Subject Section

Genome-wide Association Studies of Brain Imaging Data via Weighted Distance Correlation

Canhong Wen [1], Yuhui Yang [1], Quan Xiao [1], Meiyan Huang [2], Wenliang Pan [3,*], for the Alzheimer's Disease Neuroimaging Initiative[4]

[1] Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei, 230026, China,
[2] Guangdong Provincial Key Laboratory of Medical Image Processing, School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China,
[3] Department of Statistical Science, School of Mathematics, Sun Yat-Sen University, Guangzhou, 510275, China.

* To whom correspondence should be addressed.

Associate Editor: Jinbo Xu

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Imaging genetics is mainly used to reveal the pathogenesis of neuropsychiatric risk genes and understand the relationship between human brain structure, functional and individual differences. Increasingly, the brain-wide imaging phenotypes in voxels are available to test the association with genetic markers. A challenge with analyzing such data is their high dimensionality and complex relationships.

Results: To tackle this challenge, we introduce a weighed distance correlation ($wdCor$) that can assess the association between genetic markers and voxel-based imaging data. Importantly, the $wdCor$ test takes the voxel-based data as a whole multivariate phenotype, which preserves the spatial continuity and might enhance the power. Besides, an adaptive permutation procedure is introduced to determine the p-values of the $wdCor$ test and also alleviate the computational burden in GWAS. In extensive simulation studies, $wdCor$ achieves much better performances compared to the original distance correlation. We also successfully apply $wdCor$ to conduct a large-scale analysis on data from the Alzheimer's disease neuroimaging project (ADNI).

Availability: Our $wdCor$ method provides new research directions and ideas for multivariate analysis of high-dimensional data, it can also be used as a tool for scientific analysis of imaging genetics research in practical applications. The R package **wdcor**, and the code for reproducing all results in this paper is available in Github: https://github.com/yangyuhui0129/wdcor

Contact: panwliang@mail.sysu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

As research in developmental and clinical sciences has progressed in recent decades, there has been many significant advances in technology and methods of molecular genetics and neuroimaging. At the forefront of imaging genetics is an experimental strategy that effectively integrates molecular genetics and neuroimaging techniques (Munoz et al., 2009). The technology is mainly used to reveal the pathogenic mechanism of neuropsychiatric risk genes, and to understand the field of human brain structure, functional and individual differences in connections. Such understanding is critical for diagnosis, prevention, and treatment of numerous complex brain-related disorders (e.g.,schizophrenia and

[4] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List:pdf.

Alzheimer's disease). This also makes imaging genetics a useful tool for discovering genes for mental illness risk (Hashimoto et al., 2015).

The main method currently of gene-hunting in imaging genetics is genome-wide association study (GWAS). Several studies used a mass-univariate linear modeling (MULM) approach. Stein et al., 2010 proposed a voxel-based genome-wide association study (vGWAS), for pair analysis of each single nucleotide polymorphism (SNP) and each voxel. But treating each single voxel as a phenotype ignores the spatial continuity between the imaging data, i.e. the strong structural connections between voxel-based phenotypes, which are expected to share some common genetic variations. Besides, a large number of multiple comparisons make the vGWAS computationally intensive and they failed to find any significant result in their analysis. To reduce the number of tests, Hibar et al., 2011 proposed the voxel-based gene-wide association study (vGeneWAS), a method reduces the dimensionality of genome by considering voxel wise association with each gene. In order to reduce the computational burden, Huang et al., 2015 proposed a fast voxelwise genome wide association analysis (FVGWAS) framework to efficiently carry out vGWAS analysis.

Other potential whole-brain, genome-wide association studies are based on penalized and sparse regression techniques. The regression model based on the L_1 norm constraint has been successfully applied to multivariate genetic data analysis (Kohannim et al., 2012; Yang et al., 2015). However, these methods do not fully consider the structural relationship between the characteristic variables. After, Silver et al., 2012 and Yuan and Lin, 2006 proposed a group version of sparse models to solve the imaging genetic problems. Related studies are Kohannim et al., 2011 using ridge regression and Kohannim et al., 2012 using $L_1 + L_2$ regularization. Besides, Vounou et al., 2010a proposed the sparse reduced-rank regression (sRRR) model for the detection of genetic associations in imaging genetics studies involving high dimensional phenotypes. This method can simultaneously selects SNP variants and regions of association leveraging signal sparsity, but it is limited to linear correlation.

Another method named sparse canonical correlation analysis (SCCA) models, which is a powerful bi-multivariate analysis technique, have been used for imaging genetic association analysis. Witten et al., 2009 developed a penalized matrix decomposition (PMD) method and applied it to solve CCA with lasso and fuse lasso penalties. After that, Chen and Liu, 2012 developed an algorithm for solving CCA with overlapping group-lasso penalty and network-based fusion penalty. Besides, Lin et al., 2014 integrated the the prior knowledge with group lasso regularizer and SCCA model, to explore the correlation between genetic variation and brain activity. Du et al., 2014 proposed S2CCA using group lasso, and incorporated both the covariance matrix information and the priori knowledge information to discover group-level bi-multivariate associations.

Recently, a class of nonparametric approaches have been carried out to resolve the problem of the correlation analysis between two multivariate variables, with the distance covariance/correlation ($dCor$, SzéKely et al., 2007) as the most prominent one. Distance covariance/correlation is introduced to measure both linear and non-linear dependence between two random vectors in arbitrary dimension without relying any model assumption, making it more applicable for processing data in imaging genetics. SzéKely and Rizzo, 2009 introduce the concept of covariance of stochastic process, and then they extended the distance correlation to the problem of testing the independence of random vectors in high dimensions (SzéKely and Rizzo, 2013). *Energy statistics* as a statistical distance, proposed on the basis of distance, was more general and powerful against classical statistics (Székely and Rizzo, 2013). Besides, Székely et al., 2014 defined the partial distance correlation statistics with the help of Hilbert space, and develop a test for zero partial distance correlation. For real-valued variables, the complexity of calculating distance covariance by definition is $O(n^2)$, Huo and Székely, 2016 proposed an $O(nlog(n))$ algorithm and reduced the complexity of $dCov$ calculation. Geerligs et al., 2016 proposed a new method based on distance correlation to study the brain-wide functional connectivity and structural covariance. Wen et al., 2018 applied the distance covariance test to assess the dependence between SNPs and diffusion tensor imaging phenotypes. Throughout these articles, $dCor$ performs well, but it only illustrates that $dCor$ performs well in low-dimensional scenarios. For example, Wen et al., 2018 considered the ROI-based phenotypes and the dimension of phenotypes is 42; Geerligs et al., 2016 analyzed the voxel-based measurements within each ROI, which contains dozens of voxels on average. Yet for the brain-wide imaging phenotype, the number of voxels could be thousands or even higher. Under such high-dimensional settings, $dCor$ might be invalid. As shown in Reddi et al., 2015, the power of the $dCor$ falls sharply as the dimension increases, which implies that $dCor$ is not suitable for brain-wide imaging genetic data. To overcome the noted limitations of the existing methods, we proposed a novel framework for association analysis in imaging genetics based on a weighted distance correlation measure. In a high-dimensional voxel-based phenotype, it is often assumed that there is only a small amount of voxels are associated with the genetic marker(s) that we are examined (Vounou et al., 2010b). The weighted distance correlation ($wdCor$) is conducted by assigning positive weights to the true dependent voxels and negligible weights to the remaining independent voxels, and evaluating distance correlation based on the weighted voxels. In this way, we can alleviate the power loss of $dCor$ caused by the curse of dimensionality. We summarize our contributions as following:

1. We propose an efficient method to solve the association problem between the brain-wide voxel phenotypes and genome-wide genetic markers.
2. Our method takes the voxel-based data as a whole multivariate phenotype, which preserves the spatial continuity and might enhance the detection power.
3. An adaptive permutation procedure is introduced to make the proposed test feasible for GWAS.

