2014

Multiple Comparison Procedures for Neuroimaging Genomewide Association Studies

Wen-Yu Hua
Thomas E Nichols, *University of Warwick*
Debashis Ghosh, *University of Colorado Denver*

Biostatistics (2012), **0**, 0, *pp.* 1–25
doi:10.1093/biostatistics/biostatistic˙v3

Multiple Comparison Procedures for Neuroimaging Genomewide Association Studies

Wen-Yu Hua[†], Thomas E. Nichols[‡], Debashis Ghosh[†], for the Alzheimer's Disease

Neuroimaging Initiative

[†] *Department of Statistics, Penn State University,*

[‡] *Department of Statistics, University of Warwick*

wxh182/ghoshd@psu.edu

Summary

Recent research in neuroimaging has been focusing on assessing associations between genetic variants measured on a genomewide scale and brain imaging phenotypes. Many publications in the area use massively univariate analyses on a genomewide basis for finding single nucleotide polymorphisms that influence brain structure. In this work, we propose using various dimensionality reduction methods on both brain MRI scans and genomic data, motivated by the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. We also consider a new multiple testing adjustments inspired from the idea of local false discovery rate of Efron *and others* (2001). Our proposed procedure is able to find associations between genes and brain regions at a better significance level than in the initial analyses.

Key words: Genomewide association studies; Distance covariance/correlation; Empirical null estimation; Local false discovery rate; Multiple comparisons; Positive false discovery rate.

1. Introduction

Genomewide association studies (GWAS) involve studying a large number of genetic variants (single nucleotide polymorphisms, or SNPs) to find associations with a disease phenotype (Risch and Merikangas (1996)). This type of study is very important in discovering the genetic features that may indicate the possible future onset of illness such as diabetes, heart abnormalities, Parkinson disease, and hypertension.

A case-control design is the standard way to analyze data from a GWAS. In particular, allele frequencies for each SNP are compared between cases and controls, and significance is evaluated by the corresponding p-value from chi-square test (Balding (2006)). In contrast to the binary disease outcome, quantitative outcomes provide more accurate results as well as increased statistical power for the association study. The common statistical model for analyzing quantitative traits is linear regression, where the model controls the linear relationship between the mean value of the continuous outcomes and the allele genotype, and determines the existence of dependencies upon significant p-values.

With the availability of Magnetic Resonance Imaging (MRI), researchers are able to utilize this new form of digital information in studying the association between biological structures to diseases, and genome to biological structures. In the case of Alzheimer's Disease Neuroimaging Initiative (ADNI) study, (ADNI (2003)), one of the goals is to determine the genetic variants that influence brain structures (Stein *and others* (2010), Shen *and others* (2010), Vounou *and others* (2010), Hibar *and others* (2011)), where brain structures are volumetrically represented as 3D voxels from head MRIs of different subjects. Stein *and others* (2010) conducted a genomewide study using individual voxels from brain scans across all subjects, in which every voxel is evaluated with a regression at each SNP and the corresponding number of minor alleles, and using demographic variables as features with quantitative trait as responses. In their experiment, no significant loci were found after a multiple testing adjustment at level 0.05. In a later study, Stein *and others* (2010) performed a voxelwise search on fewer SNP markers based on the results from Stein *and others* (2010), and they found that 2 novel SNPs, i.e., $rs10845840$ and $rs11055612$ located in an intron of the $GRIN2B$ gene, were significantly associated with bilateral temporal lobe volume.

Instead of analyzing on individual voxels, Shen *and others* (2010) conducted a hierarchical search by considering groups of voxels that represent different regions of the brain. At the highest level, they performed a global voxel-based analysis on the entire brain using Voxel-based morphometry (VBM). After finding the SNPs of interest, they conducted two-way multivariate ANOVAs with main effects of baseline diagnosis and genotype, as well as the interaction effect of baseline diagnosis×genotype for each candidate SNPs in the second level (regions of the brain). Since the mass-univariate linear regression approach in Stein *and others* (2010) did not find any promising variants, Vounou *and others* (2010) proposed a sparse reduced rank regression on the association study and reduced the number of both SNPs and brain scan voxels of the test, resulted in a better power performance. Hibar *and others* (2011) performed a gene-based voxelwise association search, which greatly lowered the number of tests (from 437,607 SNPs down to 18,044 genes). Their analysis provided greater power in detecting associations, which also identified possible gene variants for further study.

For any mass univariate approach to analysis, multiple testing procedures should be employed as there are many statistical tests being considered simultaneously. The goal of multiple testing adjustment is to control the desired Type I error rate (false positive rate) over a set of m tests while attempting to maximize power. The most commonly used approach for multiple testing correction is a Bonferroni correction, which maintains the appropriate familywise error rate (FWER). In some situations, such an FWER-controlling procedure is substantially conservative. An alternative error quantity called false discovery rate (FDR) was proposed for the multiple comparisons problem by Benjamini and Hochberg (1995). In that paper, they also introduced a sequential p-value algorithm to control the FDR. Later, Storey (2002) and Storey (2003) defined the positive false discovery rate (pFDR), which is the conditional expectation of false positive findings given at least one positive identifications has occurred. He also provided a definition of q-value based on the rejection region. Efron *and others* (2001) defined a local false discovery rate (fdr), a Bayesian version of FDR. They fit the inverse cumulative Gaussian distribution of all p-values for estimating local false discovery rate and further controlled the statistical significance. To relate the frequentist and Bayesian versions of FDR, Efron *and others* (2001), Storey (2002), Efron *and others* (2001) and Efron (2004) proved that the FDR controlled by the Benjamini and Hochberg procedure is equivalent to empirical Bayesian FDR given the rejection regions. Furthermore, Newton *and others* (2004) proposed a hierarchical mixture of Gammas for determining local false discovery rate and Muralidharan (2010) showed that the local FDR estimation controls FDR/pFDR over the entire exponential distribution family.

