

*Collection of Biostatistics Research Archive*  
COBRA Preprint Series

---

*Year 2009*

*Paper 62*

---

Two-stage Decompositions for the Analysis of  
Functional Connectivity for fMRI With  
Application to Alzheimer's Disease Risk

Brian S. Caffo\*      Ciprian M. Crainiceanu†      Guillermo Verduzco‡  
Stewart H. Mostofsky\*\*      Susan Spear-Bassett††      James J. Pekar‡‡

\*Johns Hopkins University, bcaffo@jhsph.edu

†Bloomberg School of Public Health, Department of Biostatistics, Johns Hopkins,  
ccrainic@jhsph.edu

‡gverduzco@jhmi.edu

\*\*Johns Hopkins Kennedy Krieger Institute, mostofsky@kennedykrieger.org

††Johns Hopkins Department of Psychiatry, sbassett@jhmi.edu

‡‡Kennedy Krieger Institute, jpekar@kennedykrieger.org

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/cobra/art62>

Copyright ©2009 by the authors.

# Two-stage Decompositions for the Analysis of Functional Connectivity for fMRI With Application to Alzheimer's Disease Risk

Brian S. Caffo, Ciprian M. Crainiceanu, Guillermo Verduzco, Stewart H. Mostofsky, Susan Spear-Bassett, and James J. Pekar

## Abstract

Functional connectivity is the study of correlations in measured neurophysiological signals. Altered functional connectivity has been shown to be associated with numerous diseases including Alzheimer's disease and mild cognitive impairment. In this manuscript we use a two-stage application of the singular value decomposition to obtain data driven population-level measures of functional connectivity in functional magnetic resonance imaging (fMRI). The method is computationally simple and amenable to high dimensional fMRI data with large numbers of subjects. Simulation studies suggest the ability of the decomposition methods to recover population brain networks and their associated loadings. We further demonstrate the utility of these decompositions in a case-control functional logistic regression model. The method is applied to a novel fMRI study of Alzheimer's disease risk under a verbal paired associates task. We found empirical evidence of alternative connectivity in clinically asymptomatic at-risk subjects when compared to controls. The relevant brain network loads primarily on the temporal lobe and overlaps significantly with the olfactory areas and temporal poles.

# Two-stage decompositions for the analysis of functional connectivity for fMRI with application to Alzheimer's disease risk

Brian S. Caffo

Ciprian M. Crainiceanu

Guillermo Verduzco

Suresh Joel

Stewart H. Mostofsky

Susan Spear Bassett

James J. Pekar

December 9, 2009

## **Abstract**

Functional connectivity is the study of correlations in measured neurophysiological signals. Altered functional connectivity has been shown to be associated with numerous diseases including Alzheimer's disease and mild cognitive impairment. In this

manuscript we use a two-stage application of the singular value decomposition to obtain data driven population-level measures of functional connectivity in functional magnetic resonance imaging (fMRI). The method is computationally simple and amenable to high dimensional fMRI data with large numbers of subjects. Simulation studies suggest the ability of the decomposition methods to recover population brain networks and their associated loadings. We further demonstrate the utility of these decompositions in a case-control functional logistic regression model. The method is applied to a novel fMRI study of Alzheimer's disease risk under a verbal paired associates task. We found empirical evidence of alternative connectivity in clinically asymptomatic at-risk subjects when compared to controls. The relevant brain network loads primarily on the temporal lobe and overlaps significantly with the olfactory areas and temporal poles.

## 1 Introduction

Functional connectivity is the study of correlations in measured neurophysiological signals. Disruptions in functional connectivity have been shown to be associated with many clinical sequelæ. However, methods for evaluating covariate-adjusted population level differences in functional connectivity associated with high throughput imaging modalities remains under current development. Matrix decompositions are common methods to summarize single-subject connectivity in functional magnetic resonance imaging (fMRI) for subsequent use in regression modeling. In this manuscript, we follow this approach and investigate a generalization of functional principal components for analyzing popu-

lation fMRI-based connectivity data. We focus our analysis on distinguishing risk-status between subjects at high familial risk for Alzheimer's disease and matched controls.

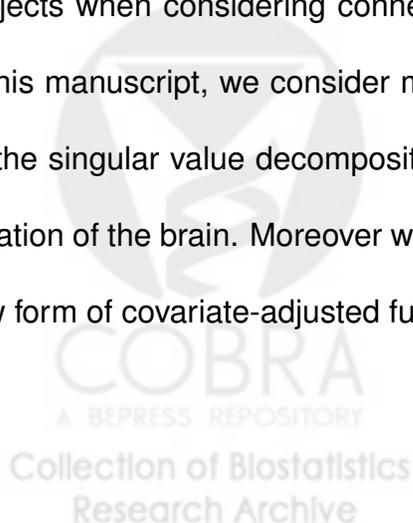
Functional connectivity is formally defined as “statistical dependencies among spatially remote neurophysiological events” (Friston et al., 2007). In practice, the study of functional connectivity is inherently tied to the methods used to evaluate the dependencies and the technology used for measurement (Horwitz, 2003). We focus entirely on functional connectivity as measured by BOLD (blood oxygen level dependent) fMRI using a two-stage singular value decomposition (SVD). The SVD is useful for summarizing the enormous number of correlations available into major directions of variation. We use the SVD to find major directions of both subject-specific and population-level variation in fMRI measurements and relate these directions to familial risk status in Alzheimer's disease using a functional logistic regression model.

The SVD has been used frequently to study connectivity in fMRI. Friston (1994) states that “[the] SVD and equivalent devices are simple and powerful ways of decomposing a neuroimaging time-series into a series of orthogonal patterns that embody, in a stepdown fashion, the greatest amounts of functional connectivity”. Unlike seed voxel or ROI-based techniques, SVD based approaches do not require specifying a-priori anatomical regions or seeds. Moreover, as shown below, the SVD can be implemented quickly on modest computing infrastructures.

Below, we apply a nested application of the singular value decomposition to evaluate group differences in functional connectivity. This method is complimentary to existing factor-analytic group decompositions, such as independent components analysis and its tensor extensions (ICA Calhoun et al., 2001; Lukic et al., 2002; Svensén et al.,

2002; Beckmann and Smith, 2005). Our approach produces orthogonal bases in time and space. These orthogonal bases permits us to connect these results to functional logistic regression. The method follows four steps: *i.*, a subject-specific SVD, *ii.* a population level decomposition of aggregated subject-specific eigenvectors, *iii.* projecting the subject level data onto the population eigenvectors to obtain subject-specific loadings, *iv.* using the subject-specific loadings in a case-control functional logistic regression model. This results in a direct approach for covariate adjustments when relating functional connectivity to group status. We apply these methods to a data set of subjects at high familial risk for Alzheimer's disease and matched controls.

Our analysis of the example dataset builds on extensive existing research demonstrating anatomical, functional and effective connectivity differences between subjects with Alzheimer's disease or cognitive impairment and non-diseased populations. Our study differs from others by considering subjects at high familial risk for Alzheimer's disease that are clinically asymptomatic and matched controls (Bassett et al., 2006). In earlier studies (see Bowman et al., 2008; Caffo et al., 2009), we found group differences in these subjects when considering connectivity associated with regional task-related activation. In this manuscript, we consider more classical voxel-based connectivity using variations on the singular value decomposition. These methods do not rely on an anatomical parcellation of the brain. Moreover we connect the group SVD loadings to risk-status using a new form of covariate-adjusted functional regression.

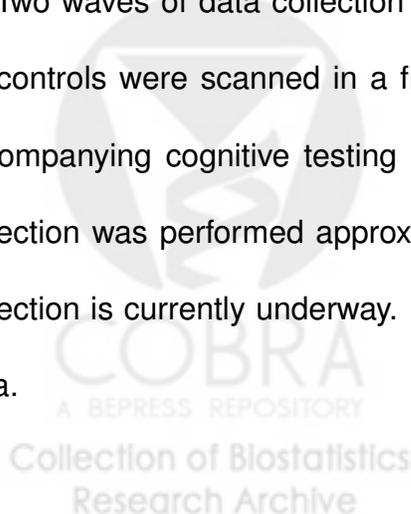


## 2 Data

### 2.1 Study population

The data derive from an ongoing study of Alzheimer's disease risk and biomarkers (Bassett et al., 2006; Yassa et al., 2008). The data compare subjects at high familial risk for Alzheimer's disease and controls, usually low-risk spouses. Subjects were declared at risk if at least one parent had an autopsy-confirmed diagnosis of AD and at least one first-degree relative with a clinical diagnosis of probable AD. Control subjects had no affected or diagnosed first degree relatives, screened negative on the Alzheimer Dementia Risk Questionnaire (Breitner and Folstein, 1984) and the Dementia Questionnaire (Silverman et al., 1986). Both control and at-risk subjects had no clinical AD symptoms. Specifically, all subjects were free of self reported memory complaints or treatments and scored in a normal range on the Telephone Interview for Cognitive Status (Brandt et al., 1988). At-risk subjects were an average of 11 years younger than the age of diagnosis for the affected parent. All subjects were over 50.

Two waves of data collection have been completed. Ninety five at-risk subjects and 90 controls were scanned in a first wave along with collection of important covariates, accompanying cognitive testing and blood for genetic typing. A second wave of data collection was performed approximately four years after baseline. A third wave of data collection is currently underway. In our example data set, we consider the second wave data.



