Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Distance canonical correlation analysis with application to an imaging-genetic study

Wenxing Hu
Aiying Zhang
Biao Cai
Vince Calhoun
Yu-Ping Wang

SPIE.

Wenxing Hu, Aiying Zhang, Biao Cai, Vince Calhoun, Yu-Ping Wang, "Distance canonical correlation analysis with application to an imaging-genetic study," *J. Med. Imag.* **6**(2), 026501 (2019), doi: 10.1117/1.JMI.6.2.026501.

Distance canonical correlation analysis with application to an imaging-genetic study

Wenxing Hu,[a] Aiying Zhang,[a] Biao Cai,[a] Vince Calhoun,[b] and Yu-Ping Wang[a,*]
[a]Tulane University, Department of Biomedical Engineering, New Orleans, Louisiana, United States
[b]University of New Mexico, Mind Research Network and Department of ECE, Albuquerque, New Mexico, United States

Abstract. Distance correlation is a measure that can detect both linear and nonlinear associations. However, applying distance correlation to imaging genetic studies often needs multiple testing correction due to the large number of multiple inferences. As a result, the sensitivity of its detection may be low. We propose a new model, distance canonical correlation analysis (DCCA), which overcomes this problem by searching a combination of features with the highest distance correlation. This is achieved by constructing a distance kernel function followed by solving a subsequent optimization problem. The ability to detect both linear and nonlinear associations makes DCCA suitable for analyzing complex multimodal and imaging-genetic associations. When applied to a brain imaging-genetic study from the Philadelphia Neurodevelopmental Cohort (PNC), DCCA detected several mental disorder-related gene pathways and brain networks. Experiments on brain connectivity found that the default mode network had strong nonlinear connections with other brain networks. When applied to the study of age effects, DCCA revealed that the connections of brain networks were relatively weak in younger groups but became stronger at older age stages. It indicates that adolescence is a vital stage for brain development. DCCA thus reveals a number of interesting findings and demonstrates a powerful new approach for analyzing multimodal brain imaging data. [DOI: 10.1117/1.JMI.6.2.026501]

Keywords: distance correlation; nonlinear; multimodal; functional magnetic resonance imaging; imaging genetics; brain networks.

Paper 18225R received Oct. 2, 2018; accepted for publication Mar. 22, 2019; published online Apr. 11, 2019.

1 Introduction

The brain is a complex organ and investigating its development and relationship with genomics is of great interest. Advances in neuroimaging, e.g., functional magnetic resonance imaging (fMRI), and sequencing of genetic variations, e.g., singular nucleotide polymorphism (SNP), have facilitated the analysis of the relationship between brain regions and genetic variations. FMRI detects changes in functional brain activity at each voxel, which can be clustered into regions of interest (ROI). SNPs are important genetic factors underlying differences in phenotypes among human beings. Association analyses, e.g., canonical correlation analysis (CCA),[1] have been conducted to study brain connectivity and how genetic factors and endophenotypes interact.[2] However, these methods typically use Pearson correlation which only captures linear relationships while nonlinear correlations may exist among brain regions.[3]

To address the limitation of Pearson correlation-based methods, Székely et al.[4] proposed a correlation measurement, distance correlation, which evaluates the dependence between two single variables or two sets of variables. The property that distance correlation equals 0 if and only if two variables are independent enables it to detect both linear and nonlinear associations. Besides the ability to detect nonlinear correlations, the flexibility to detect both single-single feature correlations and set–set feature correlations also help distance correlation find many applications in imaging genetic and brain connectivity study. Geerligs et al.[5] investigated the dependence between different ROIs using multivariate distance correlation and the results tended to be more robust than using Pearson correlation. Fang et al.[6] investigated complex imaging genetics associations using projected distance correlation, which was more accurate and fast.

Székely and Rizzo[7] constructed a statistic to evaluate the statistical significance of the distance correlation between two single or two sets of variables. Despite the well-constructed theoretical work, a challenge for applying distance correlation exists in multiple testing correction. Large-scale simultaneous inference testing, e.g., genome wide association study (GWAS), needs multiple testing correction, e.g., Bonferroni correction,[8] in order to prevent erroneous inferences. For distance correlation, the scale of simultaneous inference is $p \times q$ (p, q are variable sizes of two datasets), which is much larger than that of GWAS, i.e., p. As a result, it might be difficult to detect significant variable–variable distance correlations due to the harsher testing correction. For testing the distance correlation between two subsets of variables, the scale of multiple inference testing is even larger, i.e., 2^{p+q}, and consequently the detection of significant associations becomes even more difficult.

To address the challenge, we propose a new framework, distance canonical correlation analysis (DCCA), which overcomes the problem by searching a combination of original features with the highest distance correlation. It is achieved by first constructing a distance kernel function and then solving a subsequent optimization problem. In this way, DCCA can detect both linear and nonlinear correlations and can identify a subset of features that are significantly correlated.

This work is an expansion of a preliminary work, "A hybrid correlation analysis with application to imaging genetics,"[9] which was published in the proceedings of SPIE Medical Imaging 2018. This work refines the conference paper by adding more detailed procedures about the method and more applications on both the fusion of imaging genetics data and

*Address all correspondence to Yu-Ping Wang, E-mail: wyp@tulane.edu

the fusion of multiple brain imaging data. The rest of this paper is organized as follows. Section 2 first introduces distance correlation with pros and cons and then discusses how the proposed model, DCCA, can overcome the limitation. Section 3 presents a simulation experiment test to verify the performance of DCCA. Section 4 presents the collection and preprocessing of data as well as the experiments of applying DCCA to detecting imaging genetic associations and brain connectivity study. Discussion and conclusions are in Sec. 5.

2 Methods

2.1 *Distance Correlation*

Distance correlation, proposed by Székely et al.,[4] measures the dependence between two single variables or two sets of variables. Suppose we have two sets of random variables $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$ (where p, q represent the feature sizes of x, y, respectively) with characteristic functions f_x and f_y. Variable dimensionality p, q can either be 1 (two single variable case) or greater than 1 (two sets of variables case). The distance covariance between x and y is defined as

$$d\mathrm{Cov}^2(x,y) := \int_{R^{p+q}} |f_{x,y}(t,s) - f_x(t)f_y(s)|^2 (|t|_p^{p+1}|s|_q^{q+1})^{-1} \mathrm{d}t\,\mathrm{d}s, \quad (1)$$

where $|*|_p$, $|*|_q$ denote the Euclidean norm in space $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively; and $f_{x,y}$ denotes the joint characteristic function of x and y.

The distance correlation between x and y is defined as

$$d\mathrm{Cor}(x,y) := \begin{cases} \left[\frac{d\mathrm{Cov}^2(x,y)}{\sqrt{d\mathrm{Cov}^2(x,x)d\mathrm{Cov}^2(y,y)}}\right]^{\frac{1}{2}}, & \text{if } d\mathrm{Cov}^2(x,y) > 0 \\ 0, & \text{otherwise} \end{cases}. \quad (2)$$

It has been proved that distance correlation gets 0 iff x and y are independent, i.e.,

$$d\mathrm{Cor}(x,y) = 0 \Leftrightarrow x \perp\!\!\!\perp y. \quad (3)$$

Distance correlation outperforms conventional Pearson correlation in that it can detect both linear and nonlinear associations due to Eq. (3).

For sample data $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$, where n denotes sample size, the empirical distance covariance between X and Y can be estimated as follows. First, we calculate the Euclidean distance between each sample pair

$$a_{i,j} = \sqrt{\sum_{k=1}^{p}(X_{ik} - X_{jk})^2}, \quad i,j = 1,2,\cdots,n;$$

$$b_{i,j} = \sqrt{\sum_{k=1}^{q}(Y_{ik} - Y_{jk})^2}, \quad i,j = 1,2,\cdots,n.$$

Second, U-centering is applied to the Euclidean distance $a_{i,j}$ as

$$A_{i,j} = \begin{cases} a_{i,j} - \frac{1}{n-2}\sum_{l=1}^{n} a_{i,l} - \frac{1}{n-2}\sum_{k=1}^{n} a_{k,j} \\ \quad + \frac{1}{(n-1)(n-2)}\sum_{k,l=1}^{n} a_{k,l}, & i \neq j \\ 0, & i = j \end{cases}. \quad (4)$$

The U-centered $B_{i,j}$ can be calculated similarly, i.e., applying U-centering to Euclidean distance $b_{i,j}$. Then, the empirical distance correlation can be calculated as

$$d\mathrm{Cov}^2(x,y) = \frac{1}{n(n-3)}\sum_{i,j=1}^{n} A_{i,j}B_{i,j}. \quad (5)$$

A statistic following a t-distribution provided by Székely and Rizzo[7] is used to evaluate the significance of distance correlation as

$$\sqrt{\frac{n(n-3)}{2} - 1}\,\frac{d\mathrm{Cor}^2}{\sqrt{1 - d\mathrm{Cor}^4}} \to t\left[\frac{n(n-3)}{2} - 1\right], \quad p,q \to \infty. \quad (6)$$

2.2 *Kernel Methods*

Kernel methods are also widely used when data have nonlinear relationships. Kernel methods map original variable space $\mathbb{R}^p$ to a higher dimensional space $\mathbb{R}^P$ (P can be either ∞ or a number greater than p) via a mapping function ϕ as

$$\phi: x \in \mathbb{R}^p \mapsto \phi(x) \in \mathbb{R}^P. \quad (7)$$

In order to reduce computational complexity and to avoid computing in $\mathbb{R}^\infty$, kernel trick is used to compute with a kernel function instead of an explicit mapping function. A kernel function is defined as

$$k(x_1,x_2) = \langle \phi(x_1), \phi(x_2)\rangle_{\mathbb{R}^P}, \quad (8)$$

where $x_1, x_2 \in \mathbb{R}^p$ are two samples and $\langle\cdot,\cdot\rangle_{\mathbb{R}^P}$ denotes the inner product in $\mathbb{R}^P$ space.