The previously discussed ADNI analyses were able to find associated SNPs or genes that are likely to be related to some specific voxels of the brain scans. However, neighboring structures of the brain were not being considered, and this neighborhood interaction could play an important role in associations with disease risk. In this work, we address this issue by proposing to use distance covariance measure (Szekely *and others* (2007)) in finding the association between disease risk and brain MRI scans using ADNI data. The distance covariance computes an association measure between two random vectors, where the two random vectors are the allowances from arbitrary dimensions. Szekely *and others* (2007) and Szekely and Rizzo (2009) showed that under certain conditions, the asymptotic distribution of distance covariance statistics converges to a quadratic form of centered Gaussian variables.

We make two contributions to the analysis of the ADNI neuroimaging genomewide study. First, we utilize distance covariance for the analysis of genomewide association study. This framework is able to establish the relationships between genomic variants and brain structural imaging, while using the spatial information from neighboring brain regions. By considering a multivariate response variable, we reduce the number of tests being done relative to an approach as in Stein *and others* (2010) that should lead to increased power. Second, we propose a FDR modeling method which is to fit a two-component mixture model on the distance covariance statistics per SNP. One probabilistic output of this model is the local false discovery rate. This leads to a decision-theoretic rule for selecting significant SNPs that is related to the approach of Newton *and others* (2004). Based on our simulation and real data analysis using ADNI, our proposed method is able to control FDR even in situations where the data generating model does not match the mixture model being fit to the data. The structure of this paper is as follows. Section 2 first reviews the details of ADNI data and then introduces our method in formulating distance covariance. We present the results based on applying our method to both simulated and ADNI data in section 3. Finally, the applicabilities of our results are discussed in section 4.

2. Materials and methods

2.1 *ADNI study*

Alzheimer's disease (AD) is an illness that progresses overtime, and the brain gradually regresses in affected individuals which causes continuous worsening in reasoning, memory, and language. Most people with AD start to notice symptoms after the age of 60, and AD is irreversible and has has no curative treatment so far (ADNI (2003)).

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a $60 million, 5 year study that was initiated in 2003. It was conducted by the National Institute on Aging in conjunction with other federal agencies and private-sector corporations and organizations. The aims of the study include the following: (1). develop a set of standardized methods for collecting imaging and biomarker data; (2). collect and organize all related data into a common database; (3). identify brain and other biological changes associated with memory decline; and (4). use ADNI methods in clinical trials for slowing the progression or even preventing this devastating disease, (Weiner (2006)). The study recruited approximately 200 cognitively normal older people, 400 people with mild cognitive impairment, and 200 people with early Alzheimer's Disease across the U.S. and Canada (ADNI (2003)). To measure the degree of progression, the study provided magnetic resonance imaging (MRI) brain scans (both 1.5T and 3T scanning protocol), and Positron Emission Tomography (PET) to measure brain glucose metabolism. ADNI profiled 620,901 SNPs using Illumina 610 Quad array. The study also provided the clinical information of each participant including demographics, physical examinations, cognitive assessments, biospecimen collection, medications, diagnostic summary and lumbar puncture (ADNI (2003)). The ADNI dataset released 852 subjects, but we discarded 111 of the subjects that did not seem to be correctly registered against the reference scan (Newton *and others* (2004)). This results in a total of 741 subjects, with 206 normal older controls (NC), 358 with mild cognitive impairment (MCI), and 117 Alzheimer's disease subjects (AD). This matches with the data analyzed in Stein *and others* (2010).

With respect to the SNPs, only a subset of the original 620,901 was used. SNPs that did not fulfil the following quality control criteria were excluded: genotype call rate smaller than 95%, significant deviation from Hardy-Weinberg equilibrium where p-values $< 5.7 \times 10^{-7}$, allele frequency smaller than 0.10, and a quality control score of smaller than 0.15, (Stein *and others* (2010)). After applying this list of quality criteria, we obtain a total of 448,244 SNPs for the analysis. The number of SNPs measured on each chromosome is in Table 1.

2.2 *Brain MRI scans*

The 3D T1-weighted baseline brain MRI scans were analyzed using tensor-based morphometry (TBM) Stein *and others* (2010) and Hua *and others* (2008). The MRI scans were acquired at 58 different ADNI sites, all with 1.5T MRI scanners using a sagittal 3D MP-RAGE sequence for

Chromosome	Number of SNPs	Chromosome	Number of SNPs
1	33,850	13	16,605
2	36,384	14	14,501
3	30,765	15	13,355
4	27,072	16	13,460
5	27,396	17	11,721
6	30,054	18	13,198
7	24,446	19	7,895
8	24,768	20	11,169
9	21,283	21	6,582
10	23,089	22	6,757
11	21,796	23	10,637
12	21,461	*Total*	448,244

Table 1. Number of SNPs distributed on each chromosome

across-site consistency (Jack *and others* (2008)), and all images were calibrated with phantom-based geometric corrections. The scans were linearly registered with 9 parameters to the International Consortium for Brain Image template, (Mazziotta *and others* (2001)), to adjust for differences in brain position and scaling. Each subject's MRI scan was registered against a template scan which is averaged from the healthy subjects (minimal deformation template), using a non-linear inverse-consistent elastic intensity-based registration method (Leow *and others* (2005)). Furthermore, voxel size variation from registration is shown as the voxel intensity, in which the voxel intensity represents the volumetric difference between the subject and the reference template, calculated from taking the determinant of the Jacobian matrix of the deformation. Finally, each brain scan volume is down-sampled to 1/4 of its original size (using trilinear interpolation to $4 \times 4 \times 4\ mm^3$, which results into $31,622$ total voxels per scan for faster experimental processing. Stein *and others* (2010) pointed out that this volumetric difference representation of MRI scans is used as a quantitative measure of brain tissue volume difference and genomewide association.