The paper is organized as follows. In Section 2, we introduce the $wdCor$ and show its applicability in the independence test with two permutation procedures. Section 3 evaluates the finite sample performance of $wdCor$ including two parts : the Type-I error and the power of $wdCor$. In Section 4, we apply $wdCor$ to analyse a large-scale data from the Alzheimer's disease neuroimaging (ADNI) project, and compare its performance with $dCor$.

2 Methods

2.1 Overview of distance correlation

First, we will introduce the background of distance correlation, which was proposed for measuring and testing dependence between two random vectors X and Y. Distance correlation, which generalizes the idea of Pearson correlation, enjoys two remarkable properties: (1) $dCor(X, Y)$ is defined for X and Y of arbitrary dimensions. (2) $dCor(X, Y) = 0$ if and only if X and Y are independent. Thus, the proposed test based on distance correlation is sensitive to all types of departures from independence, including nonlinear or nonmonotone dependence structure.

The formal definitions of population coefficients $dCor$ and $dCov$ are given in SzéKely et al., 2007. Assume $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$, let $\phi_X(t)$ and $\phi_Y(s)$ be the respective characteristic functions of X and Y, and $\phi_{X,Y}(t, s)$ be the joint characteristic function of (X, Y). Distance covariance between X and Y with finite first moments is given by

$$dCov^2(X, Y) = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{\|\phi_{X,Y}(t, s) - \phi_X(t)\phi_Y(s)\|^2}{\|t\|_p^{1+p} \|s\|_q^{1+q}} dtds,$$

where $\|\cdot\|_p, \|\cdot\|_q$ stand for the Euclidean norm and c_p, c_q are some constants. Analogous to Pearson correlation, distance correlation is defined as

$$dCor(X,Y) = \begin{cases} \frac{dCov(X,Y)}{\sqrt{dVar(X)dVar(Y)}}, & dVar(X)dVar(Y) > 0, \\ 0, & dVar(X)dVar(Y) = 0. \end{cases}$$

where $dVar(X) = dCov(X,X)$.

SzéKely et al., 2007 proposed to estimate $dCor$ and $dCov$ through the usual moment estimation. To be precise, for an observed random sample $(\mathbf{X},\mathbf{Y}) = \{(X_i,Y_i) : i = 1,\ldots,n\}$, define

$$a_{kl} = \|X_k - X_l\|_p,\ b_{kl} = \|Y_k - Y_l\|_q$$

for $k,l = 1,\ldots,n$. Then compute the double-centering to the Euclidean distance matrices (a_{kl}) and (b_{kl}):

$$A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot},$$
$$B_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot},$$

where $\bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^n a_{kl}$ is the k-th row mean, $\bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^n a_{kl}$ is the l-th column mean and $\bar{a}_{\cdot\cdot}$ is the grand mean of the distance matrix of X. Analogously, $\bar{b}_{k\cdot}$, $\bar{b}_{\cdot l}$ and $\bar{b}_{\cdot\cdot}$ can be also defined.

The empirical distance covariance and empirical distance variance are defined as

$$\widehat{dCov^2}(\mathbf{X},\mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^n A_{kl}B_{kl},$$

$$\widehat{dVar^2}(\mathbf{X}) = \frac{1}{n^2}\sum_{k,l=1}^n A_{kl}^2, \quad \widehat{dVar^2}(\mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^n B_{kl}^2.$$

Accordingly, the empirical distance correlation can be defined as

$$\widehat{dCor}(\mathbf{X},\mathbf{Y}) = \begin{cases} \frac{\widehat{dCov}(\mathbf{X},\mathbf{Y})}{\sqrt{\widehat{dVar}(\mathbf{X})\widehat{dVar}(\mathbf{Y})}}, & \widehat{dVar}(\mathbf{X})\widehat{dVar}(\mathbf{Y}) > 0, \\ 0, & \widehat{dVar}(\mathbf{X})\widehat{dVar}(\mathbf{Y}) = 0. \end{cases}$$

2.2 Weighted distance correlation ($wdCor$)

Distance covariance or distance correlation provides a new approach to testing the independence of two random vectors. However, Reddi et al., 2015 showed that the power of $dCor$ drops polynomially with increasing dimension. That is to say, *dCor* does not perform well in high-dimensional cases. This may be caused by the increasing noise with the increase dimension of X and/or Y. As the dimension of *Y* increases with those dependent parts of *Y* remain unchanged, the proportion of the dependent parts will become smaller and smaller. Conversely, the independent parts turn to occupy an increasing proportion in calculating *dCor*, resulting in a low power in high dimensions.

We let $X \in \mathbb{R}$ be the genetic marker of interest and $Y \in \mathbb{R}^q$ be the imaging phenotype. In order to avoid the power of *dCor* falling sharply as the dimension of *Y* increases, it is necessary to highlight the dependent parts of *Y* when calculating *dCor* and to minimize the influence of the independent parts. Therefore, we are motivated to assign different weights according to their dependence. Specifically, when calculating *dCor*, the dependent parts of *Y* are given a larger weight, and those independent parts are assigned a smaller weight to relieve their influ/ence. Hence, we make the following improvements based on the original *dCor* and define this new statistic as weighted distance correlation (*wdCor*):

$$wdCor(X,Y;\omega) = dCor(X,(\omega_1 Y_1, \omega_2 Y_2, \ldots, \omega_q Y_q)),$$

such that $\sum_{i=1}^q \omega_i^2 = 1$. With various values of $\omega = (\omega_1,\omega_2,\ldots,\omega_q)$, we obtain a class of distance correlation, including the original distance correlation as the special case that $\omega = (1/\sqrt{q}, 1/\sqrt{q}, \ldots, 1/\sqrt{q})$. It is a key issue to reasonably determine ω making the weights of the dependent parts are non-zero and the weights of the independent parts are as small as possible. However, optimizing the objection function directly to obtain the weight function ω is very difficult due to distance correlation for the weight function ω is nonconvex. Notice that distance correlation is a general dependence measure of X and Y_j, which can detect either linear or nonlinear dependence. Moreover, distance correlation is defined in a finite interval with $0 \le dCor \le 1$, and kind of computationally fast. So here we consider a function of $dCor(X,Y_j)$ as weight function.