## 2.2 Imaging protocol

The fMRI images were obtained via a 1.5 T Philips Intera-NT scanner (Philips Medical Systems, Best, The Netherlands) at the F.M. Kirby Functional Imaging Research Center (Kennedy Krieger Institute, Baltimore, MD). The system utilizes a Galaxy gradient (66 mT/m at 110 mT/m/s). A standard head coil was used in image acquisition. A sagittal localizer scan was collected for orientation. Two functional scans were acquired using echo-planar imaging (EPI) and a blood oxygenation level-dependent (BOLD) technique with repetition time (TR) = 1000 ms, echo time (TE) = 39 ms, flip angle = 90 degrees, field of view (FOV) = 230 mm in the x-y plane and matrix size =  $64 \times 64$  reconstructed to  $128 \times 128$ . Eighteen coronal slices were acquired with a 4.5 mm thickness and an interslice gap of 0.5 mm, oriented perpendicular to the anteriorposterior commissure (ACPC) line. Slices were acquired along the z-axis, yielding a total coverage of 90 mm. Two sessions were performed, each with 370 time points. The data in this analysis considers only the first session. Total fMRI acquisition time was 12 minutes and 20 seconds.

The paradigm, programmed in E-prime 1.1 (Psychology Software Tools, Inc., Pittsburgh, PA, USA), was an auditory word-pair association task consisting of two six minute and ten second sessions. Each session consisted of six sets of three blocks. The types of blocks included encoding, recall, and rest. In the encoding block, subjects were presented with seven unrelated word pairs. In the recall block, subjects were presented with the first word of each pair and instructed to silently recall the second. In the baseline block, subjects were presented with an asterisk.

Data pre-processing was performed using Statistical Parametric Mapping (SPM99,

Wellcome Department of Imaging Neuroscience, University College, London, UK) under MATLAB 7.0 (The Mathworks, Sherborn, MA, USA). Images were motion corrected by a six-parameter rigid-body realignment with the mean image across sessions. This was followed by re-slicing using a windowed-sinc interpolation. Non-linear normalization using  $7 \times 8 \times 7$  basis functions was used to warp each individual's data into standard stereotaxic space. Template space was defined by SPM's standard EPI template (Montreal Neurologic Institute, McGill University, Montreal, Canada). The template was manually cut to fit each individual scan in order to improve the quality of normalization on the partial-brain scans. Normalized scans were re-sliced to isotropic voxels ( $2 \text{ mm}^3$ ), using trilinear interpolation and spatially smoothed with a full-width at half-maximum (FWHM) Gaussian kernel of 5 mm.

### 3 Methods

Let  $Y_i(v, t)$  represent the fMRI data for voxel  $v = 1, \dots, V$  and scan  $t = 1, \dots, T$ . Our goal is to obtain a parsimonious decomposition

$$Y_i(v, t) = \sum_j \sum_k \psi_j(v) \xi_k(t) \lambda_{ijk},$$

where  $\psi_j(v)$  and  $\xi_k(t)$  represent orthonormal functional bases in space and time, respectively. Notice that  $\psi_j(v)$  and  $\xi_k(t)$  are population level bases that do not vary by subject. In contrast, the loadings,  $\lambda_{ijk}$ , are subject-specific. We show how to use the  $\lambda_{ijk}$  in subsequent analyses as summaries of functional connectivity that achieve a great deal of data reduction.

Our approach utilizes two stages, subject-specific SVDs followed by population-level principal components analysis. This approach is particularly well suited to high-dimensional neuroimaging data and we further demonstrate how calculations can be performed on very modest computing resources. In the first stage, we obtain subject-specific decompositions

$$Y_i(v, t) = \sum_j \sum_k \tilde{\psi}_{ij}(v) \tilde{\xi}_{ik}(t) \tilde{\lambda}_{ijk},$$

where

$$\tilde{\lambda}_{ijk} = \int \int \tilde{\psi}_{ij}(v) \tilde{\xi}_{ik}(t) dv dt.$$

In the second stage, we retain a small number (say  $L$ ) of  $\tilde{\psi}_{ij}(v)$  and consider the populations of spatial functions  $\mathcal{A} = \{\tilde{\psi}_{ij}(v)\}_{i=1, \dots, N, j=1, \dots, L}$  and time series  $\mathcal{B} = \{\tilde{\xi}_{ik}(t)\}_{i=1, \dots, N, k=1, \dots, L}$ . These collections of functions are then decomposed using functional principal components. In specific, we obtain decompositions

$$\tilde{\psi}_{ij}(v) = \sum_l \delta_{ijl} \psi_j(v) \quad \text{and} \quad \tilde{\xi}_{ik}(t) = \sum_l \gamma_{ijl} \xi_k(t).$$

Here,  $\psi_j(v)$  are the eigenfunctions associated with  $\mathcal{A}$  and  $\xi_k(t)$  are the eigenfunctions associated with  $\mathcal{B}$  and  $\delta_{ijl}$  and  $\gamma_{ijl}$  are associated eigenvalues. The eigenfunctions are then used to obtain the subject-specific loadings:  $\lambda_{ijk} = \int \int Y_i(v, t) \psi_j(v) \xi_k(t) dv dt$ , where  $\lambda_{ijk}$  is the subject-specific loading onto the left eigenfunction  $j$  and right eigenfunction  $k$ .

The benefits of this approach for summarizing connectivity information over related methods are numerous. Firstly, only standard matrix decompositions are needed to estimate the  $\lambda_{ijk}$ . This is in contrast with full tensor-based SVD methods (see Leibovici and Sabatier, 1998). Secondly, the process is performed iteratively in two stages. Thus, it mirrors standard two-stage random effect analyses of fMRI data and computing can be

parallelized. Moreover, because of the two stage process, the method can be applied on very low memory systems. Thirdly, the parameters are uniquely interpretable. The  $\psi_j(v)$  are population-level eigenimages, summarizing areas of temporal synchronization across subjects. The  $\xi_k(t)$  are population-level eigenvariates, summarizing times of spatial synchronization. The  $\lambda_{ijk}$  represent the loading of subject  $i$  onto population eigenimages  $j$  and eigenvariates  $k$ . Hence we hypothesize that these loadings will be a useful summary of connectivity, that may be useful as predictors. Moreover, we demonstrate how their use in regression models connects to functional regression. The loadings achieve a great deal of dimension reduction; in our example, we demonstrate interesting findings using only 25 of the loadings.

### 3.1 Implementation

Here we discuss implementation issues in dealing with high dimensional fMRI data. First, a brain mask is applied across subjects and only those voxels represented in all subjects are retained. This removes both background voxels as well as boundary voxels with incomplete data across subjects due to inexact registration. Let  $\mathbf{Y}_i$  be the  $V \times T$  data matrix for subject  $i$ . We assume that  $\mathbf{Y}_i$  is centered in both time and space; i.e.  $\mathbf{Y}_i = \{\mathbf{I} - \mathbf{1}'(\mathbf{1}\mathbf{1}')^{-1}\mathbf{1}'\}\tilde{\mathbf{Y}}_i\{\mathbf{I} - \mathbf{1}'(\mathbf{1}\mathbf{1}')^{-1}\mathbf{1}'\}$  where  $\tilde{\mathbf{Y}}_i$  is the uncentered data matrix. We define the global connectivity matrix as the  $V \times V$  matrix  $\mathbf{Y}_i\mathbf{Y}_i'/T$ , which has  $(v_1, v_2)$  element  $\frac{1}{T} \sum_{t=1}^T Y_i(v_1, t)Y_i(v_2, t)$ . This matrix completely summarizes temporally synchronous behavior in the fMRI data. However, having  $\binom{V}{2}$  unique elements, it necessarily must be summarized and is difficult to work with computationally. In contrast,  $\mathbf{Y}_i'\mathbf{Y}_i$  is  $T \times T$ ,

where  $T$  is usually on the order of 500 or fewer. Consider the eigenvalue decomposition of  $\mathbf{Y}'_i \mathbf{Y}_i = \mathbf{V}_i \mathbf{D}_i^2 \mathbf{V}'_i$  where  $\mathbf{V}'_i \mathbf{V}_i = \mathbf{I}$  and  $\mathbf{D}_i^2$  is a diagonal matrix of eigenvalues. Here, the columns of  $\mathbf{V}_i$  contain the subject-specific eigenvariates. Let  $\mathbf{U}_i = \mathbf{Y}_i \mathbf{V}_i \mathbf{D}_i^{-1}$ . Notice that, performed in this order,  $\mathbf{U}_i, \mathbf{D}_i$  and  $\mathbf{V}_i$  can be obtained quickly, without having to reserve memory or perform operations on the  $V \times V$  global connectivity matrix. Defining  $\mathbf{U}_i$  as such implies  $\mathbf{Y}_i = \mathbf{U}_i \mathbf{D}_i \mathbf{V}'_i$  and  $\frac{1}{T} \mathbf{Y}_i \mathbf{Y}'_i = \frac{1}{T} \mathbf{U}_i \mathbf{D}_i^2 \mathbf{U}'_i$ .