Our analyses studied genomewide associations with brain activity using regions of interest (ROIs). In order to conduct the regionwise experiment, we extracted corresponding voxels, and

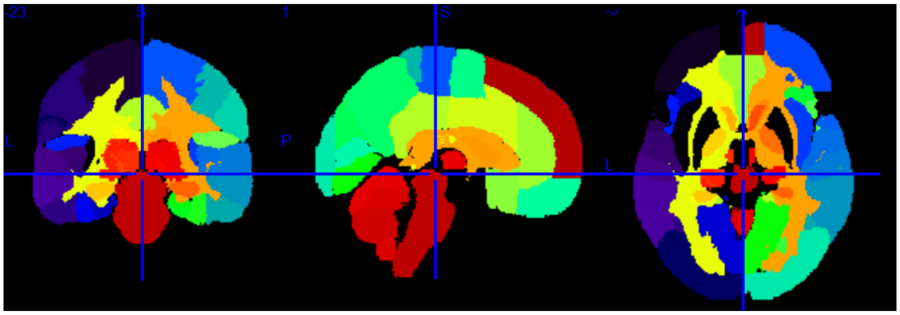

Fig. 1. left to right: Coronal, Sagittal and Axial views of GSK CIC Atlas, color coded by the 119 different regions.

computed the average Jacobian scores that make up the 119 different brain regions from the GSK CIC Atlas as shown in Fig. 1 (Tziortzi *and others* (2011)), which is based on the Harvard-Oxford atlas with a 6-level hierarchy. To extract the corresponding voxels for each of the brain regions in the atlas, we used FSL's FLIRT linear registration tool (Jenkinson and Smith (2001), Jenkinson *and others* (2002), Smith *and others* (2004), and Woolrich *and others* (2009)) to register the brain atlas to our template scan as described previously, which allows us to find voxels of different brain regions from both the subject's scan and the registered atlas by direct comparison. We then applied the 119 brain regions as the response into genomewide association and denoted it as the regionwide genomewide association study.

2.3 *Distance covariance*

We now define the distance covariance statistic that was proposed by Szekely *and others* (2007). Let ϕ_X and ϕ_Y be the characteristic functions of X and Y, where $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ are two random vectors from two arbitrary dimensions p and q, respectively. The distance covariance $\mathrm{dCov}^2(X, Y)$ between random vectors X and Y is a non-negative value with finite first moments:

$$\begin{aligned} \mathrm{dCov}^2(X,Y) &= \|\phi_{X,Y}(x,y) - \phi_X(x)\phi_Y(y)\|^2 \\ &= \int_{\mathbb{R}^{p+q}} |\phi_{X,Y}(x,y) - \phi_X(x)\phi_Y(y)|^2 w(x,y)\,\mathrm{d}x\mathrm{d}y, \end{aligned} \tag{2.1}$$

where $w(x,y)$ is a positive weight function for which the integral in Eq. 2.1 exists. Similarly, distance variance is defined as the square root of $\mathrm{dCov}^2(X,X)$. Also, the distance correlation $\mathrm{dCorr}^2(X,Y)$ between random vectors X and Y with finite first moments is given by:

$$\mathrm{dCorr}^2(X,Y) = \frac{\mathrm{dCov}^2(X,Y)}{\sqrt{\mathrm{dCov}^2(X,X)\mathrm{dCov}^2(Y,Y)}}, \qquad \mathrm{dCov}^2(X,X)\mathrm{dCov}^2(Y,Y) > 0, \tag{2.2}$$

and $\mathrm{dCorr}^2(X,Y) = 0$ if $\mathrm{dCov}^2(X,X)\mathrm{dCov}^2(Y,Y) = 0$.

The sample distance covariance and correlation statistics are given by taking two Euclidean distance matrices $a_{ij} = |X_i - X_j|_p$ and $b_{ij} = |Y_i - Y_j|_q$, (Szekely *and others* (2007) and Szekely and Rizzo (2009)) as:

$$A_{ij} = a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}, \qquad i,j = 1,...,n \tag{2.3}$$

where

$$a_{ij} = |X_i - X_j|_p, \quad \bar{a}_{i.} = \frac{1}{n}\sum_{j}^{n} a_{ij} \quad \bar{a}_{.j} = \frac{1}{n}\sum_{i}^{n} a_{ij}, \text{ and } \quad \bar{a}_{..} = \frac{1}{n^2}\sum_{i,j}^{n} a_{ij}. \tag{2.4}$$

Similarly, $b_{ij} = |Y_i - Y_j|_p$ and $B_{ij} = b_{ij} - \bar{b}_{i.} - \bar{b}_{.j} + \bar{b}_{..}$ for $i,j = 1,...,n$.. Therefore, the sample distance covariance dCov_n^2 is:

$$\mathrm{dCov}_n^2(X,Y) = \frac{1}{n^2}\sum_{i,j}^{n} A_{ij}B_{ij} \tag{2.5}$$

and the sample distance correlation $\mathrm{dCorr}_n^2(X,Y)$ is defined by:

$$\mathrm{dCorr}_n^2(X,Y) = \begin{cases} \frac{\mathrm{dCov}_n^2(X,Y)}{\sqrt{\mathrm{dCov}_n^2(X,X)\mathrm{dCov}_n^2(Y,Y)}} & ,\mathrm{dCov}_n^2(X,X)\mathrm{dcov}_n^2(Y,Y) > 0 \\ 0 & ,\mathrm{dCov}_n^2(X,X)\mathrm{dCov}_n^2(Y,Y) = 0 \end{cases} \tag{2.6}$$

In addition, the sample distance covariance $\mathrm{dCov}_n(X,Y)$ is defined in Eq. 2.5 is equivalent to Eq. 2.7 that takes the following form:

$$\mathrm{dCov}_n^2(X,Y) = T_1 + T_2 - 2T_3 \tag{2.7}$$

where

$$T_1 = \frac{1}{n^2}\sum_{i,j}^{n} a_{ij}b_{ij} \qquad T_2 = \frac{1}{n^2}\sum_{i,j}^{n} a_{ij}\frac{1}{n^2}\Sigma_{i,j}^{n} b_{ij} \qquad T_3 = \frac{1}{n^3}\sum_{i}^{n}\sum_{j,k}^{n} a_{ij}b_{ik}. \tag{2.8}$$

