot of Percent Variance explained and Number of Principal components

Introduction

Alzheimer's Disease is a neurodegenerative disorder that is suspected to affect 50-75% of individuals over the age of 65 who have dementia, approximately some 48 million individuals worldwide (Duthey 2013, pg 11). It is symptomatically characterized by short term memory issues in early stages, and later commonly by disorientation, problems with speech, aggression or agitation, mood swings, difficulty remembering or thinking and understanding, and depression (Wenk 2003). The disease is also associated with senile plaques and neurofibrillary tangles in the brain tissue (Ballard et al. 2011), as well as other changes in the physiology of the brain (Wenk 2003). However, there is no definitive manner with which to diagnose Alzheimer's Disease except post-mortem with a brain dissection (Ballard et al. 2011)

Due to this, there is a litany of cognitive tests that have been developed that purport to accurately differentiate dementia from Alzheimer's Disease, even at the earliest stages (Tombaugh, & McIntyre, 1992). A major issue with these cognitive tests is that the tests are extremely lengthy and individuals who are affected by dementia or Alzheimer's cannot pay attention or remain cognizant for the duration of the entire exam (Grundman et al. 2004). For this, a simpler diagnostic tool is needed to accurately and concisely provide feedback to clinicians regarding the mental state of patients.

The Alzheimer's Disease Neuroimaging Initiative (ADNI), a group of researchers who collect, validate and utilize various types of data in the ongoing study of Alzheimer's Disease, are currently collecting vast amounts of diagnostic and clinical data on individuals who have normal cognitive function, have mildly impaired cognitive function, or have been diagnosed with Alzheimer's Disease. Though ADNI also collects biochemical, genetic, PET, neurological,

and clinical data on those subjects enrolled, they also collect MRI (magnetic resonance imaging) scans of the brains of those subjects participating. MRI is known for their diagnostic value in the diagnosis of Alzheimer's Disease since the characteristic loss of volume and change in physiology is abundantly apparent in the scan (Frisoni, Fox, Jack, Scheltens, and Thompson 2010). Due to this reason, MRI has been previously used with some success to differentiate between subjects with normal cognitive function, those with mild cognitive impairment, and those with Alzheimer's Disease (Desikan et al. 2009), but using brain matter volume as the differentiator between the three groups.

It has been previously suggested that a multivariate analysis may be superior to univariate techniques of analysis in analyzing brain scan data, due to the ability to interpret the results in a manner similar to the natural neural network structure of the brain (Habeck et al. 2010), an example of this would be Principal Component Analysis (PCA). Principal Component Analysis is a statistical methodology in which dimensions of a large dataset are reduced by taking a linear combination of the original variables such that the output (called principal components) captures the maximum variance of the original data, thus curtailing information loss (Qi and Luo 2015). However, there are significant drawbacks to PCA, especially for high dimensional data and large data. In high dimensions, classic PCA does not return consistent estimates, with loadings returning as zero making it difficult to both interpret the components produced and the variables included in each component (Qi, Luo, and Zhou 2013). Due to this, we will utilize Sparse Principal Component Analysis, which selects linear combinations of subsets of variables that explain the most variance of the data with the fewest variables (Zou, Hastie, and Tibshirani 2006).

In order to address this issue, yet still take advantage of multivariate analysis for brain scan data, we look to modified methodologies for PCA, namely sparse PCA (SPCA). Sparse PCA is a type of PCA wherein the principal components are formed such that they are a linear combination of a small subset of the variables but still explain a high percentage of the variance within the data (Qi, Luo, and Zhou 2013). This method utilizes a penalty in order to select variables, such that only those variables that are necessary for the model remain (Zou, Hastie, and Tibshirani 2006).

Here we apply both classic PCA and sparse PCA to brain scan data in order to develop a predictive model that can differentiate between individuals with mild cognitive impairment and those with Alzheimer's disease. Using this technique, we demonstrate that a model utilizing sparse PCA can be built with a small error.

Methods & Materials

Data was retrieved from the ADNI database, specifically from the phase of the project known as ADNI1, wherein individuals with normal cognitive function, mild cognitive impairment, and Alzheimer's Disease were enrolled and data on their cognitive function, biomarkers, clinical data, and MRI and PET scans were collected. From this sample, we selected only those with mild cognitive impairment and Alzheimer's Disease, for a total sample size of 1433 subjects, 532 with Alzheimer's Disease and 901 with mild cognitive impairment. For each of these subjects, we utilized the MR Imaging Analysis dataset, which pre-analyzed the data of the brain scans into numerical values for each subject by region of the brain via voxel based morphometry. A total of 119 brain region variables were included for analysis.

Inclusion/Exclusion Criteria

Participants were included in the ADNI study if they were: aged between 55 years and 90 years old, were on medications (not psychological medications) for more than 4 weeks prior to study participation onset, were not depressed or otherwise suffering from a psychological disorder, were able to speak English of Spanish and had the visual and auditory ability to complete neurological exams, were not in any other ongoing study, were willing and able to join a 3 year study, able to agree to DNA and ApoE sample banking as well as blood and urine testing, and were otherwise in good health.

Exclusion criteria also included: not having neurodegenerative events or diseases such as Parkinson's Disease or a stroke, heart attack, or other brain trauma history, other memory

complaints, taking anti-neuroleptics, other medications with significant central nervous system anticholinergic activity, or discontinuation of current permitted medications during the study.