The columns of  $\mathbf{U}_i$  are referred to as eigenimages (Friston et al., 2007) and approximate the  $\psi_i(v)$ . Areas of the brain that load heavily within a column jointly explain variation in the fMRI images and hence are often thought to represent brain networks (Friston, 1994). This is further evidenced by noting that, the global connectivity matrix satisfies  $\frac{1}{T} \mathbf{Y}_i \mathbf{Y}'_i = \frac{1}{T} \sum D_{ij} \mathbf{U}_{ij} \mathbf{U}'_{ij}$  where  $D_{ij}$  is the  $j^{th}$  diagonal entry of  $\mathbf{D}_i$  and  $\mathbf{U}_{ij}$  is column (eigenimage)  $j$  from  $\mathbf{U}_i$ . That is, the global connectivity matrix decomposes into a weighted sum of the outer products of the brain networks. The columns of  $\mathbf{V}_i$ , referred to as eigenvariates, estimate  $\xi_i(t)$ . One can think of these as representing how the brain networks mix over time.

Slightly abusing notation, suppose that  $\mathbf{V}_i$  and  $\mathbf{U}_i$  contain relatively few columns (say  $L = 5$ ). Let  $\mathbf{E}$  be the  $(NL) \times T$  matrix obtained by stacking the  $\mathbf{V}'_i$  across subjects and  $\mathbf{H}$  be the  $(NL) \times V$  matrix obtained by stacking the  $\mathbf{U}'_i$  across subjects. Let  $\hat{\Sigma}_E$  be the sample variance matrix:  $\mathbf{E}' \{ \mathbf{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \} \{ \mathbf{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \} \mathbf{E}$  and  $\hat{\Sigma}_H = \mathbf{H}' \{ \mathbf{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \} \{ \mathbf{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1} \mathbf{1}' \} \mathbf{H}$  be the corresponding matrix for  $\mathbf{H}$ . We consider the eigenvalue decomposition of  $\hat{\Sigma}_E = \mathbf{V} \mathbf{D}_E^2 \mathbf{V}'$  and  $\hat{\Sigma}_H = \mathbf{U} \mathbf{D}_H^2 \mathbf{U}'$ . Now the columns of  $\mathbf{V}$  represent population eigenvariates and the columns of  $\mathbf{U}$  represent population eigenimages; hence estimating  $\psi_j(v)$  and  $\xi_j(v)$  respectively. In practice, calculating  $\mathbf{U}$  requires a similar technique as

outlined above to avoid creating the  $V \times V$  matrix  $\Sigma_H$ .

We then project original subject specific data onto these bases. In specific, Let  $\Lambda_i = [\lambda_{ijk}]_{j,k} = \mathbf{U}\mathbf{Y}_i\mathbf{V}'$ . Here element  $(j, k)$  of  $\Lambda_i$  represents the subject-specific loading onto population eigenimage  $j$  and eigenvariate  $k$ . Hence, it represents the loading onto the specific brain network given by column  $j$  of  $\mathbf{U}$  for the particular time series represented by column  $k$  of  $\mathbf{V}$ . These are the proposed estimates for  $\lambda_{ijk}$ .

### 3.2 Functional regression

We consider a retrospective-style analysis with case-status as the outcome and the fMRI data and covariates as predictors. Let  $D_i \in \{0, 1\}$  represent the risk status for subject  $i$  with covariate values  $\mathbf{X}_i$ . Consider the functional regression model:

$$\text{logit}\{P(D_{ij} = 1)\} = \int \int Y_i(v, t)\beta(v, t)dvdt + \mathbf{X}_i\boldsymbol{\gamma}.$$

Let  $\phi_j(v)$  and  $\xi_k(t)$  be eigenfunctions. Then we have

$$\int \int Y_i(v, t)\beta(v, t)dvdt = \int \int \left\{ \sum_j \sum_k \phi_j(v)\xi_k(t)\lambda_{ijk} \right\} \beta(v, t)dvdt = \sum_j \sum_k \lambda_{ijk}\tau_{jk},$$

where  $\tau_{jk} = \int \int \phi_j(v)\xi_k(t)\beta(v, t)$ . Hence our model becomes

$$\text{logit}\{P(D_{ij} = 1)\} = \sum_j \sum_k \lambda_{ijk}\tau_{jk} + \mathbf{X}_i'\boldsymbol{\gamma}.$$

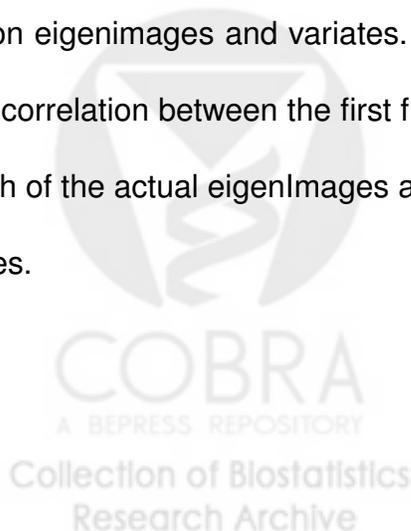
Here  $\tau_{jk}$  represents the change in the log-odds for risk status for brain network  $j$  and time series  $k$ .

To summarize, logistic regression models having the loadings as covariates result in a form of functional regression involving the entire subject-specific fMRI time series

integrated over the bases derived by the population eigenimages and variates. In this way, covariate-adjusted regression relationships associated with connectivity can be explored easily. Moreover, to account for matching generalized linear mixed effect models (see McCulloch and Searle, 2004) with a fixed effect design matrix comprised of the  $\lambda_{ijk}$  and  $X_i$  and pair-specific random effects can be used.

### 3.3 Simulation study

To evaluate the ability of the two-stage decomposition method to recover population eigenimages and variates, we conducted a simulation study. Figure 1 displays five population eigenimages and eigenvariate time series. For the eigenimages, each gray scale represents a different network. The eigenvariates were a set of orthogonal cosine functions with different periods. Let  $\mathbf{U}$  and  $\mathbf{V}$  be matrices with columns representing the eigenimages and variates, respectively. We then simulated 200 a  $5 \times 5$  matrices  $\tilde{\Lambda}_i = [\lambda_{ijk}]_{j,k}$  with each  $\lambda_{ijk} \sim N(0, \frac{216}{2^{i3^j}})$ . Then we defined subject-specific data matrices as  $\mathbf{Y}_i = \mathbf{U}'\tilde{\Lambda}_i\mathbf{V}$ . The two-stage decomposition method was then applied to obtain estimates of the population eigenimages and variates. We then calculated the maximum of absolute value of the correlation between the first five estimated population eigenimages and variates with each of the actual eigenimages and variates. This entire process was then repeated 100 times.



## 4 Results

### 4.1 Simulation data

Table 1 displays the median, maximum and range of the maximum absolute value of the correlation between the estimated population eigenimages and eigenvalues and the ones used for simulation. That is, we found the correlation between the estimated eigenimage and variate that agrees best with the true values. The results show that the method recovers the eigenvariates extremely well and recovers the eigenimages fairly well. The results were similar over a range of simulation settings. In the discussion we illustrate settings where the method obtains mixtures of the eigenimages and was unable to separate them.

### 4.2 Analysis of the AD data set

Table 2 displays demographic data for the AD At-risk data set. In this second wave of study there are 81 at-risk subjects and 68 controls. The groups are well matched on gender, age and education level. Unsurprisingly, there is a significant difference in the number of  $\epsilon 4$  alleles of the Apolipoprotein E gene, as the number of such alleles has been associated with late onset Alzheimer's disease (see Strittmatter and Roses, 1996, for a review).

We then applied the two-stage decompositions outlined above with one caveat. The population eigenimages and eigenvariates were heavily dominated by overall subject-specific signal changes from mean shifts being the first component of the subject-specific decompositions. This heterogeneity represent technological variation, such as scanner

gain, that is not of interest when comparing group connectivity. Subtracting the row means of  $E$  and  $H$  above (i.e. forcing the first level eigenimages and variates to have mean zero) removed any overall shifts from the population level eigenimages and variates.

Figure 2 displays the percentage of the population variation in the first level eigenimages and eigenvariates explained by the second level principal components decomposition. Both of these curves have a fairly slow rate of decay, suggesting a large degree of population-level heterogeneity in the subject-specific decompositions.

Figure 3 displays three-D renderings of the first ten population eigenimages. For context, Figure 4 gives regions of interest (based on the anatomical parcellation given in Tzourio-Mazoyer et al., 2002) that have over 20% of their area overlapping with the eigenimage. Recall, eigenimages and variates are unique only up to scalings and therefore positive and negative values could be reversed with no loss of information.

We summarize a subset of the population eigenimages. The first loads primarily on the superior portion of the temporal lobe. The second covers the majority of the imaging area. The third loads heavily on the temporal lobe and limbic substructures, such as the para-hippocampal gyrus. The fourth covers temporal and limbic areas and intersects with the small portion of the cerebellum in the imaging area. The eighth, which we will see is one of the more important eigenimages, loads specifically on temporal and limbic areas, especially covering olfactory areas. This is of interest as deficits in olfaction have been hypothesized to be connected with neurodegenerative disorders and AD in particular (see Meshulam et al., 1998, for a meta analysis and review).