The sample distance covariance estimators (Eq. 2.3 through Eq. 2.7) from Szekely *and others* (2007) and Szekely and Rizzo (2009) require that there be no missing values among observations $X_i's$ and $Y_j's$ for $i,j = 1,...,n$. In the following, we propose a modified version which relaxes the assumption for non-missing values of sample distance covariance. We define δ as an indicator which indicates if a variable is missing or present:

$$\delta_k = \begin{cases} 1, & \text{if variable } k \text{ is presence} \\ 0, & \text{if variable } k \text{ is missing} \end{cases}. \tag{2.9}$$

By introducing the indicator δ for observations $X_i's$ and $Y_j's$ according to Eq. 2.9, it basically reweight the samples, and put the larger weights on observations with no missing values and smaller weights on observations with missing values. Therefore, the modified sample distance dependence statistics are defined as:

$$A'_{ij} = a'_{ij} - \bar{a}'_{i.} - \bar{a}'_{.j} + \bar{a}'_{..}; \tag{2.10}$$

where

$$a'_{ij} = \frac{|X_i - X_j|_p \delta_i \delta_j}{P(\delta_i = 1)P(\delta_j = 1)}, \qquad \bar{a}'_{i.} = \frac{1}{n}\sum_{j}^{n} a'_{ij} \qquad \bar{a}'_{.j} = \frac{1}{n}\sum_{i}^{n} a'_{ij}, \text{ and } \qquad \bar{a}'_{..} = \frac{1}{n^2}\sum_{i,j}^{n} a'_{ij}.$$

Similarly, we define

$$b'_{ij} = \frac{|Y_i - Y_j|_q \delta_i \delta_j}{P(\delta_i = 1)P(\delta_j = 1)}$$

and $B'_{ij} = b'_{ij} - \bar{b}'_{i.} - \bar{b}'_{.j} + \bar{b}'_{..}$ for $i,j = 1,...,n$.

The modified sample distance covariance $\widetilde{\mathrm{dCov}}_n^2(X,Y)$ is then given by $\widetilde{\mathrm{dCov}}_n^2(X,Y) = n^{-2}\sum_{i,j}^n A'_{ij}B'_{ij}$. Correspondingly, the sample distance correlation $\widetilde{\mathrm{dCorr}}_n^2(X,Y)$ is then modified as:

$$\widetilde{\mathrm{dCorr}}_n^2(X,Y) = \begin{cases} \frac{\widetilde{\mathrm{dCov}}_n^2(X,Y)}{\sqrt{\widetilde{\mathrm{dCov}}_n^2(X,X)\widetilde{\mathrm{dCov}}_n^2(Y,Y)}} & ,\widetilde{\mathrm{dCov}}_n^2(X,X)\widetilde{\mathrm{dCov}}_n^2(Y,Y) > 0 \\ 0 & ,\widetilde{\mathrm{dCov}}_n^2(X,X)\widetilde{\mathrm{dCov}}_n^2(Y,Y) = 0 \end{cases} \tag{2.11}$$

There are two properties of distance covariance that motivate us to utilize it for testing for associations between SNPs and brain images. The first property is the independent property. Szekely *and others* (2007) and Szekely and Rizzo (2009) prove that $\mathrm{dCorr}(X,Y) = 0$ if and only if X and Y are independent. This property can identify non-linear or non-monotone dependencies between X and Y (Szekely and Rizzo (2009)).

Having proposed a modified empirical distance covariance/correlation for situations when there are missing values, we now wish to study its asymptotic properties. The expectation of a'_{ij} in Eq. 2.10 is:

$$\begin{aligned} E(a'_{ij}) &= E\{\frac{|X_i - X_j|_p\delta_i\delta_j}{p(\delta_i = 1)p(\delta_j = 1)}\} \\ &= E(|X_i - X_j|_p) \\ &= E(a_{ij}) \end{aligned} \tag{2.12}$$

and similarly, $E(b'_{ij}) = E(b_{ij})$. Arguing as in Szekely *and others* (2007), we have that if $E|X|_p < \infty$ and $E|Y|_q < \infty$, then $\widetilde{\mathrm{dCov}}_n \rightarrow_{a.s} \mathrm{dCov}$. Similarly, it can be shown that $n \times \widetilde{\mathrm{dCov}}_n^2/T'_2$ converges in distribution to Q, a positive semidefinite quadratic form of centered Gaussian random variables with $E(Q) = 1$. This is a computationally intractable distribution, and in their work, Szekely *and others* (2007) proposed a permutation test as an alternative approach to inference for the distance covariance statistic.

Finally, what inspired us with the use of Gamma mixture models is from work done by Szekely and Bakirov (2003), in which they proved that for any $x > 0$:

$$\inf_{Q} P(Q \leqslant x) = P(\frac{\chi_n^2}{n} \leqslant x) \tag{2.13}$$

$$\tag{2.14}$$

where χ_n^2 is a chi-square random variable with degree of freedom $n(x)$. This asymptotic property gave us the idea to fit the mixture of Gammas on the statistics of distance covariance for False Discovery Rate to address the issue of multiple comparisons.