2.3 *Distance Canonical Correlation Analysis*

Distance correlation provides a way to evaluate the dependence between two single variables or two sets of variables. Given two datasets $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$, it is of interest to identify which two single variables $x_1 \in \mathbb{R}^{n\times 1}$ and $y_1 \in \mathbb{R}^{n\times 1}$ are significantly dependent by computing their distance correlation. However, it may be difficult to detect significant distance correlations due to multiple testing correction. Multiple testing correction, e.g., Bonferroni correction,[8] is used to counteract the problem of multiple comparisons when conducting a large scale of statistical inference simultaneously, e.g., GWAS. For GWAS study, the scale of simultaneous inference is the variable/feature size p. For univariate distance correlation (distance correlation between two single variables), the scale of simultaneous inference is $p \times q$, which is much larger than that of GWAS, i.e., p.

In data application, it is usually of interest to study groups of variables rather than a single feature. For examples, complex phenotypes and diseases may be regulated by a group of genes and pathways. For brain imaging data, different brain regions function and harmonize in a connected network when performing a specific brain function.[10] Therefore, it is of interest

to identify two subsets/groups of variables $X_{\text{sub}} \in \mathbb{R}^{n\times r}$ $(1 \le r \le p)$ and $Y_{\text{sub}} \in \mathbb{R}^{n\times s}$ $(1 \le s \le q)$ which are significantly dependent. However, the scale of simultaneous inference in this case is very large, i.e., 2^{p+q}, making it more difficult to detect significantly dependent subsets.

Motivated by the problem in detecting significant distance correlation, we develop a multivariate approach, namely DCCA, to seek the optimal combination of original variables with the highest distance correlations. Given two datasets $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$, distance CCA first projects original samples to a higher dimensional space as in the following procedure.

For any two single features $x_1, x_2 \in \mathbb{R}^{n\times 1}$ from data X, a distance kernel is defined as

$$k(x_1, x_2) \coloneqq \sum_{i,j=1}^{n} |x_{1,i} - x_{1,j}||x_{2,i} - x_{2,j}|, \tag{9}$$

where $x_{*,i}$, $x_{*,j}$ denote the i'th and j'th elements of $x_*(* = 1,2)$, respectively; and the corresponding mapping function is

$$\phi : x_1 \mapsto \phi(x_1) = [\phi_1(x_1), \phi_2(x_1), \cdots, \phi_{n^2}(x_1)], \tag{10}$$

where $\phi_m(x_1) = |x_{1,i} - x_{1,j}|$, $(i = m/n,\ j = m \bmod n)$.

It is easy to check that Eq. (9) is a well-defined inner product in a reproducing kernel Hilbert space. With distance kernel constructed, a multivariate method is used to find the optimal combination of original features/variables with the highest distance correlation by solving the optimization problem as

$$\begin{aligned}(\hat{\beta}_1, \hat{\beta}_2) &= \underset{\beta_1,\beta_2}{\text{argmax}} \left[\frac{\beta_1' k(X,Y)\beta_2}{\sqrt{\beta_1' k(X,X)\beta_1}\sqrt{\beta_2' k(Y,Y)\beta_2}}\right]^{\frac{1}{2}}, \\ &= \underset{\beta_1,\beta_2}{\text{argmax}} \frac{\beta_1' k(X,Y)\beta_2}{\sqrt{\beta_1' k(X,X)\beta_1}\sqrt{\beta_2' k(Y,Y)\beta_2}},\end{aligned} \tag{11}$$

where $X \in \mathbb{R}^{n\times p}$, $Y \in \mathbb{R}^{n\times q}$, $\beta_1 \in \mathbb{R}^{p\times 1}$, $\beta_2 \in \mathbb{R}^{q\times 1}$, $K(X,Y)_{i,j} \coloneqq K(x_i, y_j)$; $K(X,Y)_{i,j}$ denotes the (i,j)'th element of $K(X,Y)$; and x_i, y_i denote the i'th column of X, Y, respectively.

The detailed algorithm for the proposed model, DCCA, and the detailed procedures of solving the optimization problem [Eq. (11)] are described in Algorithm 1.

The framework of distance CCA is similar to that of kernel CCA, which is another nonlinear methods, and therefore we call the constructed Gram matrix [Eqs. (9) and (10)] "distance kernel." However, it is noteworthy that distance CCA differs from conventional kernel CCA and cannot be regarded as kernel CCA with a newly defined kernel function. For kernel CCA, there are a number of options for kernel functions, e.g., Gaussian radial basis function kernel, polynomial kernel, etc. The choice of kernel function depends on data distributions and the hidden relationship pattern within the data. Distance kernel function [Eqs. (9) and (10)] differs from conventional kernel function in that distance kernel retains the original feature information [for $X \in \mathbb{R}^{n\times p}$, distance kernel operation $K(X,X) \in \mathbb{R}^{p\times p}$] while conventional kernel function breaks the original feature structure [for $X \in \mathbb{R}^{n\times p}$, kernel operation $K(X,X) \in \mathbb{R}^{n\times n}$]. The retaining of original feature structure enables distance CCA to perform feature selection which can facilitate subsequent result interpretation. In comparison, it is difficult to interpret the result of kernel CCA since the original feature information is lost after kernel mapping.

Algorithm 1 Algorithm for DCCA.

Input $X \in \mathbb{R}^{n\times p}$, $Y \in \mathbb{R}^{n\times q}$, initial loading vectors $\beta_1^0 \in \mathbb{R}^{p\times 1}$, $\beta_2^0 \in \mathbb{R}^{q\times 1}$

Output Optimal loading vectors $\hat{u}_1$, $\hat{u}_2$

Construct distance kernel Gram matrices

$K(X,Y)_{i,j} \leftarrow \sum_{c,d=1}^{n} |x_{i,c} - x_{i,d}||y_{j,c} - y_{j,d}|$

$K(X,X)_{i,j} \leftarrow \sum_{c,d=1}^{n} |x_{i,c} - x_{i,d}||x_{j,c} - x_{j,d}|$

$K(Y,Y)_{i,j} \leftarrow \sum_{c,d=1}^{n} |x_{i,c} - x_{i,d}||x_{j,c} - x_{j,d}|$

U-centering: $K_{i,j} \leftarrow K_{i,j} - \frac{n}{n-2}\overline{K}_{i,\cdot} - \frac{n}{n-2}\overline{K}_{\cdot,j} + \frac{n^2}{(n-1)(n-2)}\overline{K}_{\cdot,\cdot}$

Solve optimization problem [Eq. (11)]

$\hat{u}_1 \leftarrow$ the eigenvector of $K(X,X)^{-\frac{1}{2}}K(X,Y)K(Y,Y)^{-1}K(Y,X)$ $K(X,X)^{-\frac{1}{2}}$

$\hat{u}_2 \leftarrow$ the eigenvector of $K(Y,Y)^{-\frac{1}{2}}K(Y,X)K(X,X)^{-1}K(X,Y)$ $K(Y,Y)^{-\frac{1}{2}}$

return $\hat{u}_1$, $\hat{u}_2$

3 Simulation Test

To illustrate the strengths and limitations of our method, namely DCCA, we conducted a simulation study and compared the performances of DCCA to that of linear CCA. For performance comparison, two aspects were considered: correlation detection and feature selection.