Figure 5 displays the first ten population eigenvariates and their associated spectrum. The first eigenvariate represents a drift in the signal, which could represent biological

or technological trends, such as learning effects or scanner drift. We reiterate that one must remember that the signs of such analysis are arbitrary and could represent either an increase or decrease in the signal over the session. The following two population eigenvariates represent slowly varying functions. The remaining have spectra that include spikes at the same frequency as the paradigm (see Figure 6), but also include higher frequency information. We further investigated if the eigenvariates separate between the two components of the task (encoding versus recall), which have the same spectra, but different phases. The fourth eigenvariate time is more correlated with the recall paradigm rather than encoding (-.02 versus .23). Eigenvariates 6 and 9 are more correlated with encoding than recall (correlations of -.14 versus -.04 and .19 versus .02, respectively). Eigenvariate 8, which is the most obviously associated with the paradigm, was more correlated with encoding (-.45 versus -.25), but retained significant correlation with the recall blocks. To elaborate, the peaks of this eigenvariate occur between the encoding and recall blocks, though are slightly closer to the recall blocks.

We next considered the use of the subject-specific loadings in functional logistic regression models. Before fitting fixed effect regression models, we first considered random effect models that accounted for spousal matching and familial aggregation. Fitted results suggested little or no correlation due to spousal matching or family. Therefore, we omit addressing this potential correlation in subsequent analyses.

Table 3 displays P-values for predicting risk-status treating each population loading in a separate model. Figure 7 shows 25<sup>th</sup>, 50<sup>th</sup> and 75<sup>th</sup> percentiles by risk group for the ten most significant loadings. Of these, the most significant is the fourth population eigenimage and eighth eigenvariate. This appears to incorporate variation associated

with the paradigm in the temporal poles. The next most significant loads heavily on the eighth population eigenimage and sixth eigenvariate. This eigenvariate includes slower variation contrasted between the superior temporal lobe and the olfactory areas of the temporal lobe. Given that the estimates are entirely empirical without a-priori hypotheses, multiplicity issues demand that these results must be interpreted with a grain of salt. To address this issue, we refit the models and retained the smallest P-value 1,000 times, permuting risk status each time (thus breaking any potential association between risk status and loadings). A histogram estimate of the minimum P-value is given in Figure 8. These results suggests that there is a 50% chance of obtaining a minimum P-value as small as 0.01 and hence the possibility that the results may be due to chance associations can not be ruled out.

## 5 Discussion

This paper shows the utility of two-stage decompositions for the analysis of population based fMRI data. Our approach first used the singular value decomposition to obtain subject-specific eigenimages (networks) and eigenvariates (time series). A small number of these are retained and aggregated. Separate second-level eigenvalue decompositions for the collections of eigenvariates and eigenimages, respectively, are used to form population-level brain networks and time series. We project the subject-level data onto these population eigenvectors to obtain a matrix of loadings onto each network/time series combination. We further showed how these loadings can be used in a generalized functional regression. We applied then in a matched case/control style analysis of

Alzheimer's disease familial risk status.

The two-stage decomposition approach has several notable benefits as an exploratory method for discovering population brain networks and major directions of functional variation. Foremost is computational ease. Subject-specific decompositions are relatively easily obtained and, by retaining only a few of the networks and time series, the population values are similarly easily computed. We further explicitly demonstrated how calculations can be approached so that a high dimensional full connectivity matrix is never required to be loaded into memory. In addition, first level calculations can be easily made embarrassingly parallel. Thus, this methodology will scale to next-generation studies involving hundreds or thousands of subjects. Another alternative would be the use of tensor extensions of the SVD and factor analysis (Leibovici and Sabatier, 1998; Kolda and Bader, 2009; De Lathauwer et al., 2000). While these methods offer more theoretically complete alternatives, they lack the simplicity and easy execution of two-stage decompositions.

By using the SVD as the basis for the decomposition, the most variable aspects of the population of fMRI data are used in the ensuing functional regression. This is useful, as more variable predictors will have lower standard errors. Our simulation studies highlight the ability of these methods to recover population-level networks and time series and effectively incorporate them into functional regression models.

We also demonstrate the inferential potential of these population networks and time series by projecting subject level data onto these bases. We then show how functional regression modeling can be used to assess significance of the loadings. Our work is influenced by work in functional regression for non-functional neuroimaging in Reiss and Ogden (2008a), Reiss and Ogden (2008b) and Reiss et al. (2005).

A potential point of criticism is the difficulty in the choice of the number of components to be included from the subject-specific decompositions in the second level analysis. Our current approach uses visual inspection of the collection of scree plots and we stipulate that this could be improved upon. A second point of criticism is the lack of accounting for the multiple observations per subject contributing to the population level decomposition. We are less concerned with this aspect of the analysis, as this would affect inference based on the population eigenvectors more than estimation. However, in this manuscript, we focus on estimation and the use of the eigenvectors as predictors in functional regression models and do not make use of their measurement variation (though see Crainiceanu et al., 2009a). A final point of criticism is the lack of use of the variance ordering of subject-specific eigenvectors in the subsequent population analysis. That is, a subject's first and fifth eigenvectors are treated equally in the population decomposition. We hypothesize that this criticism can be addressed by a weighting using the inverse of the associated eigenvalues. However, we relegate this approach to future research.

This manuscript addresses decomposition methods to evaluate cross-sectional variation in brain networks. However, longitudinal functional imaging studies are becoming increasingly common. We have developed multilevel functional principal component methods for functions of one variable (time, for example) and are currently generalizing methods to consider hierarchical imaging data. However, the extension to extremely high dimensional imaging data remains a difficult task. Furthermore, connecting these decomposition methods with outcomes via functional regression is an area of active research (see Di et al., 2009; Crainiceanu et al., 2009b).

It is also worth noting limitations of using SVD to study brain networks. First, these

decompositions guarantee orthogonal eigenimages and eigenvariates, which may or may not reflect actual biology. Our simulation study specifically assumed orthogonal networks and time series. Moreover, our simulation study imposed a large amount of variation when mixing over the time-series and images, also creating an ideal setting for the SVD. The method would struggle if signals were mixed largely in equal parts. In contrast, other methods, such as ICA, are more robust to these assumptions and hence are popular for analyzing brain connectivity (see Calhoun et al., 2003). However, unlike ICA, this two-stage SVD does not force a distinction between spatial and temporal decompositions. In addition, our two-stage method avoids the question of whether to stack rows or columns for group analysis (Calhoun et al., 2001; Lukic et al., 2002; Svensén et al., 2002; Guo and Pagnoni, 2008). More analogous tensor versions of ICA have been proposed (Beckmann and Smith, 2005); however, it is not clear whether computations will scale to large fMRI studies. Finally, though imposing orthogonality is rigid, it is very useful for creating a basis to decompose the fMRI signal for subsequent use in regression models.

The analysis of the AD risk data set yields interesting findings on altered connectivity in subjects with high familial risk for Alzheimer's disease. The atypically large sample size for a functional imaging study and pre-clinical population including subjects at high familial risk are unique aspects of this study. This analysis corroborates differences in connectivity found using other methods on the same data (Caffo et al., 2009; Bowman et al., 2008). It is also of interest to note that, unlike the first wave (Bassett et al., 2006), group differences in paradigm-related activity were unremarkable in the second wave (see Caffo et al., 2009; Bowman et al., 2008). This change in paradigm-related activity may be due to a variety of reasons, including learning effects, differential attention to the task between the groups

across visits and so on. A benefit of the study of functional connectivity is the lack of reliance on the paradigm, and hence potential robustness to these effects.

We demonstrate evidence for altered connectivity between asymptomatic at-risk subjects. Of primary interest is group segregation for the network encompassing the temporal poles and the olfactory areas of the temporal lobe. However, we caution over-interpretation of these results, as connectivity differences were not a primary a-priori hypothesis of the study and this effect did not survive multiplicity adjustment. For future work, we are investigating the robustness of the networks over time, both in the earlier phase and the third phase currently being collected. Further, potential weakness of our study is the narrow imaging area, which ignores possible long-range connectivity. However, we note that the imaging area focused on a band surrounding the medial temporal lobe, an area believed to be associated with AD (see the discussion in Bassett et al., 2006).

Our study compliments existing research on altered anatomical and functional connectivity between mild AD and mild cognitive impairment subjects and controls. Grady et al. (2001) studied 21 health elderly subjects and 11 mildly demented subjects using rCBF PET. They reported decreased correlations for the demented group between task-related areas in the prefrontal cortex and hippocampus. Stam et al. (2007) considered small-world network hypotheses using EEG comparing 15 Alzheimer's patients and 13 control subjects. They report decreased complexity of the network for the diseased group. Greicius et al. (2004) used ICA and fMRI to study default mode network differences between 13 mild AD cases and matched controls. They found decreased activation in the default mode network for the AD group in the posterior cingulate and hippocampus. Wang et al.

(2006) studied connectivity between the hippocampus and other regions in 13 mild AD cases and matched controls and found disrupted and increased connectivity for the AD and control groups. Wang et al. (2007) considered inter-regional correlations between 14 AD subjects and matched controls in PET and found both increased and decreased inter-group connectivity differences.