Inspired by the adaptive Lasso (Zou, 2006), we can transform the original optimization as follows. Denote $\beta = (\beta_1,\beta_2,\ldots,\beta_q)$, we let

$$\beta_j = dCor(X,Y_j),\ \omega_j = \frac{\beta_j^\gamma}{\|\beta^\gamma\|_2},$$

for $j = 1,2,\ldots,q$. Under the above transform, the value of ω_j ranges from 0 to 1. As γ increases, those parts of Y which have stronger dependence with X will be highlighted and the those redundant of Y will be gradually ignored in the calculation of $wdCor$. That is because larger γ makes the weights of those dependent parts occupied more share in the whole weight vector. Of all γ choices, we denote the one that maximizes $wdCor$ as the optimal:

$$\begin{aligned}\gamma_{opt} &= \arg\max_\gamma wdCor(X,Y;\omega) \\ &= \arg\max_\gamma dCor(X,(\omega_1 Y_1, \omega_2 Y_2, \ldots, \omega_q Y_q)).\end{aligned}$$

Hence, the optimal weight is

$$\omega_{opt} = \frac{\beta^{\gamma_{opt}}}{\|\beta^{\gamma_{opt}}\|_2}. \quad (1)$$

In this way, we transform the optimization problem involving q-dimensional variable, say ω, to the univariate optimization problem of γ. This simplified strategy highly reduces the complexity of the original problem and keeps the efficiency of $wdCor$. Notice that in practice, we cannot exhaust all possible values of γ to find the ω_{opt}, which may increase the computational burden. Also, we need to guarantee the efficiency of the $wdCor$ test. Considering these two points, we extract some SNPs from our ADNI data set in advance and calculate the optimal γ. The testing result shows that the optimal γ is relatively concentrated and stable. Using these results as a reference, we define a positive integer set $\Gamma = \{1,3,5\ldots,15\}$, after γ traverses the set, the optimal gamma is selected. Finally, we obtain the weighted distance correlation with the optimal weight:

$$wdCor(X,Y;\omega_{opt}) = dCor(X,(\omega_{opt,1}Y_1,\ldots,\omega_{opt,q}Y_q)),$$

such that $\sum_{i=1}^q \omega_{opt,i}^2 = 1$.

Similar to the empirical distance correlation, we can also obtain the empirical weighted distance correlation as follows

$$\begin{aligned}\widehat{wdCor}(\mathbf{X},\mathbf{Y};\hat{\omega}_{opt}) &= \max_{\gamma\in\Gamma} \widehat{wdCor}(\mathbf{X},\mathbf{Y};\omega) \\ &= \max_{\gamma\in\Gamma} \widehat{dCor}(\mathbf{X},(\omega_1\mathbf{Y}_1,\ldots,\omega_q\mathbf{Y}_q)).\end{aligned}$$

2.3 Permutation procedure for the $wdCor$ test

As shown in SzéKely et al., 2007, the asymptotic distribution of distance covariance under null hypothesis is hard to compute in practice and a permutation procedure is introduced to estimate it. Motivated by this, we here introduce a permutation procedure to determine the null distribution

of the proposed *wdCor* statistic. In specific, for any significance level $\alpha \in (0, 1)$, the α-level test is conducted in the following manner:

Algorithm 1. Permutation procedure for the *wdCor* test

1. Given a sample $(\mathbf{X}, \mathbf{Y}) = \{(X_i, Y_i), i = 1, \ldots, n\}$, calculate the test statistic $\widehat{WD} \triangleq \widehat{wdCor}(\mathbf{X}, \mathbf{Y}; \hat{\omega}_{opt})$ on the given sample.
2. For a sample $(\mathbf{X}, \mathbf{Y})$, we randomly permute the indices $\{1, \ldots, n\}$, denoted the permuted indices as $\{i_1, \ldots, i_n\}$, and determine the permuted sample by $(\mathbf{X}, \mathbf{Y}^*)$, where $\mathbf{Y}^* = \{Y_{i_1}, \ldots, Y_{i_n}\}$. Compute the permutation test statistic $\widehat{WD}^* \triangleq \widehat{wdCor}(\mathbf{X}, \mathbf{Y}^*; \hat{\omega}^*_{opt})$.
3. Repeat step 2 for T times to get $\widehat{WD}^{*(1)}, \ldots, \widehat{WD}^{*(T)}$. For a pre-specified significant level α, calculate the empirical p-value

$$\hat{p} = \frac{1}{T} \sum_{i=1}^{T} I(\widehat{WD}^{*(i)} \geq \widehat{WD}),$$

   where $I(\cdot)$ is the indicator function.
4. Reject H_0 if $\hat{p} \leq \alpha$.

For data in a large-scale GWAS, the permutation time in the above procedure should be larger than millions, i.e., $T \approx 10^6$, to avoid failure in the Type-I error control. Moreover, the procedure is applied for each SNP, which aggravates the computation burden and might be infeasible in GWAS. Thus a computationally efficient algorithm is urgent. In GWAS, it is assumed that the vast majority of SNPs are non-causal/independent to the phenotype. Identifying these SNPs early in permutation testing could reduce the total number of permutation times. An intuitive idea is that we may terminate the permutation at an early stage if there is little or even no evidence towards alternative based on the *wdCor* statistic, while perform exhaustive permutations for strongly dependent features (Besag and Clifford, 1991).

Here, we propose an adaptive permutation strategy by using an appropriate permutation time T bases on the dependence level. Specifically, we first apply **Algorithm 1** with $T = T_0$ (say 10) to calculate a rough p-value for each SNP, and screen out those SNPs with larger p-values. Then we apply **Algorithm 1** with $T = T_1 (> T_0)$ to update the p-values for the remaining SNPs. The above procedure is applied with gradually increasing permutation times, say $\{T_0, T_1, \ldots, T_k\}$, and stops when there is no SNPs left. We summarize the adaptive permutation procedure for testing independence in high-dimensional studies having a large number of tests as follows:

Algorithm 2. Adaptive permutation procedure for the *wdCor* statistic

1. Determine the number of independent tests (m) and the remaining set $(RS = \{1, \ldots, m\})$. Initialize the number of permutations T_0 and the threshold α_0.
2. While RS is not empty,
   - For each SNP, run **Algorithm 1** with $T = T_0$ and denote the output p-value as $\{\hat{p}_1, \ldots, \hat{p}_m\}$.
   - Screen out the SNPs with their p-values larger than α_0 and update RS. That is, $RS = \{j, \hat{p}_j \leq \alpha_0\}$.
   - Determine $\alpha_0 = \alpha_0/10$ and $T_0 = 10 \times T_0$.

In practice, we can initialize as $T_0 = 10$ and $\alpha_0 = 1/T_0$. Notice that the adaptive permutation is much faster than the standard permutation and provides good estimates of p-values, and thus it is computationally feasible for GWAS.

3 Simulation

In this section, we compare the performance of our proposed tests (*wdCor*) with the *dCor* test proposed by SzéKely et al., 2007. To be fair, we use the permutation procedure to calculate the p-values for both methods. Here, we set $T = 200$, $m = 500$, the nominal level of significance at 0.05.