Demographics

The demographics data collected in the study was largely focused on variables that did not inform this study, such as marital status, occupation, retired or not (and if yes, date), type of residence, year of onset symptoms or of diagnosis, and primary language spoken or used in the testing process. These variables, while certainly informative in other settings are not additive to this project. The demographic variables reviewed were: disease status, gender, age, education level (in total years, with though high school = 12, through college = 16, and through a graduate degree =20), ethnicity, and race. Majority of the sample was affected by mild cognitive impairment (MCI), and were male. Similarly, clear majority of the sample had some college and post-undergraduate education. Majority were Non-Hispanic, and identified their race as white. However, this should not affect the analysis as it is well known in the field that the demographics, such as race and gender, do not affect the disease (Ballard et al. 2011; Tombaugh, & McIntyre, 1992)

Statistical Analysis

Statistical analysis began after first recoding the response variable (disease status) such that 0 was representative of mild cognitive impairment and 1 representative of those with Alzheimer's Disease. From there logistic regression was run on the data using all the brain region variables as predictors, and the error rate was calculated. This determined that a simple logistic regression model including all the variables was not the best model for the project, but

also determined how well using such a model would be in regards to the given data. Then, classical principal component analysis was performed on the data so that the produced principal components coefficients could be determined. These components were double checked against eigenvalues produced from the same data. From the principal components, a scree plot was produced to determine which components contained the majority of the variance of the data, and from this information, a numerically chronologic set of components were selected to analyze going forward. These selected components were then used as predictor variables for logistic regression using chronologic decreasing numbers of components. From these logistic regression equations, we then determined error rate for each number of components included by comparing the determined model with the original components coefficients.

We then progressed to selecting a sparse PCA model. To begin, we set up a series of loops within R such that the selection process was automated. Within the loops the data were partitioned, the data run through the sparse PCA call in the elasticnet package and then scaled, and finally, tested via logistic regression then prediction error calculated. In order to optimize the sparse model, a tuning parameter would have to be selected, as well as the optimal number of components to include in the model. A tuning parameter is an externally selected value that is prespecified in model selection such that the lowest possible error is achieved. Both tuning parameter and optimal number of components selection was achieved by creating iterative loops where certain values of tuning parameters were tested against the data using *k*-fold cross validation techniques.

In order to perform this cross validation, the data was into 10 equally sized folds so that we could then perform 10-fold cross validation. Cross validation is a model evaluation methodology that is used over residuals because it can determine how well the training data will be able to predict for data it does not have. This specific type of cross validation, called *k*-fold cross validation (with *k* equally the number of folds partitioning the data), splits the data into *k* folds, and uses one fold as the test data and the *k-1* fold rest as training data. This means that every data point gets to be in the test set once and in a training set *k*-1 times. This is then used to predict the output value for the test data (data the training data which have been used to build the model have not previously utilized nor does it have the output values for).

This is achieved by running sparse PCA on the partitioned training data, and then using the loadings derived from this to scale the test and training data. Then the training data are run through logistic regression for numerically chronologic amounts of components – that is to say, first running the first component, then the first two components, then the first three components, so on until all ten components had been run. After this has been done, the error is calculated and used to evaluate the models created through the loops, with the lowest error indicating the better model. The tuning parameter and number of components included in this best model were then carried forward and used as parameters for a second sparse principal component analysis wherein the original data was utilized in order to determine the variables included in the final model. This model includes only the most significant variables for determining the difference between those with mild cognitive impairment and those with Alzheimer's Disease.

Results

All Variable Logistic Regression

The subset of data only including those with mild cognitive impairment and those with Alzheimer's Disease determined that there were 901 subjects with mild cognitive impairment and 534 subjects with Alzheimer's Disease, for a grand total of 1435 subjects. From there, the logistic regression run with all the variables as predictors were run. Based on this model, the error rate was calculated as being 0.556045 with a cutoff value of .37, or the probability of the predicted disease status not matching the observed disease status based on the model was 0.5560.

Classic Principal Component Analysis & Scree Plot

The principal components that were determined (Appendix A), which returns how strongly each variable is correlated with each component (loadings), with the larger magnitude (either positive or negative) indicating stronger correlation. These concurred with the calculated eigenvalues of the same data. Based on the results, all values above .1 were considered significant, with 112 variables having a correlation value above the threshold.

The components were then used to create a scree plot to determine how many components to include in the models. The scree plot indicated that given the first ten components, the majority of the variance was concentrated in the first component (Figure 1) and that the first ten components captured about 87% of the total variation (Table 2). These first ten components were then taken forward as the focus of further exploration.

Logistic Regression using Chronologic Numerical Components

Logistic regression was run on the principal components obtained from the traditional PCA, where first the first component was evaluated, then the first two components, iteratively adding the next chronological component until all ten components were included in the model. For each, the error rate was calculated (Table 3). Based on simply utilizing the components, the error rate is lowest for inclusion of 10 components and 6 components, and highest when including only one component.

Cross Validation For Tuning Parameter and Number of Components Inclusion Selection

The sparse PCA method has a tuning parameter controlling the sparsity of components. The training and test data were evaluated at several different tuning parameter values at different numbers of component inclusion, ranging from .1 to 9 for tuning parameter values and from 1 to 10 for number of components. These were evaluated for classification error (Table 4), wherein the model with the lowest error was selected for final model building and variable selection. This is the model containing 8 components with a tuning parameter of 6, with an error of 0.165657.

Final Model & Variable Selection via Sparse Principal Component Analysis

Based on the results of the cross validation, the tuning parameter was set at 6 and the number of components included was 8. From this information, sparse principal component analysis was run on the original dataset in order to build the final model and select which variables would be included in that model. The variables included in this model were: middle occipital gyrus (right hemisphere), gyrus rectus (left hemisphere), orbital part of the superior

frontal gyrus (left hemisphere), middle frontal gyrus (both hemispheres), orbital portion of the middle frontal gyrus (both hemispheres), Rolandic operculum (both hemispheres), olfactory cortex (both hemispheres), medial superior frontal gyrus (right hemisphere), medial orbital superior frontal gyrus (both hemispheres), insula (both hemispheres), median cingulate and paracingulate (right hemisphere), hippocampus (both hemispheres), amygdala (right hemisphere), calcarine fissure (both hemisphere), cuneus cortex (both hemispheres), inferior occipital gyrus (both hemispheres), fusiform gyrus (both hemispheres), superior parietal gyrus (both hemispheres), inferior parietal gyrus (right hemisphere), supramarginal gyrus (left hemisphere), angular gyrus (right hemisphere), lenticular nucleus - putamen (both hemispheres), lenticular nucleus – pallidum (right hemisphere), heschl gyrus (both hemispheres), middle temporal gyrus (right hemisphere), temporal pole – middle temporal gyrus (both hemispheres), temporal inferior gyrus (right hemisphere), hemispheric lobule II (both hemispheres), hemispheric lobule III (both hemispheres), hemispheric lobule VI (left hemisphere), hemispheric lobule VIIb (both hemispheres), hemispheric lobule VIII (both hemispheres), hemispheric lobule IX (both hemispheres), vermic lobule I/II, paracentral lobule (right hemisphere), vermic lobule VI, vermic lobule VIII, vermic lobule X, and estimated total intercranial volume.