2.4 *Multiple testing procedure*

In this section, we review the multiple testing problem and define FDR. To begin this discussion, suppose there are m tests for the genomewide study, and the goal is to identify significant SNP variants given a certain type 1 error. Table 2 shows the possible outcomes of conducting m tests simultaneously, and the null hypothesis is true for m_0 of them. This hypothesis indicates that the SNP is not significantly associated with the decrease in brain volume. There is a total of W hypotheses that are failed to be rejected, and R rejected null hypothesis. Benjamini and Hochberg (1995) introduced a new measure called the False Discovery Rate (FDR), defined as:

	Accept null hypothesis	Reject null hypothesis	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

Table 2. m hypotheses

$$FDR = \begin{cases} E(\frac{V}{R}); & R > 0 \\ 0; & R = 0 \end{cases} \tag{2.15}$$

$$= E\left[\frac{V}{R}|R > 0\right] P(R > 0)$$

Storey (2002) and Storey (2003) proposed another measure, positive FDR, which is the expected false-positive rate conditionally on positive finding ($P(R = 0) > 0$). The positive FDR, pFDR, takes the following form:

$$p\text{FDR} = E[\frac{V}{R}|R > 0] \tag{2.16}$$

Furthermore, Storey (2002) and Storey (2003) provided a Bayesian interpretation for pFDR, and it is related to the local false discovery rate of Efron *and others* (2001).

We propose using a new approach based on distance covariance measure mentioned in Section 2.3 for multiple testing adjustment. The traditional multiple correction methods are based on p-values, while our proposed FDR controlling method models the test statistics directly. This is in the vein of what is done in Efron *and others* (2001). Assume that we have the statistics $T_1, ..., T_m$ for testing m hypotheses and each statistics T_i has a corresponding unknown indicator variable H_i, where H_i is defined as:

$$H_i = \begin{cases} 0, & \text{if the null hypothesis is true} \\ 1, & \text{if the alternative hypothesis is true.} \end{cases} \tag{2.17}$$

Therefore, the likelihood distribution of observed testing statistics can be represented as the conditional distribution which is conditioned on $H_i = 0$ and $H_i = 1$, such that $T_i|H_i = 0$ follows f_0 and $T_i|H_i = 1$ follows f_1. With Eq. 2.17, we then model $H_1, ..., H_m$ as our prior which is a random sample from a Bernoulli distribution with $P(H_i = 0) = p_0$; $i = 1, ..., m$, and $P(H_i = 1) = 1 - p_0 = p_1$. So, we can set up the mixture of model and use EM to estimate the model parameters. Then we formulate the posterior distribution of $H_i = 0$ and $H_i = 1$ into $P(H_i = 0|T)$ and $P(H_i = 1|T)$ and estimate the local false discovery rate (fdr) by the following equation as the definition in Efron *and others* (2001).

$$\begin{aligned} \widehat{fdr}(t) &= \frac{\hat{p}_0 \hat{f}_0(t)}{\hat{f}(t)} \\ &= P(H = 0|T = t), \end{aligned} \tag{2.18}$$

where $\hat{f}(t) = \hat{p}_0 \hat{f}_0(t) + \hat{p}_1 \hat{f}_1(t)$. On the other hand, we can also estimate pFDR from the definition of Storey (2002) which is the posterior distribution of $H = 0$ given $T \in \Gamma$:

$$\begin{aligned}\widehat{p\text{FDR}} &= \hat{p}_0\hat{F}_0(t \in \Gamma)/\hat{p}_0\hat{F}_0(t \in \Gamma) + \hat{p}_1\hat{F}_1(t \in \Gamma) \qquad (2.19)\\ &= P(H = 0|T \in \Gamma)\end{aligned}$$

where Γ is the rejection region and $\hat{F}_0$ and $\hat{F}$ are the cumulative distribution function of $\hat{f}_0$ and $\hat{f}$, respectively. Furthermore, Efron *and others* (2001), Efron *and others* (2001), and Efron (2004) give the following theorem:

Theorem 1 Average theorem: pFDR is the conditional expectation of $fdr(t)$ given $t \in \Gamma$.

Proof.

$$\begin{aligned}E(fdr(t)|t \in \Gamma) &= \frac{\int_{t\in\Gamma} fdr(t)f(t)\,\mathrm{d}t}{\int_{t\in\Gamma} f(t)\,\mathrm{d}t}\\ &= \frac{p_0P(T \in \Gamma|H = 0)}{p_0F_0(t \in \Gamma) + p_1F_1(t \in \Gamma)}\\ &= p\text{FDR}\end{aligned}$$

□

The threshold $\tilde{t}$ can be obtained by solving $\tilde{t} = \arg\max_t\{\widehat{p\text{FDR}}(t) \leqslant \alpha\}$, and rejecting all tests with $t_i \geqslant \tilde{t}$ which gives an expected rate of false discoveries no greater than α. Notice that if we replace the denominator by the empirical cdf of t_i's $\#\{t_i \in \Gamma\}/n$, then the estimator $\widehat{\text{FDR}}(t) = \hat{p}_0\widehat{F}_0(t \in \Gamma)/\#\{t \in \Gamma\}$ is equivalent to Benjamini and Hochberg's algorithm for controlling the FDR given the rejection region Γ. The following algorithm details the step by step procedure of our FDR modeling method.

Algorithm: FDR modeling:

Input: m hypotheses with statistics $T_1, ..., T_m$.

1. Fit a 2 component mixture model to $T_1, ..., T_m$.

2. Estimate the parameters using EM algorithm.

3. Compute the local fdr (fdr) as defined in Efron *and others* (2001) by $\widehat{fdr}(t) = P(H = 0|T = t) = \hat{p}_0\hat{f}_0(t)/\hat{f}(t)$, where $\hat{f}(t) = \hat{p}_0\hat{f}_0(t) + \hat{p}_1\hat{f}_1(t)$, and $\hat{p}_1 = 1 - \hat{p}_0$.

4. $\widehat{p\text{FDR}}$(t) is the conditional expectation of $\widehat{fdr}(t)$ given $t \in \Gamma$ based on theorem 1 (the average theorem in Efron *and others* (2001)).

5. Define $\tilde{t} = \arg\max_t\{\widehat{p\text{FDR}}(t) \leqslant \alpha\}$, and reject all tests with $t_i \geqslant \tilde{t}$. In theory, this gives an pFDR no greater than α.