3.1 *Synthetic Data*

We employed a latent variable model,[11] also used in works,[12,13] to simulate two correlated data $X \in \mathbb{R}^{n\times p}$, $Y \in \mathbb{R}^{n\times q}$, where n represents sample/subject size, and p, q represent feature size. Suppose we have two latent variables $u_1 \in \mathbb{R}^{n\times 1}$, $u_2 \in \mathbb{R}^{n\times 1}$, and u_1, u_2 are correlated. The correlation between data X and Y can be generated by loading u_1 and u_2 as follows:

$$X = E_X + u_1\alpha_1', \quad Y = E_Y + u_2\alpha_2', \tag{12}$$

where $E_X \in \mathbb{R}^{n\times p}$ and $E_Y \in \mathbb{R}^{n\times q}$ are background Gaussian noise, and $\alpha_1 \in \mathbb{R}^{p\times 1}$ and $\alpha_2 \in \mathbb{R}^{q\times 1}$ are loading vectors of latent variables.

3.2 *Three Types of Data Dependence Scenarios*

In order to perform a comprehensive comparison, three types of data dependence scenarios were considered, including independence, linear dependence, and nonlinear dependence, as shown in Figs. 1(a)–1(c). The correlation between data X and Y originates from the correlation between latent variables u_1, u_2. Therefore, the three types of correlation scenarios can be generated by enforcing different relationship patterns on u_1, u_2. Three relation patterns were used, i.e., independence, sine function, and linear function, as shown in Figs. 1(a)–1(c), respectively.

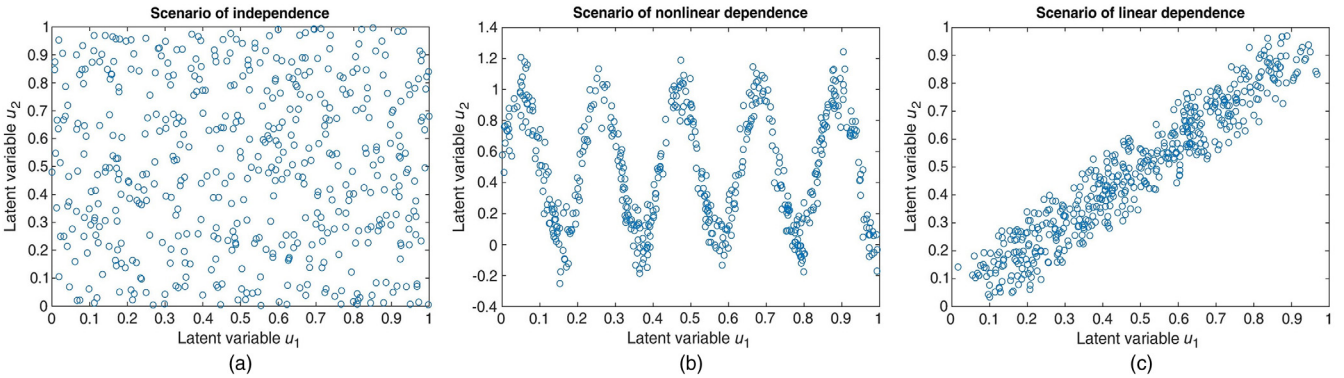

Fig. 1 Three scenarios of data dependence: (a) independence, (b) nonlinear dependence, and (c) linear dependence.

3.3 *Results of Simulation Test*

In each scenario, we implemented both CCA and our model, DCCA, to detect both correlation and true correlated features between two datasets. Note that loading vectors α_1, α_2 were sparse vectors in our experiment setting, i.e., most of the elements were zeros. The numbers of features, i.e., p, q, were 100 in our setting, among which only 20 features were set as true correlated features. That is to say, the length of loading vectors $\alpha_1 \in \mathbb{R}^{p\times 1}$ and $\alpha_2 \in \mathbb{R}^{q\times 1}$ was 100, and only 20 of their elements were nonzeros, as shown at the top of Fig. 2. An ideal method should be able to accurately detect the cross-data correlation and also the 20 true correlated features.

The results are shown in Figs. 2, 3, 4, for the three scenarios, respectively. In each figure, the top two subfigures represent the ground truth of the true correlated features, and the bottom four subfigures represent the identified features by CCA and DCCA, respectively. From Fig. 2, when two data are independent, both CCA and DCCA detect a weak correlation (CCA: 0.0739 versus DCCA: 0.1019) and neither method can identify true correlated features. From Fig. 4, when two data follow a linear relationship, both CCA and DCCA can detect a strong correlation (CCA: 0.9807 versus DCCA: 0.9525) and both methods can accurately identify the true correlated features. From Fig. 3, when two data follow a nonlinear relationship, CCA cannot detect the correlation (CCA: 0.0886) and cannot identify the true correlated features. In comparison, DCCA can detect the nonlinear correlation (DCCA: 0.6772) and also the true correlated features. The results in the three scenarios, i.e., Figs. 1–4, verified the superior performance of DCCA over conventional CCA in terms of detecting both complex correlations and true correlated features.

4 Application to Brain Imaging Data

4.1 *Brain Imaging Data and Brain Connectivity*

The DCCA was then applied to a brain development study focused on two experiments. One experiment is to study the imaging-genetic associations (Sec. 4.3) and the other one is to study the connections between different brain subnetworks or subdomains, e.g., default mode network (DMN), and how the connections change across different age stages (Sec. 4.5). Imaging-genetic study analyzes the correlation between fMRI data, which detects the change of the brain functional activity at voxel level and SNPs data. SNPs are important genetic factors underlying differences in phenotypes among human beings. Genetic factors may function as a complicated group, e.g., protein–protein interaction network, gene pathway, when regulating a certain phenotype or disease. Similarly, neurons and brain regions also function and harmonize in a connected network when performing a specific brain function.[10] Therefore, distance CCA, which seeks the optimal combinations of features with the strongest cross-data associations, might be superior in detecting group–group nonlinear associations between brain imaging scans and genetic factors.

4.2 *Data Collection and Preprocessing*

The Philadelphia Neurodevelopmental Cohort (PNC)[15] is a large-scale collaborative study between the Brain Behavior Laboratory at the University of Pennsylvania and the Children's Hospital of Philadelphia. The data include fMRI and SNPs data of adolescents aged from 8 to 21 years. The fMRI data were collected during a resting state from 857 subjects. After the collection of raw fMRI data, SPM12[16] was used to conduct motion correction, spatial normalization, spatial smoothing with a 3×3-mm Gaussian kernel, and multiple regression to mitigate the influence of motion. Finally, 264 ROIs (containing 21,384 voxels) were extracted based on the power coordinates[17] with a sphere radius parameter of 5 mm. SNPs data were collected from 7863 subjects based on four platforms, Illumina Human610Quadv1, HumanHap550v1, HumanHap550v3, and HumanOmniExpress. SNPs with >5% missing values were deleted and the rest missing values were further imputed using Plink.[18,19] Then, the SNPs within gene bodies were kept, resulting in 95,639 SNPs.

4.3 *Imaging-Genetic Associations*

In order to implement distance CCA, the subjects having both fMRI and SNP data are further extracted, resulting in 855 subjects. For fMRI data, the stimulus-on versus stimulus-off contrast was obtained from the raw resting-state time series data. To find the interactions that are more related to mental disorders, SNPs located in genes associated with brain disorders were kept, where the brain disorders included schizophrenia, bipolar disorder, depression, attention-deficit/hyperactivity disorder, and post-traumatic stress disorder. Finally, 736 genes containing 21,487 SNPs were left for further analysis.

When applied to detect the group associations between fMRI and SNPs, distance CCA identified a subset of 45