Discussion & Conclusion

This project attempted to build a model that could be utilized to predict and differentiate between those with mild cognitive impairment and those with Alzheimer's Disease based on MRIs of brains. The approach of sparse principal component analysis has allowed the development of a diagnostic model for individuals affected by cognitive decline, but also highlighted regions of the brain that can help differentiate between cognitive impairment and Alzheimer's Disease. This approach also allowed for us to select for the least number of correlated variables that also explained the greatest amount of variance in the data.

Examination of the first model, as a simple logistic regression equation which included all the variables, produced a model that had very small coefficients for all values, but also the highest error rate of all the models run. The error rate was calculated at 55% misclassification. From here it was increasingly apparent that a simple logistic regression model calculated in this manner would not be appropriate.

The classic principal component analysis was determined, and the identification of values above .1 in the correlation matrix produced indicated that there were 112 variables that had a strong correlation with the data (given only the first 10 components for clarities sake). This is most of the variables included in the dataset, and also significantly, most of these values fall within the first two components. This is also supported by the scree plot, which similarly shows that the first two components contain most of the variance in the data, with the percent of variance contained in each component decreasing until it evens out around the 10^{th} component. Thus, though these 10 components only prescribe 87% of the total variation, they were the components carried forward in this investigation.

The stepwise addition of components via logistic regression and subsequent error calculations indicate that the number of components do affect the error rate, with a general trend of fewer components meaning a higher error rate (though there is a shift in this trend for inclusion 6 components where the error rate drops, but it rises again for 7-9 components. Inclusion of 10 components also drops the error rate to equal that of inclusion of 6 components in the model). For all of these models, however, the error rate remains between 47%-60%. For inclusion of 3-1 components, the error rate was higher than the model with inclusion of all the variables.

Comparatively, the error rate for the sparse principal component analysis, with the tuning parameter set at 6 and the number of components being 8 was 0.1656566, or 16.5%, demarcating this model as less erroneous than the preceding models.

Similarly, from this model we can extract those variables that were used to model the SPCA model, some 59 variables. Of these variables, the following were also referenced in literature as having a significant association with Alzheimer's Disease: middle frontal gyrus , orbital portion of the middle frontal gyrus, hippocampus, amygdala, cuneus cortex, inferior occipital gyrus, superior parietal gyrus, inferior parietal gyrus, middle temporal gyrus, temporal pole – middle temporal gyrus, and temporal inferior gyrus (Desikan et al., 2010). This large difference between the cited literature and the findings of this study can be cited to the article, as the article only referenced these specific regions with no appendix with further regions listed (it was cited as the only found study with such detailed regions explicitly stated- largely due to using the same dataset; most others simply name regions in broad strokes, such as the Medial Temporal Gyrus, which is in fact several subregions grouped together.)

For this, the regions found in this study were cross referenced to the regions broadly mentioned, and for that the following variables concurred: gyrus rectus, orbital part of the superior frontal gyrus, middle frontal gyrus, orbital portion of the middle frontal gyrus, Rolandic operculum, olfactory cortex, medial superior frontal gyrus, medial orbital superior frontal gyrus, insula, median cingulate and paracingulate, hippocampus, amygdala , cuneus cortex, fusiform gyrus, superior parietal gyrus, inferior parietal gyrus, supramarginal gyrus, angular gyrus, lenticular nucleus - putamen, lenticular nucleus – pallidum, heschl gyrus , middle temporal gyrus, temporal pole – middle temporal gyrus, temporal inferior gyrus, and intercranial volume (Poinier, & Weiner, 2017).

With the advent and dissemination of high dimensional data analysis techniques, information about large datasets can be reduced and synthesized with greater speed and efficiency than before. Of the available techniques, sparse principal component analysis is able to produce results that take into account the highly correlated nature of brain scan data, but also to synthesize and extract only the most important variables. The variables selected are also supported by the data, and thus validate the found results. Of note, very few of the regions found were singularly associated with only one hemisphere of the brain, rather, majority of the variables were found to be significantly associated for both hemispheres. There is currently no literature available as to why this might be so, and is an avenue for further inquiry in the field.

Similarly, though the gold standard of diagnosis of Alzheimer's Disease is a post mortem autopsy, the data used here were based of brain scans – that is to say, still not 100% certain that the classification was correct. Though the results were validated to an extent within the

data collection step with cognitive tests and genetic and other biomarker data, it is still not completely certain, and thus this methodology should be repeated on multiple different datasets, as well as having the dataset cross validated with post mortem autopsies to be certain of the disease state.

A final limitation is the fact that the model does not account for the variation in disease status that lies between mild cognitive impairment and full Alzheimer's Disease, nor the stages of severity for Alzheimer's Disease. A more robust inquiry should be explored for multiple levels of cognitive impairment, not simply a binary outcome as in this project. Future studies would attempt to utilize available brain scan data to build a model to differentiate between these stages, and most especially the severity of Alzheimer's Disease.

Table 1. Demographic Data

Demographic Variable	Count n (%)
Disease Status	
MCI	901 (62%)
AD	532 (37%)
Gender	
Male	794 (55%)
Female	584 (41%)
Missing	55 (4%)
Education	
Did not Complete HS	168 (12%)
HS	203 (14%)
College	612 (43%)
Post Undergraduate	459 (39%)
Ethnicity	
Hispanic	34 (2%)
Non-Hispanic	1315 (92%)
Oher	11 (1%)
Missing	73 (5%)
Race	
American Indian or Alaskan Native	1 (<1%)
Asian	21 (1%)
Native Hawaiian or Other Pacific Islander	0 (0%)
Black or African American	73 (5%)
White	1266 (88%)
More than one race	5 (<1%)
Unknown	2 (<1%)
Missing	65 (4%)

Majority of the sample was affected by mild cognitive impairment (MCI), and were male. Similarly, vast majority of the sample had some college and post-undergraduate education. Majority were Non-Hispanic, and identified their race as white.