The genetic marker $X \in \mathbb{R}$ consists of n i.i.d. random variables from Binomial distribution with size 2 and probability p. The noise $E = (E_{ij})_{n \times q} = (E_1, \ldots, E_q)$ is generated from multivariate normal distribution $N(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{ij})$ and $\sigma_{ij} = \rho^{|i-j|}$. For the imaging phenotype $Y = (Y_1, Y_2, \ldots, Y_q)$, we consider the following five cases:

Case 1: $Y_j = E_j, \quad j = 1, 2, \ldots, q.$

Case 2:

$$Y_j = \begin{cases} \beta X + E_j, & j = 1, 2, \ldots, q_1, \\ E_j, & \text{otherwise.} \end{cases}$$

Case 3:

$$Y_j = \begin{cases} \beta X^2 + E_j, & j = 1, \\ \beta X + E_j, & j = 2, \ldots, q_1, \\ E_j, & \text{otherwise.} \end{cases}$$

Case 4:

$$Y_j = \begin{cases} \sin(\pi X/6) + E_j, & j = 1, \\ \beta X + E_j, & j = 2, \ldots, q_1, \\ E_j, & \text{otherwise.} \end{cases}$$

Case 5:

$$Y_j = \begin{cases} \beta X^2 + E_j, & j = 1, \\ \sin(\pi X/6) + E_j, & j = 2, 3, \\ \beta X, & j = 4, \ldots, q_1, \\ E_j, & \text{otherwise.} \end{cases}$$

Case 6:

$$Y_j = \begin{cases} \tan(X/2) + E_j, & j = 1, \\ \ln(X + 1) + E_j, & j = 2, \\ 0.2^X + E_j, & j = 3, \\ 0.1 + 0.1X + 0.1X^2 + E_j, & j = 4, \\ \beta X + E_j, & j = 5, \ldots, q_1, \\ E_j, & \text{otherwise.} \end{cases}$$

Case 1 considers the situation when the genetic marker X and the phenotype Y are independent. In **Cases 2-6**, the phenotype Y is dependent with the genetic marker X. There is linear dependence between X and the first q_1 coordinates of Y in **Case 2**. In **Case 3-4**, there are also linear dependence between X and the $2 - q_1$ coordinates of Y, but there are nonlinear dependence between X and the first coordinate of Y compared to **Case 2**. In particular, there is quadratic and sine dependence in **Case 3** and **Case 4**, respectively. In **Case 5**, the dependence contains both quadratic and trigonometric function. The most complicated dependence is shown in **Case 6**, with tangent, logarithmic, exponential function and polynomial are applied.

3.1 Proper definition of weights

In this section, we assess whether the weights defined in Equation (1) can identify the true dependent parts of the phenotype Y. We fix the coefficient β to be 0.3 and $\rho = 0$. The sample size n is 200 and the number of true dependent coordinates in Y is $q_1 = 10$. We choose the dimension of Y as $q = 10$ for **Case 1** and $q = 10, 20, 50, 100, 200$ for **Cases 2-6**. The average of the optimal weights between X and the first 10 coordinates of Y are presented in Supplementary Figure 1.

For **Case 1** with $q = 10$, it can be seen from the left panel of Supplementary Figure 1 a that the weights stay almost the same for all dimensions in Y. Besides, the statistic values from *wdCor* is not significant

different from those from *dCor*. This is expected because X is independent with Y and *wdCor* reduces to *dCor* in this case.

From the top panels of Supplementary Figure 1 b, we can see that the weights of the true dependent indexes, say $1, \ldots, q_1$, are basically identical and considerably larger than those of independent indexes. Furthermore, their gaps increase as the dimension of Y increases, or the sparsity of Y increases. When all dimensions in Y are dependent with X ($q = 10$), the *wdCor* statistic is slightly larger than the *dCor* statistic. This indicates the superiority of our proposed *wdCor* over *dCor* in the high-dimensional scenarios.

Supplementary Figure 1 c presents the average weights for the first q_1 dimension of Y in **Cases 3-6**. Although the first q_1 weights is not the same, their values are much larger than the remaining values. It is not surprising but reassuring that the weights from those independent indexes become close to 0 as the sparsity of Y increases regardless of linear or nonlinear cases. It suggests that our proposed weights allocation strategy is able to distribute non-zero weights to the true important indexes.

3.2 Type-I error

In **Case 1**, X and Y are independent and thus the type-I error is calculated. We choose the sample size as $n = 100, 200, 300$, and the dimension of Y as $q = 10, 50, 100, 200, 400, 600$. For ρ, we set it to be 0 and 0.5 for uncorrelated and correlated scenarios, respectively.

The empirical p-values are given in Supplementary Table 1. It can be seen from Supplementary Table 1 that the estimated p-values of the two tests are controlled fairly well around 0.05 for all cases. This implies that the null distributions of the *wdCor* statistic and the *dCor* statistic are well approximated from the permutation procedure.

3.3 Statistical power

In this section, we will compare the power of $dCor$ and $wdCor$ especially in the ultra-high dimension situations. **Cases 2-6** consider the genetic marker X that has linear or nonlinear dependence with the phenotype Y. We choose the sample size $n = 100, 200$, set the dimension of Y as $q = 100, 200, 400, 600, 800, 1000$, and use the same setting ρ as in Section 3.2. For the coefficient β, we set it to be a fixed value 0.3 or a random value generated from $N(0, 0.4)$.

The power for $dCor$ and $wdCor$ under different settings were presented in Supplementary Table 2. It can been seen that the empirical p-values of the $wdCor$ is larger than those of the $dCor$ in all cases, which suggests the necessity of introducing weights into the distance correlation.

To better portray the decays tendency of the empirical p-values as dimension increases, Figure 1 plots the power versus the dimension q for both the $dCor$ and $wdCor$ methods. It can be seen that the gaps in the empirical p-values between the two tests increases with the dimension q increase. What's more, when given sufficient sample (in our simulation $n = 200$), the declining speed is much slower for the $wdCor$ test compared to the $dCor$ test. For both tests, the performance in **Cases 3** and **5** is much better than those in **Cases 2** and **4**. This is because the distance correlation tends to perform better in detecting the quadratic dependence compared to the trigonometric dependence. Yet even in handling such weak association, our proposal still has detection power no less than 0.6 in **Case 2** and 0.7 in **Case 4**. **Case 6** includes some nonlinear dependence that are not contained in **Case 2-5**. Among them, the strongest dependence is from the tangent function, and the weakest is from the polynomial function.Even though, the polynomial function still have stronger dependence than the linear function.

4 ADNI data analysis

Alzheimer's disease (AD) is a common degenerative disease of the central nervous system. It mainly occurs in the elderly over 65 years old. Clinical manifestations are often memory impairment, cognitive deficit, language decline and so on, which seriously affect people's life and health.

The genome-wide association study has greatly promoted the study of the genetic traits of AD and detected a large number of candidate genes of AD. Labeling AD-related genes with SNPs has become a new breakthrough in the study of AD. At the same time, imaging techniques are also applied to the study of brain structure and function. Thus by combining neuroimaging and genomics, it is more likely to discover the underlying biological pathogenesis of the disease, and to identify genes associated with AD.