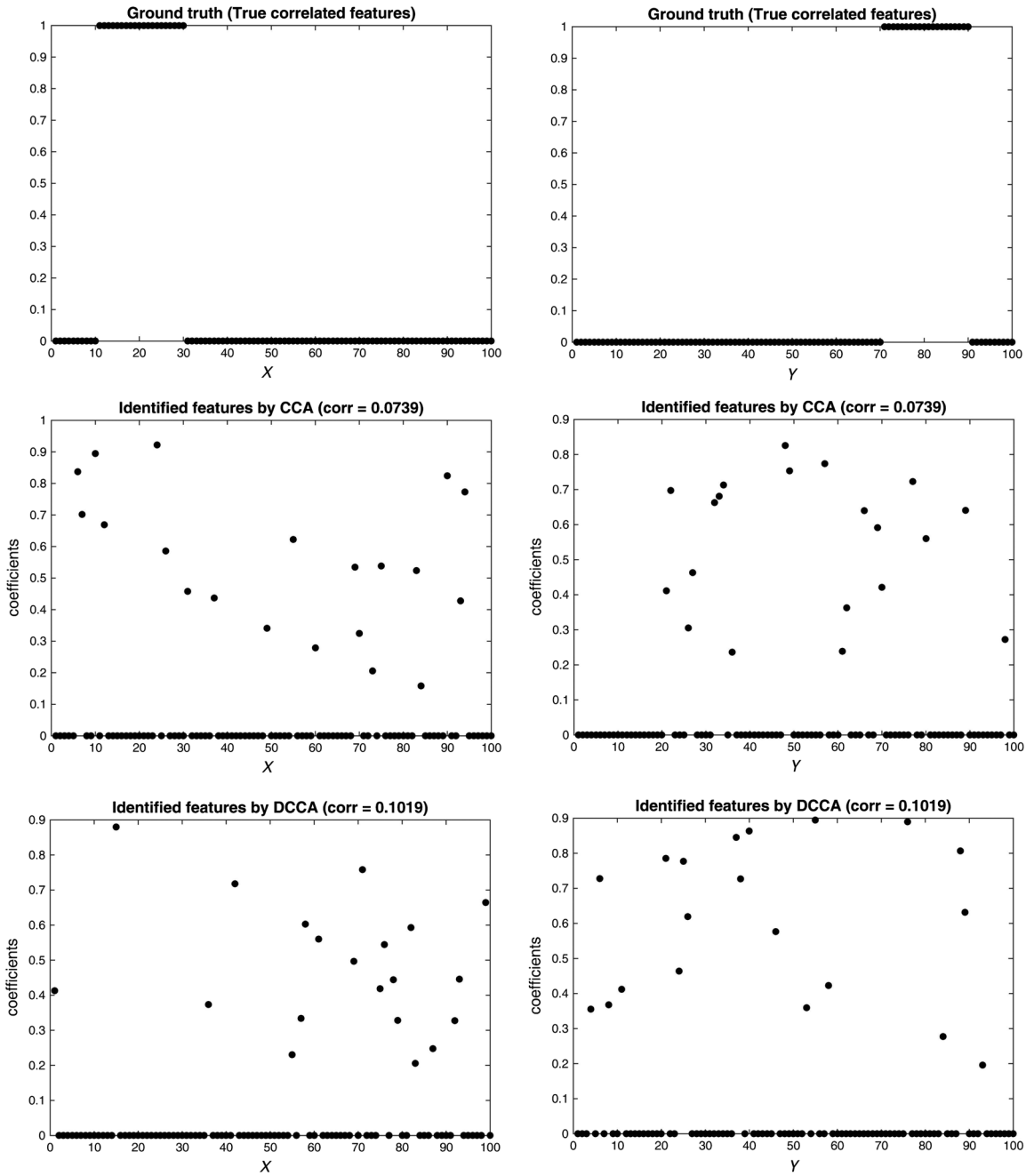

Fig. 2 Performance comparison between CCA and DCCA [independence scenario: Fig. 1(a)].

genes and a subset of 15 ROIs that were strongly correlated. The distance correlation between the identified ROIs and genes was 0.2047 with p-value of $6.58e-30$ (calculated based on Eq. 6). In comparison, the largest single ROI–gene distance correlation is 0.1759 with p-value of $1.29e-18$. This demonstrated that distance covariance-based CCA can find a pair of variable groups with an enhanced distance correlation and significance level. The lists of the identified genes and ROIs are in Tables 1 and 2, respectively. The locations of the identified ROIs are further visualized in Fig. 5 using the BrainNet Viewer toolbox.[14,20]

After that, gene enrichment analysis was conducted to reveal the underlying biological functions of the identified genes. Ten pathways were selected with a screening of q-value <0.05 (q-value represents the multiple testing corrected p-value), and the pathways together with their corresponding q-values were listed in Table 3. P-values are calculated using the hypergeometric test based on the numbers of genes in the particular biological pathway and the identified gene set. The q-values are then calculated by correcting the p-values using multiple testing correction, e.g., Bonferroni correction,[8] based on the false discovery rate method. Among the identified pathways, pathways "neurodegenerative diseases," "oxidative damage," and "deregulated CDK5 triggers multiple neurodegenerative pathways in Alzheimer's disease models" have been reported to be related to neuron activities and brain development. Pathway "neurodegenrative diseases" is related to the death of neurons and corticobasal degeneration, which might further lead to the progressive

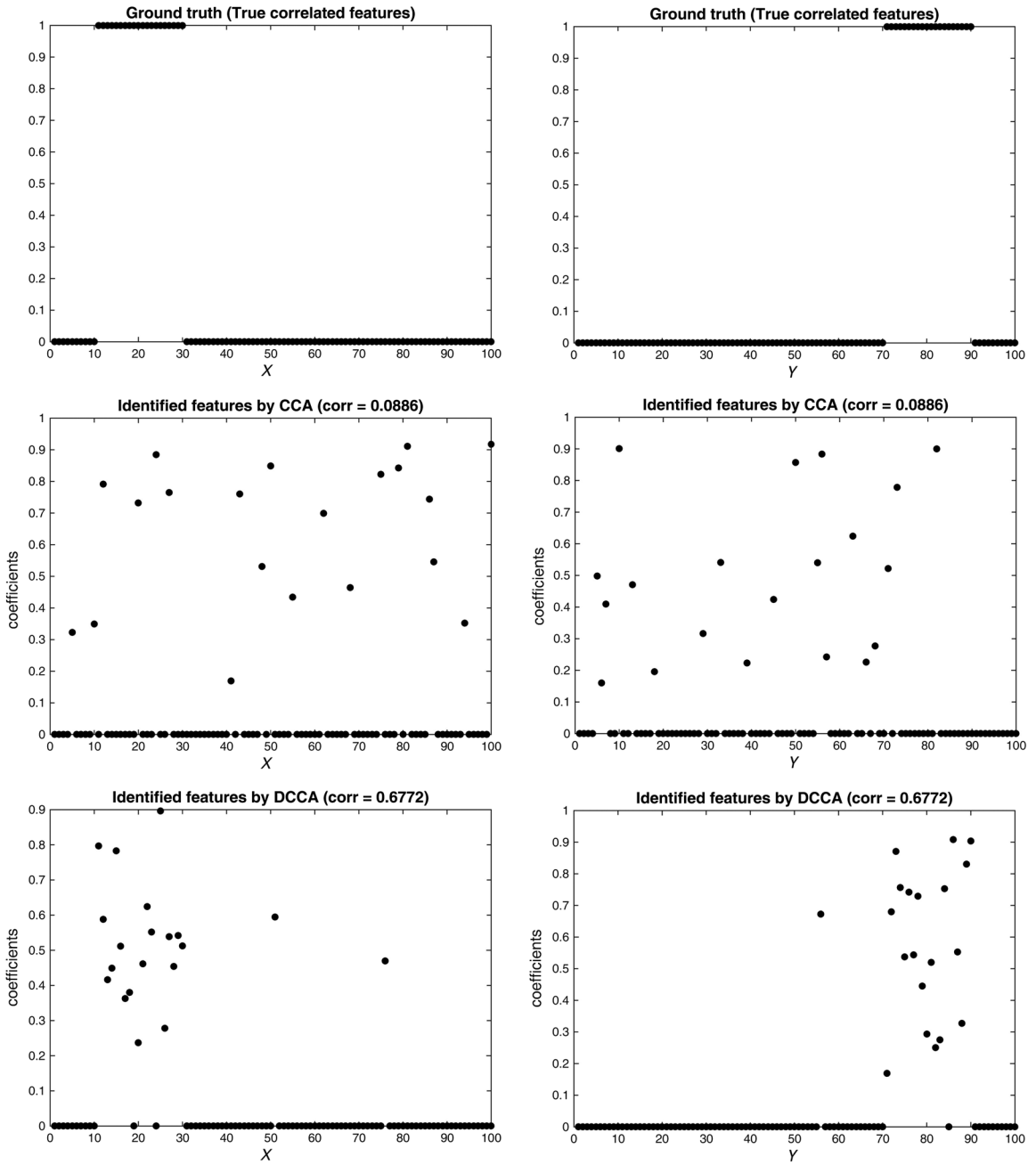

Fig. 3 Performance comparison between CCA and DCCA [nonlinear dependence scenario: Fig. 1(b)].

dysfunction in the brain and a number of mental disorders.[21] Pathway "oxidative damage," which is related to cell signaling, may lead to damage of cell and the death of neurons.[22] It may be related to the pathogenesis of several neural degenerative diseases, including Parkinson's disease,[23] depression,[24] and Alzheimer's disease.[25] For pathway "deregulated CDK5 triggers multiple neurodegenerative pathways in Alzheimer's disease models," abnormal CDK5 may result in unregulated activation of the cycle of cell,[26] which might further lead to the death of neurons.[27] Mental disorders, such as Alzheimer's disease, may occur if CDK5 is deregulated.[28] The interactions of pathway "neurodegenerative diseases" and "deregulated CDK5 triggers multiple neurodegenerative pathways in Alzheimer's disease models" are visualized in Figs. 6 and 7, respectively. Figure 6 was plotted using Cytoscape software,[29] which was an open source platform for visualizing complex networks. Figure 7 was generated using reactome pathway database.[30]