Table 2. Percent of Total Variance Explained by Principal Component

PC	Total % Variance Explained
1	0.6945843
2	0.7364264
3	0.770626
4	0.8029463
5	0.8204928
6	0.8350852
7	0.8471534
8	0.857879
9	0.8665157
10	0.8743729
11	0.8812432
12	0.88738
13	0.8931174
14	0.8982999
15	0.9030644

The percent of total variance explained by each component (PC) is detailed, with each iterative value being the previous chronologic numerical component's percent variance value plus whatever percent variance explained by the component.

Table 3. Error Rate by Total Number of Components Included In Logistic Regression Model

Number of Components	Error Rate
10	0.4822299
9	0.5003484
8	0.4989547
7	0.4919861
6	0.4822299
5	0.5066202
4	0.5198606
3	0.5595819
2	0.5554007
1	0.5944251

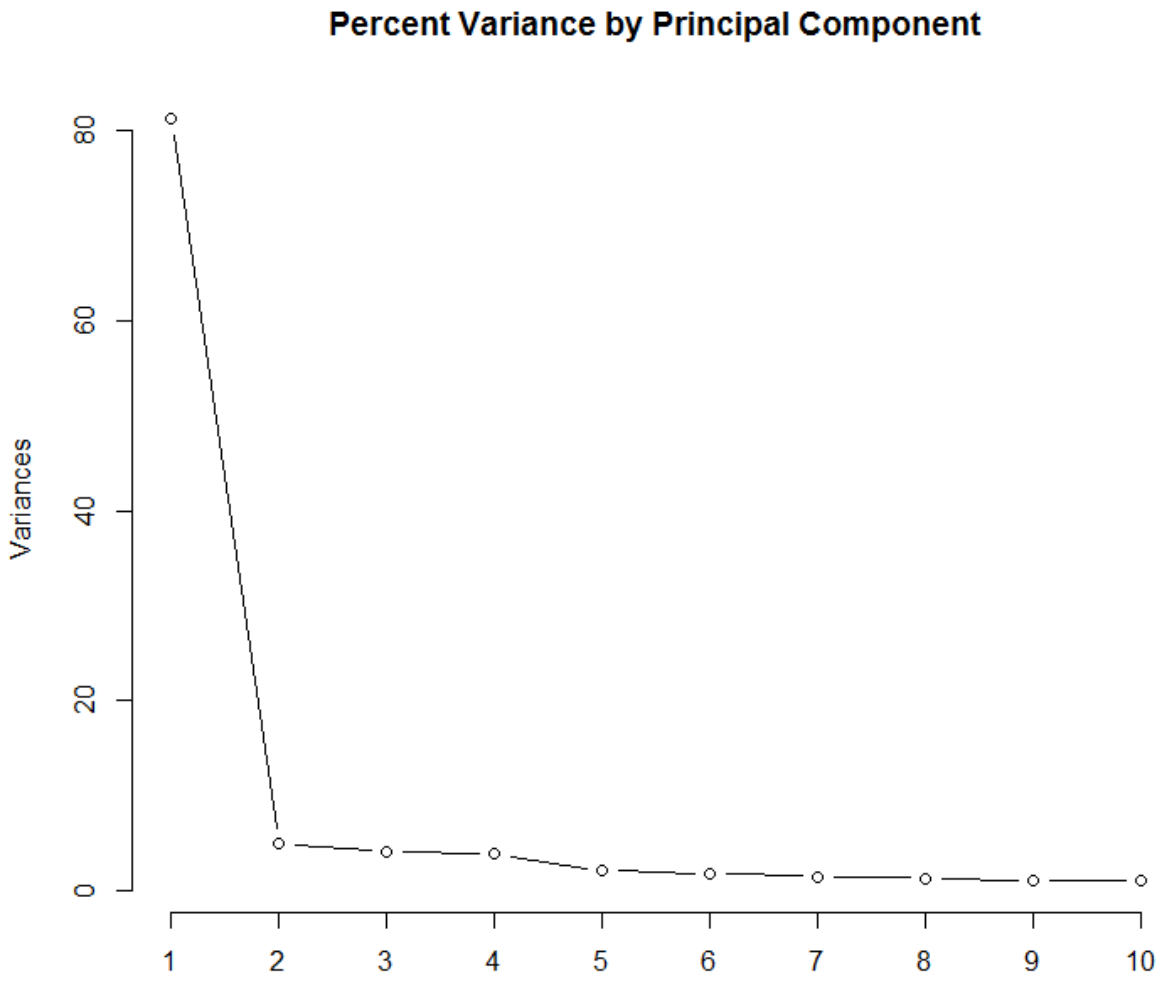
For each logistic regression model, the error rate was calculated and reported. The lowest were for including all 10 components and the first 6 components; the highest was for only including the first component in the model.

Table 4. Error Rate by Total Number of Components Included and Tuning Parameter Value

	Number of Principal Components Included									
	1	2	3	4	5	6	7	8	9	10
Tuning Parameter										
0.1	0.24172	0.224306	0.235485	0.230628	0.225772	0.22788	0.223699	0.225806	0.234164	0.226471
0.5	0.24172	0.224306	0.235485	0.229929	0.225078	0.227181	0.223699	0.225806	0.234164	0.226471
1	0.24172	0.223611	0.23548	0.229929	0.225078	0.227181	0.223699	0.225107	0.234164	0.226471
2	0.24172	0.22431	0.23548	0.229929	0.225078	0.227875	0.223699	0.225107	0.233465	0.226471
3	0.24172	0.22501	0.236179	0.229235	0.225777	0.228574	0.224398	0.225801	0.233465	0.227171
4	0.24172	0.22501	0.236878	0.232032	0.226467	0.228579	0.224393	0.225102	0.232765	0.22787
5	0.241026	0.226404	0.23479	0.233431	0.232027	0.229283	0.227176	0.223004	0.227894	0.229254
6	0.190676	0.176044	0.183727	0.177472	0.177477	0.169153	0.169833	**0.165657**	0.169148	0.170513
7	0.189977	0.178142	0.186519	0.180954	0.174689	0.169852	0.170537	0.165661	0.168444	0.169814

For the selection of tuning parameter and number of components to include in the model, these are the error rates produced via testing each model. The lowest error rate was when the first 8 components were included and the tuning parameter was set at 6.