We note in passing that this rule is similar to one proposed by Newton *and others* (2004) in a different genomics setting.

3. Simulation study and real data analysis

We implemented and analyzed our numerical experiments in R and Matlab, where we used the MIXTOOLS package from R for mixture of Gammas. All the analyses were accomplished by using the Penn State University computing cluster, which consists of 128 Intel Xeon E5450 nodes, each with 8 cores and 32 GB of memory.

3.1 *Simulation Design*

In this section, we generated the simulated data to assess the performance of our proposed method described in section 2.3 and 2.4. Our designed the simulation procedure is based on the structure of real data (ANDI dataset), and we considered 2 cases as followed: in the first case, we assumed the data are from two populations i.e., highly positively correlated population, and low to no correlation populations. From the high correlated population, we generated feature pairs (SNPs vs. voxels) that each consists of a SNP vector of 100 samples by 1 SNP, and a corresponding voxel feature matrix that is 100 samples by 30 voxels. For this highly correlated population feature pair, half of the voxel samples were generated with linear, exponential, and quadratic transformation from their corresponding SNP samples, while the remaining half of SNP and voxel sample pairs were generated as single standard Normal and multivariate Normal random variables (mean 0 and identity covariance matrix), respectively. We separately generated several 100 sample SNP-and-voxel pairs from low to no correlated populations by keeping just 0% to 30% of the 100 sample pairs linearly correlated with the remaining portion generated by single and multivariate Normal random variables. For a single run of this simulation, 100 feature pairs were from the highly positively correlated population, 900 feature pairs were from low to no correlated populations, and they were generated such that it totals to 1000 pairs per run. In addition, we ran the second case that is similar to the first one, but differed only in replacing the identity covariance matrix from multivariate Normal with a covariance matrix that had its diagonal terms equal to 1, and the off diagonal terms equal to 0.5.

We computed the distance covariance statistics for each SNP and voxels pair separately per simulation setting. After collecting the resulting distance covariance statistics over both simulations, we performed multiple testing adjustment using our FDR modeling algorithm by repeating the above procedures 1000 times per setting.

3.2 *Simulation Results*

In figure 2, the histogram illustrates the empirical null fit ($\hat{f}_0$) in the red curve and the empirical alternative fit ($\hat{f}_1$) using the green curve. From the middle plot of figure 2, it shows that although the local fdr estimates of the Skew normal distribution is slightly off from Normal and Gamma distributions, their overall patterns are similar since the local fdr estimates of Normal and Gamma are fitted well onto each other. This implies that our FDR modeling method offers robust estimates of different empirical null distributions. In the following analysis, we chose mixture of Gammas for the density fits $\hat{f}_0$ and $\hat{f}_1$, since the asymptotic distribution of the distance covariance measure is a quadratic form of centered Gaussian random variable, and the quadratic centered Gaussian is Chi-squared. Therefore, we fitted Gamma distribution on f_0 which is the parent family of Chi-square distribution, Gretton *and others* (2008). The right plot in figure 2 illustrates the relationship between the estimates of local fdr and pFDR, based on mixture of Gammas fits that are from the dataset of a single run. We can see that the estimated local fdr is small, the estimates of pFDR is close to zero, while the estimates of both local fdr and pFDR approaches 1.0. Also, the curves between the estimated local fdr and estimated pFDR are one-to-one, and this pattern supports the average theorem from Efron *and others* (2001), that is pFDR is the conditional expectation of $fdr(t)$ given t in the tail area.

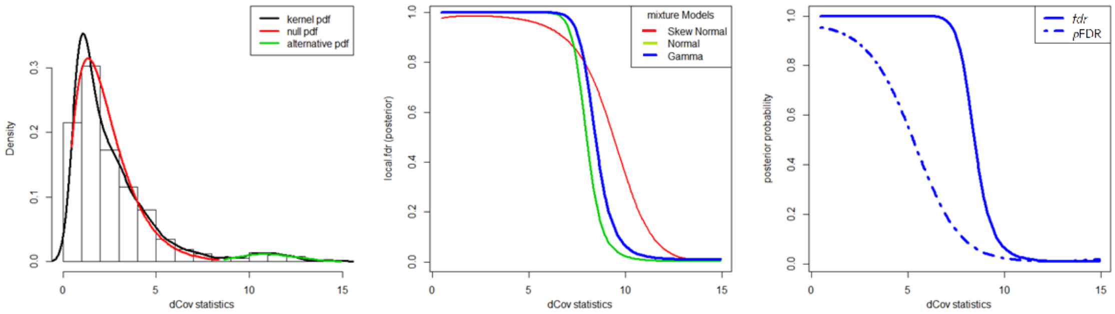

Fig. 2. Figure 2 shows the histogram on the left, 3 mixture of models local fdr estimates in the middle, and the local fdr v.s. pFDR estimates on the right. The histogram in figure 2 presents the density fits of the statistics. The black line represents the overall density fits using kernel density estimation, the red line is the estimates of the null density $\hat{f}_0$, and the green line represents the density fit of $\hat{f}_1$. The middle figure shows the local fdr estimates using 3 mixture of models, Skew Normal, Normal and Gamma, based on the posterior probability of $P(H = 0|T)$. The right plot in figure 2 illustrates the relationship between estimates of the local fdr and pFDR, based on mixture of Gammas fits.

On the other hand, from our simulation design, we can obtain the true class labels of each run and compute the final predicted class labels from the association estimates in section 2.3 and the multiple testing correction in section 2.4.