For brain imaging data, as shown in Table 2, the majority (13/15) of the detected ROIs are from three brain subdomains: sensorimotor network (SM), DMN, and visual network (VIS). SM is related to the coordination of the body when performing motor tasks.[31] DMN is the dominant network when subjects are in resting state, mind-wandering, or not involved in a specific task. Dysfunction within the DMN has been associated with several mental disorders,[32,33] e.g., schizophrenia, depression, autism, etc. Associations between DMN and genetic factors exist according to a multivariate study of schizophrenia subjects scanned during the resting state.[34]

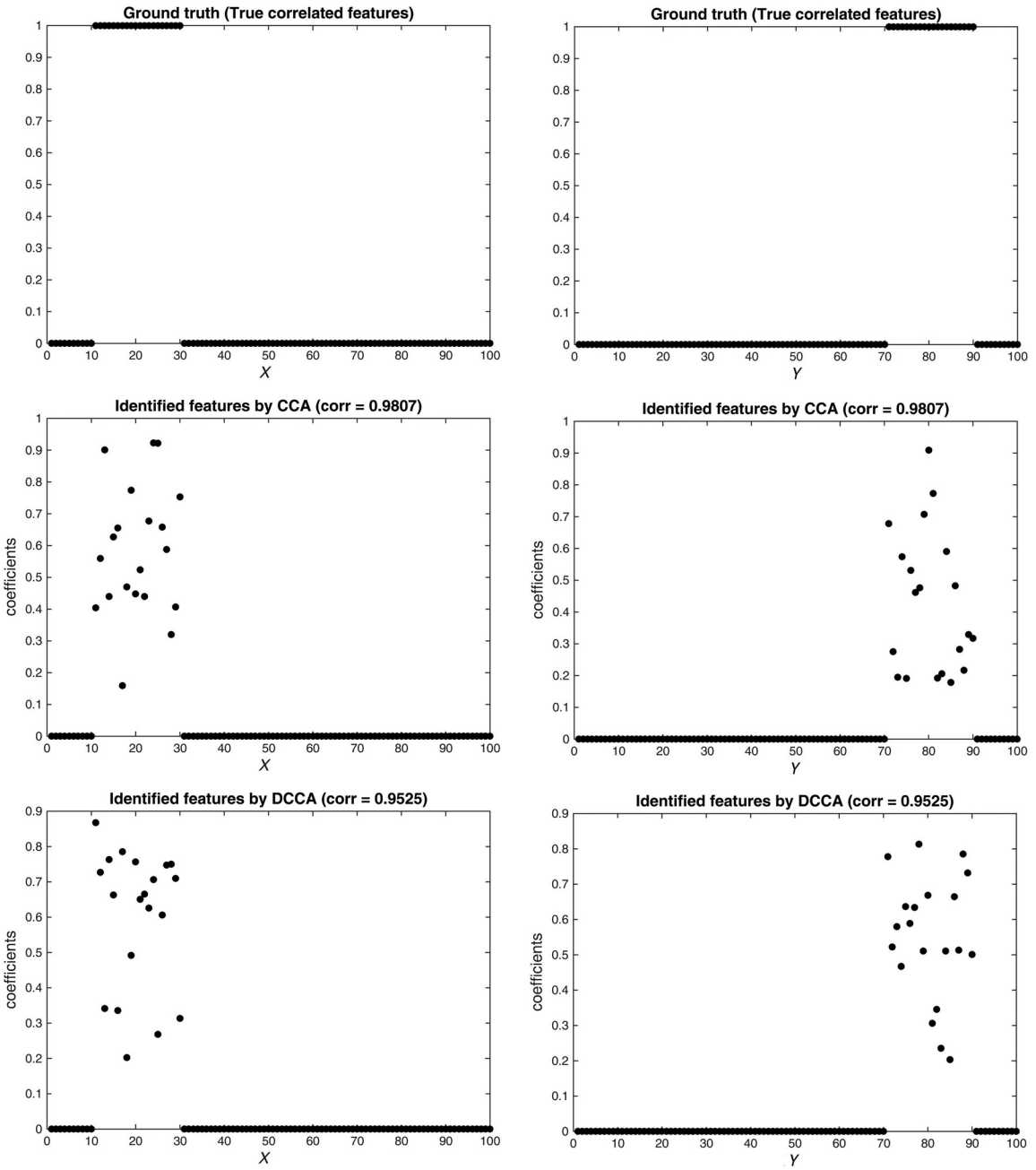

Fig. 4 Performance comparison between CCA and DCCA [linear dependence scenario: Fig. 1(c)].

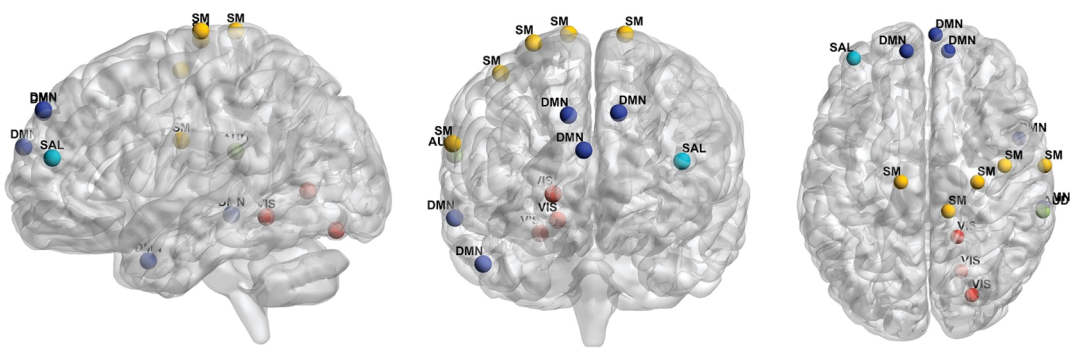

Fig. 5 The sagittal, coronal, and axial views of the identified brain ROIs. Figures were drawn using the BrainNet viewer toolbox.[14]

Table 1 The genes identified by DCCA.

Gene	Gene	Gene	Gene	Gene
RERE	OPRD1	FAF1	MIR137HG	CADM3
AKT3	LOC101929452	CYP26B1	SLC4A5	INPP4A
HAT1	CIR1	ZNF330	CLCN3	CTC-436P18.1
LOC102467655	MEF2C-AS1	CDC25C	CAP2	SP4
FAM126A	ATP6V1B2	BNIP3L	LETM2	C8orf87
NAPRT1	NT5C2	EIF3F	MARK2	TRPT1
ZNF202	SIAE	NRGN	PAWR	C12orf76
MAP3K9	TRAF3	CALB2	YWHAE	SRR
PRRG2	DNMT3B	ARHGAP40	KCNS1	YPEL1

4.4 *Functional Connectivity Between Brain Subnetworks*

For brain FC study, we selected five brain subnetworks or subdomains and then applied distance CCA to investigate the connections between each subnetwork pair and to study the age effects on the connections. Resting-state fMRI was used in this experiment and data were preprocessed using group ICA of fMRI toolbox[35] for independent component analysis (ICA).[36] The five brain subnetworks include SM, VIS, cognitive control network (CCN), auditory network (AUD), and DMN, and the corresponding locations in the brain are shown in Fig. 8.

In order to investigate both linear and nonlinear connections of the brain, we applied both linear CCA and distance CCA to the PNC data and the results are shown in Fig. 9. The results were based on a 10-fold cross-validation, in which each time five folds were used as training data and the rest five folds were used as testing data. It is noteworthy that the metric of distance correlation is different from that of linear Pearson correlation, e.g., distance correlation $=0.4 \not\Leftrightarrow$ Pearson correlation $=0.4$. Nevertheless, distance correlation reflects the relative

Table 2 The identified brain ROIs. *X*, *Y*, *Z* represent ROI coordinates in the Montreal Neurological Institute (MNI) space.

X	*Y*	*Z*	ROI name	Suggested system
13	−33	75	Postcentral gyrus	Sensory/somatomotor hand
29	−17	71	Precentral gyrus	Sensory/somatomotor hand
44	−8	57	Precentral gyrus	Sensory/somatomotor hand
−13	−17	75	Precentral gyrus	Sensory/somatomotor hand
66	−8	25	Precentral gyrus	Sensory/somatomotor mouth
65	−33	20	Superior temporal gyrus	Auditory
13	55	38	Superior frontal gyrus	Default mode
−10	55	39	Superior frontal gyrus	Default mode
6	64	22	Medial frontal gyrus	Default mode
65	−31	−9	Middle temporal gyrus	Default mode
52	7	−30	Middle temporal gyrus	Default mode
18	−47	−10	Parahippocampa gyrus	Visual
20	−66	2	Lingual gyrus	Visual
26	−79	−16	Lingual gyrus	Visual
−39	51	17	Superior frontal gyrus	Salience

Table 3 Gene enrichment analysis of the identified genes. *Q*-values represent multiple testing corrected *p*-value.