Figure 1. Screeplot of Percent Variance explained and Number of Principal components



Scree plot of the percent variance explained by each component. This depicts the amount of variance by each component, starting at the total variance for all of the components (87%) with the first component, but dropping to 18% for the second component, indicating that the first component only explained 69% of the variance.

References

Ballard, C., Gauthier, S., Corbett, A., Brayne, C., Aarsland, D., & Jones, E. (2011). Alzheimer's disease. *The Lancet*, *377*(9770), 1019-1031. http://dx.doi.org/10.1016/s0140-6736(10)61349-9

Desikan, R., Cabral, H., Hess, C., Dillon, W., Glastonbury, C., & Weiner, M. et al. (2009). Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. *Brain*, *132*(8), 2048-2057. http://dx.doi.org/10.1093/brain/awp123

Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., & Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nature Reviews. Neurology*, *6*(2), 67–77. http://doi.org/10.1038/nrneurol.2009.215

Habeck, C., Stern, Y., & the Alzheimer's Disease Neuroimaging Initiative. (2010). Multivariate Data Analysis for Neuroimaging Data: Overview and Application to Alzheimer's Disease. *Cell Biochemistry and Biophysics*, *58*(2), 53–67. http://doi.org/10.1007/s12013-010-9093-0

Grundman M, Petersen RC, Ferris SH, Thomas RG, Aisen PS, Bennett DA, Foster NL, Jack, Jr CR, Galasko DR, Doody R, Kaye J, Sano M, Mohs R, Gauthier S, Kim HT, Jin S, Schultz AN, Schafer K, Mulnard R, van Dyck CH, Mintzer J, Zamrini EY, Cahn-Weiner D, Thal LJ, for the Alzheimer's Disease Cooperative Study. Mild Cognitive Impairment Can Be Distinguished From Alzheimer Disease and Normal Aging for Clinical Trials. *Arch Neurol.* 2004;61(1):59-66. doi:10.1001/archneur.61.1.59

Latest Alzheimer's Facts and Figures. (2017). *Latest Facts & Figures Report | Alzheimer's Association*. Retrieved 5 February 2017, from http://www.alz.org/facts/overview.asp

Poinier, A., & Weiner, M. (2017). *Areas of the Brain Affected by Alzheimer's and Other Dementias*. *WebMD*. Retrieved 16 April 2017, from http://www.webmd.com/alzheimers/areas-of-the-brain-affected-by-alzheimers-and-other-dementias

Qi, X., Luo, R., & Zhao, H. (2013). Sparse principal component analysis by choice of norm. *Journal Of Multivariate Analysis*, *114*127-160. doi:10.1016/j.jmva.2012.07.004

Qi, X., & Luo, R. (2015). Sparse Principal Component Analysis in Hilbert Space. *Scandinavian Journal Of Statistics*, *42*(1), 270-289. doi:10.1111/sjos.12106

Tombaugh, T., & McIntyre, N. (1992). The Mini-Mental State Examination: A Comprehensive Review. *Journal Of The American Geriatrics Society*, *40*(9), 922-935. http://dx.doi.org/10.1111/j.1532-5415.1992.tb01992.x.

Wenk GL. Neuropathologic Changes in Alzheimer's Disease. *The Journal of Clinical Psychiatry*. 2003;64 Suppl 9:7–10. PMID 12934968.

Zou H., Hastie T., Tibshirani R. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15 (2006), pp. 265–286

Appendix A

Correlation Matrix For the Principal Component Analysis (Restricted to the first 10 components for clarity)