Table 3 shows the average estimated FDR, the average estimated power, and their standard deviation based on 1000 runs at nominal type 1 levels 0.025, 0.050, 0.075, 0.1, 0.15 and 0.2 for both cases in section 3.1. The results show that the average estimated FDR are all less than nominal type 1 errors and the average estimated power are all higher than 70%. This implies our proposed procedure controlled FDR at each nominal type 1 level while offered powerful findings of true alternative samples. In addition, the average estimated power of the second case is close to the power of the first case at each level. This shows the results of our proposed method are slightly affected by the noise of the dependent covariance structure between voxel features but the overall performances are quite robust.

type 1 error	FDR ($\times 10^3$) (std)	Power (std)
0.025	1.6520 (0.0049)	0.7766 (0.0659)
0.050	5.2320 (0.0082)	0.8967 (0.0471)
0.075	9.7740 (0.0124)	0.9397 (0.0347)
0.100	14.3980 (0.0120)	0.9603 (0.0269)
0.150	25.4670 (0.0216)	0.9793 (0.0135)
0.200	35.4220 (0.0209)	0.9853 (0.0079)
0.025	0.9630 (0.0030)	0.7252 (0.0729)
0.050	4.2210 (0.0072)	0.8654 (0.0557)
0.075	7.3910 (0.0077)	0.9185 (0.0420)
0.100	12.4860 (0.0136)	0.9465 (0.0315)
0.150	20.6700 (0.0136)	0.9716 (0.0195)
0.200	32.6180 (0.0204)	0.9813 (0.0121)

Table 3. Upper table and lower table correspond the first and the second case in section 3.1 : the average estimated FDR (std) and the average estimated power (std) for nominal type 1 errors at the 0.025, 0.050, 0.075, 0.1, 0.15 and 0.2 levels over 1000 runs for the method described in section 2.3 and section 2.4

3.3 *Application to ADNI data sets*

The genomewide association was conducted using the ADNI dataset. For each test, the independent variable is a single SNP across the whole genome (448,244 SNPs), and the responses are the entire brain imaging voxels (31,622 voxels) or the 119 brain regions based on the GSK CIC Atlas, (Tziortzi *and others* (2011)).

The number of significant SNPs are shown in table 4. The results in the first row of table 4 is from Stein *and others* (2010), where they modeled the association between one single voxel and

Method	Multiple testing method	# SNPs selected
Single voxel and single SNP	q-value (0.50)	8,212
*All voxels and single SNP	FDR modeling (0.05)	20,635
*All regions and single SNP	FDR modeling (0.05)	23,128

Table 4. Comparison of ADNI study results. The first line is from Stein *and others* (2010), while the second and third line represents the proposed approaches in the paper.

one single SNP for each test, and applied the q-value method for multiple correction at the type 1 error 0.5 for all the tests that found 8,212 significant SNPs. Comparing to our work, 2 analyses are presented here marked by * in table 4, the dependency between all voxels v.s. single SNP, and all regions v.s. single SNPs, are measured by distance covariance. The FDR modeling method is then implemented for both analyses by controlling the type 1 error at 0.05, and the significant SNPs of 20,635 and 23,128 are found for the whole image genomewide association study and the region-wide genomewide association study, respectively. Therefore, we can conclude that our FDR modeling method provide more powerful findings since the significant SNPs we found were more than 20,000 for both analyses, and this number is much greater than 8,212 reported in Stein *and others* (2010) using the ADNI dataset.

We also performed the analysis to find the biological terms that are more likely to be associated in our significant SNP lists. The functional annotation clustering analysis was done using DAVID v6.7 DAVID (2003). Table 5 listed the top 10 clusters for the whole image genomewide association study and the regionwide genomewide association study. Notice that the enriched terms that we obtained from the annotation clustering analysis, their adjusted p-values using the Benjamini and Hochberg (1995) procedures are all less than 0.05, which indicates the association between the enriched terms and the SNPs are significant. This provided promising results that these enriched terms may point out the important pathways to Alzheimer's disease.

Top functions	Enrichment scores
Insoluble fraction, membrane fraction, and cell fraction	6.30
ABC transporter-like, transmembrane region, conserved site ABC transporter integral membrane type 1	5.94
Regulation of apoptosis, programmed cell death, and cell death	5.48
Vesicle, cytoplasmic membrane-bounded vesicle, cytoplasmic vesicle, and membrane-bounded vesicle	4.85
ATPase activity, coupled to transmembrane movement of substances, hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances, P-P-bond-hydrolysis-driven transmembrane transporter activity, and primary active transmembrane transporter activity	4.63
Cellular homeostasis, cellular ion homeostasis, ion homeostasis, and cellular chemical homeostasis	4.39
Positive regulation of apoptosis, programmed cell death, and cell death	5.42
ABC transmembrane type-1 2, and type-1 1	4.28
Coagulation, blood coagulation, hemostasis wound healing, and regulation of body fluid levels	4.19
Positive regulation of biosynthetic process, cellular biosynthetic process, and nitrogen compound metabolic process	4.04

Table 5. Top enriched terms of the results from the regionwide genomewide study

4. Discussion and Conclusion

In this work, we used the distance covariance measure for the genomewide association study. The is able to identify the dependency between the SNP variants and the brain volume decline, and use the neighborhood information from the brain regions in the same time. We also proposed a novel FDR modeling strategy. From both the simulation study and real data analysis, our procedures, distance covariance plus FDR modeling, are able to control FDR at the proper level and provided the powerful results relative to previous analyses of the ADNI data.

However, for both of the generalizations, there remain many open questions that could lead to important further developments. We utilized the distance covariance to measure the relationship in the first stage, this representation can be applied to capture the non-linear dependency between 2 variable vectors with arbitrary dimensions, but it also suffered a substantial bias when the number of dimensionality is much greater then the sample size (Cope (2009)). We plan to study regularization approaches to the dependency measure to reduce this bias in future work. On this other hand, we adjusted the multiple comparisons in the second stage, 2 components mixture of models were considered for modeling the empirical null and alternative distribution, but the number of components affects the performance of type 1 error and power in our FDR modeling method. Therefore, we are planing to propose an adaptive FDR modeling method which combines the model selection techniques and FDR modeling method for multiple testing procedure in the future.

5. Supplementary Material

Supplementary material is available online at http://www.adni-info.org/.

