

# Identify Consistent Cross-Modality Imaging Genetic Patterns via Discriminant Sparse Canonical Correlation Analysis

Meiling Wang<sup>1</sup>, Wei Shao<sup>1</sup>, Xiaoke Hao<sup>1</sup>, Li Shen<sup>2</sup>, and Daoqiang Zhang<sup>1</sup>

**Abstract**—Sparse canonical correlation analysis (SCCA) is a bi-multivariate technique used in imaging genetics to identify complex multi-SNP-multi-QT associations. However, the traditional SCCA algorithm has been designed to seek a linear correlation between the SNP genotype and brain imaging phenotype, ignoring the discriminant similarity information between within-class subjects in brain imaging genetics association analysis. In addition, multi-modality brain imaging phenotypes are extracted from different perspectives and imaging markers from the same region consistently showing up in multimodalities may provide more insights for the mechanistic understanding of diseases. In this paper, a novel multi-modality discriminant SCCA algorithm (MD-SCCA) is proposed to overcome these limitations as well as to improve learning results by incorporating valuable discriminant similarity information into the SCCA algorithm. Specifically, we first extract the discriminant similarity information between within-class subjects by the sparse representation. Second, the discriminant similarity information is enforced within SCCA to construct a discriminant SCCA algorithm (D-SCCA). At last, the MD-SCCA algorithm is adopted to fully explore the relationships among different modalities of different subjects. In experiments, both synthetic dataset and real data from the Alzheimer's Disease Neuroimaging Initiative database are used to test the performance of our algorithm. The empirical results have demonstrated that the proposed algorithm not only produces improved cross-validation performances but also identifies consistent cross-modality imaging genetic biomarkers.

**Index Terms**—Imaging genetics, sparse canonical correlation analysis, multi-modality, multi-SNP, Alzheimer's disease

## 1 INTRODUCTION

ALZHEIMER'S disease (AD) is the most common form of dementia. Early diagnosis and effective prevention are important research topics for AD [1], [2], [3]. At present, brain imaging genetics is an emerging field of study in brain research [4], [5], [6], [7], [8], [9], [10], [11], [12]. It aims to examine the association between genetic markers such as single nucleotide polymorphisms (SNPs) [13] and quantitative traits (QTs) extracted from multimodal neuroimaging data (e.g., structured, functional, and molecular imaging scans).

An increasing amount of high-dimensional biomedical data, such as genome sequencing or brain imaging scans, are collected every day. Bridging these two factors and exploring their connections have the potential to assist a better mechanistic understanding of normal or disordered brain functions. The complexity of these data, however, has presented critical bioinformatics challenges requiring new enabling tools. In

early imaging genetic studies, pairwise univariate analyses [14], [15], [16], [17], [18] were often performed to identify the associations between SNPs and neuroimaging QTs. In recent studies, taking into account the inherent structure among genotype or phenotype data, regression analysis and bi-multivariate analyses have achieved promising results for revealing complex multi-SNP-multi-QT associations [19], [20], [21], [22], [23], [24], [25], [26]. Canonical correlation analysis (CCA) is a common multivariate approaches to integrate two or more data types [27]. It has been applied to imaging genetics applications. The basic idea is to find the best linear transformations for imaging and genetics features so that the highest correlation between imaging and genetic components can be achieved. Based on the assumption that a real imaging genetic signal typically involves a small number of SNPs and QTs, sparse canonical correlation analysis (SCCA) has also been applied in several imaging genetic studies by imposing the Lasso regularization term to yield sparse results [22], [23], [24], [25]. Some extensions of SCCA are subsequently proposed to take advantage of the specific structure and group information [28], [29], [30], [31], [32], [33]. However, most of the existing SCCA algorithms are utilized to seek linear correlation of two data, do not consider the discriminant similarity information between different subject groups in feature extraction for brain imaging genetics association analysis. How to mine the potential information in a large amount of multi-modal data remains a challenging problem. Ignoring the discriminant similarity information in the data will inevitably limit the capability of yielding optimal results.

- M. Wang, W. Shao, and D. Zhang are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 211106, China. E-mail: {mely, dqzhang}@nuaa.edu.cn, 527606857@qq.com.
- X. Hao is with the School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China. E-mail: robinhc@163.com.
- L. Shen is with the Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia 19104 USA. E-mail: Li.Shen@perinmedicine.upenn.edu.

Manuscript received 19 Feb. 2019; revised 18 Sept. 2019; accepted 26 Sept. 2019. Date of publication 1 Oct. 2019; date of current version 6 Aug. 2021.

(Corresponding author: Daoqiang Zhang.)

Digital Object Identifier no. 10.1109/TCBB.2019.2944825

In this paper, we propose a novel multi-modality discriminant SCCA algorithm (MD-SCCA) to overcome this limitation as well as improve learning results by incorporating valuable discriminant similarity information. Specifically, our proposed method contains three steps: 1) Extract the discriminant similarity information between within-class subjects. The discriminant similarity information is extracted between within-class subjects by the sparse representation. 2) Construct a discriminant SCCA algorithm (D-SCCA). The traditional SCCA algorithm has been designed to seek a linear correlation between the SNP genotype and brain imaging phenotype data, ignoring the discriminant similarity information between within-class subjects in brain imaging genetics association analysis. The discriminant similarity information is enforced within the SCCA algorithm to construct a discriminant SCCA algorithm. 3) Present a multi-modality discriminant SCCA model. As multi-modality brain imaging phenotypes are extracted from different perspectives, imaging markers consistently showing up in multi-modality may provide more insights into the mechanistic understanding of diseases (i.e., Alzheimer's disease). A novel multi-modality discriminant SCCA algorithm is adopted to fully explore the relationships among different modalities of different subjects. We demonstrate the effectiveness of our algorithm with both synthetic and real data. For real data, two modalities of phenotypes, voxel-based measures extracted from structural MRI (VBM-MRI) and fluorodeoxyglucose positron emission tomography (FDG-PET) scans are used to evaluate the associations between the imaging data (VBM-MRI and FDG-PET) and the APOE SNP data. All the data are downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. The empirical results have demonstrated that our proposed algorithm produces improved cross-validation performances as well as biologically meaningful results.

The rest of this paper is organized in the following fashion. Section 2 Introduces the multi-modality discriminant SCCA algorithm. Related simulation and experimental results are included in Section 3. Section 4 further studies the identified SNP loci, the identified brain regions, regularization influence, and high-dimensional in simulations, as well as presents the limitations of our method and possible research directions. Conclusions are given in Section 5.

## 2 METHODOLOGY

We first describe the relevant notation. The lowercase letters and uppercase letters are respectively used to denote the vectors and matrices. Let  $X = [x_1, \dots, x_n, \dots, x_N]^T \in \mathbb{R}^{N \times p}$  be the SNP genotype data,  $Y = [y_1, \dots, y_n, \dots, y_N]^T \in \mathbb{R}^{N \times r}$  be the phenotype data, where  $N$  is the number of subjects, and  $p, r$  are the feature numbers of SNPs and imaging data, respectively.

### 2.1 Sparse Canonical Correlation Analysis (SCCA)

For detecting complex multi-SNP-QT associations, sparse canonical correlation analysis [22], [23], [24], [25] is a powerful association method that seeks linear transformations of two data sets  $X$  and  $Y$  to achieve the maximal correlation between  $Xu$  and  $Yv$ , which can be formulated as

$$\begin{aligned} & \max_{u,v} u^T X^T X^T Y v \\ \text{s.t. } & u^T X^T X u \leq 1, v^T Y^T Y v \leq 1, \|u\|_1 \leq c_1, \|v\|_1 \leq c_2, \end{aligned} \quad (1)$$

where  $u$  and  $v$  are canonical loadings or weights, reflecting the contribution of each feature in the identified canonical correlation. Note that  $u^T X^T X u \leq 1$  and  $v^T Y^T Y v \leq 1$  are used to embrace the covariance structure of the data in the model.  $\|u\|_1 \leq c_1$  and  $\|v\|_1 \leq c_2$  are constraints for controlling the sparsity so that only a small number of relevant features will be selected automatically from the SNP and imaging data. The optimization function of SCCA [34] can also be formulated as follows:

$$\begin{aligned} & \max_{u,v} u^T \sum_{i,j=1}^N (x_i - x_j)^T (y_i - y_j) v \\ \text{s.t. } & u^T \sum_{i,j=1}^N (x_i - x_j)^T (x_i - x_j) u \leq 1, \\ & v^T \sum_{i,j=1}^N (y_i - y_j)^T (y_i - y_j) v \leq 1, \\ & \|u\|_1 \leq c_1, \|v\|_1 \leq c_2. \end{aligned} \quad (2)$$

### 2.2 Discriminant Similarity Information Between Subject Classes

From the above Eqs. (1) and (2), the SCCA method only considers the correlation between  $X$  and  $Y$ , ignoring the relationship between the subjects ( $x_i$  and  $x_j$ , or  $y_i$  and  $y_j$ ). The weight matrix of the sparse representation can reflect the intrinsic geometric properties of the data [35]. Otherwise, in order to preserve the class information for the SNP genotype data, we only use the samples with the same label to reconstruct  $x_i$ , for  $c$  classes samples,  $S^x = [x_1^{(1)}, \dots, x_{n_1}^{(1)}, \dots, x_1^{(d)}, \dots, x_{n_d}^{(d)}, \dots, x_1^{(c)}, \dots, x_{n_c}^{(c)}]$ , where  $S^x \in \mathbb{R}^{p \times N}$ .  $c$  is the number of classes.  $x_i^{(d)}$  denotes the  $i$ th sample in the  $d$ th class. Then, for the sample  $x_i^{(d)}$  from  $d$ th class, a sparse reconstructive weight vector  $(W_i^x)^{(d)}$  is computed as

$$\min_{(W_i^x)^{(d)} \geq 0} \frac{1}{2} \left\| x_i^{(d)} - (S^x)^{(d)} (W_i^x)^{(d)} \right\|_2^2 + \lambda \left\| (W_i^x)^{(d)} \right\|_1, \quad (3)$$

where  $\lambda$  is a regularized parameter.  $(W_i^x)^{(d)}$  is the sparse reconstructive weight vector where the element in the position of  $x_i^{(d)}$  is zero. The optimization problem for Eq. (3) can be solved by the SLEP toolbox [36]. When we get  $(W_i^x)^{(1)}, \dots, (W_i^x)^{(d)}, \dots, (W_i^x)^{(c)}$  which are the optimal solution of Eq. (3), the discriminant similarity matrix  $D^x$  can be defined as

$$D^x = [(W_i^x)^{(1)}, \dots, (W_i^x)^{(c)}] + [(W_i^x)^{(1)}, \dots, (W_i^x)^{(c)}]^T. \quad (4)$$