4.1 Data processing

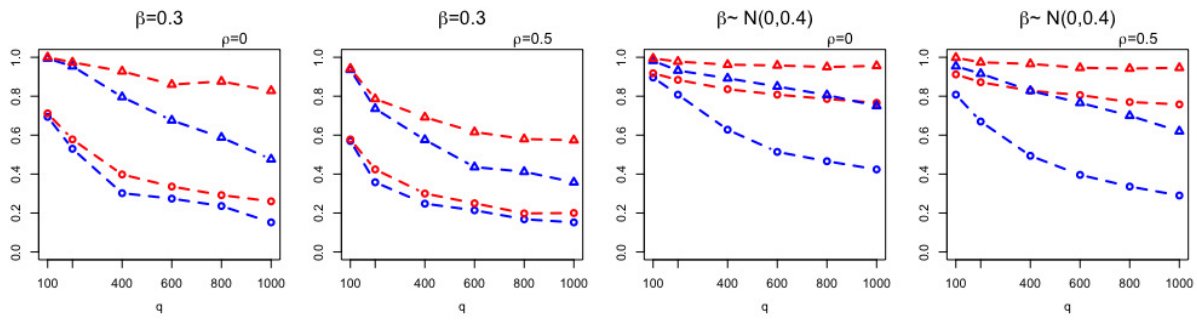
To illustrate the usefulness of the proposed method, the brain MRI data were used as phenotypes in GWAS and were obtained from ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging, National Institute of Biomedical Imaging and Bioengineering, Food and Drug Administration, private pharmaceutical companies and non-profit organizations as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians in developing new treatments and monitoring their effectiveness, as well as lessening the time and cost of clinical trials. The principal investigator of this initiative is Michael W. Weiner, MD, at the VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The goal was to recruit 800 subjects, but the initial study (ADNI-1) has been followed by ADNI-GO and ADNI-2. To date, these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow-up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI- GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www. adni-info.org.

In this study, 362 MRI scans obtained from ADNI database were used. The scans from 164 AD and 198 healthy controls were performed on a 1.5T MRI scanners with some individual protocols. The typical protocol includes the following parameters: repetition time (TR) = 2400 ms, inversion time (TI) = 1000 ms, flip angle = 8^o, and field of view (FOV) = 24 cm with a $256 \times 256 \times 170$ acquisition matrix in the $x-$, $y-$, and $z-$ dimensions, which yields a voxel size of $1.25 \times 1.26 \times 1.2$ mm^3.

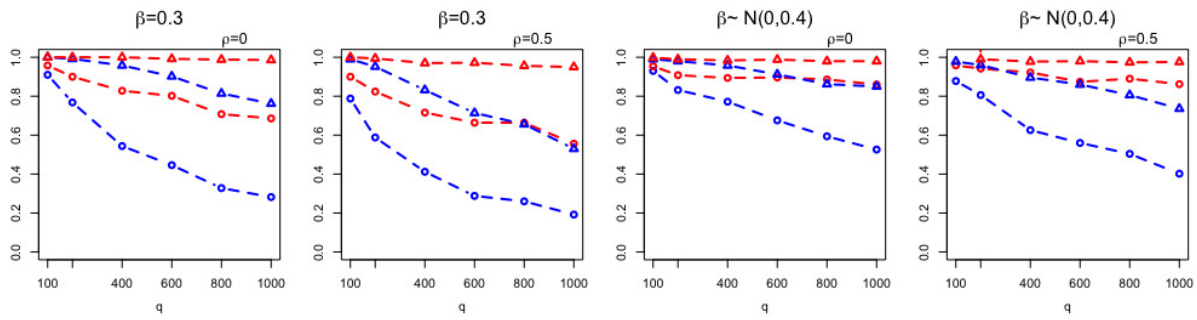
We process the MRI data by using the following steps: anterior commissure and posterior commissure correction, skull-stripping, cerebellum removal, intensity inhomogeneity correction, segmentation, and registration (Shen and Davatzikos, 2004). After segmentation, we segment the brain data into four different tissues: gray matter (GM), white matter (WM), ventricle (VN), and cerebrospinal fluid (CSF). We use the deformation field to generate RAVENS maps (Davatzikos et al., 2001a) to quantify the local volumetric group differences for the whole brain and each of the segmented tissue type (GM, WM, VN, and CSF), respectively. Moreover, we automatically label 93 ROIs on the template and transferred the labels following the de-formable registration of subject images (Davatzikos et al., 2001b). To simplify the analysis of each ROI, we first select the labeled ROIs and their corresponding voxel sites on the template, and extract the voxels on the identical sites as the template on the register image of each subject to obtain the same voxel number for each subject.

The SNP data were genotyped using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA). Quality control and SNP screening were