Pathway name	Source	*p*-value	*q*-value
Chk1/Chk2(Cds1)-mediated inactivation of cyclin B:Cdk1	Reactome	0.00032	0.012
Activation of BAD and translocation to mitochondria	Reactome	0.00044	0.013
Deregulated CDK5 triggers multiple neurodegenerative	Reactome	0.00063	0.013
Pathways in Alzheimer's disease models			
Neurodegenerative diseases	Reactome	0.00063	0.013
TNFalpha	NetPath	0.0013	0.021
Activation of BH3-only proteins	Reactome	0.0018	0.023
Class I PI3K signaling events mediated by Akt	PID	0.0024	0.027
Oxidative damage	Wikipathways	0.0031	0.029
LKB1 signaling events	PID	0.0036	0.029
Intrinsic pathway for apoptosis	Reactome	0.0036	0.029

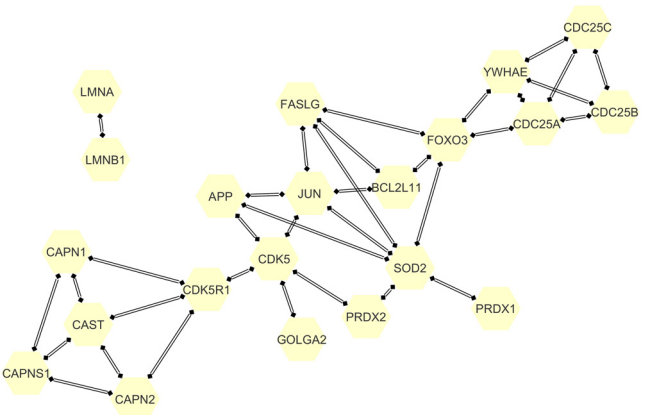

Fig. 6 The interaction mechanisms of the pathway "neurodegenerative diseases."

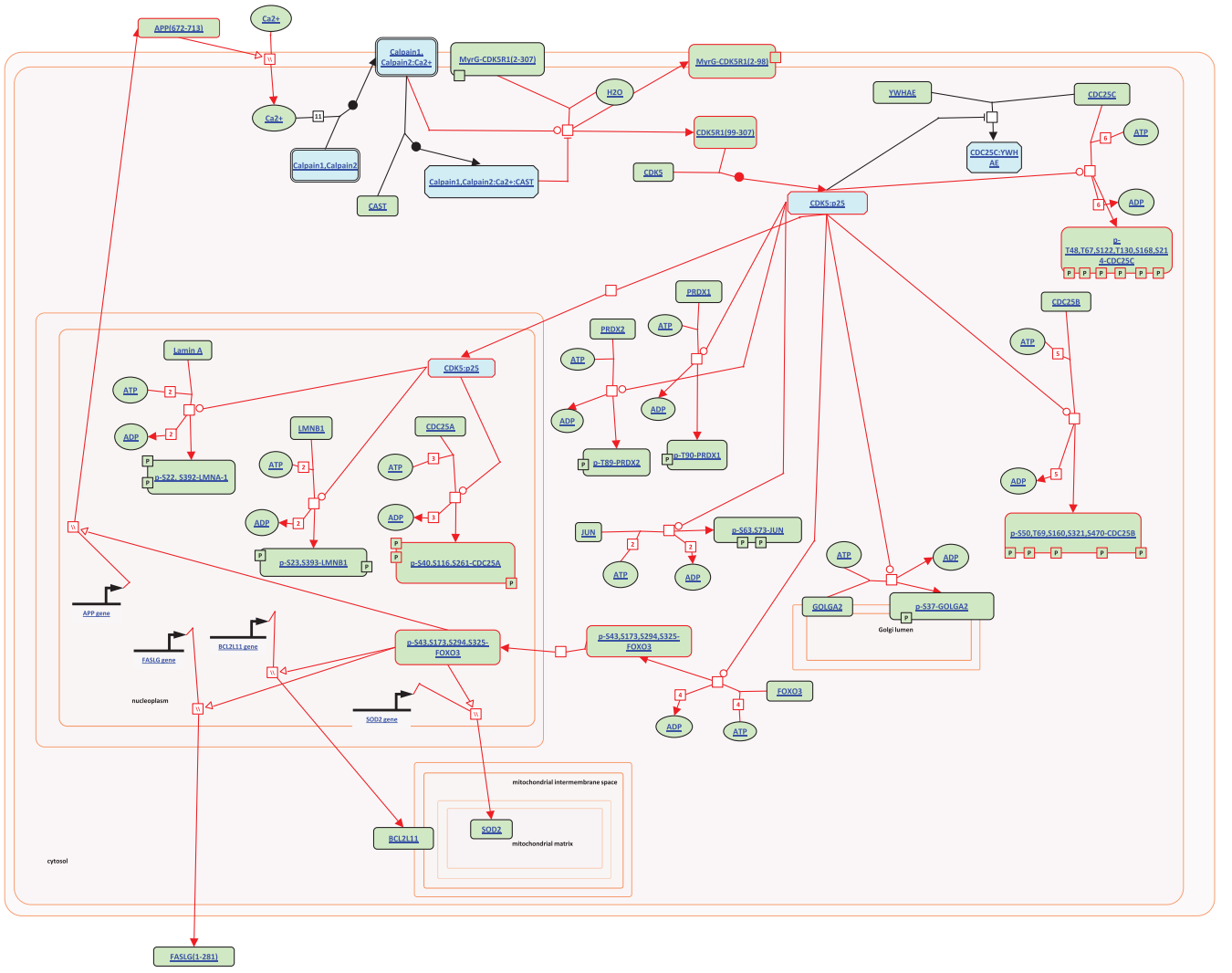

Fig. 7 The interaction mechanisms of the pathway "deregulated CDK5 triggers multiple neurodegenerative pathways in Alzheimer's disease models."

strength of the dependence between two variables. From Fig. 9, strong linear connections are detected between each pair of SM, VIS, CCN, and AUD networks, while the linear connections between DMN and other networks are weak. Research[32] has shown that DMN may have strong intrinsic connections while the connections between DMN and the rest networks are weak in the resting state, which is consistent with the result of linear CCA. In comparison, distance CCA detected stronger DMN-SM, DMN-CCN, and DMN-AUD connections, which might be a new discovery.

4.5 *Ages Effects on Brain FC*

It is of interest to investigate how brain connectivity changes during adolescence and how it changes across different age stages, e.g., children and young adults, which may further contribute to the study of normal and pathological brain development. Three age groups, 8 to 11 years, 13 to 16 years, and 18 to 22 years, were selected and then distance CCA was applied to each age group to analyze brain network connections. Subjects aged 12 and 17 years were not included in the experiments in order to get a clear boundary between different age groups. The connections between brain subnetworks for each age group are shown in Fig. 10. From Fig. 10, the patterns of the connections are different between different age groups. For instance, the connections between different brain networks are relatively weaker at age 8 to 11 but become relatively stronger at age 13 to 16 and age 18 to 22. It demonstrates that different brain regions become more and more connected during adolescence, which may be a result of the training and development of the brain during multiple types of brain activities. Moreover, it seems that the connections between CCN and SM are weak across all three age groups, which indicates that the connection between CCN and SM may be weak at the adolescent stage.