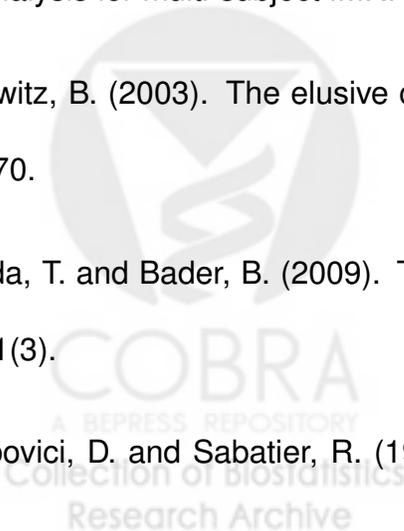
In summary, the two-stage applications of the singular value decomposition along with functional logistic regression can shed considerable light on group fMRI studies. Estimates are easily calculated and computations scale to large studies. The functional logistic regression model allows for easy consideration of covariate effects, subject-level matching. This method of analysis, though exploratory, raises the possibility of novel associations between altered connectivity and Alzheimer's disease risk status.

## References

- Bassett, S., Yousem, D., Cristinzio, C., Kusevic, I., Yassa, M., Caffo, B., and Zeger, S. (2006). Familial risk for Alzheimer's disease alters fMRI activation patterns. *Brain*, 129(5):1229.
- Beckmann, C. and Smith, S. (2005). Tensorial extensions of independent component analysis for multisubject fMRI analysis. *Neuroimage*, 25(1):294–311.
- Bowman, F., Caffo, B., Bassett, S., and Kilts, C. (2008). A Bayesian hierarchical framework for spatial modeling of fMRI data. *NeuroImage*, 39(1):146–156.

- Brandt, J., Spencer, M., and Folstein, M. (1988). The telephone interview for cognitive status. *Neuropsychiatry Neuropsychol Behav Neurol*, 1(2):111–117.
- Breitner, J. and Folstein, M. (1984). Familial Alzheimer dementia: a prevalent disorder with specific clinical features. *Psychological medicine*, 14(1):63.
- Caffo, B., Bowman, D., Eberly, L., and Bassett, S. (2009). A Markov chain Monte Carlo based analysis of a multilevel model for functional MRI data. *The Handbook of Markov Chain Monte Carlo edited by Steve Brooks and Andrew Gelman and Galin Jones and Xiao-Li Meng*.
- Calhoun, V., Adali, T., Hansen, L., Larsen, J., and Pekar, J. (2003). ICA of functional MRI data: an overview. In *Fourth International Symposium on Independent Component Analysis and Blind Source Separation*, pages 281–288.
- Calhoun, V., Adali, T., Pearlson, G., and Pekar, J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, 14(3):140–151.
- Crainiceanu, C. M., Caffo, B., Di, C., and Punjabi, N. M. (2009a). Nonparametric signal extraction and measurement error in the analysis of electroencephalographic activity during sleep. *Journal of the American Statistical Association*, 486:541–555.
- Crainiceanu, C. M., Staicu, A.-M., and Di, C. (2009b). Generalized multilevel functional regression. *Journal of the American Statistical Association to appear*.
- De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278.

- Di, C., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics*, 3:458–488.
- Friston, K. (1994). Functional and effective connectivity in neuroimaging: a synthesis. *Human Brain Mapping*, 2.
- Friston, K., Ashburner, J., Stefan, K., Nichols, T., and Penny, W., editors (2007). *Statistical Parametric Mapping The Analysis of Functional Brain Images*. Academic Press.
- Grady, C., Furey, M., Pietrini, P., Horwitz, B., and Rapoport, S. (2001). Altered brain functional connectivity and impaired short-term memory in Alzheimer's disease. *Brain*, 124(4):739.
- Greicius, M., Srivastava, G., Reiss, A., and Menon, V. (2004). Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proceedings of the National Academy of Sciences*, 101(13):4637–4642.
- Guo, Y. and Pagnoni, G. (2008). A unified framework for group independent component analysis for multi-subject fMRI data. *NeuroImage*, 42(3):1078–1093.
- Horwitz, B. (2003). The elusive concept of brain connectivity. *NeuroImage*, 19(2):466 – 470.
- Kolda, T. and Bader, B. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3).
- Leibovici, D. and Sabatier, R. (1998). A singular value decomposition of a k-way array



- for a principal component analysis of multiway data, PTA-k. *Linear Algebra and its Applications*, 269(1-3):307–329.
- Lukic, A., Wernick, M., Hansen, L., Anderson, J., and Strother, S. (2002). A spatially robust ICA algorithm for multiple fMRI data sets. In *IEEE International Symposium on Biomedical Imaging*, volume 839, page 842. Citeseer.
- McCulloch, C. and Searle, S. (2004). *Generalized, linear, and mixed models*. Wiley-Interscience.
- Meshulam, R., Moberg, P., Mahr, R., and Doty, R. (1998). Olfaction in neurodegenerative disease a meta-analysis of olfactory functioning in Alzheimer's and Parkinson's diseases.
- Reiss, P. and Ogden, R. (2008a). Functional generalized linear models with applications to neuroimaging. *Philip T. Reiss*, page 10.
- Reiss, P., Ogden, R., Mann, J., and Parsey, R. (2005). Functional logistic regression with PET imaging data: A voxel-level clinical diagnostic tool. *Journal of Cerebral Blood Flow & Metabolism*, 25(1s):S635.
- Reiss, P. and Ogden, T. (2008b). Functional generalized linear models with images as predictors. *Biometrika*, 64:1–24.
- Silverman, J., Breitner, J., Mohs, R., and Davis, K. (1986). Reliability of the family history method in genetic studies of Alzheimer's disease and related dementias. *American Journal of Psychiatry*, 143(10):1279–1282.

- Stam, C., Jones, B., Nolte, G., Breakspear, M., and Scheltens, P. (2007). Small-world networks and functional connectivity in Alzheimer's disease. *Cerebral Cortex*, 17(1):92.
- Strittmatter, W. and Roses, A. (1996). Apolipoprotein E and Alzheimer's disease. *Annual review of neuroscience*, 19(1):53–77.
- Svensén, M., Kruggel, F., and Benali, H. (2002). ICA of fMRI group study data. *NeuroImage*, 16(3PA):551–563.
- Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289.
- Wang, K., Liang, M., Wang, L., Tian, L., Zhang, X., Li, K., and Jiang, T. (2007). Altered functional connectivity in early alzheimer's disease: a resting-state fmri study. *Human brain mapping*, 28(10):967–978.
- Wang, L., Zang, Y., He, Y., Liang, M., Zhang, X., Tian, L., Wu, T., Jiang, T., and Li, K. (2006). Changes in hippocampal connectivity in the early stages of Alzheimer's disease: evidence from resting state fMRI. *Neuroimage*, 31(2):496–504.
- Yassa, M., Verduzco, G., Cristinzio, C., and Bassett, S. (2008). Altered fMRI activation during mental rotation in those at genetic risk for Alzheimer disease. *Neurology*, 70(20):1898.

## A Figures

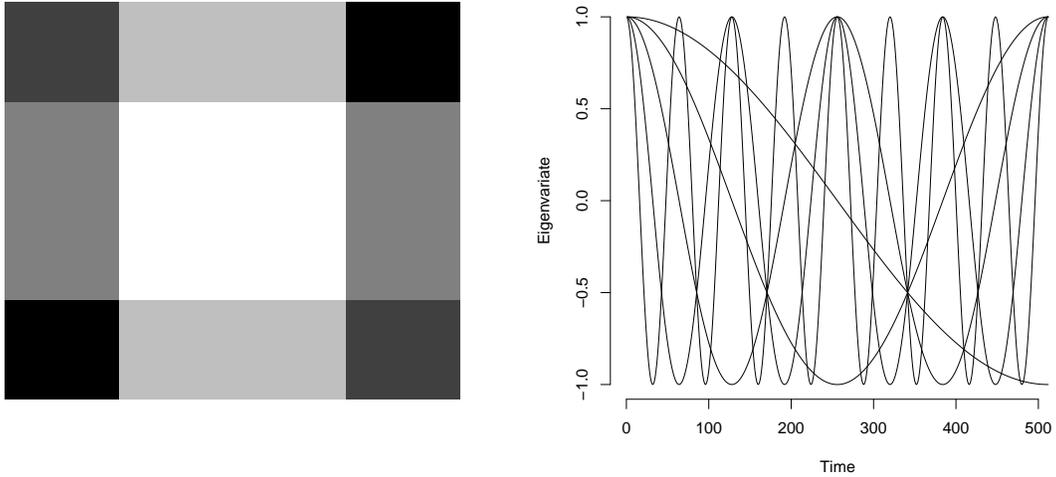


Figure 1: Eigenimages (left) and eigenvariates (right) used for the simulation study.