According to the above process, for phenotype data, we can also get the the discriminant similarity matrix  $D^y$  as follows:

$$D^y = [(W_i^y)^{(1)}, \dots, (W_i^y)^{(c)}] + [(W_i^y)^{(1)}, \dots, (W_i^y)^{(c)}]^T. \quad (5)$$

where the discriminant similarity matrices  $D^x$  and  $D^y$  are the  $N \times N$  symmetric matrices,  $D_{ij}^x$  and  $D_{ij}^y$  are the elements

of discriminant similarity matrices  $D^x$  and  $D^y$ , which represent the contribution of each  $\tilde{x}_i$  and  $\tilde{y}_i$  to reconstruct  $x_i$  and  $y_i$ . Generally speaking, if the elements  $D_{ij}^x$  and  $D_{ij}^y$  are larger, the samples  $\tilde{x}_i$  and  $\tilde{y}_i$  are more important to reconstruct  $x_i$  and  $y_i$ .

It is worth noting that we only use the samples with the same label to reconstruct  $x_i$  and  $y_i$  for preserving the class information. The sparse weight matrices  $D^x$  and  $D^y$  can reflect the intrinsic geometric properties of the data to some extent and naturally preserve potential discriminant information. It observes the discriminant similarity information between different diagnostic groups (A specific diagnostic group is formed by all the subjects with the same diagnostic label (i.e., NC, SMC, EMCI, LMCI, or AD)).

### 2.3 Discriminant Sparse Canonical Correlation Analysis (D-SCCA)

The SCCA method is not able to capture relationships across subjects. To address this issue, we consider the information between subjects ( $x_i$  and  $x_j$ , or  $y_i$  and  $y_j$ ) as previously mentioned in Section 2.2. The local structure information and the within-class cross correlations are here incorporated into Eq. (2). The discriminant sparse canonical correlation analysis can be expressed as

$$\begin{aligned} \max_{u,v} u^T \left( \sum_{i,j=1}^N D_{ij}^x (x_i - x_j)^T D_{ij}^y (y_i - y_j) + \sum_{i,j=1}^N (D_{ij}^x + D_{ij}^y) x_i^T y_j \right) v \\ \text{s.t. } u^T \sum_{i,j=1}^N D_{ij}^x (x_i - x_j)^T D_{ij}^x (x_i - x_j) u \leq 1, \\ v^T \sum_{i,j=1}^N D_{ij}^y (y_i - y_j)^T D_{ij}^y (y_i - y_j) v \leq 1, \\ \|u\|_1 \leq c_1, \|v\|_1 \leq c_2, \end{aligned} \quad (6)$$

where  $D_{ij}^x$  and  $D_{ij}^y$  are elements of the discriminant similarity matrices  $D^x$  and  $D^y$ . As can be seen, there are some differences between Eqs. (2) and (6). First, the discriminant similarity matrices  $D^x$  and  $D^y$  are obtained by sparse reconstruction instead of the standard euclidean distance. The sparse weight matrix can reflect the intrinsic geometric properties of the data to some extent and naturally preserve potential discriminant information. Second, we emphasize the cross correlations between two modalities (genotype and phenotype) from within-class samples by adding the correlation term  $\sum_{i,j=1}^N (D_{ij}^x + D_{ij}^y) x_i^T y_j$ , which can capture the local structure information in the cross correlation among different subjects of different modalities. In summary, the optimization problem of our proposed D-SCCA method in Eq. (6) can be written in the equivalent form as

$$\begin{aligned} \max_{u,v} u^T X^T R^{xy} Y v \\ \text{s.t. } u^T X^T D^{xx} X u \leq 1, v^T Y^T D^{yy} Y v \leq 1, \|u\|_1 \leq c_1, \|v\|_1 \leq c_2, \end{aligned} \quad (7)$$

where  $R^{xy} = 2D^{xy} + D^x + D^y$ ,  $D^{xy} = \text{diag}(D^x * D^y) - [D^x * D^y]$ ,  $D^{xx} = \text{diag}(D^x * D^x) - [D^x * D^x]$ ,  $D^{yy} = \text{diag}(D^y * D^y) - [D^y * D^y]$ , the symbol  $*$  denotes an operator such that  $[D^x * D^y] = [D_{ij}^x D_{ij}^y]$ ,  $\text{diag}(D^x * D^y)$  is a diagonal matrix with  $D_{ij}^x D_{ij}^y$  main diagonal elements. The model not only

preserves the relationships from the SNP and imaging data, but also concerns the correlation within-class subjects. More details can be found in the *Supplementary Materials*, which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TCBB.2019.2944825>.

### 2.4 Multi-Modality Discriminant Sparse Canonical Correlation Analysis (MD-SCCA)

A common practice in coupled datasets is extracting relationships from different modalities, which may provide essential complementary information for the association study. Given  $N$  training subjects or samples with  $M$  modalities of phenotypes  $Y^m = [y_1^m, \dots, y_n^m, \dots, y_N^m]^T \in R^{N \times r}$  ( $m, q = 1, \dots, M$ ). Then, the following 'multi-modality D-SCCA' formulation is proposed

$$\begin{aligned} \max_{u,v} \sum_{m=1}^M u^T X^T R^{xY^m} Y^m v_m + \sum_{m,q=1, m < q}^M v_m^T (Y^m)^T R^{Y^m Y^q} Y^q v_q \\ \text{s.t. } u^T X^T D^{xx} X u \leq 1, v_m^T (Y^m)^T D^{Y^m Y^m} Y^m v_m \leq 1, \\ \|u\|_1 \leq c, \|v_m\|_1 \leq c_m, \end{aligned} \quad (8)$$

where  $R^{xY^m} = 2D^{xY^m} + D^x + D^{Y^m}$ ,  $R^{Y^m Y^q} = 2D^{Y^m Y^q} + D^{Y^m} + D^{Y^q}$ ,  $D^{xY^m} = \text{diag}(D^x * D^{Y^m}) - [D^x * D^{Y^m}]$ ,  $D^{xx} = \text{diag}(D^x * D^x) - [D^x * D^x]$ ,  $D^{Y^m Y^m} = \text{diag}(D^{Y^m} * D^{Y^m}) - [D^{Y^m} * D^{Y^m}]$ . In Eq. (8), the model incorporates multi-SNP-multi-QT associations (the first term) and multi-QT-QT (the second term) associations together within the learning framework. For convenience, assuming  $M = 2$ , the second term in Eq. (8) is expressed as follows:

$$\max_{v_1, v_2} v_1^T (Y^1)^T R^{Y^1 Y^2} Y^2 v_2, \quad (9)$$

the Eq. (9) can extract the correlated feature between different modalities as a regularization term, which provides essential complementary information for this association study.

As mentioned above, we introduce multi-modality discriminant-similarity sparse canonical correlation analysis, which incorporates valuable discriminant similarity information into the SCCA algorithm in order to improve the learning results in brain imaging genetics association analysis. As Eq. (8) depicted, the correlation between the within-class neighborhoods is maximized, and also the correlation among between-class neighborhoods approaches to zero. In theory, as a generalization of the existing work bi-convex, the objective function Eq. (8) is a multi-convex problem, so the global optimum can be guaranteed [37]. Moreover, when  $M = 1$ , Eq. (8) will degenerate into the D-SCCA model, which shows that this equation has a more general application scene.

### 2.5 Optimization

In this section, an algorithm is designed to solve the optimization problem defined in Eq. (8). We use the Lagrange multiplier and write the penalties into the objective function to get the following expression:

$$\begin{aligned}
L(u, v_1, \dots, v_M) &= \min_{u, V} \sum_{m=1}^M \|Xu - R^{xY^m} Y^m v_m\|_2^2 \\
&+ \sum_{m,q=1, m < q}^M \|Y^m v_m - R^{Y^m Y^q} Y^q v_q\|_2^2 + \frac{\alpha}{2} \|D^{xx} Xu\|_2^2 \\
&+ \sum_{m=1}^M \frac{\alpha_m}{2} \|D^{Y^m Y^m} Y^m v_m\|_2^2 + \beta \|u\|_1 + \sum_{m=1}^M \beta_m \|v_m\|_1,
\end{aligned} \tag{10}$$

where  $(\alpha, \alpha_1, \dots, \alpha_M)$  and  $(\beta, \beta_1, \dots, \beta_M)$  are the parameters. The parameters  $(\alpha, \alpha_1, \dots, \alpha_M)$  and  $(\beta, \beta_1, \dots, \beta_M)$  need to be tuned to control the global sparsity. Otherwise, in order to explain that these results are insensitive to  $(\alpha, \alpha_1, \dots, \alpha_M)$  settings, we show the performances of our method on simulated datasets. Here, we implement our experiments on the simulated datasets introduced in Section 3.2, and the regularization parameters  $(\alpha, \alpha_1, \dots, \alpha_M)$  are tuned from  $\{10^{-3}, 3 \times 10^{-3}, 10^{-2}, 3 \times 10^{-2}, 10^{-1}, 3 \times 10^{-1}, 1, 3, 10, 30, 100\}$ . *Supplementary Fig. S1*, available online, and *Supplementary Fig. S2*, available online, show all the test performances of MD-SCCA ( $\alpha = 1$ ) and MD-SCCA (grid search) in the simulations, which demonstrate that these results are insensitive to  $(\alpha, \alpha_1, \dots, \alpha_M)$  settings. Following our experiments, assuming  $\alpha = \alpha_1 = \dots = \alpha_M = 1$  for the sake of simplicity. Since the L1 norm introduced as the regularization term is not differentiable at 0, a similar optimization has been used in [38] for the solution of the Eq. 10. The solution for  $u, v_1, \dots$ , and  $v_M$  in each iteration step is as follows:

$$u = (X^T D^{xx} X + \beta H)^{-1} X^T \left( \sum_{m=1}^M R^{xY^m} Y^m v_m \right), \tag{11}$$

$$\begin{cases}
v_1 = ((Y^1)^T D^{Y^1 Y^1} Y^1 + \beta_1 H_1)^{-1} (Y^1)^T (R^{xY^1} Xu \\
\quad + \sum_{q=2}^M R^{Y^1 Y^q} Y^q v_q) \\
\vdots \\
v_M = ((Y^M)^T D^{Y^M Y^M} Y^M + \beta_M H_M)^{-1} (Y^M)^T (R^{xY^M} Xu \\
\quad + \sum_{q=1}^{M-1} R^{Y^M Y^q} Y^q v_q),
\end{cases} \tag{12}$$

where  $H$  is a diagonal matrix with the  $k$ th element as  $\frac{1}{2\|u^k\|_1}$  ( $k \in [1, p]$ ),  $H_1$  is a diagonal matrix with the  $k_1$ th element as  $\frac{1}{2\|v_1^{k_1}\|_1}$  ( $k_1 \in [1, r]$ ),  $\dots$ , and  $H_M$  is a diagonal matrix with the  $k_M$ th element as  $\frac{1}{2\|v_M^{k_M}\|_1}$  ( $k_M \in [1, r]$ ). Here, when  $|u| = 0, |v_1| = 0, \dots$ , and  $|v_M| = 0$ , the  $k$ th,  $k_1$ th,  $\dots$ , and  $k_M$ th elements  $\frac{1}{2\|u^k\|_1}, \frac{1}{2\|v_1^{k_1}\|_1}, \dots$ , and  $\frac{1}{2\|v_M^{k_M}\|_1}$  can not be computed. So we consider  $\frac{1}{2\|u^k\|_1}, \frac{1}{2\|v_1^{k_1}\|_1}, \dots$ , and  $\frac{1}{2\|v_M^{k_M}\|_1}$  as  $\frac{1}{2\sqrt{u^{k^2} + \xi}}, \frac{1}{2\sqrt{v_1^{k_1^2} + \xi}}, \dots$ , and  $\frac{1}{2\sqrt{v_M^{k_M^2} + \xi}}$ ,  $\xi$  is a very small real number.

Since  $H, H_1, \dots, H_M$  rely on  $u, v_1, \dots, v_M$ , the iterative procedure is introduced to solve this objective. In each iteration, we first fix  $v_1, \dots, v_M$  to solve  $u$ , then, fix  $u, v_2, \dots, v_M$

to solve  $v_1$ , finally, fix  $u, v_1, \dots, v_{M-1}$  to solve  $v_M$ . The procedure stops until it satisfies a predefined stopping criterion. The general optimization procedure of the proposed algorithm is described in Algorithm 1.

---

### Algorithm 1. MD-SCCA Method

---

- 1 **Input:** SNP genotype  $X = [x_1, \dots, x_n, \dots, x_N]^T \in R^{N \times p}$ ; Multi-modality imaging phenotype  $[y_1^m, \dots, y_n^m, \dots, y_N^m]^T \in R^{N \times r}$ ; Subjects with label information (i.e., NC, SMC, EMCI, LMCI or AD)
  - 2 **Output:**  $u, v_1, \dots, v_M$
  - 3 **Optimization:**
    - 1: Construct  $D^x, D^{Y^1}, \dots, D^{Y^m}$  with the discriminant information by the sparse representation;
    - 2: Get  $R^{xY^1}, \dots, R^{xY^m}, R^{Y^m Y^q}$  ( $m \neq q$ ),  $D^{xx}, D^{Y^1 Y^1}, \dots, D^{Y^m Y^m}$ ;
    - 3: Initialization:  $u \in R^{p \times 1}, v_1 \in R^{r \times 1}, \dots, v_m \in R^{r \times 1}$ ;
    - 4: While not converged regarding to  $u, v_1, \dots, v_m$  do
    - 5: Calculate the diagonal matrix  $T$ , where the  $k$ th element is  $\frac{1}{2\|u^k\|_1}$ ;
    - 6: Update  $u = (X^T D^{xx} X + \beta T)^{-1} X^T (\sum_{m=1}^M R^{xY^m} Y^m v_m)$ ;
    - 7: Scale  $u$  so that  $\|Xu\|_2 = 1$ ;
    - 8: Calculate the diagonal matrix  $T_1$ , where the  $k_1$ th element is  $\frac{1}{2\|v_1^{k_1}\|_1}$ ;
    - 9: Update  $v_1 = ((Y^1)^T D^{Y^1 Y^1} Y^1 + \beta_1 T_1)^{-1} (Y^1)^T (R^{xY^1} Xu + \sum_{q=2}^M R^{Y^1 Y^q} Y^q v_q)$ ;
    - 10: Scale  $v_1$  so that  $\|Y^1 v_1\|_2 = 1$ ;
    - 11:  $\vdots$
    - 12: Calculate the diagonal matrix  $T_M$ , where the  $k_M$ th element is  $\frac{1}{2\|v_M^{k_M}\|_1}$ ;
    - 13: Update
    - 14:  $v_M = ((Y^M)^T D^{Y^M Y^M} Y^M + \beta_M T_M)^{-1} (Y^M)^T (R^{xY^M} Xu + \sum_{q=1}^{M-1} R^{Y^M Y^q} Y^q v_q)$ ;
    - 15: Scale  $v_M$  so that  $\|Y^M v_M\|_2 = 1$ ;
    - 16: End while.
- 

## 3 EXPERIMENT

In this section, we evaluate the performances of our method on both simulated and real datasets.

### 3.1 Experimental Settings

In our experiments, the 5-fold cross-validation strategy is adopted to evaluate the effectiveness of our proposed method. The regularization parameters  $(\beta, \beta_1, \dots, \beta_M)$  in Eq. (10) are tuned using a grid search from the range of  $\{10^{-3}, 3 \times 10^{-3}, 10^{-2}, 3 \times 10^{-2}, 10^{-1}, 3 \times 10^{-1}, 1, 3, 10, 30, 100\}$ . The performance on each dataset is assessed with the correlation coefficient ( $CC$ ) between actual and predicted response values, which is widely used in measuring performances of regression and association analysis.

We choose the existing SCCA based methods for comparison in this study, including the state-of-the-art method AGN-SCCA [31], and the benchmark algorithm SCCA (CCA with lasso) used in [22], [23], [24], [25]. We do not compare our methods with KG-SCCA [25] since it requires predefined group and network structure. We also do not compare our methods with TGSCCA [29] because it only focuses on the time-point features in brain from longitudinal phenotypes. In addition, T-SCCA [38], S2CCA [39], and ssCCA [40], and

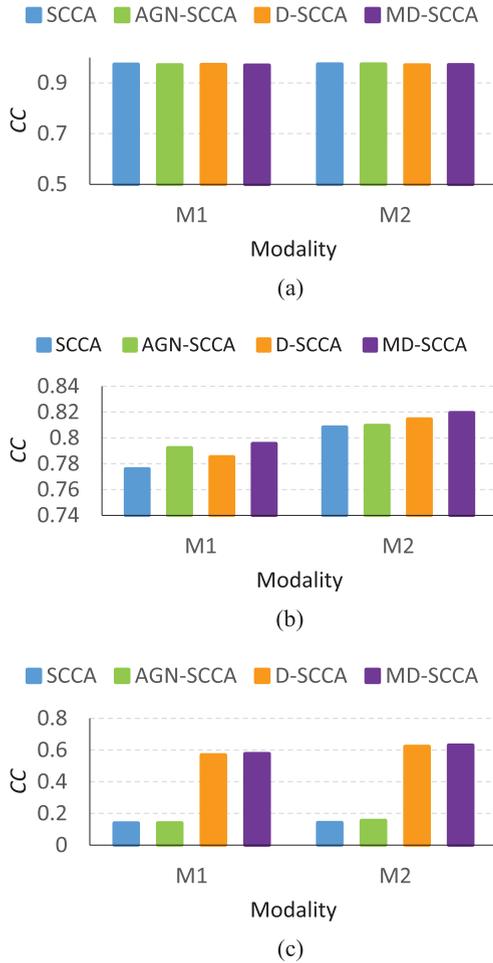


Fig. 1. The averaged correlation coefficients on 5-fold test data using different methods on different simulation datasets. (a) Results on simulation dataset 1. (b) Results on simulation dataset 2. (c) Results on simulation dataset 3.

CCA-SG [24] are opted out, since they are knowledge-guided methods and applicable only when priori knowledge is available.