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
PRECENTL	-0.097655	-0.021379	**0.145325**	-0.02066	0.010248	-0.053436	8.91E-02	-0.069326	0.027525	-0.05642
PRECENTR	-0.095458	-0.001155	**0.164821**	-0.050262	0.014286	-0.046749	**1.22E-01**	-0.066888	0.052192	**-0.102004**
FRONTSUPL	**-0.102029**	0.060415	0.096643	0.005434	-0.036125	0.081047	-1.54E-02	**-0.10071**	-0.083918	0.01074
FRONTSUPR	**-0.100383**	0.077435	**0.112905**	-0.005856	-0.022202	0.080207	-1.51E-03	**-0.107657**	-0.07014	-0.023068
FRONTSORBL	-0.099749	**0.138207**	0.017292	0.014615	-0.034804	0.085411	-4.34E-02	-0.030361	**-0.118609**	-0.038982
FRONTSORBR	-0.09934	**0.138943**	0.024605	0.032324	-0.055615	0.078791	-3.96E-02	-0.023358	**-0.119796**	-0.051245
FRONTMIDL	**-0.104348**	0.049642	0.05298	0.016517	0.024001	0.055704	-4.37E-02	-0.085129	-0.097055	0.015566
FRONTMIDR	**-0.104408**	0.085169	0.063531	0.004535	0.007339	0.058879	-5.48E-03	-0.070022	-0.072785	-0.014905
FRTMIDORBL	**-0.101686**	**0.113162**	-0.013347	0.050975	-0.04442	0.085241	-4.74E-02	-0.019996	-0.082424	0.017247
FRTMIDORBR	**-0.100485**	**0.125538**	-0.0049	0.046805	-0.043508	0.091103	-3.51E-02	-0.007895	-0.088419	0.008487
FRONTINOPL	**-0.101468**	0.01905	-0.03575	0.031207	0.054516	0.043549	-8.00E-02	-0.055998	-0.006192	-0.000484
FRONTINOPR	**-0.103007**	0.017703	-0.019112	0.040837	0.020852	0.00601	-4.92E-02	-0.076524	0.016295	-0.00488
FRONTINTRL	**-0.101497**	-0.011647	-0.060112	0.059594	0.074801	0.026807	-9.23E-02	-0.07507	-0.023716	-0.000855
FRONTINTRR	**-0.104102**	0.039045	-0.022943	0.032812	0.06888	0.028471	-3.51E-02	-0.065534	-0.026488	-0.047874
FRONTINOBL	**-0.103518**	0.048732	-0.041139	0.058479	0.004132	0.057031	-5.14E-02	-0.04716	0.029281	-0.030974
FRONTINOBR	**-0.104242**	0.07328	-0.018391	0.052098	-0.006962	0.03619	-1.94E-02	-0.019504	0.016781	-0.060121
ROLANDOPL	-0.093461	-0.041073	**-0.10971**	**0.138242**	-0.061267	-0.076626	-7.79E-02	-0.022716	**0.118293**	0.0548
ROLANDOPR	**-0.100753**	0.0051	-0.07957	0.090072	-0.013035	-0.063189	-2.38E-02	-0.010162	0.050041	-0.010766
SUPMOTORL	-0.098313	0.004887	**0.123757**	-0.026362	-0.065577	0.028521	6.54E-02	-0.11484	0.078696	0.049313
SUPMOTORR	-0.089963	-0.078441	**0.149537**	-0.007583	**-0.123409**	0.027647	5.04E-02	**-0.176286**	**0.148746**	**0.119053**
OLFACTL	-0.09655	0.038734	-0.077614	0.097497	-0.082069	-0.072235	8.12E-02	-0.022503	0.068531	0.018632
OLFACTR	-0.092986	0.033443	-0.084788	**0.114**	-0.080698	-0.080925	3.70E-02	-0.021054	**0.118698**	-0.00201
FRONTSMEDL	**-0.101396**	0.090526	0.007917	0.055377	-0.051494	0.095601	-2.52E-02	**-0.123284**	-0.032469	0.027945
FRONTSMEDR	**-0.100126**	0.089316	0.002097	0.054136	-0.068745	0.10979	-2.38E-02	**-0.124379**	-0.022755	0.047663
FRTMEDORBL	-0.09953	0.092314	-0.026606	0.080816	-0.044079	0.063727	-4.84E-02	-0.059254	-0.085754	-0.051623
FRTMEDORBR	-0.099225	0.078821	-0.036001	0.094401	-0.04015	0.04281	-6.30E-02	-0.071621	-0.067969	-0.047715
RECTUSL	-0.098859	0.075716	-0.030094	0.082307	-0.058293	0.062613	-7.04E-02	-0.054529	-0.054966	-0.046059
RECTUSR	-0.099897	0.085041	-0.027638	0.081918	-0.047298	0.044348	-5.54E-02	-0.057793	-0.064286	-0.041823
INSULAL	**-0.100274**	0.015217	**-0.114599**	0.095191	-0.030852	-0.022441	-4.72E-03	-0.015725	0.082217	0.09085
INSULAR	**-0.104228**	0.044338	-0.066011	0.050088	0.00928	0.010641	5.80E-03	-0.01626	0.082747	-0.000289
CINGANTL	**-0.100634**	0.034804	-0.016965	0.066169	0.019727	0.029013	-7.33E-02	-0.148069	-0.022309	-0.064205
CINGANTR	-0.092893	-0.057815	-0.059403	**0.116238**	0.014038	-0.016862	**-1.24E-01**	**-0.201422**	0.017208	-0.052746