References

ADNI. (2003). Alzheimer!—s disease neuroimaging initiative. http://www.loni.ucla.edu/ADNI/.

Balding, David J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7**, 781–791.

Benjamini, Yoav and Hochberg, Yosef. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* **57**, 289–300.

Cope, Leslie. (2009). Discussion of: Brownian distance covariance. *The Annals of Applied Statistics* **3**, 1279–1281.

DAVID. (2003). The database for annotation, visualization and integrated discovery ((david)). http://david.abcc.ncifcrf.gov/.

Efron, Bradley. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association* **99**, 96–104.

Efron, Bradley, Storey, John D. and Tibshirani, Robert. (2001*a*). Microarrays, empirical bayes methods, and false discovery rates. *Genetic Epidemiology* **23**, 70–86.

Efron, Bradley, Tibshirani, Robert, Storey, John D. and Tusher, Virginia. (2001*b*). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

Gretton, A., Fukumizu, K., Teo, C.H., Song, L., Schölkopf, B. and Smola, A. J. (2008). A kernel statistical test of independence. In: Platt, J.C., Koller, D., Singer, Y. and Roweis, S. (editors), *Advances in Neural Information Processing Systems 20*. Cambridge, MA: MIT Press.

Hibar, Derrek P., Stein, Jason L., Kohannim, Omid, Jahanshad, Neda, Saykin, Andrew J., Shen, Li and et al. (2011). Voxelwise gene-wide association study (vgenewas): Multivariate gene-based association testing in 731 elderly subjects. *Neuroimage* **56**, 1875–1891.

Hua, Xue, Leow, Alex D., Parikshak, Neelroop, Lee, Suh, Chiang, Ming-Chang, Toga, Arthur W. and et al. (2008). Tensor-based morphometry as a neuroimaging biomarker for alzheimer's disease: An mri study of 676 ad, mci, and normal subjects. *Neuroimage* **43**, 458–469.

Jack, Clifford R., Bernstein, Matt A., Fox, Nick C., Thompson, Paul, Alexander, Gene, Harvey, Danielle and et al. (2008). The alzheimer's disease neuroimaging initiative (adni): Mri methods. *J Magn Reson Imaging* **7**, 685!V691.

Jenkinson, Mark, Bannister, Peter, Brady, Michael and Smith, Stephen. (2002).

Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841.

Jenkinson, Mark and Smith, Stephen M. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis* **5**, 143–156.

Leow, Alex D., Huang, Sung-Cheng, Geng, Alex, Becker, James T., Davis, Simon, Toga, Arthur W. and Thompson, Paul M. (2005). Inverse consistent mapping in 3d deformable image registration: Its construction and statistical properties. *Information Processing in Medical Imaging* **3565**, 493–503.

Mazziotta, John, Toga, Arthur, Evans, Alan, Fox, Peter, Lancaster, Jack, Zilles, Karl and et al. (2001). A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm). *Philosophical Transactions of the Royal Society of London - Series B, Biological Sciences* **356**, 1293–1322.

Muralidharan, Omkar. (2010). An empirical bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics* **4**, 422–438.

Newton, Michael A., Noueiry, Amine, Sarkar, Deepayan and Ahlquist, Paul. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155–176.

Risch, Neil and Merikangas, Kathleen. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517.

Shen, Li, Kim, Sungeun, Risacher, Shannon L., Nho, Kwangsik, Swaminathan, Shanker, West, John D. and et al. (2010). Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in mci and ad: A study of the adni cohor. *Neuroimage* **53**, 1051–1063.

Smith, Stephen M., Jenkinson, Mark, Woolrich, Mark W., Beckmann, Christian F., Behrens, Timothy E.J., Johansen-Berg, Heidi and et al. (2004). Advances in functional and structural mr image analysis and implementation as fsl. *Neuroimage* **23**, 208–219.

Stein, Jason L., Hua, Xue, Lee, Suh, Ho, April J., Leow, Alex D., Toga, Arthur W. and et al. (2010*a*). Voxelwise genome-wide association study (vgwas). *Neuroimage* **53**, 1160–1174.

Stein, Jason L., Hua, Xue, Morra, Jonathan H., Lee, Suh, Hibar, Derrek P., Ho, April J. and et al. (2010*b*). Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in alzheimer's disease. *Neuroimage* **51**, 542–554.

Storey, John D. (2002). A direct approach to false discovery rates. *Journal of, the Royal Statistical Society. Series B* **64**, 479–498.

Storey, John D. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. *The Annals of Statistics* **31**, 2013–2035.

Szekely, Gabor J. and Bakirov, Nail K. (2003). Extremal probabilities for gaussian quadratic forms. *Probability Theory and Related Fields* **126**, 184–202.

Szekely, Gabor J. and Rizzo, Maria L. (2009). Brownian distance covariance. *The Annals of Applied Statistics* **3**, 1236–1265.

Szekely, Gabor J., Rizzo, Maria L. and Bakirov, Nail K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35**, 2769–2794.

Tziortzi, Andri C., Searlea, Graham E., Tzimopouloua, Sofia, Salinasa, Cristian, Beavera, John D., Jenkinsonb, Mark and et al. (2011). Imaging dopamine receptors in humans with [11c]-(+)-phno: dissection of d3 signal and anatomy. *Neuroimage* **54**, 264–277.

Vounou, Maria, Nichols, Thomas E. and Montana, Giovanni. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *Neuroimage* **53**, 1147–1159.

Weiner, Michael. (2006). A special newsletter for participants in the alzheimer's disease neuroimaging initiative. http://www.adcs.org/Research/PDF/ADNIExclusiveSummer06.pdf.

Woolrich, Mark W., Jbabdi, Saad, Patenaude, Brian, Chappell, Michael, Makni, Salima, Behrens, Timothy and et al. (2009). Bayesian analysis of neuroimaging data in fsl. *Neuroimage* **45**, 173–186.