### 3.2 Experimental Results on Simulation Data

For our first test, a simulation study is presented, which is similar to the procedure of simulation generation in Hao et al. [29] and Fang et al. [30]. First, one canonical vector  $a$  with  $s'$  non-zero entries and successive canonical vector  $b$  with  $l'$  non-zero entries were generated, where  $a_{k+1} = a_k + \Delta a$ ,  $\Delta a \sim N(0, 0.1)$ , and  $k = 1$ . Each non-zero variable in  $a_1$  and  $b$  is sampled independently from a uniform distribution in the range of  $[-2, -0.5] \cup [0.5, 2]$ . Then, we consider the data belonging to two classes in all simulations. Each class consists of  $N = 500$  samples. Specifically, one latent variable  $t_1$  with normal distribution  $N(0, \sigma_t)$  for 300 samples is randomly generated. For the data matrices  $X$  and  $Y$ , the features are simulated from Gaussian distributions  $N(bt_1, \sigma_w I_l)$  and  $N(a_k t_1, \sigma_w I_s)$ , respectively; the other latent variable  $t_2$  with normal distribution  $N(1, \sigma_t)$  for 200 samples is randomly generated. For the data matrices  $X$  and  $Y$ , the features are simulated from Gaussian distributions  $N(bt_2, \sigma_w I_l)$  and  $N(a_k t_2, \sigma_w I_s)$ , respectively. Set  $l = 100$ ,  $s = 50$ ,  $s' = 30$ ,  $l' = 20$ , and  $\sigma_t = 0.1$ . In this paper, the noise levels  $\sigma_w = 0.1, 0.3, 0.5$  are given to generate three different simulation datasets.

We compare our proposed methods (including MD-SCCA and D-SCCA) with SCCA and AGN-SCCA algorithms, respectively. As shown in Fig. 1, all methods yield stable results on the simulation datasets with low noise, however, our proposed MD-SCCA consistently outperforms SCCA, AGN-SCCA, and D-SCCA in the metric of correlation coefficients on the simulation datasets with high noise. Furthermore, we show the estimated canonical weights from different methods. As shown in Fig. 2, the overall profiles of the estimated  $u$  and  $v$  values from MD-SCCA are consistent with the ground truth on simulation datasets, whereas D-SCCA, AGN-SCCA, and SCCA are only capable of identifying regionwise-inconsistent signals at different modalities. In addition, MD-SCCA is comparable with D-SCCA, AGN-SCCA, and SCCA due to the low noise, while MD-SCCA is more robust to the datasets with high noise.

### 3.3 Experimental Results on Imaging Genetic Data

1) *Data acquisition*: Data used in the preparation of this article were obtained from the Alzheimer's disease Neuroimaging Initiative database ([adni.loni.usc.edu](http://adni.loni.usc.edu)). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease. For up-to-date information, see [www.adni-info.org](http://www.adni-info.org).

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 US and Canada. The initial goal of ADNI was to recruit 800 adults, aged 55-90, to participate in the research—approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years, and 200 people with an early AD to be followed for 2 years. In current studies, a total of 913 non-Hispanic Caucasian participants with both imaging and genotyping data available are studied, which include 211 normal control (NC), 82 significant memory concern (SMC), 273 early mild cognitive impairment (EMCI), 187 late mild cognitive impairment (LMCI) and 160 AD. Table 1 lists the demographics of all these subjects.

Since genetic risk factors can help scientists focus on relevant biological pathways and form an effective hypothesis for drug design, identifying risk genetic markers associated with brain imaging assists to understand the underlying biological mechanisms. We downloaded the ADNI-GO/2 genotyping data and performed quality control and population stratification using the approach described in the previous study. To limit the potential effects of population stratification, this study focuses only on analyzing non-Hispanic white participants. As the best-known genetic risk factor in the AD, APOE (located on chromosome 19) has a key role in coordinating the mobilization and redistribution of cholesterol, phospholipids, and fatty acids, and it is implicated in mechanisms such as neuronal development, brain plasticity, and repair functions. Thus, we concern our analysis on all SNPs within 20 k base pairs of the APOE gene boundary based on the ANNOVAR (<http://annovar.openbioinformatics.org>) annotation, which include a total number of 85 SNPs as candidates. For the input in the models, each SNP value was coded in an additive fashion as 0, 1, or 2, indicating the number of minor alleles.

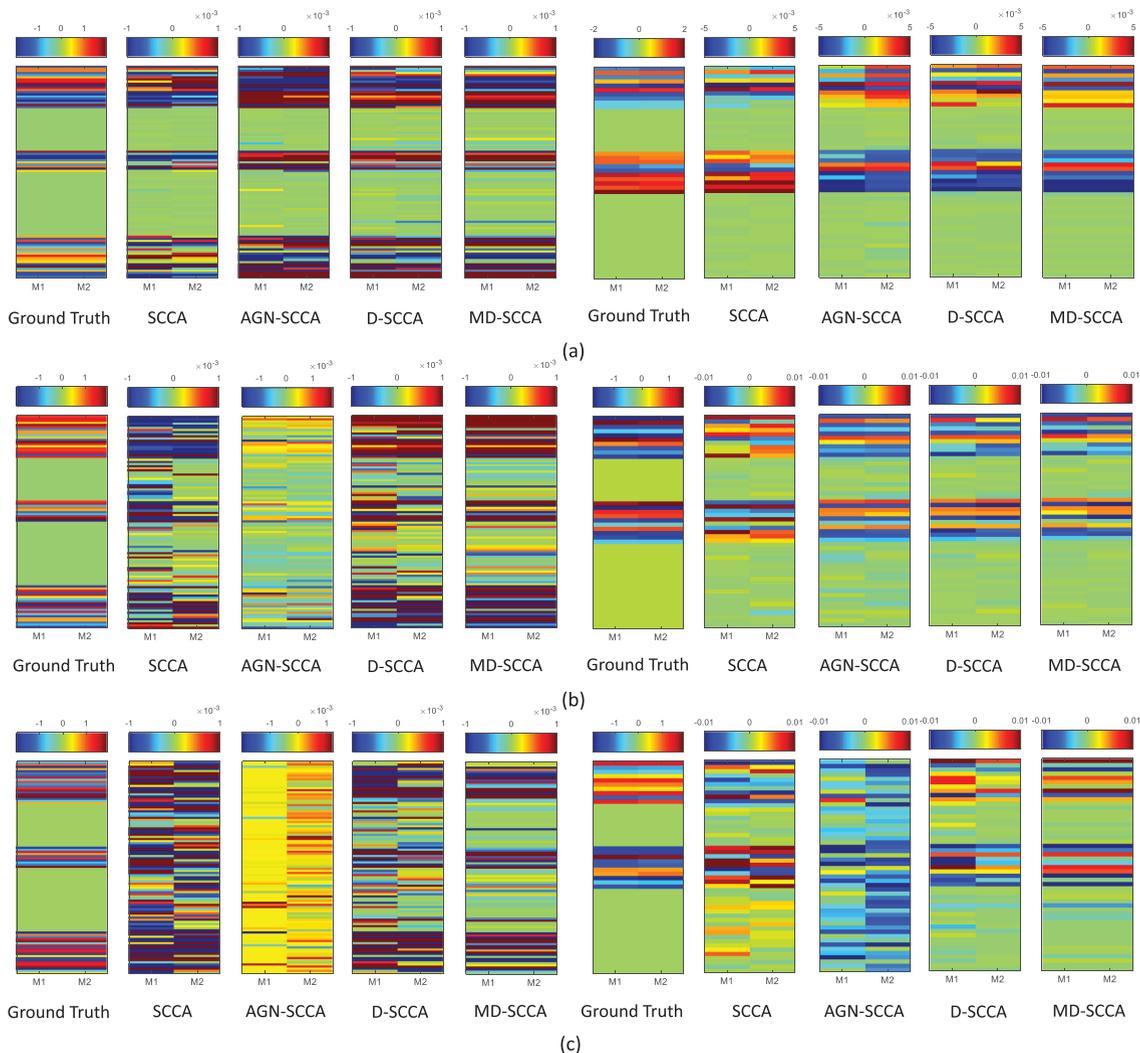


Fig. 2. The estimated weights of  $u$  and  $v$  from average 5-fold cross-validation test on different simulation datasets are shown in the left four panels and right four panels, corresponding to different methods. (a) Results on simulation dataset 1. (b) Results on simulation dataset 2. (c) Results on simulation dataset 3.

TABLE 1

Characteristics of the Subjects, Note: NC=Normal Control, SMC=Significant Memory Concern, EMCI=Early Mild Cognitive Impairment, LMCI=Late Mild Cognitive Impairment, and AD=Alzheimer's Disease

Subjects	NC	SMC	EMCI	LMCI	AD
Number	211	82	273	187	160
Gender(Male/Female)	109/102	33/49	153/120	108/79	95/65
Age(mean $\pm$ std)	76.14 $\pm$ 6.53	72.45 $\pm$ 5.67	71.48 $\pm$ 7.12	73.86 $\pm$ 8.44	75.18 $\pm$ 7.88
Education(mean $\pm$ std)	16.45 $\pm$ 2.62	16.78 $\pm$ 2.67	16.08 $\pm$ 2.62	16.38 $\pm$ 2.81	15.86 $\pm$ 2.75

The VBM-MRI and FDG-PET data used in this paper were also obtained from the ADNI database (adni.loni.usc.edu). We aligned the preprocessed imaging data (including VBM and FDG) to each participant's same visit scan, and then created normalized gray matter density maps from VBM data in the standard Montreal Neurological Institute (MNI) space as  $2 \times 2 \times 2 \text{ mm}^3$  voxels, registered the FDG scans into the same space by SPM software package. 116 ROI level measurements of mean gray matter densities and FDG glucose utilization were further extracted based on the MarsBaR AAL atlas [41]. The imaging measures on each modality (VBM and FDG) of 116 ROIs were used as QTs in our experiments. All the measures were pre-adjusted for age, gender, and education.