CINGMIDL	-0.096506	-0.071199	**0.101617**	0.019156	-0.03643	-0.030512	-6.26E-02	**-0.15264**	0.022128	0.045708
CINGMIDR	-0.093478	**-0.116066**	0.062379	0.083163	-0.094735	-0.040904	-6.84E-02	-0.144806	0.078835	0.084473
CINGPOSTL	-0.082699	-0.029771	**0.125437**	-0.032664	**0.15936**	-0.064885	-8.11E-02	-0.07174	-0.037269	**-0.150896**
CINGPOSTR	-0.090339	**0.012724**	**0.111938**	-0.018418	0.059413	-0.029296	-8.76E-03	-0.010604	0.023316	**-0.162175**
HIPPL	-0.083818	-0.026227	**-0.179047**	**0.158549**	-0.03683	**-0.123678**	7.18E-02	**0.121607**	0.073648	**0.138909**
HIPPR	-0.085311	-0.005526	**-0.154154**	**0.159336**	-0.046585	**-0.125545**	8.70E-02	**0.136262**	0.075726	0.081595
PARAHIPPL	**-0.101261**	0.022903	-0.036623	0.023393	0.046247	-0.042594	5.01E-02	0.025521	**0.157632**	**-0.102882**
PARAHIPPR	-0.099342	0.005139	-0.053846	-0.004286	**0.105799**	-0.031662	3.29E-02	0.007346	**0.159706**	**-0.104771**
AMYGDL	-0.097158	**0.158557**	0.010795	-0.021616	0.025538	0.052858	9.05E-02	0.063817	0.044894	-0.015175
AMYGDR	-0.098034	**0.135874**	-0.00032	-0.002624	0.041411	0.035889	9.38E-02	0.059914	0.060402	-0.031283
CALCARINEL	**-0.101244**	-0.051809	-0.006937	0.012123	-0.024327	-0.054591	1.02E-02	**0.175207**	0.053009	-0.090765
CALCARINER	**-0.101315**	-0.028339	-0.008427	-0.022402	0.040897	-0.041906	2.66E-02	**0.161063**	0.036753	**-0.128493**
CUNEUSL	-0.096046	-0.097438	0.08918	0.023217	-0.077411	-0.028709	5.81E-03	**0.188879**	-0.004056	-0.017424
CUNEUSR	-0.098437	-0.077458	0.073016	0.011064	-0.036763	-0.044285	2.51E-02	**0.178962**	-0.001516	-0.05696
LINGUALL	**-0.100765**	-0.039061	0.015388	-0.044108	**0.102431**	-0.063255	1.86E-02	0.103408	0.066526	-0.154533
LINGUALR	-0.09822	-0.075403	0.014204	-0.041005	0.045463	-0.059611	3.68E-02	**0.128214**	**0.128237**	**-0.182263**
OCCSUPL	-0.091453	-0.06651	**0.145121**	-0.005166	-0.032247	-0.074037	6.08E-03	**0.217454**	**-0.116967**	-0.028335
OCCSUPR	-0.093162	-0.069192	**0.120734**	-0.009886	0.012576	-0.05968	-9.09E-06	0.2253	**-0.115905**	-0.03256
OCCMIDL	**-0.101244**	-0.025074	0.06086	-0.000542	0.045924	-0.05087	-6.62E-02	0.196669	**-0.115897**	-0.024249
OCCMIDR	**-0.10187**	-0.00549	0.073376	-0.015055	0.034334	-0.019578	-5.50E-02	0.190275	**-0.104658**	-0.014977
OCCINFL	-0.097654	0.006854	0.021539	-0.019336	**0.113185**	-0.046261	-3.78E-02	0.156924	-0.09039	-0.095473
OCCINFR	-0.094614	0.078231	0.051548	-0.056762	**0.133739**	-0.017726	2.02E-02	0.142753	-0.088188	-0.159241
FUSIFORML	**-0.103707**	-0.061278	-0.041528	0.015698	0.030554	-0.017182	-1.44E-02	0.06271	0.090778	-0.03169
FUSIFORMR	**-0.105854**	-0.016848	-0.025779	0.014107	0.019695	-0.002988	1.16E-02	0.053155	0.081585	-0.065109
POSTCENTL	-0.095787	-0.095609	**0.135464**	-0.004224	-0.024667	-0.058592	1.84E-02	-0.072885	0.056825	-0.005096
POSTCENTR	-0.096043	-0.038499	**0.186105**	-0.045164	-0.014413	-0.02481	5.71E-02	-0.048281	0.032138	-0.027015
PARIETSUPL	-0.079386	**-0.111833**	**0.23253**	0.005898	**-0.121794**	-0.041359	1.93E-02	0.029603	-0.06152	**0.13309**
PARIETSUPR	-0.080289	-0.08349	**0.235575**	-0.00455	**-0.118994**	-0.015882	4.28E-02	0.016586	-0.043446	**0.137062**
PARIETINFL	-0.095939	-0.090166	**0.121411**	0.043085	-0.040291	-0.047377	-8.41E-02	-0.017723	-0.095359	**0.124786**
PARIETINFR	-0.095551	-0.067004	**0.123302**	0.043872	-0.058355	-0.033652	-5.94E-02	-0.000443	-0.07533	**0.137191**
SUPRAMARGL	-0.097285	-0.044897	-0.053332	0.056503	0.069362	-0.06179	**-1.59E-01**	-0.0208	-0.049253	0.043454
SUPRAMARGR	-0.098858	-0.01567	0.067556	0.004575	**0.10874**	-0.036548	**-1.14E-01**	-0.048482	-0.057856	-0.026433
ANGULARL	-0.096783	-0.05641	0.066482	0.046611	0.041918	-0.074705	**-1.46E-01**	0.08469	**-0.125834**	0.097784
ANGULARR	-0.096481	-0.043926	**0.120802**	-0.011924	0.084908	-0.028689	**-1.06E-01**	0.067008	**-0.15379**	0.032229
PRECUNEUSL	-0.097236	-0.091497	**0.167455**	-0.008275	-0.029125	-0.028608	-2.57E-02	0.000606	**-0.011628**	0.065742
PRECUNEUSR	-0.095718	**-0.125868**	**0.117768**	0.041108	-0.091862	-0.035898	-6.21E-02	-0.014399	0.034707	0.086354
PARCENTLBL	-0.07764	-0.065713	**0.209151**	-0.091365	-0.03938	-0.076213	**2.07E-01**	-0.086536	**0.147221**	-0.101071
PARCENTLBR	-0.076671	**-0.111365**	0.193757	-0.036272	**-0.189057**	0.011991	2.94E-02	**-0.128745**	**0.17371**	**0.124618**