5 Discussion and Conclusion

In this work, we proposed a new model, DCCA, which overcomes the limitation of distance correlation in detecting significant associations when feature size is large. Conventional distance correlation analysis needs large-scale multiple testing when testing feature–feature association simultaneously. The proposed model, DCCA, addresses the problem by searching a combination of original features with the highest distance correlation. It is achieved by first constructing a distance kernel

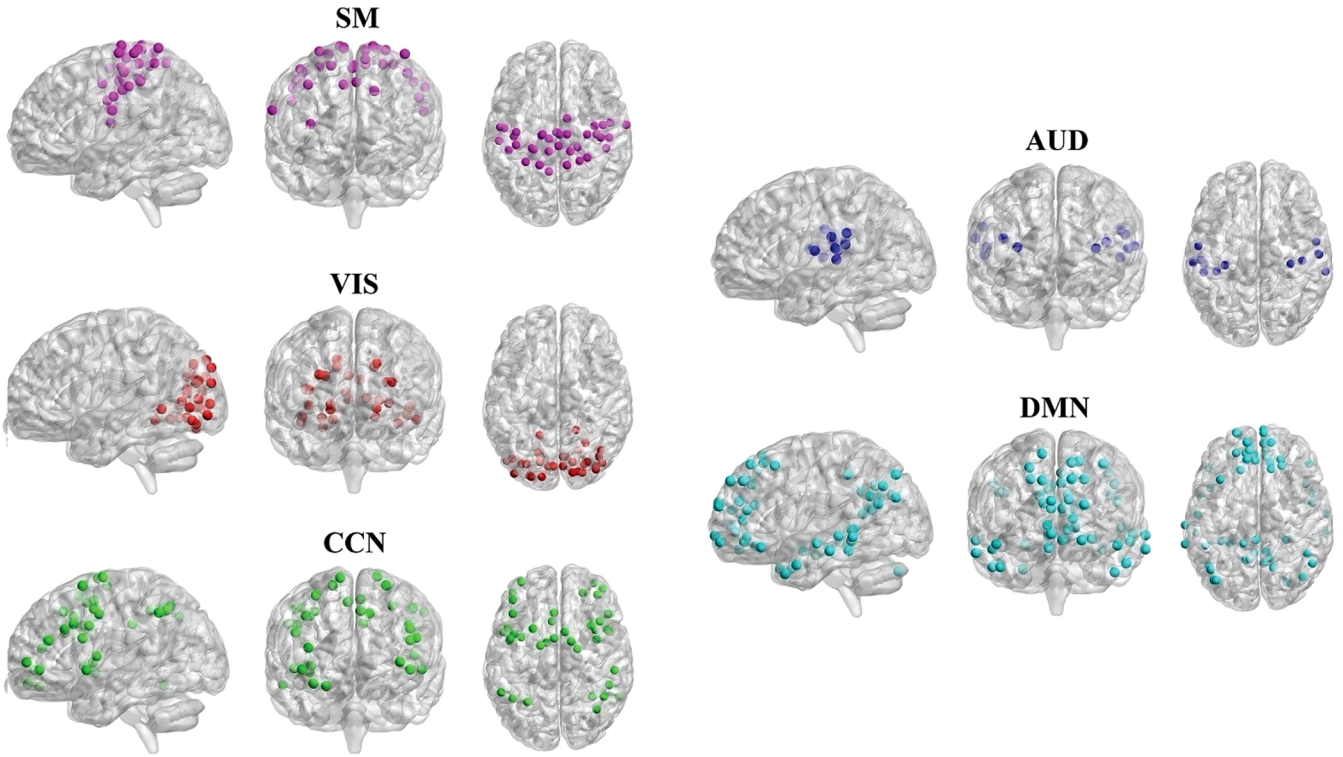

Fig. 8 The sagittal, coronal, and axial views of brain functional network domains extracted via group ICA. The names of the brain network domains are: SM, AUD, VIS, DMN, CCN, and salience network (SAL).

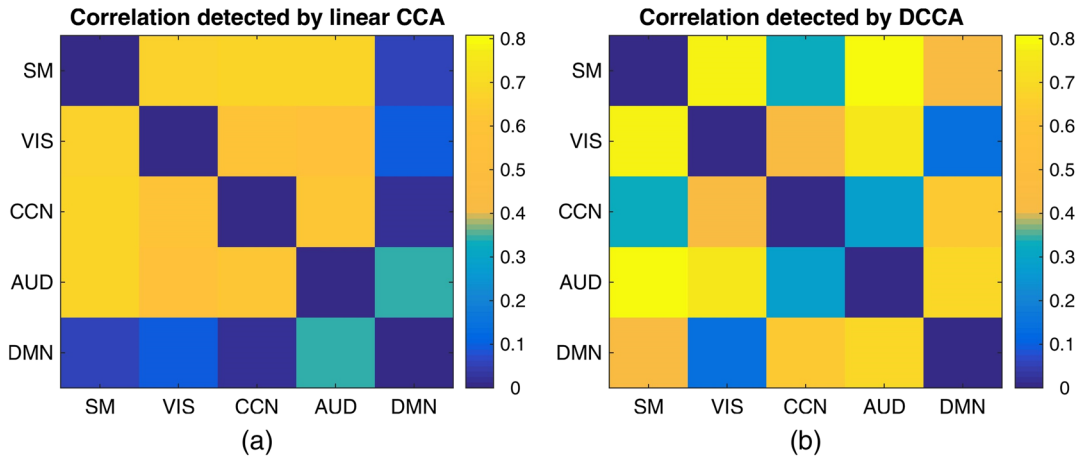

Fig. 9 The heatmap showing the correlations between brain networks. (a) The results by linear CCA and (b) the results by DCCA. The color bar indicates the value of detected correlations.

function and then solving an optimization problem. The ability to detect nonlinear group–group associations makes DCCA more suitable for analyzing complex multi-omics and imaging-genetic associations, in which both genetic factors and brain ROIs may work as groups when regulating a phenotype or performing a specific brain function.

When applied to imaging-genetic association study, DCCA detected a strong correlation between a subset of genes and a subset of brain ROIs with an improved significance level. Several neuron degeneration and mental disorder related pathways were enriched from the identified genes after gene enrichment analysis, which demonstrated the biological significance of our findings. In addition, DCCA found several mental disorder-related brain networks which had been reported by existing literature. Experiments on brain connectivity study also found several new discoveries using DCCA. Brain network DMN, which is considered to be distinct from other brain domains/networks, may have strong nonlinear connections with other brain networks according to the results of DCCA. When applied to analyzing each age groups, DCCA reveals that younger groups (8 to 11 years) exhibit weak connections of brain networks while the connections become strong at an older age stage (13 to 16 and 18 to 22) which may a result of brain development. The discoveries of imaging genetic associations and brain connections verified the performance of DCCA. Besides the examples in this study, it may find more

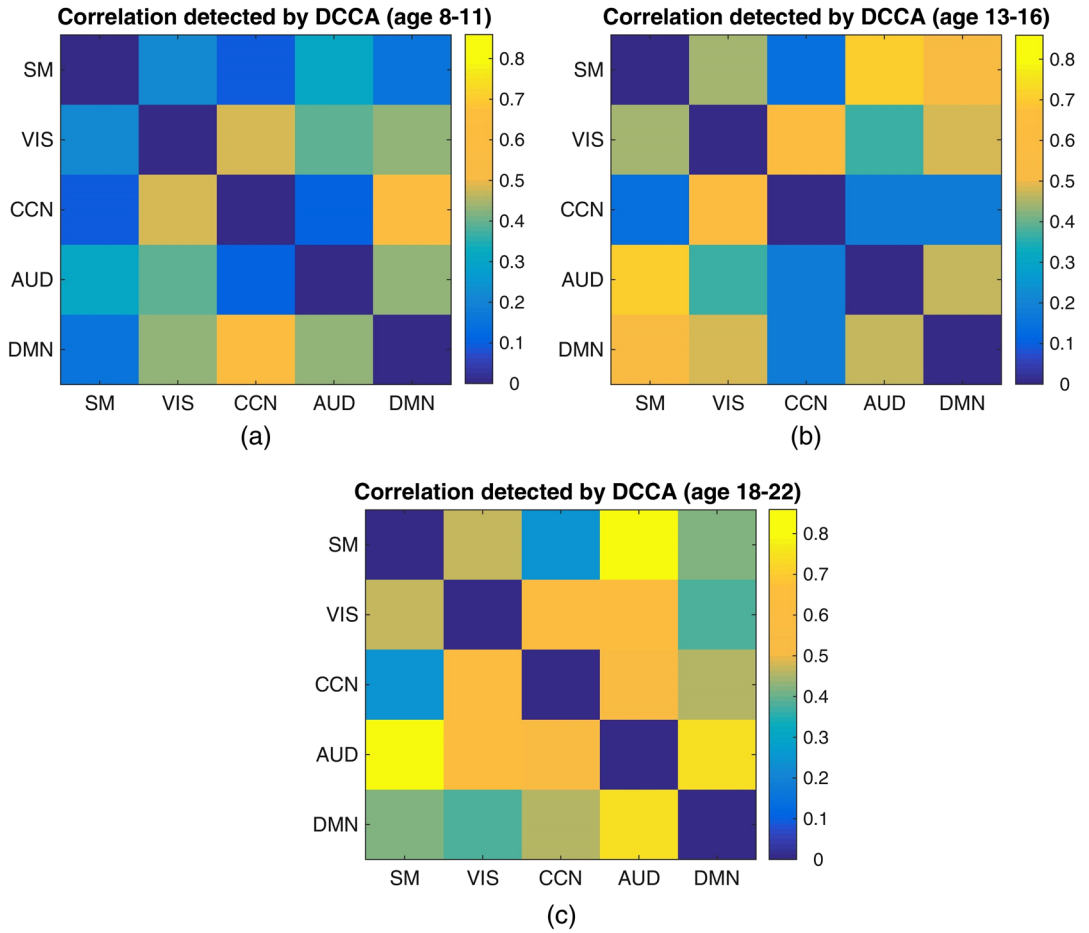

Fig. 10 The heatmap showing age differences in brain connectivity links in resting state. (a) The network connection for age group 8 to 11 years; (b) the network connection for age group 13 to 16 years; and (c) the network connection for age group 18 to 22 years. The color bar indicates the value of detected correlations.

applications in multiimaging and multi-omics studies, where identifying correlations between multiple datasets is a common challenge.