2) *Quantitative analysis*: In the real data experiments, we also compare our proposed methods (including MD-SCCA, D-SCCA) with SCCA and AGN-SCCA algorithms, respectively. Similar to the previous analysis, 5-fold cross-validation is utilized to optimally tune the parameters. Ten experiments are performed with ten different partitions to eliminate the bias. For each single experiment, the same partition is used for the MD-SCCA, D-SCCA, AGN-SCCA, and SCCA. Tables 2 and 3 exhibit both training and testing correlation coefficients where each individual result and its mean and standard deviation are shown. Obviously, MD-SCCA and D-SCCA are observed to outperform the AGN-SCCA and SCCA in every single experiment on both training and test performance for

TABLE 2  
Five-Fold Cross Validation Results on FDG Imaging Data: The Models Learned from the Training Data are used to Estimate the Correlation Coefficients between Canonical Components for Both Training and Testing Sets

Method		Train					mean $\pm$ std	Test					mean $\pm$ std
		$f_1$	$f_2$	$f_3$	$f_4$	$f_5$		$f_1$	$f_2$	$f_3$	$f_4$	$f_5$	
SCCA	exp1	0.4471	0.4329	0.4449	0.4434	0.4487	0.4434 $\pm$ 0.0062	0.1959	0.2468	0.2289	0.1553	0.1189	0.1891 $\pm$ 0.0525
	exp2	0.4526	0.4392	0.4527	0.4480	0.4514	0.4488 $\pm$ 0.0057	0.1920	0.2466	0.2179	0.1564	0.1193	0.1864 $\pm$ 0.0501
	exp3	0.4423	0.4274	0.4383	0.4394	0.4467	0.4388 $\pm$ 0.0072	0.1968	0.2447	0.2348	0.1513	0.1210	0.1897 $\pm$ 0.0532
	exp4	0.4471	0.4329	0.4449	0.4434	0.4487	0.4434 $\pm$ 0.0062	0.1959	0.2468	0.2289	0.1553	0.1189	0.1891 $\pm$ 0.0525
	exp5	0.4471	0.4329	0.4449	0.4434	0.4487	0.4434 $\pm$ 0.0062	0.1959	0.2468	0.2289	0.1553	0.1189	0.1891 $\pm$ 0.0525
	exp6	0.4526	0.4392	0.4527	0.4480	0.4514	0.4488 $\pm$ 0.0057	0.1920	0.2466	0.2179	0.1564	0.1193	0.1864 $\pm$ 0.0501
	exp7	0.4526	0.4392	0.4527	0.4480	0.4514	0.4488 $\pm$ 0.0057	0.1920	0.2466	0.2179	0.1564	0.1193	0.1864 $\pm$ 0.0501
	exp8	0.4471	0.4329	0.4449	0.4434	0.4487	0.4434 $\pm$ 0.0062	0.1959	0.2468	0.2289	0.1553	0.1189	0.1891 $\pm$ 0.0525
	exp9	0.4471	0.4329	0.4449	0.4434	0.4487	0.4434 $\pm$ 0.0062	0.1959	0.2468	0.2289	0.1553	0.1189	0.1891 $\pm$ 0.0525
	exp10	0.4526	0.4392	0.4527	0.4480	0.4514	0.4488 $\pm$ 0.0057	0.1920	0.2466	0.2179	0.1564	0.1193	0.1864 $\pm$ 0.0501
AGN-SCCA	exp1	0.2932	0.3004	0.2711	0.2834	0.2900	0.2876 $\pm$ 0.0111	0.1439	0.1683	0.3169	0.1543	0.2017	0.1970 $\pm$ 0.0705
	exp2	0.2919	0.3004	0.2712	0.2834	0.2900	0.2874 $\pm$ 0.0109	0.1422	0.1682	0.3162	0.1550	0.2019	0.1967 $\pm$ 0.0704
	exp3	0.2908	0.2983	0.2680	0.2812	0.2883	0.2853 $\pm$ 0.0114	0.1422	0.1665	0.3018	0.1599	0.2010	0.1943 $\pm$ 0.0638
	exp4	0.2932	0.3004	0.2711	0.2834	0.2900	0.2876 $\pm$ 0.0111	0.1439	0.1683	0.3169	0.1543	0.2017	0.1970 $\pm$ 0.0705
	exp5	0.2911	0.2988	0.2674	0.2818	0.2859	0.2850 $\pm$ 0.0117	0.1427	0.1680	0.2973	0.1594	0.1903	0.1966 $\pm$ 0.0616
	exp6	0.2911	0.2988	0.2674	0.2818	0.2859	0.2850 $\pm$ 0.0117	0.1427	0.1680	0.2973	0.1594	0.1903	0.1966 $\pm$ 0.0616
	exp7	0.2919	0.3004	0.2712	0.2834	0.2900	0.2874 $\pm$ 0.0109	0.1422	0.1665	0.3018	0.1599	0.2010	0.1943 $\pm$ 0.0638
	exp8	0.2859	0.2964	0.2668	0.2783	0.2876	0.2830 $\pm$ 0.0111	0.1434	0.1649	0.3084	0.1670	0.1982	0.1964 $\pm$ 0.0656
	exp9	0.2871	0.2986	0.2702	0.2808	0.2892	0.2852 $\pm$ 0.0105	0.1433	0.1674	0.3226	0.1624	0.1994	0.1990 $\pm$ 0.0719
	exp10	0.2874	0.2989	0.2705	0.2812	0.2893	0.2854 $\pm$ 0.0105	0.1440	0.1690	0.3231	0.1615	0.1999	0.1995 $\pm$ 0.0720
D-SCCA	exp1	0.3126	0.3146	0.2780	0.3283	0.3074	0.3082 $\pm$ 0.0185	0.2330	0.2192	0.3341	0.1238	0.2111	0.2242 $\pm$ 0.0749
	exp2	0.3126	0.3146	0.2780	0.3283	0.3074	0.3082 $\pm$ 0.0185	0.2330	0.2192	0.3341	0.1238	0.2111	0.2242 $\pm$ 0.0749
	exp3	0.3006	0.3052	0.2644	0.3210	0.2978	0.2978 $\pm$ 0.0207	0.2387	0.2185	0.3173	0.1200	0.2138	0.2216 $\pm$ 0.0705
	exp4	0.3131	0.3157	0.2795	0.3296	0.3102	0.3096 $\pm$ 0.0184	0.2380	0.2234	0.3362	0.1255	0.2066	0.2259 $\pm$ 0.0754
	exp5	0.3126	0.3146	0.2780	0.3283	0.3074	0.3082 $\pm$ 0.0185	0.2330	0.2192	0.3341	0.1238	0.2111	0.2242 $\pm$ 0.0749
	exp6	0.3126	0.3146	0.2780	0.3283	0.3074	0.3082 $\pm$ 0.0185	0.2330	0.2192	0.3341	0.1238	0.2111	0.2242 $\pm$ 0.0749
	exp7	0.3131	0.3157	0.2795	0.3296	0.3102	0.3096 $\pm$ 0.0184	0.2380	0.2234	0.3362	0.1255	0.2066	0.2259 $\pm$ 0.0754
	exp8	0.2948	0.2973	0.2565	0.3168	0.2916	0.2914 $\pm$ 0.0218	0.2403	0.2200	0.3032	0.1166	0.2118	0.2184 $\pm$ 0.0672
	exp9	0.3006	0.3052	0.2644	0.3210	0.2978	0.2978 $\pm$ 0.0207	0.2387	0.2185	0.3173	0.1200	0.2138	0.2216 $\pm$ 0.0705
	exp10	0.2953	0.2985	0.2585	0.3182	0.2950	0.2931 $\pm$ 0.0216	0.2446	0.2244	0.3063	0.1189	0.2053	0.2195 $\pm$ 0.0681
MD-SCCA	exp1	0.3200	0.3189	0.2854	0.3372	0.3337	0.3190 $\pm$ 0.0205	0.2322	0.2382	0.3496	0.1063	0.2038	0.2260 $\pm$ 0.0870
	exp2	0.3200	0.3189	0.2854	0.3372	0.3337	0.3190 $\pm$ 0.0205	0.2322	0.2382	0.3496	0.1063	0.2038	0.2260 $\pm$ 0.0870
	exp3	0.3265	0.3211	0.2867	0.3393	0.3324	0.3212 $\pm$ 0.0204	0.2283	0.2320	0.3522	0.1094	0.2006	0.2245 $\pm$ 0.0869
	exp4	0.3265	0.3211	0.2867	0.3393	0.3324	0.3212 $\pm$ 0.0204	0.2283	0.2320	0.3522	0.1094	0.2006	0.2245 $\pm$ 0.0869
	exp5	0.3265	0.3211	0.2867	0.3393	0.3324	0.3212 $\pm$ 0.0204	0.2283	0.2320	0.3522	0.1094	0.2006	0.2245 $\pm$ 0.0869
	exp6	0.3200	0.3189	0.2854	0.3372	0.3337	0.3190 $\pm$ 0.0205	0.2322	0.2382	0.3496	0.1063	0.2038	0.2260 $\pm$ 0.0870
	exp7	0.3200	0.3189	0.2854	0.3372	0.3337	0.3190 $\pm$ 0.0205	0.2322	0.2382	0.3496	0.1063	0.2038	0.2260 $\pm$ 0.0870
	exp8	0.3240	0.3201	0.2862	0.3380	0.3322	0.3201 $\pm$ 0.0202	0.2296	0.2337	0.3505	0.1085	0.2019	0.2249 $\pm$ 0.0865
	exp9	0.3200	0.3189	0.2854	0.3372	0.3337	0.3190 $\pm$ 0.0205	0.2322	0.2382	0.3496	0.1063	0.2038	0.2260 $\pm$ 0.0870
	exp10	0.3265	0.3211	0.2867	0.3393	0.3324	0.3212 $\pm$ 0.0204	0.2283	0.2320	0.3522	0.1094	0.2006	0.2245 $\pm$ 0.0869

FDG and VBM imaging data. And, MD-SCCA is more stable than D-SCCA, AGN-SCCA and SCCA. Moreover, the paired t-test is performed to compare the performance across ten experiments. The MD-SCCA outperforms SCCA and AGN-SCCA significantly in both training ( $p < 0.05$  (FDG),  $p < 0.05$  (VBM)) and test cases ( $p < 0.05$  (FDG),  $p < 0.05$  (VBM)). The D-SCCA outperforms SCCA and AGN-SCCA significantly in both training ( $p < 0.05$  (FDG),  $p < 0.05$  (VBM)) and test cases ( $p < 0.05$  (FDG),  $p < 0.05$  (VBM)). The resulting p-value ( $p < 0.05$ ) shows that the improvement for our methods are statistically significant.