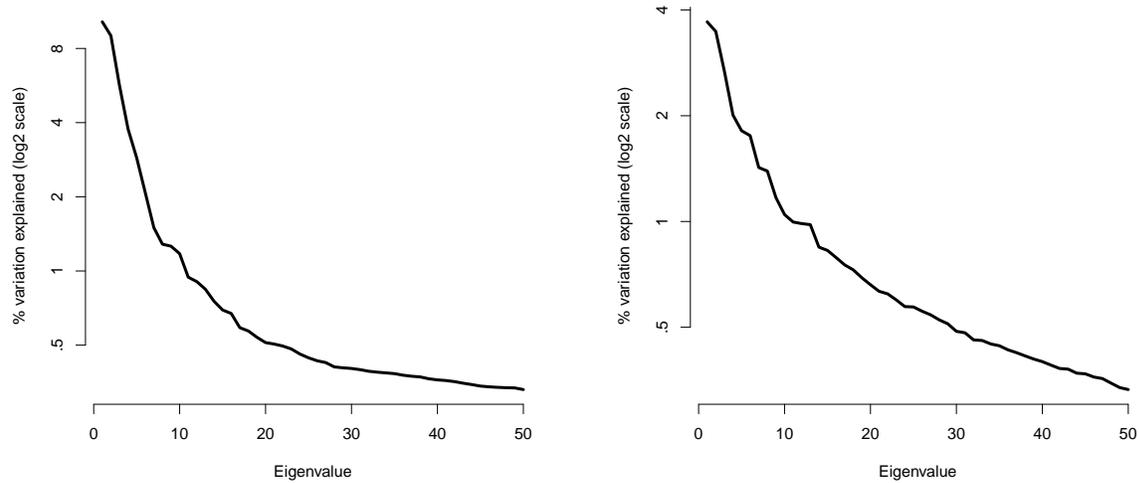


Figure 2: Percent of the population variation in first level-eigenvariates (left) and eigenimages (right) explained by second-level eigenvalue decomposition. The horizontal axis has been transformed by log base 2 with the natural scale percentages displayed on the tick marks.



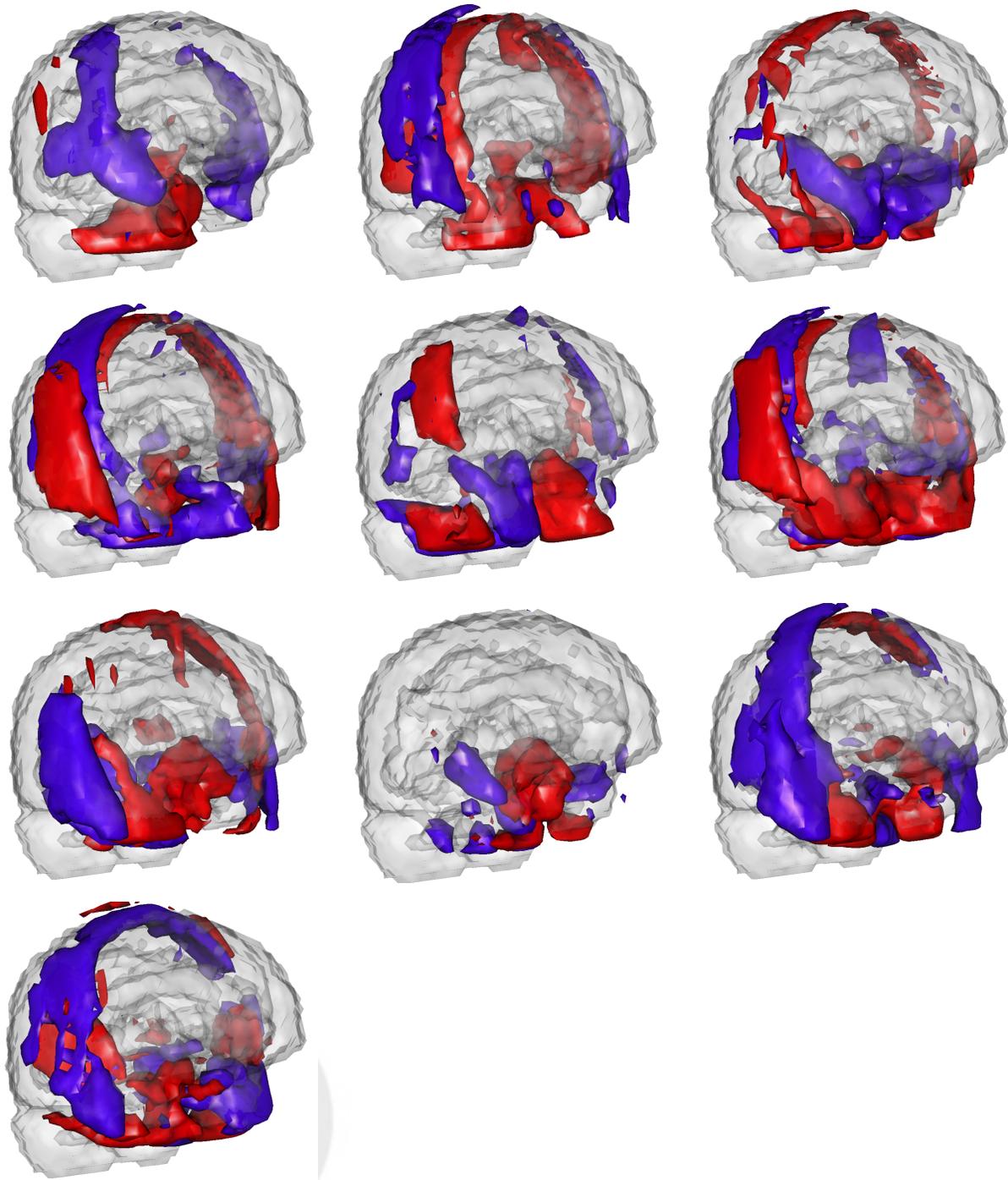


Figure 3: Three-D rendering of thresholded versions of the first six eigen images overlaid on a template. Red areas load positively while blue areas load negatively. The upper left is the first eigen image, the upper middle is the second, and so on.

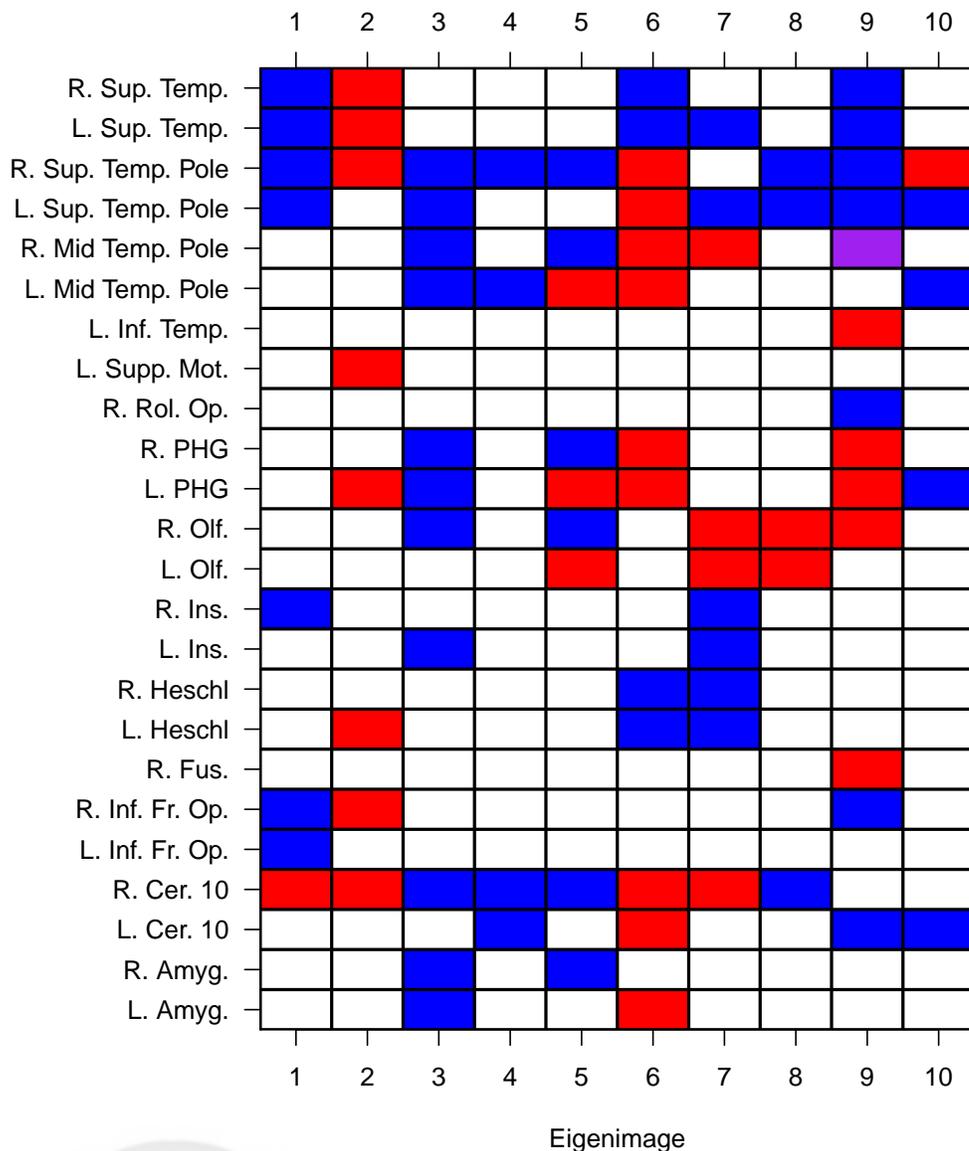


Figure 4: Regions with over 20% overlap with the specified eigenimage. Red areas load positively, blue negatively, purple have partial volumes loading positively and negatively. Abbreviations: Amyg. = Amygdala, Cer. = Cerebellum, Fr. = Frontal, Fus. = Fusiform gyrus, Inf. = Inferior, Ins. = Inusla, L. = Left, Mot. = Motor Area, Olf. = Olfactory, Op. = Opercular part, PHG = Para-Hippocampal Gyrus R. = Right, Rol. = Rolandic, Sup. = Superior, Supp. = Supplementary, Temp. = Temporal .