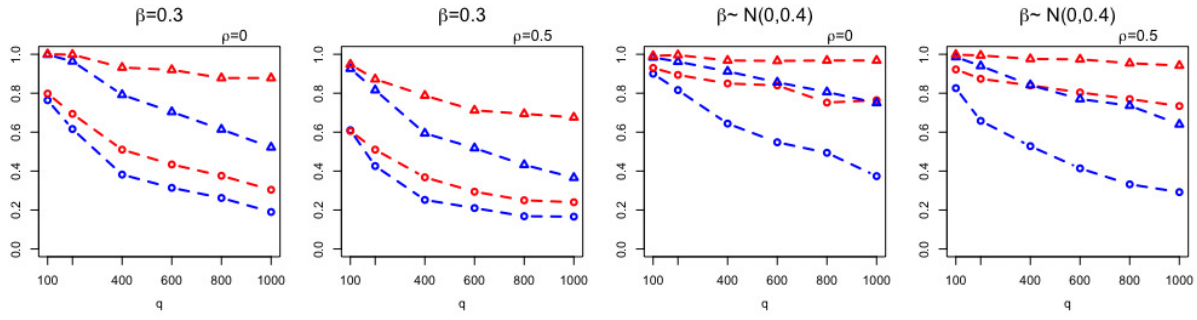
a **.Case 2** $(X \sim (0, 1, 2))$



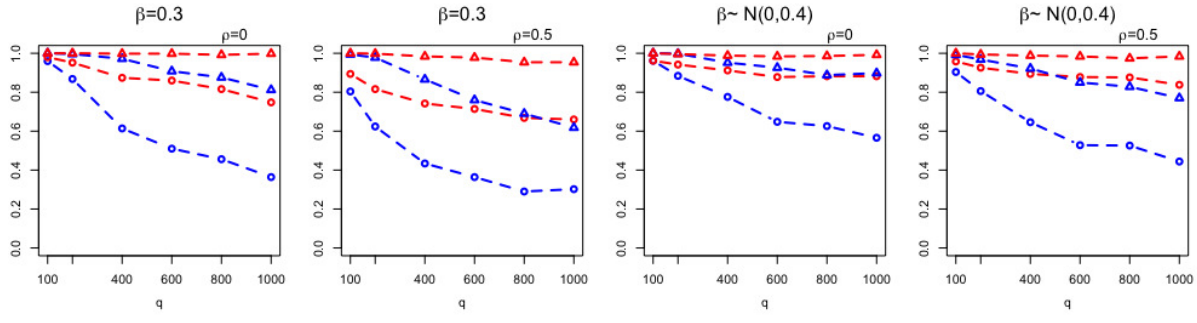
b **.Case 3** $(X \sim (0, 1, 2))$



c **.Case 4** $(X \sim (0, 1, 2))$



d **.Case 5** $(X \sim (0, 1, 2))$



e **.Case 6** $(X \sim (0, 1, 2))$

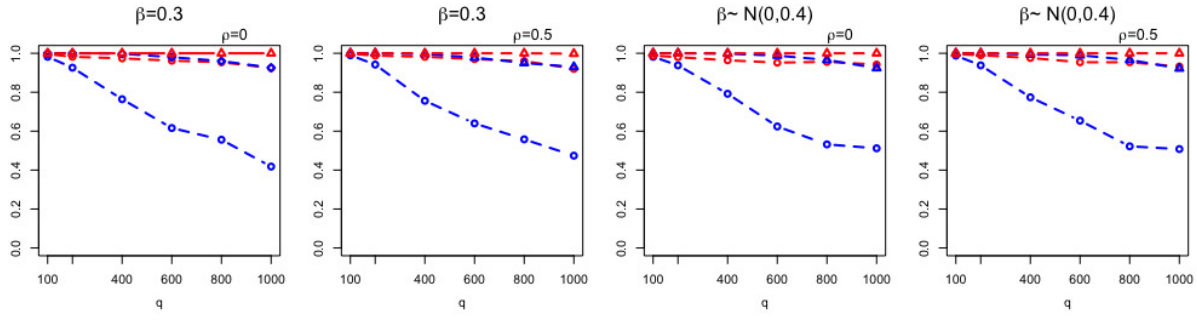


Table 1. Top 10 SNP-ROI paris identified by wdCor and dCor

ROI	SNP	CHR	BP	p.value	$wdCor$	$dCor$	Gene
			wdCor				
posterior limb of internal capsule	rs12721364	12	48231430	2.0e-08	0.340	0.229	LINC02354
posterior limb of internal capsule	rs2765588	13	103863278	4.0e-08	0.317	0.253	None
right amygdala	rs2075650	19	45395619	4.0e-07	0.305	0.275	TOMM40
right medial occipitotemporal gyrus	rs8095770	18	53093724	4.0e-07	0.346	0.251	TCF4
left precuneus	rs7933176	11	100052221	5.0e-07	0.314	0.226	CNTN5
right medial frontal gyrus	rs924598	5	35267238	5.0e-07	0.342	0.236	None
left superior parietal lobule	rs11201974	10	88067925	6.0e-07	0.326	0.206	GRID1
left parahippocampal gyrus	rs2622927	1	47861163	7.0e-07	0.319	0.237	LINC01389
left superior occipital gyrus	rs9864595	3	4420255	8.0e-07	0.323	0.189	SUMF1
left uncus	rs2075650	19	45395619	8.0e-07	0.299	0.247	TOMM40
			dCor				
right amygdala	rs2075650	19	45395619	2.0e-07	0.305	0.275	TOMM40
right occipital pole	rs8076012	17	47101989	2.0e-07	0.317	0.276	IGF2BP1
left parietal lobe white matter	rs17565737	10	93361825	3.0e-07	0.308	0.283	HECTD2-AS1
right lateral ventricle	rs713532	16	71371993	5.0e-07	0.284	0.263	None
right temporal lobe white matter	rs7088910	10	6030194	5.0e-07	0.281	0.281	None
right temporal lobe white matter	rs713532	16	71371993	6.0e-07	0.291	0.282	None
left anterior limb of internal capsule	rs17565737	10	93361825	7.0e-07	0.277	0.262	HECTD2-AS1
left subthalamic nucleus	rs1157531	4	13882785	2.0e-06	0.299	0.266	LINC01182
left amygdala	rs2075650	19	45395619	2.0e-06	0.285	0.266	TOMM40
right caudate nucleus	rs12077089	1	85298014	3.0e-06	0.285	0.271	LPAR3

conducted on the SNP data as introduced in a previous study (Huang et al., 2015). Moreover, the remaining missing genetic data were imputed as the modal value. After these procedures, we retain 362 subjects, and each subject had 501675 SNPs.

4.2 Data Analysis and Results

We treat the 93 ROIs as different functional phenotypes, and perform chromosome-wide tests separately. Before analysis, we perform multiple linear regression analysis to adjust the confounding effects from covariates including gender, age, whole brain volume, and the top 5 principal component scores in SNPs, and the resulting residual matrix is stored for the next step.

For each ROI, there only exist a few strong-associated SNPs, screening out those weakly associated SNPs can greatly reduce unnecessary calculation. Since the SNP having larger value of $wdCor$ tends to have stronger dependence with the ROI, we calculate the empirical weighted distance correlation for each SNP with the ROI, denoted by $\widehat{WD}_k$ for $k = 1, 2, \ldots, p$, where p is the total number of SNPs. Then we keep the stronger associated SNPs and let the active set $\mathscr{A}$ be

$$\mathscr{A} = \{k : \widehat{WD}_k \geq \gamma, k = 1, 2, \ldots, p\},$$

where the thresholding parameter γ is to distinguish the active SNPs from the inactive ones. Then we borrow the idea of random decoupling in Barut et al., 2016 to determine the γ, the procedure is described as follows.

We generate a set of pseudo-predictors $\mathbf{Z}$ by randomly permuting the rows of each SNP, and compute the empirical weighted distance correlation for $\mathbf{Z}$ and the ROI, denoted by $\widehat{WD}_{p+k}$ for $k = 1, 2, \ldots, p$. The $\hat{\gamma}_{max} = max_{k=1,2,\ldots,p}\widehat{WD}_{p+k}$ can be seen as a cutoff value to distinguish the active SNPs from the inactive ones. According to Barut et al. (2016), a more practical approach is to choose the $\hat{\gamma}_{(q)}$ as cutoff value, $\hat{\gamma}_{(q)}$ is the the q_{th} quantile of $\{\widehat{WD}_{p+k}, k = 1, 2, \ldots, p\}$, where $0 \leq q \leq 1$. Thus we repeat this procedure B times and compute the mean of $\{\hat{\gamma}^b_{(q)}\}^B_{b=1}$, denoted by $\hat{\gamma}^*_{(q)}$. Consequently, the active set $\mathscr{A}$ is

$$\hat{\mathscr{A}} = \{k : \widehat{WD}_k \geq \hat{\gamma}^*_{(q)}, k = 1, 2, \ldots, p\}.$$

Following from Barut et al. (2016), the useful range for q is [0.95,1]. In our method, we take $B = 10$ and $q = 0.99$, which exhibit reasonable performance of the thresholding rule.