3) *Identification of risk SNP loci and imaging ROI markers:* Besides improving association performance, one major goal of this study is to identify some vital SNP loci and imaging phenotypic markers for disease progression in AD research. Therefore, finding genetic risk factors and imaging ROIs helps scientists better understand how the disease develops and identify possible treatments to study. Fig. 3 compares the average regression coefficients of different methods.

Here, regression coefficients are the combinations of genotype for risk SNP loci and brain phenotype for the ROIs. This figure demonstrates the canonical loadings tested from 5-fold cross-validation in one experiment, suggesting relevant genetic (top three panels) and imaging (bottom three panels) markers. As expected, our proposed methods (including MD-SCCA, D-SCCA) can select significant ROIs and SNP loci. Although it may affect different sets of SNP loci and ROIs while using different imaging data (including FDG and VBM) as phenotype, our proposed methods (including MD-SCCA, D-SCCA) can jointly select consistent and clear SNP loci (rs429358-C) and ROIs (right hippocampus) association with AD for different imaging data, which are in accordance with prior findings [42], [43]. In addition, our proposed MD-SCCA method also can jointly select other consistent and clear SNP loci and ROIs for different imaging data. These results demonstrate the effectiveness of our proposed method and show great potential for further investigation.

TABLE 3  
Five-Fold Cross Validation Results on VBM Imaging Data: The Models Learned from the Training Data are used to Estimate the Correlation Coefficients between Canonical Components for Both Training and Testing Sets

Method		Train					Test						
		$f_1$	$f_2$	$f_3$	$f_4$	$f_5$	mean $\pm$ std	$f_1$	$f_2$	$f_3$	$f_4$	$f_5$	mean $\pm$ std
SCCA	exp1	0.4260	0.4338	0.4290	0.4069	0.4260	0.4244 $\pm$ 0.0103	0.1633	0.1496	0.0958	0.2041	0.2698	0.1565 $\pm$ 0.0394
	exp2	0.4046	0.4287	0.3945	0.3992	0.4196	0.4093 $\pm$ 0.0144	0.2337	0.1679	0.0674	0.2364	0.2037	0.1818 $\pm$ 0.0697
	exp3	0.4053	0.4303	0.4101	0.4039	0.4231	0.4145 $\pm$ 0.0116	0.2830	0.1798	0.0804	0.2519	0.1842	0.1959 $\pm$ 0.0782
	exp4	0.4023	0.4267	0.3865	0.3974	0.4192	0.4064 $\pm$ 0.0163	0.2660	0.1731	0.0658	0.2452	0.1992	0.1899 $\pm$ 0.0785
	exp5	0.3994	0.4233	0.3824	0.3999	0.4193	0.4048 $\pm$ 0.0166	0.2760	0.1746	0.0603	0.2523	0.1941	0.1915 $\pm$ 0.0842
	exp6	0.4492	0.4661	0.4542	0.4353	0.4571	0.4524 $\pm$ 0.0114	0.2691	0.1526	0.0917	0.2564	0.1428	0.1825 $\pm$ 0.0769
	exp7	0.4023	0.4267	0.3865	0.3974	0.4192	0.4064 $\pm$ 0.0163	0.2660	0.1731	0.0658	0.2452	0.1992	0.1899 $\pm$ 0.0785
	exp8	0.4642	0.4715	0.4640	0.4395	0.4658	0.4610 $\pm$ 0.0124	0.1408	0.1420	0.0643	0.2452	0.1543	0.1493 $\pm$ 0.0643
	exp9	0.4492	0.4661	0.4542	0.4353	0.4571	0.4524 $\pm$ 0.0114	0.2691	0.1526	0.0917	0.2564	0.1428	0.1825 $\pm$ 0.0769
	exp10	0.4046	0.4287	0.3945	0.3992	0.4196	0.4093 $\pm$ 0.0144	0.2337	0.1679	0.0674	0.2364	0.2037	0.1818 $\pm$ 0.0697
AGN-SCCA	exp1	0.3784	0.3551	0.3487	0.3798	0.3388	0.3602 $\pm$ 0.0182	0.1904	0.2149	0.1978	0.1166	0.2397	0.1919 $\pm$ 0.0461
	exp2	0.3784	0.3551	0.3487	0.3798	0.3388	0.3602 $\pm$ 0.0182	0.1904	0.2149	0.1978	0.1166	0.2397	0.1919 $\pm$ 0.0461
	exp3	0.3784	0.3551	0.3487	0.3798	0.3388	0.3602 $\pm$ 0.0182	0.1904	0.2149	0.1978	0.1166	0.2397	0.1919 $\pm$ 0.0461
	exp4	0.3687	0.3136	0.3284	0.4067	0.3505	0.3536 $\pm$ 0.0314	0.1899	0.1375	0.1766	0.2242	0.1460	0.1748 $\pm$ 0.0349
	exp5	0.4076	0.3932	0.4619	0.3186	0.3638	0.3890 $\pm$ 0.0531	0.1716	0.2118	0.1355	0.1295	0.1960	0.1689 $\pm$ 0.0362
	exp6	0.3784	0.3551	0.3487	0.3798	0.3388	0.3602 $\pm$ 0.0182	0.1904	0.2149	0.1978	0.1166	0.2397	0.1919 $\pm$ 0.0461
	exp7	0.4505	0.4022	0.4373	0.2268	0.4141	0.3862 $\pm$ 0.0911	0.2391	0.2112	0.1615	0.1395	0.1615	0.1826 $\pm$ 0.0411
	exp8	0.3784	0.3551	0.3487	0.3798	0.3388	0.3602 $\pm$ 0.0182	0.1904	0.2149	0.1978	0.1166	0.2397	0.1919 $\pm$ 0.0461
	exp9	0.3784	0.3551	0.3487	0.3798	0.3388	0.3602 $\pm$ 0.0182	0.1904	0.2149	0.1978	0.1166	0.2397	0.1919 $\pm$ 0.0461
	exp10	0.4505	0.4022	0.4373	0.2268	0.4141	0.3862 $\pm$ 0.0911	0.2391	0.2112	0.1615	0.1395	0.1615	0.1826 $\pm$ 0.0411
D-SCCA	exp1	0.2978	0.3130	0.3320	0.2929	0.3086	0.3088 $\pm$ 0.0152	0.3499	0.1770	0.1790	0.2567	0.2306	0.2186 $\pm$ 0.0383
	exp2	0.2786	0.2910	0.3081	0.2734	0.2846	0.2872 $\pm$ 0.0134	0.2284	0.1784	0.1662	0.2658	0.2360	0.2150 $\pm$ 0.0416
	exp3	0.2786	0.2910	0.3081	0.2734	0.2846	0.2872 $\pm$ 0.0134	0.2284	0.1784	0.1662	0.2658	0.2360	0.2150 $\pm$ 0.0416
	exp4	0.2978	0.3130	0.3320	0.2929	0.3086	0.3088 $\pm$ 0.0152	0.3499	0.1770	0.1790	0.2567	0.2306	0.2186 $\pm$ 0.0383
	exp5	0.2820	0.2945	0.3115	0.2773	0.2886	0.2908 $\pm$ 0.0133	0.2320	0.1787	0.1699	0.2647	0.2357	0.2162 $\pm$ 0.0404
	exp6	0.2897	0.3025	0.3199	0.2849	0.2980	0.2990 $\pm$ 0.0136	0.2411	0.1780	0.1770	0.2620	0.2347	0.2186 $\pm$ 0.0388
	exp7	0.2978	0.3130	0.3320	0.2929	0.3086	0.3088 $\pm$ 0.0152	0.3499	0.1770	0.1790	0.2567	0.2306	0.2186 $\pm$ 0.0383
	exp8	0.2937	0.3073	0.3256	0.2886	0.3031	0.3037 $\pm$ 0.0143	0.2459	0.1775	0.1786	0.2592	0.2333	0.2189 $\pm$ 0.0384
	exp9	0.2937	0.3073	0.3256	0.2886	0.3031	0.3037 $\pm$ 0.0143	0.2459	0.1775	0.1786	0.2592	0.2333	0.2189 $\pm$ 0.0384
	exp10	0.2978	0.3130	0.3320	0.2929	0.3086	0.3088 $\pm$ 0.0152	0.3499	0.1770	0.1790	0.2567	0.2306	0.2186 $\pm$ 0.0383
MD-SCCA	exp1	0.3028	0.3061	0.3021	0.2788	0.3007	0.2981 $\pm$ 0.0110	0.1861	0.1833	0.2532	0.2417	0.2511	0.2231 $\pm$ 0.0353
	exp2	0.3028	0.3061	0.3021	0.2788	0.3007	0.2981 $\pm$ 0.0110	0.1861	0.1833	0.2532	0.2417	0.2511	0.2231 $\pm$ 0.0353
	exp3	0.2914	0.3001	0.2936	0.2714	0.2916	0.2896 $\pm$ 0.0108	0.2090	0.1761	0.2397	0.2426	0.2458	0.2226 $\pm$ 0.0299
	exp4	0.3127	0.3218	0.3221	0.2912	0.3209	0.3137 $\pm$ 0.0132	0.2043	0.1956	0.2567	0.2465	0.2673	0.2341 $\pm$ 0.0322
	exp5	0.3028	0.3061	0.3021	0.2788	0.3007	0.2981 $\pm$ 0.0110	0.1861	0.1833	0.2532	0.2417	0.2511	0.2231 $\pm$ 0.0353
	exp6	0.3028	0.3061	0.3021	0.2788	0.3007	0.2981 $\pm$ 0.0110	0.1861	0.1833	0.2532	0.2417	0.2511	0.2231 $\pm$ 0.0353
	exp7	0.3127	0.3218	0.3221	0.2912	0.3209	0.3137 $\pm$ 0.0132	0.2043	0.1956	0.2567	0.2465	0.2673	0.2341 $\pm$ 0.0322
	exp8	0.2851	0.2945	0.2881	0.2665	0.2858	0.2840 $\pm$ 0.0105	0.2250	0.1732	0.2246	0.2453	0.2457	0.2228 $\pm$ 0.0296
	exp9	0.2914	0.3001	0.2936	0.2714	0.2916	0.2896 $\pm$ 0.0108	0.2090	0.1761	0.2397	0.2426	0.2458	0.2226 $\pm$ 0.0299
	exp10	0.2969	0.3120	0.3097	0.2800	0.3071	0.3011 $\pm$ 0.0132	0.2448	0.1855	0.2304	0.2505	0.2592	0.2341 $\pm$ 0.0291

## 4 DISCUSSION

In this section, we first show the most prominent SNP loci and brain regions identified by the proposed method, then discuss regularization influence of our method, show high-dimensional in simulations, and present the limitations of our method as well as possible future research directions.