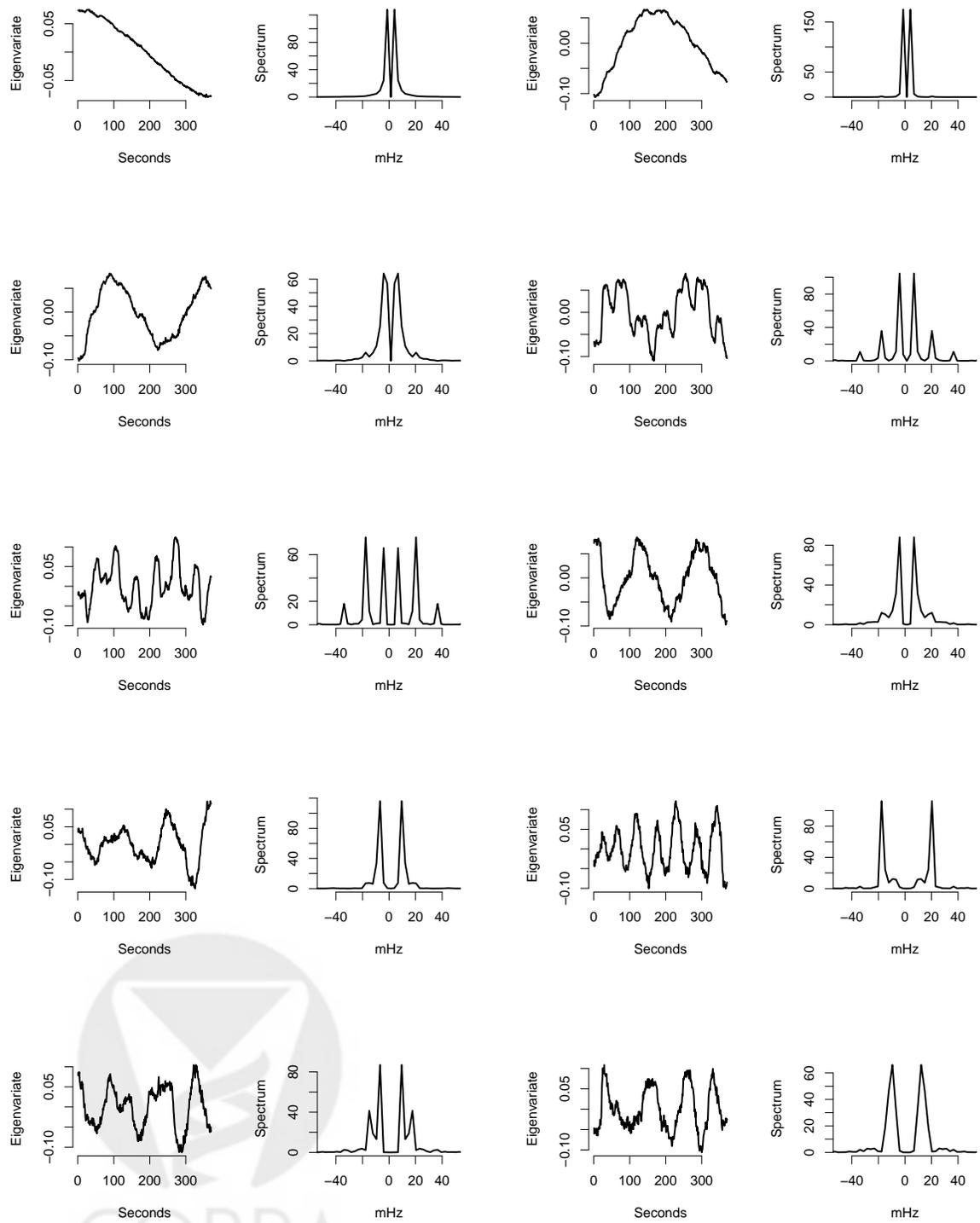


Figure 5: First ten population eigenvariates for the at-risk AD data set. To the right of each plot is the associated spectrum in the -50 to 50 millihertz (mHz) range.

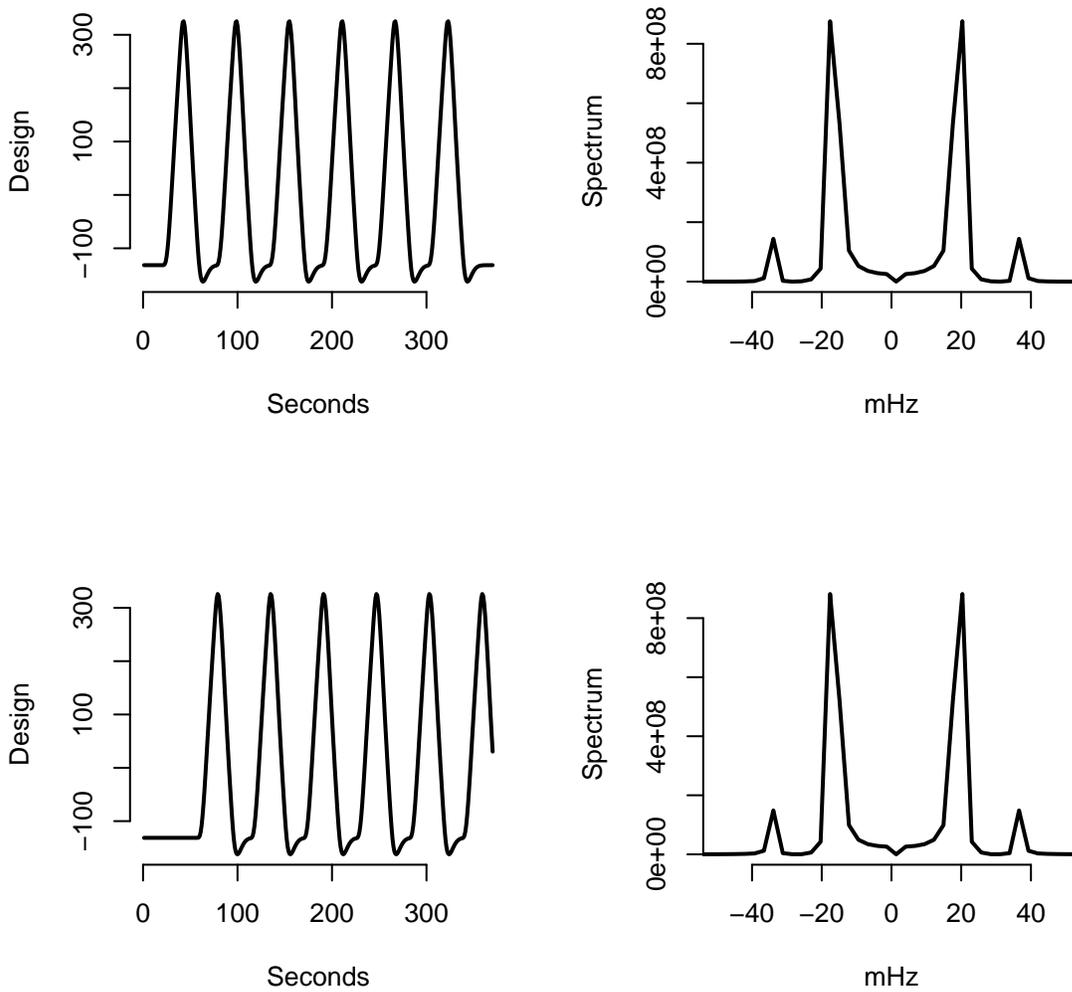
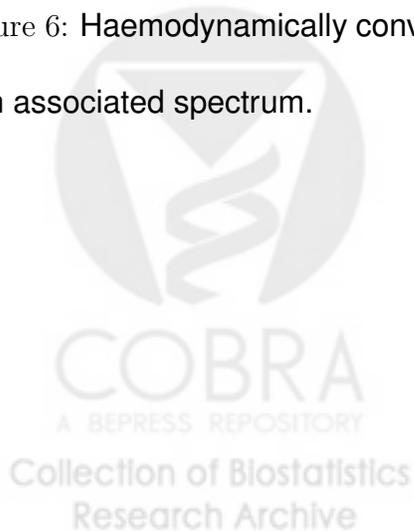


Figure 6: Haemodynamically convolved encoding (top) and recall (bottom) design vectors with associated spectrum.