CAUDATEL	-0.089935	**0.168847**	-0.043403	0.05556	-0.041486	-0.05147	**1.84E-01**	-0.026417	-0.098115	**0.143548**
CAUDATER	-0.087918	**0.107582**	-0.089884	**0.105771**	-0.06273	-0.074038	**1.82E-01**	-0.016721	-0.062646	**0.22153**
PUTAMENL	-0.078069	0.096877	0.055971	**-0.108747**	**0.333639**	0.023784	**1.61E-01**	-0.119758	-0.07364	0.069896
PUTAMENR	-0.062639	0.043991	0.067393	**-0.102588**	**0.414965**	-0.024206	**1.94E-01**	-0.169378	-0.082645	0.090789
PALLIDUML	0.063462	**-0.264518**	-0.013279	-0.004934	**0.259925**	-0.074851	-9.44E-02	-0.181314	0.036186	0.042834
PALLIDUMR	0.072909	**-0.266483**	-0.063059	0.023789	**0.167419**	-0.080985	**-1.02E-01**	**-0.137673**	0.050778	0.065172
THALAMUSL	-0.068186	0.051879	-0.030412	0.071498	**0.182427**	**-0.258394**	**3.66E-01**	-0.050018	**-0.104956**	**0.121977**
THALAMUSR	-0.079322	0.051878	-0.052599	**0.113209**	0.059851	**-0.241021**	**3.48E-01**	0.020667	-0.041388	**0.12881**
HESCHLL	-0.084953	-0.052905	-0.093307	0.069701	0.073142	**-0.112368**	-2.14E-02	-0.002856	0.02565	0.013553
HESCHLR	-0.075357	**-0.103833**	**-0.127611**	**0.146238**	-0.00159	**-0.153491**	-7.83E-03	-0.027622	**0.114859**	0.019292
TEMPSUPL	-0.093785	-0.088187	-0.090173	0.062168	**0.139517**	-0.092142	**-1.33E-01**	-0.039079	0.037454	-0.015494
TEMPSUPR	**-0.102632**	-0.047697	-0.03383	0.04782	**0.108502**	-0.059867	-8.85E-02	-0.021711	0.012161	-0.035056
TEMPPLSUPL	-0.097713	0.039165	-0.075489	0.08375	0.001748	0.01152	-7.65E-03	-0.020407	**0.133486**	0.005679
TEMPPLSUPR	**-0.100713**	0.04942	-0.054233	0.088114	-0.005975	0.012412	7.45E-03	-0.018831	**0.119298**	-0.057975
TEMPMIDL	-0.098859	-0.073792	-0.054553	0.054378	**0.116517**	-0.057272	**-1.74E-01**	0.03838	-0.024472	0.069617
TEMPMIDR	**-0.103696**	-0.011464	0.000543	**0.011352**	**0.127912**	-0.033966	-9.89E-02	0.043539	-0.036532	-0.015858
TEMPPLMIDL	**-0.100587**	**0.11195**	-0.01763	0.039463	-0.018751	0.049174	-2.17E-02	0.01547	0.008936	0.023087
TEMPPLMIDR	**-0.102396**	**0.114576**	0.002358	0.036558	-0.007789	0.041796	-2.91E-03	0.012338	-0.011414	-0.032376
TEMPINFL	**-0.101875**	-0.002441	-0.064783	0.045485	0.050887	0.006989	**-1.32E-01**	0.016101	-0.005363	0.099909
TEMPINFR	**-0.105191**	0.04475	-0.014037	0.020261	0.048612	-0.000466	-6.28E-02	0.053342	-0.001115	-0.001002
CEREBCR1L	-0.096929	-0.081364	-0.06554	-0.052851	-0.024384	0.093057	-2.90E-02	**0.111567**	0.097654	0.062359
CEREBCR1R	**-0.100734**	-0.042335	-0.031864	-0.05064	-0.046333	0.099592	3.04E-03	**0.107349**	0.071269	0.019975
CEREBCR2L	-0.097349	0.03303	-0.056208	-0.074829	-0.046689	**0.103851**	-2.63E-02	**0.118332**	0.028823	0.090452
CEREBCR2R	-0.098264	0.05309	-0.050134	-0.062618	-0.055508	0.087399	-7.48E-03	**0.105807**	0.052118	0.073184
CEREB3L	-0.081518	-0.083747	**-0.105009**	-0.085179	-0.032051	**0.141604**	**1.68E-01**	0.011556	-0.076468	-0.000734
CEREB3R	-0.079782	**-0.145603**	**-0.100333**	-0.066078	-0.032455	**0.161775**	**1.28E-01**	-0.046196	-0.093406	-0.000957
CEREB45L	-0.098268	-0.030342	-0.015488	**-0.109968**	-0.009033	**0.135046**	9.93E-02	-0.017197	0.099309	-0.045366
CEREB45R	-0.091353	**-0.101728**	-0.042155	**-0.119368**	0.016211	**0.166434**	9.05E-02	-0.057554	**0.112466**	-0.021446
CEREB6L	-0.097748	-0.088625	-0.061217	**-0.105338**	-0.008983	**0.124021**	4.33E-02	0.070162	0.069038	0.036337
CEREB6R	-0.095745	-0.099802	-0.073734	**-0.118269**	0.034966	**0.138917**	7.94E-03	0.049742	0.088938	0.050194
CEREB7BL	-0.081973	0.090919	-0.0817	**-0.260621**	-0.039874	**-0.109192**	-9.06E-02	-0.021581	0.06469	0.042604
CEREB7BR	-0.08363	**0.115788**	-0.085717	**-0.227787**	-0.083374	-0.090191	-9.67E-02	-0.00976	0.070606	0.058787
CEREB8L	-0.063724	0.084281	-0.078358	**-0.330539**	**-0.107019**	**-0.240242**	-9.11E-02	-0.081558	0.021842	0.008321
CEREB8R	-0.065385	**0.103374**	-0.071676	**-0.323922**	**-0.120802**	**-0.233841**	-8.27E-02	-0.069383	0.023593	0.00601
CEREB9L	-0.068828	0.044362	**-0.109685**	**-0.304606**	**-0.117521**	**-0.228983**	-5.07E-02	-0.074897	-0.079614	0.021213
CEREB9R	-0.072569	-0.00068	**-0.146124**	**-0.271661**	**-0.104036**	**-0.202851**	-3.58E-02	-0.050717	**-0.111029**	0.075656
CEREB10L	-0.071632	-0.088878	**-0.120857**	0.009324	**-0.189011**	-0.063941	9.44E-02	**-0.110818**	**-0.24988**	**-0.342692**
CEREB10R	**-0.044914**	**-0.192817**	**-0.155625**	0.040541	**-0.135046**	-0.0991	8.69E-02	**-0.204009**	**-0.162862**	**-0.442921**

VERMIS12	-0.078962	-0.090975	**-0.133294**	-0.072079	-0.044957	**0.143299**	**1.05E-01**	0.001714	**-0.280115**	0.066821
VERMIS3	-0.053296	**-0.238189**	**-0.125289**	-0.012053	0.00939	**0.156372**	**1.79E-01**	-0.078835	-0.047021	-0.035302
VERMIS45	-0.082785	**-0.163728**	-0.043778	-0.069615	-0.029014	**0.17535**	**1.27E-01**	0.000399	0.020125	-0.029821
VERMIS6	-0.075138	**-0.21381**	-0.07555	-0.031584	**-0.119749**	0.091376	7.69E-02	0.074731	0.042129	-0.045055
VERMIS7	-0.084798	-0.058324	-0.066785	-0.042731	0.023356	**0.109554**	4.32E-02	0.048781	0.037807	0.090749
VERMIS8	-0.091873	0.002117	-0.059331	-0.137486	0.093308	**0.171553**	9.00E-03	0.054897	-0.021493	0.081702
VERMIS9	-0.085225	**-0.019838**	-0.078693	**-0.170402**	**0.164324**	**0.163319**	-3.43E-02	-0.009297	0.01499	0.087137
VERMIS10	-0.002736	**-0.288841**	**-0.173665**	-0.011534	-0.051588	-0.008476	-1.49E-02	0.014088	-0.370613	**0.207447**
ETIV	-0.081256	-0.02689	-0.091786	**-0.133629**	**0.146869**	**0.127823**	**-1.03E-01**	-0.021786	0.13753	0.060554

Correlation values for each variable per the first 10 components. Value must be greater than 0.1. 112 variables were found to be correlated.