### 4.1 Identification of The Most Related SNP Loci

As expected, the best-known risk genetic loci rs429358 has been identified by our proposed MD-SCCA method for all imaging data, which indicates the promise of our method in terms of identifying strong imaging genetic signals that are also related to AD diagnosis. The C allele increases the risk of the AD in *APOE* e4, which is encoded by rs429358 [44].

### 4.2 Identification of The Most Related Brain Regions

For detecting brain imaging ROIs, we average the obtained sparse coefficients by 5-fold cross-validation for FDG and

VBM imaging data. Then, the top 10 maximum weight ROIs are selected as the important ROI markers. The top 10 selected imaging features as shown in Table 4, as well as their average regression coefficients across five cross-validation trials, are visualized in Fig. 4 by mapping them onto the human brain [45]. As expected, left hippocampus, right hippocampus, left amygdala, and right amygdala have been detected on top 10 ROIs associated with the risk genotype biomarker by the proposed MD-SCCA method. It's worth noting that these stable markers are in accordance with the previous studies. For example, the reduction of hippocampal gray matter has been correlated with the risk SNP locus (rs429358-C) [42], [43], [46]. It is well known that the selected right calcarine cortex is related to the metabolic alteration in the brain, which shows that our proposed method can effectively utilize FDG imaging data [47], [48]. Meanwhile, the left inferior frontal gyrus (opercular) and left lingual gyrus have potentially associated with mild cognitive impairment [13], [49]. Besides confirming the



TABLE 4  
The Top 10 ROIs Selected by the MD-SCCA Method for FDG and VBM

ID	Name
37	L. Hippocampus
41	L. Amygdala
15	L. Orbitofrontal cortex (inferior)
47	L. Lingual gyrus
13	L. Inferior frontal gyrus (triangular)
38	R. Hippocampus
44	R. Calcarine cortex
28	R. Rectus gyrus
11	L. Inferior frontal gyrus (opercular)
42	R. Amygdala

L. = left; R. = right.

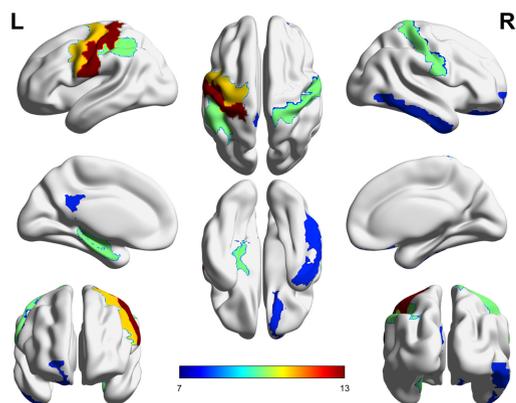


Fig. 4. Visualization of the mapping top 10 ROIs selected by the MD-SCCA method for FDG and VBM. The color represents the regression coefficients of the corresponding VBM markers.

TABLE 5  
The Averaged Correlation Coefficients on 5-Fold Test Data on Different Simulation Datasets

	Simulation data 1	
	M1	M2
MD-SCCA(no regularization)	0.9430	0.9368
MD-SCCA	0.9701	0.9719
	Simulation data 2	
	M1	M2
MD-SCCA(no regularization)	0.7697	0.7772
MD-SCCA	0.7956	0.8195
	Simulation data 3	
	M1	M2
MD-SCCA(no regularization)	0.4236	0.4752
MD-SCCA	0.5772	0.6307

(a) Results on simulation dataset 1. (b) Results on simulation dataset 2. (c) Results on simulation dataset 3. M1 and M2 are two different modalities.

### 4.3 Regularization Influence of Our Proposed Model

In the simulations, we also discuss our proposed model Eq. (10) without regularization terms since the sample size is much higher than the dimension. Table 5 and Fig. 5 show

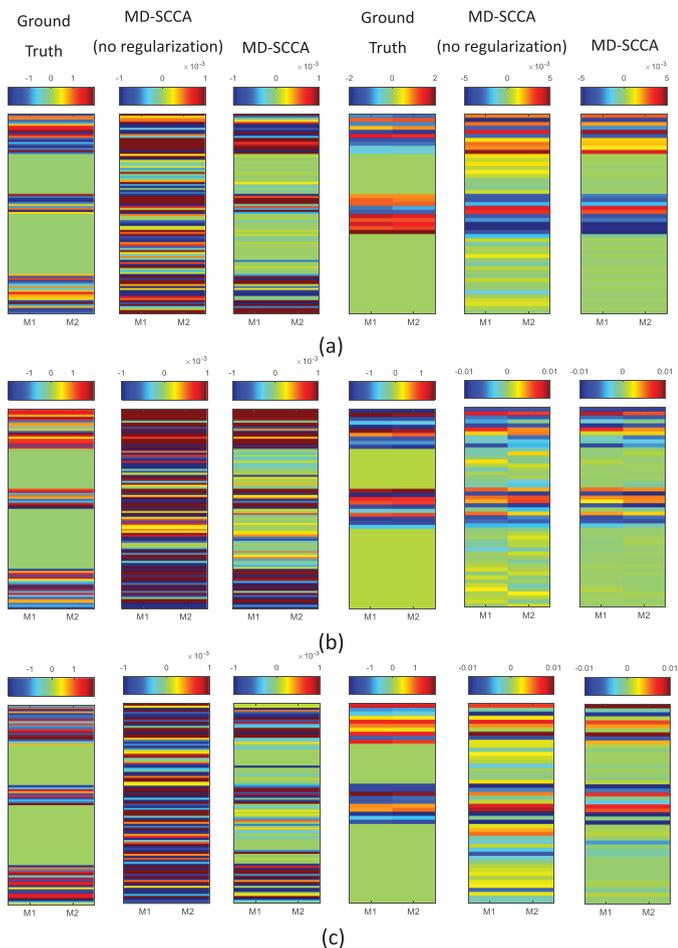


Fig. 5. The estimated weights of  $u$  and  $v$  from average 5-fold cross-validation test on different simulation datasets are shown in the left three panels and right three panels, corresponding to different methods. (a) Results on simulation dataset 1. (b) Results on simulation dataset 2. (c) Results on simulation dataset 3.

all the test performances of MD-SCCA and MD-SCCA without regularization terms in the simulations. As can be seen from Table 5 and Fig. 5, the regularization terms control the global sparsity of our proposed model, which can produce improved cross-validation performances.

### 4.4 High-Dimensional Case in Simulations

For our test, we also include the high-dimensional case into our simulations, which is similar to the procedure of simulation generation in Section 3.2. We consider the data belonging to two classes in all simulations. Each class consists of  $N = 100$  samples. Specifically, one latent variable  $t_1$  with normal distribution  $N(0, \sigma_t)$  for 80 samples is randomly generated. For the data matrices  $X$  and  $Y$ , the features are simulated from Gaussian distributions  $N(bt_1, \sigma_w I_l)$  and  $N(a_k t_1, \sigma_w I_s)$ , respectively; the other latent variable  $t_2$  with normal distribution  $N(1, \sigma_t)$  for 20 samples is randomly generated. For the data matrices  $X$  and  $Y$ , the features are simulated from Gaussian distributions  $N(bt_2, \sigma_w I_l)$  and  $N(a_k t_2, \sigma_w I_s)$ , respectively. Set  $l = 200$ ,  $s = 150$  ( $N < l < s$ ),  $s' = 30$ ,  $l' = 20$ , and  $\sigma_t = 0.1$ . In this paper, the noise levels  $\sigma_w = 0.1, 0.3, 0.5$  are given to generate three different simulation datasets.

We compare our proposed methods (including MD-SCCA and D-SCCA) with SCCA and AGN-SCCA

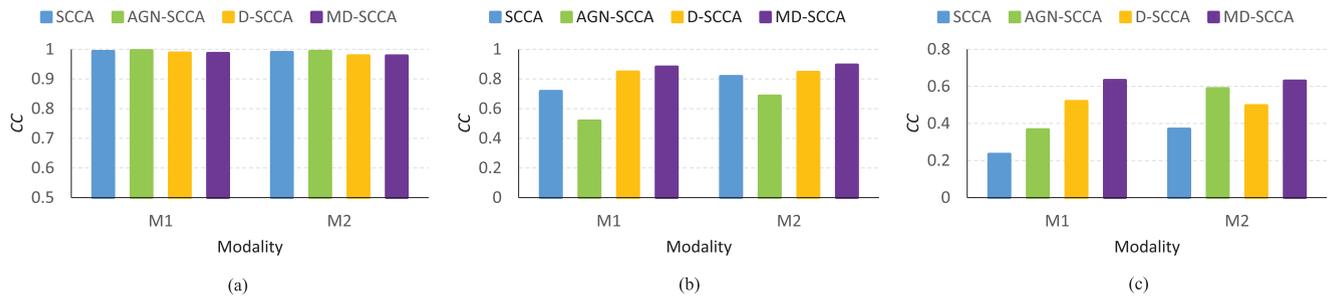


Fig. 6. The averaged correlation coefficients on 5-fold test data using different methods on different simulation datasets. (a) Results on simulation dataset 1. (b) Results on simulation dataset 2. (c) Results on simulation dataset 3.