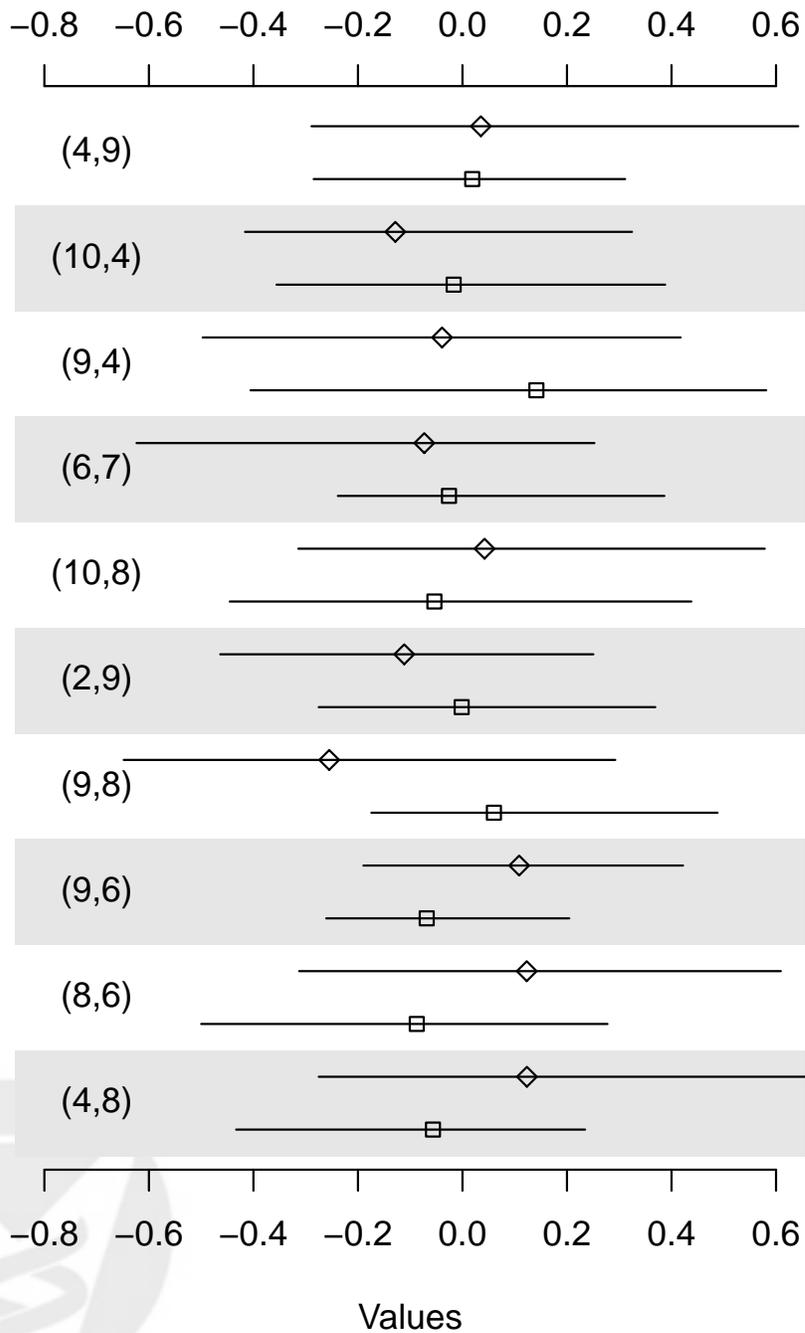


Figure 7: Normalized loadings' 25<sup>th</sup>, median and 75<sup>th</sup> percentiles by group. Diamonds are at-risk, controls are squares. The (eigenimage, eigenvalue) pair are depicted to the left and gray bars are used to highlight groups.

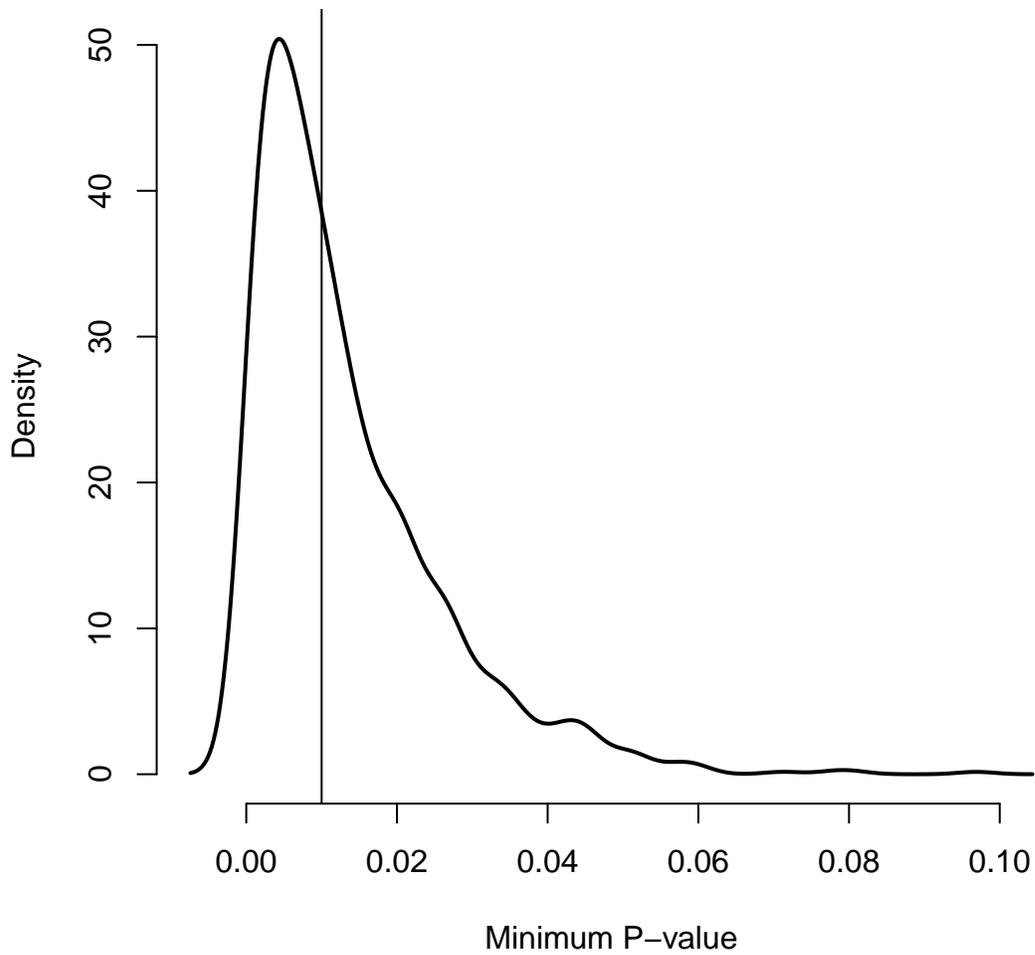


Figure 8: Resampled distribution of minimum P-values with a reference line at 0.01.



## B Tables

	Eigenimages				
	1	2	3	4	5
Med	0.819	0.633	0.625	0.917	0.917
Min	0.990	0.889	0.897	0.985	0.994
Max	1.000	0.998	0.999	0.999	0.999

	Eigenvariates				
	1	2	3	4	5
Med	0.945	0.917	0.914	0.964	0.980
Min	0.996	0.991	0.990	0.992	0.996
Max	1.000	1.000	1.000	1.000	1.000

Table 1: Summary of maximum absolute correlation between true population eigenimages and eigenvariates and the first five estimated eigenimages and variates. Numbers are Median (Range).



		At-risk	Control	P-value
Count		81	68	
Gender	No. Male (%)	33 (41%)	36 (53%)	.19
Age	Mean (SD)	62 (6.68)	62 (7.5)	.90
APOE	No. Any 4 (%)	28* (35%)	12 (18%)	.04
Years of Educ.	No. < 12 (%)	5 (6%)	2 (3%)	
	No. 12 (%)	18 (22%)	12 (18%)	
	No. (12, 16) (%)	16 (20%)	11 (16%)	
	No. 16 (%)	16 (20%)	12 (18%)	
	No. > 16 (%)	26 (32%)	31 (46%)	

.51

Table 2: Demographic data by risk status. \* One at-risk and one control subjects missing APOE status. Age P-values based on two group t-test while remaining were based on Chi-squared tests.



EigIm	Eigenvariate									
	1	2	3	4	5	6	7	8	9	10
1	0.822	0.779	0.264	0.791	0.850	0.235	0.210	0.200	0.379	0.987
2	0.759	0.734	0.792	0.326	0.329	0.699	0.265	0.735	0.076	0.692
3	0.774	0.528	0.361	0.579	0.507	0.189	0.240	0.981	0.512	0.183
4	0.710	0.603	0.900	0.549	0.696	0.166	0.953	0.010	0.153	0.186
5	0.819	0.754	0.767	0.774	0.381	0.417	0.162	0.525	0.512	0.849
6	0.735	0.721	0.716	0.483	0.941	0.303	0.091	0.931	0.715	0.398
7	0.582	0.686	0.706	0.818	0.996	0.314	0.910	0.713	0.560	0.474
8	0.305	0.930	0.165	0.968	0.743	0.050	0.354	0.681	0.262	0.299
9	0.684	0.742	0.675	0.097	0.718	0.052	0.822	0.053	0.348	0.674
10	0.945	0.678	0.529	0.145	0.845	0.574	0.996	0.078	0.158	0.828

Table 3: P-values comparing At-risk and control subject for each loading from functional linear models with a covariate term indicating the presence of any four APOE alleles.