Next, we apply the $wdCor$ statistic to test the association between each SNP and each voxel-based ROI respectively. We also include the results based on the *dCor* statistic for comparison. To be fair, the p-values from both tests are obtained via the adaptive permutation procedure as in **Algorithm 2**.

The top 10 SNP-ROI pairs identified by the *wdCor* and *dCor* tests for all ROIs and chromosomes are reported in Tables 1. It is hard to choose the threshold due to the dependence of the tests, thus we consider the threshold $5.0e^{-8}$ commonly used in GWAS. SNP rs12721364 (p-value=$2.0e^{-8}$) and SNP rs2755588 (p-value=$4.0e^{-8}$) are significantly associated with the posterior limb of internal capsule (PLIC) via $wdCor$ test. Hall et al., 2016 showed the structural damage of PLIC is related to executive dysfunction, attention deficit, visual space defect, memory loss and cognitive impairment, and most of these symptoms are consistent with the clinical symptoms of AD.

Besides, $wdCor$ test detects several SNP-ROI pairs with strong association. The SNP rs7933176 on gene CNTN5 is highly associated with the left precuneus (p-value=$5.0e^{-7}$). The precuneus is involved in source memory, and Karas et al., 2007 found the disproportionate atrophy in the precuneus of patients with early-onset AD. The gene CNTN5 was also identified to be of heightened interest with AD (Biffi and Alessandro, 2010). The SNP rs2075650 on TOMM40 was detected to have a strong correlation with 2 ROIs. The p-values obtained by $wdCor$ test are $4.0e^{-7}$ for the right amygdala and $8.0e^{-07}$ for the left uncus. Moreover, the rs2075650–right amygdala pair is also detected intensively associated via $dCor$ test (p-value=$2.0e^{-7}$). Huang et al., 2016 and Michal et al., 2018 pointed out that the polymorphism of rs2075650 on TOMM40 may be an independent risk factor of developing AD. Liu et al., 2018 also explained the relationship between rs2075650 and AD. The amygdala is related to aberrant motor behavior and potentially associated with anxiety and irritability. Several studies found considerable shrinkage of amygdala of AD patients (Knafo, 2012; Poulin et al., 2011; Denys et al., 1993).

Compared with the FVGWAS proposed by Huang et al., 2015, our method can detect more SNPs with small p-values. Huang et al., 2015 carried out FVGWAS for the whole-brain data but did not detect any significant SNP with p-value < $5.0e^{-08}$, while our $wdCor$ method detects two significant SNPs. Besides, eight strong associated SNPs (p-value=e^{-07}) are detected in our method, compared with only four in Huang et al., 2015. Supplementary Figure 3 presents the spatial distribution of the optimized weights of three strong associated SNP-ROI pairs by the $wdCor$ test respectively.

5 Discussion

We present the $wdCor$ test for voxel-based imaging genetic association, and test it extensively on simulated and ADNI dataset. The $wdCor$ is much powerful than the $dCor$, in that is has comparable performance on low-dimensional data and improved performance on high-dimensional data.

The main insight $wdCor$ uses is that the independent coordinates of phenotypes are removed from the test statistic by assigning ignorable weights to the corresponding dimensions. The original $dCor$ puts equal weights to each dimension and thus might occur the curve of dimensionality in dealing with voxel-based phenotypes. To determine the p-values of the wdCor test, a traditional permutation procedure is introduced in general and an adaptive permutation procedure is proposed to alleviate the computational burden in genome-wide association analysis.

The simplicity of the $wdCor$ statistic makes it more flexible and more easily adapted to different circumstances. Inherited from $dCor$, $wdCor$ does not rely on any model or distribution assumptions, and is easy to implement. Moreover, $wdCor$ can detect the both linear and non-linear dependence between two random variables of any finite dimensions.

We show the rationality and superiority of introducing weights by extensive simulation studies. The proposed $wdCor$ test overwhelmingly improved the statistical power compared the $dCor$ test, especially when the dimension is hundreds or even higher.

We apply $wdCor$ to test the independence between 93 ROI phenotypes and SNPs from the ADNI datasets. Our proposed method detects two significant SNP-ROI pairs, while the $dCor$ test cannot detect any significant SNP.

There are two substantial issues to be addressed in our future research. First, since our $wdCor$ is still a single SNP analysis framework, there exists unobserved dependence between SNPs, such as the causal relationship between SNPs, dependence between adjacent SNPs, and the interaction between SNPs. All these unobserved SNP-SNP interactions may undermine the power of $wdCor$. Alternative methods for testing the dependence between a single SNP set and an individual phenotype have been shown to be useful for improving the efficacy of GWAS (Ge et al., 2012; Thompson et al., 2013). Therefore, it is of great importance to generalize our $wdCor$ test for multi-multi dimensional correlation analysis to map the association between a SNP set and a functional neuroimaging phenotype. Second, in order to improve the power of $wdCor$, we add the parameter γ to the weight vector and choose the optimal values from a list of candidate values. This does give us a higher statistical power, but at the same time it loses computational efficiency. To alleviate this problem, we have optimized the R code using C++ with Rcpp package, the R package –wdcor– are available on GihHub (https://github.com/yangyuhui0129/wdcor).

Acknowledgements

We thank the Editor, the Associate Editor, Jinbo Xu and two referees for their insightful comments and constructive suggestions that have greatly improved the presentation of the paper.

Funding

This work has been supported by the National Natural Science Foundation of China [11801540 to C.H.W.]; the Natural Science Foundation of Guangdong [2017A030310572 to C.H.W.]; the Fundamental Research Funds for the Central Universities [WK2040170015,WK2040000016 to C.H.W]; the National Natural Science Foundation of China [81601562 to M.Y.H]; and Science and Technology Planning Project of Guangzhou [201904010417 to M.Y.H]; the National Natural Science Foundation of China [11701590 to W.L.P]; the Natural Science Foundation of Guangdong Province of China [2017A030310053 to W.L.P]; the Young te acher program/Fundamental Research Funds for the Central Universities [17lgpy14 to W.L.P].

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Reference

Barut, E., J. Fan, and A. Verhasselt (2016). Conditional sure independence screening. *Publications of the American Statistical Association 111*(515), 1266–1277.

Besag, J. and P. Clifford (1991). Sequential monte carlo p-values. *Biometrika 78*(2), 301–304.

Biffi and Alessandro (2010). Genetic variation and neuroimaging measures in alzheimer disease. *Archives of Neurology 67*(6), 677.

Chen, X. and H. Liu (2012). An efficient optimization algorithm for structured sparse cca, with applications to eqtl mapping. *Statistics in Biosciences 4*(1), 3–26.

Davatzikos, C., A. Genc, D. Xu, and S. M. Resnick (2001a). Voxel-based morphometry using the ravens maps: Methods and validation using simulated longitudinal atrophy. *NeuroImage 14*(6), 1361–1369.

Davatzikos, C., A. Genc, D. Xu, and S. M. Resnick (2001b). Voxel-based morphometry using the ravens maps: Methods and validation using simulated longitudinal atrophy. *NeuroImage 14*(6), 1361–1369.