Disclosures

The authors have no relevant financial interests in the paper and no other potential conflicts of interest to disclose.

Acknowledgments

The authors would like to thank the NIH (P30 GM122734, R01 GM109068, R01 MH104680, R01 MH107354, P20 GM103472, R01 REB020407, and R01 EB006841) and NSF (#1539067) for partial support.

References

1. H. Hotelling, "Relations between two sets of variates," *Biometrika* **28**(3/4), 321–377 (1936).
2. J. Fang et al., "Joint sparse canonical correlation analysis for detecting differential imaging genetics modules," *Bioinformatics* **32**(22), 3480–3488 (2016).
3. J. Hlinka et al., "Functional connectivity in resting-state fMRI: is linear correlation sufficient?" *Neuroimage* **54**(3), 2218–2225 (2011).
4. G. J. Székely et al., "Measuring and testing dependence by correlation of distances," *Ann. Stat.* **35**(6), 2769–2794 (2007).
5. L. Geerligs et al., "Functional connectivity and structural covariance between regions of interest can be measured more accurately using multivariate distance correlation," *NeuroImage* **135**, 16–31 (2016).
6. J. Fang et al., "Fast and accurate detection of complex imaging genetics associations based on greedy projected distance correlation," *IEEE Trans. Med. Imaging* **37**(4), 860–870 (2018).
7. G. J. Székely and M. L. Rizzo, "The distance correlation t-test of independence in high dimension," *J. Multivar. Anal.* **117**, 193–213 (2013).
8. S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.* **6**(2), 65–70 (1979).
9. W. Hu et al., "A hybrid correlation analysis with application to imaging genetics," *Proc. SPIE* **10579**, 1057905 (2018).
10. M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: uses and interpretations," *Neuroimage* **52**(3), 1059–1069 (2010).
11. K. G. Jöreskog and A. S. Goldberger, "Estimation of a model with multiple indicators and multiple causes of a single latent variable," *J. Am. Stat. Assoc.* **70**(351a), 631–639 (1975).
12. E. Parkhomenko, D. Tritchler, and J. Beyene, "Sparse canonical correlation analysis with application to genomic data integration," *Stat. Appl. Genet. Mol. Biol.* **8**(1), 1–34 (2009).
13. D. Lin, V. D. Calhoun, and Y.-P. Wang, "Correspondence between fmri and SNP data by group sparse canonical correlation analysis," *Med. Image Anal.* **18**(6), 891–902 (2014).
14. M. Xia, J. Wang, and Y. He, "BrainNet Viewer: a network visualization tool for human brain connectomics," *PloS ONE* **8**(7), e68910 (2013).

15. T. D. Satterthwaite et al., "Neuroimaging of the Philadelphia Neuro-Developmental Cohort," *Neuroimage* **86**, 544–553 (2014).
16. K. Friston et al., "Statistical parametric mapping," http://www.fil.ion.ucl.ac.uk/spm/software/spm12/ (25 September 2018).
17. J. D. Power et al., "Functional network organization of the human brain," *Neuron* **72**(4), 665–678 (2011).
18. C. Chang et al., "PLINK," https://www.cog-genomics.org/plink/2.0/ (1 April 2019).
19. S. Purcell et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am. J. Human Genet.* **81**(3), 559–575 (2007).
20. M. Xia et al., "BrainNet Viewer," https://www.nitrc.org/projects/bnv/ (7 July 2011).
21. W. W. Seeley et al., "Neurodegenerative diseases target large-scale human brain networks," *Neuron* **62**(1), 42–52 (2009).
22. M. D. Evans and M. S. Cooke, "Factors contributing to the outcome of oxidative damage to nucleic acids," *Bioessays* **26**(5), 533–542 (2004).
23. O. Hwang, "Role of oxidative stress in Parkinson's disease," *Exp. Neurobiol.* **22**(1), 11–17 (2013).
24. S. Jiménez-Fernández et al., "Oxidative stress and antioxidant parameters in patients with major depressive disorder compared to healthy controls before and after antidepressant treatment: results from a meta-analysis," *J. Clin. Psychiatry* **76**, 1658–1667 (2015).
25. M. Valko et al., "Free radicals and antioxidants in normal physiological functions and human disease," *Int. J. Biochem. Cell Biol.* **39**(1), 44–84 (2007).
26. J. Lopes, C. Oliveira, and P. Agostinho, "Cell cycle re-entry in Alzheimer's disease: a major neuropathological characteristic?" *Curr. Alzheimer Res.* **6**(3), 205–212 (2009).
27. K.-H. Chang, F. Vincent, and K. Shah, "Deregulated cdk5 triggers aberrant activation of cell cycle kinases and phosphatases inducing neuronal death," *J. Cell Sci.* **125**(21), 5124–5137 (2012).
28. Y. Yang, E. J. Mufson, and K. Herrup, "Neuronal cell death is preceded by cell cycle events at all stages of Alzheimer's disease," *J. Neurosci.* **23**(7), 2557–2563 (2003).
29. P. Shannon et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res.* **13**(11), 2498–2504 (2003).
30. D. Croft et al., "The reactome pathway knowledgebase," *Nucleic Acids Res.* **42**(D1), D472–D477 (2013).
31. S. Chenji et al., "Investigating default mode and sensorimotor network connectivity in amyotrophic lateral sclerosis," *PLoS ONE* **11**(6), e0157443 (2016).
32. R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter, "The brain's default network," *Ann. New York Acad. Sci.* **1124**(1), 1–38 (2008).
33. T. J. Akiki et al., "Default mode network abnormalities in posttraumatic stress disorder: a novel network-restricted topology approach," *NeuroImage* **176**, 489–498 (2018).
34. S. A. Meda et al., "Multivariate analysis reveals genetic associations of the resting default mode network in psychotic bipolar disorder and schizophrenia," *Proc. Natl. Acad. Sci. U. S. A.* **111**(19), E2066–E2075 (2014).
35. V. D. Calhoun et al., "Group ICA of fMRI Toolbox (GIFT)," http://mialab.mrn.org/software/gift/ (20 Feburary 2017).
36. V. D. Calhoun and T. Adali, "Multisubject independent component analysis of fMRI: a decade of intrinsic networks, default mode, and neurodiagnostic discovery," *IEEE Rev. Biomed. Eng.* **5**, 60–73 (2012).

Wenxing Hu received his BS degree in applied mathematics from Xi'an Jiaotong University, China, 2011. Now, he is a PhD student in biomedical engineering, Tulane University, USA. His research interests include machine learning and deep learning, dimension reduction, correlation analysis, and multi-omics data integration.

Aiying Zhang received her BS degree in statistics from the University of Science and Technology of China. She is now a PhD student in the Department of Biomedical Engineering, Tulane University. Her research interests mainly focus on graphical models (directed and undirected) with applications in multi-omics data integration.

Biao Cai received his BS and MS degrees in biomedical engineering from Tianjin University, China, in 2013 and 2016, respectively. Now, he is a PhD student in biomedical engineering, Tulane University, USA. His research interests include dictionary learning and time-varying graphical LASSO, dynamic function network connectivity, and brain development.

Vince Calhoun is currently president for the MRN, and a distinguished professor in the ECE Department, University of New Mexico. He has published over 600 journal articles. His work includes ICA-based fMRI analysis, and data fusion of multimodal-imaging and genetics data. He leads an NIH P20 COBRE grant on multimodal imaging of mental disorders and an NSF EPSCoR grant focused on brain imaging and epigenetics of adolescent development. He is a fellow of the American Association for the Advancement of Science, the American Institute of Biomedical and Medical Engineers, the American College of Neuropsychopharmacology, and the International Society of Magnetic Resonance in Medicine.

Yu-Ping Wang received his BS degree from Tianjin University in 1990, and his MS and PhD degrees from Xian Jiaotong University in 1993 and 1996, respectively. He is currently a professor of biomedical engineering at Tulane University. His research interests include computer vision, signal processing, and machine learning with applications to biomedical imaging and bioinformatics, where he has published about 200 publications. He has served on numerous NSF/NIH review panels, and as editor for several journals.