Denys, A., J. L. Michot, P. Jehenson, F. Forette, and F. Boller (1993). Amygdala atrophy in alzheimer's disease. *A.m.a.archives of Neurology 50*(9), 941–5.

Du, L., J. Yan, S. Kim, S. L. Risacher, H. Huang, M. Inlow, J. H. Moore, A. J. Saykin, and L. Shen (2014). A novel structure-aware sparse learning algorithm for brain imaging genetics. *Medical Image Computing and Computer-assisted Intervention 17*, 329–336.

Ge, T., J. Feng, D. P. Hibar, P. M. Thompson, and T. E. Nichols (2012). Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. *Neuroimage 63*(2), 858–873.

Geerligs, L., Camcan, and R. N. Henson (2016). Functional connectivity and structural covariance between regions of interest can be measured more accurately using multivariate distance correlation. *NeuroImage 135*, 16–31.

Hall, J. M., K. A. E. Martens, C. C. Walton, C. Ocallaghan, P. E. Keller, S. J. G. Lewis, and A. A. Moustafa (2016). Diffusion alterations associated with parkinson's disease symptomatology: A review of the literature. *Parkinsonism & Related Disorders 33*, 12–26.

Hashimoto, R., K. Ohi, H. Yamamori, Y. Yasuda, M. Fujimoto, S. Umedayano, Y. Watanabe, M. Fukunaga, and M. Takeda (2015). Imaging genetics and psychiatric disorders. *Current Molecular Medicine 15*(2), 168–175.

Hibar, D. P., J. L. Stein, O. Kohannim, N. Jahanshad, A. J. Saykin, L. Shen, S. Kim, N. Pankratz, T. Foroud, M. J. Huentelman, et al. (2011). Voxelwise gene-wide association study (vgenewas): Multivariate gene-based association testing in 731 elderly subjects. *NeuroImage 56*(4), 1875–1891.

Huang, H., J. Zhao, B. Xu, X. Ma, Q. Dai, T. Li, F. Xue, and B. Chen (2016). The tomm40 gene rs2075650 polymorphism contributes to alzheimer's disease in caucasian, and asian populations. *Neuroscience Letters 628*, 142–146.

Huang, M., T. E. Nichols, C. Huang, Y. Yu, Z. Lu, R. C. Knickmeyer, Q. Feng, and H. Zhu (2015). Fvgwas: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. *NeuroImage 118*, 613–627.

Huo, X. and G. J. Székely (2016). Fast computing for distance covariance. *Technometrics 58*(4), 435–447.

Karas, G., P. Scheltens, S. Rombouts, R. V. Schijndel, M. Klein, B. Jones, W. V. D. Flier, H. Vrenken, and F. Barkhof (2007). Precuneus atrophy in early-onset alzheimer's disease: a morphometric structural mri study. *Neuroradiology 49*(12), 967–976.

Knafo, S. (2012). *Amygdala in Alzheimer's Disease*.

Kohannim, O., D. P. Hibar, J. L. Stein, N. Jahanshad, X. Hua, P. Rajagopalan, A. Toga, C. R. Jack Jr, M. W. Weiner, G. I. De Zubicaray, et al. (2012). Discovery and replication of gene influences on brain structure using lasso regression. *Frontiers in neuroscience 6*, 115.

Kohannim, O., D. P. Hibar, J. L. Stein, N. Jahanshad, C. R. Jack, M. W. Weiner, A. W. Toga, and P. M. Thompson (2011). Boosting power to detect genetic associations in imaging using multi-locus, genome-wide scans and ridge regression. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1855–1859.

Lin, D., V. D. Calhoun, and Y. Wang (2014). Correspondence between fmri and snp data by group sparse canonical correlation analysis. *Medical Image Analysis 18*(6), 891–902.

Liu, C., J. Chyr, W. Zhao, Y. Xu, Z. Ji, H. Tan, C. Soto, and X. Zhou (2018). Genome-wide association and mechanistic studies indicate that immune response contributes to alzheimer's disease development. *Frontiers in Genetics 9*, 410.

Michal, Prendecki, Jolanta, Florczak-Wyspianska, Marta, Kowalska, Jan, Ilkowski, Teresa, and G. and (2018). Biothiols and oxidative stress markers and polymorphisms of tomm40 and apoc1 genes in alzheimer's disease patients. *Oncotarget 9*(81), 35207–35225.

Munoz, K. E., L. W. Hyde, and A. R. Hariri (2009). Imaging genetics. *48*(4), 6.

Poulin, S. P., R. Dautoff, J. C. Morris, L. F. Barrett, and B. C. Dickerson (2011). Amygdala atrophy is prominent in early alzheimerš disease and relates to symptom severity. *Psychiatry Research 194*(1), 7–13.

Reddi, S. J., A. Ramdas, B. Poczos, A. Singh, and L. Wasserman (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. *arXiv: Machine Learning*, 3571–3577.

Shen, D. and C. Davatzikos (2004). Measuring temporal morphological changes robustly in brain mr images via 4-dimensional template warping. *NeuroImage 21*(4), 1508–1517.

Silver, M. J., G. Montana, and A. D. N. Initiative (2012). Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. *Statistical Applications in Genetics and Molecular Biology 11*(1), 1–43.

Stein, J. L., X. Hua, S. Lee, A. J. Ho, A. D. Leow, A. W. Toga, A. J. Saykin, L. Shen, T. Foroud, N. Pankratz, et al. (2010). Voxelwise genome-wide association study (vgwas). *NeuroImage 53*(3), 1160–1174.

SzéKely, G. J. and M. L. Rizzo (2009). Brownian distance covariance. *The Annals of Applied Statistics 3*(4), 1236–1265.

SzéKely, G. J. and M. L. Rizzo (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis 117*, 193–213.

Székely, G. J. and M. L. Rizzo (2013). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference 143*(8), 1249–1272.

Székely, G. J., M. L. Rizzo, et al. (2014). Partial distance correlation with methods for dissimilarities. *The Annals of Statistics 42*(6), 2382–2412.

SzéKely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics 35*(6), 2769–2794.

Thompson, P. M., T. Ge, D. C. Glahn, N. Jahanshad, and T. E. Nichols (2013). Genetics of the connectome. *Neuroimage 80*, 475–488.

Vounou, M., T. E. Nichols, and G. Montana (2010a). Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *NeuroImage 53*(3), 1147–1159.

Vounou, M., T. E. Nichols, and G. Montana (2010b). Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *NeuroImage 53*(3), 1147–1159.

Wen, C., C. M. Mehta, H. Tan, and H. Zhang (2018). Whole genome association study of brain-wide imaging phenotypes: A study of the ping cohort. *Genetic Epidemiology 42*(3), 265–275.

Witten, D., R. Tibshirani, and T. Hastie (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics 10*(3), 515–534.

Yang, T., J. Wang, Q. Sun, D. P. Hibar, N. Jahanshad, L. Liu, Y. Wang, L. Zhan, P. M. Thompson, and J. Ye (2015). Detecting genetic risk factors for alzheimer's disease in whole genome sequence data via lasso screening. *2015*, 985–989.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B-statistical Methodology 68*(1), 49–67.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.