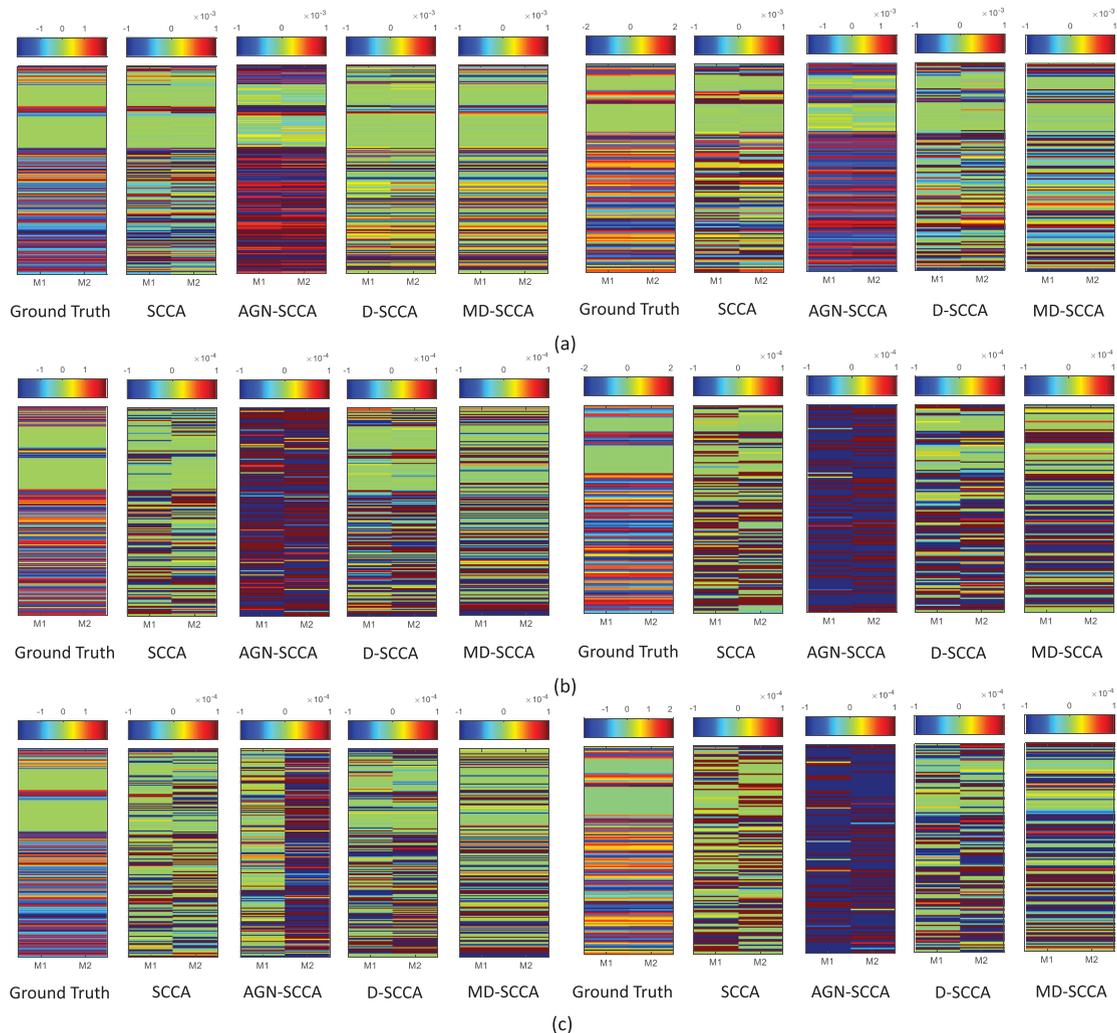


Fig. 7. The estimated weights of  $u$  and  $v$  from average 5-fold cross-validation test on different simulation datasets are shown in the left four panels and right four panels, corresponding to different methods. (a) Results on simulation dataset 1. (b) Results on simulation dataset 2. (c) Results on simulation dataset 3.

algorithms, respectively. As shown in Fig. 6, all methods yield stable results on the simulation data with low noise, however, our proposed MD-SCCA consistently outperforms SCCA, AGN-SCCA, and D-SCCA in the metric of correlation coefficients on the simulation data with high noise. Furthermore, in Fig. 7, we show the estimated canonical weights from different methods. As can be observed, the overall profiles of the estimated  $u$  and  $v$  values from MD-SCCA are consistent with the ground truth on simulation dataset.

#### 4.5 Limitations and Future Work

Although the proposed method achieves good results in brain imaging genetics association analysis, there are still several limitations to be considered in this study. First, when the number of samples exceeds the number of total features, the MD-SCCA model can be successfully applied for association discovery coupled with feature selection. However, when the datasets contain far more features (e.g., SNPs at the genome-wide magnitude), it will greatly increase the computational complexity and memory requirement. It is interesting to

further investigate the genome-wide analysis and build a scalable model to apply to more complex imaging genetic tasks, which will be one of our future works. Second, the normal equation in the optimization contains matrix inversion operations. We can consider developing a more efficient solution to further improve the performance of our method. Third, in this study, we have explored the information within-class subjects in feature extraction for brain imaging genetics association analysis. However, the effect of the subject gender is another important topic, which may affect the identifications in multivariate associations due to the potential bias introduced by different subjects in a study. In this case, the subject gender should be considered in the study.

## 5 CONCLUSION

In this paper, a brain imaging genetics study has been performed to explore the relationship between two modalities of imaging phenotypes (including VBM-MRI, FDG-PET) and genetic variations in the APOE gene. Because most of the existing SCCA algorithms have been designed to seek linear correlation of two data, which cannot yield optimal results owing to ignoring the disease-specific information in feature extraction for brain imaging genetics association analysis. We have proposed a novel MD-SCCA algorithm, which not only overcomes this above limitation, but also can incorporate valuable discriminant similarity information into SCCA algorithm to produce improved learning results. A comparative study has been performed between our proposed algorithm and the two competing SCCA algorithms on both synthetic and real data. The promising empirical results demonstrated that our proposed algorithm significantly outperforms the SCCA and AGN-SCCA algorithms in both cases. Furthermore, our proposed algorithm could accurately recover the true signals from the synthetic data, as well as yield improved canonical correlation performances and biologically meaningful findings from real data. Specifically, the main contribution of this work is to identify a compact set of robust and consistent genetic-imaging markers across the multi-modality phenotypes (i.e., MRI-VBM and FDG-PET) to have a better mechanistic understanding of AD biology.

## ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (Nos. 61876082, 61861130366, 61806071, 61732006), the National Key Research and Development Program of China (2018YFC2001600, 2018YFC2001602), the Royal Society-Academy of Medical Sciences Newton Advanced Fellowship (NAF\R1\180371), the Nanjing University of Aeronautics and Astronautics PHD Short-Term Visiting Scholar Project (56450081901), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) of China (201900043).

## REFERENCES

- [1] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, et al., "Forecasting the global burden of Alzheimer's disease," *Alzheimers Dementia*, vol. 3, no. 3, pp. 186–191, 2007.
- [2] A. M. Winklern, P. Kochunov, J. Blangero, et al., "Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies," *NeuroImage*, vol. 53, no. 3, pp. 1135–1146, 2010.
- [3] J. C. Lambert, "Meta-analysis in more than 74,000 individuals identifies 11 new susceptibility loci for Alzheimer's disease," *Alzheimers Dementia*, vol. 9, no. 4, pp. 1452–1458, 2013.
- [4] D. C. Glahn, P. M. Thompson, and J. Blangero, "Neuroimaging endophenotypes: Strategies for finding genes influencing brain structure and function," *Human Brain Mapping*, vol. 28, no. 6, pp. 488–501, 2007.
- [5] T. Ge, G. Schumann, and J. Feng, "Imaging genetics-towards discovery neuroscience," *Quantitative Biol.*, vol. 1, no. 4, pp. 227–245, 2013.
- [6] Y. Fu, Z. Ma, C. Hamilton, et al., "Genetic influences on resting-state functional networks: A twin study," *Human Brain Mapping*, vol. 36, no. 10, pp. 3959–3972, 2015.
- [7] X. Hao, X. Yao, J. Yan, et al., "Identifying multimodal intermediate phenotypes between genetic risk factors and disease status in Alzheimer's disease," *Neuroinformatics*, vol. 14, no. 4, pp. 1–14, 2016.
- [8] A. Song, J. Yan, S. Kim, et al., "Network-based analysis of genetic variants associated with hippocampal volume in Alzheimer's disease: A study of ADNI cohorts," *Neuroinformatics*, vol. 9, no. 1, pp. 1–8, 2016.
- [9] X. Yao, J. Yan, K. Liu, et al., "Tissue-specific network-based genome wide study of amygdala imaging phenotypes to identify functional interaction modules," *Bioinf.*, vol. 33, no. 20, pp. 3250–3257, 2017.
- [10] B. Consortium, V. Anttila, B. Bulik-Sullivan, et al., "Analysis of shared heritability in common disorders of the brain," *Sci.*, vol. 360, no. 6395, 2018, Art. no. eaap8757.
- [11] M. Wang, X. Hao, J. Huang, et al., "Discovering network phenotype between genetic risk factors and disease status via diagnosis-aligned multi-modality regression method in Alzheimer's disease," *Bioinf.*, vol. 35, no. 11, pp. 1948–1957, 2019.
- [12] W. Shao, Z. Han, J. Cheng, et al., "Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis," *IEEE Trans. Med. Imag.*, early access, Jun. 3, 2019, doi: 10.1109/TMI.2019.2920608.
- [13] T. G. P. Consortium, "A global reference for human genetic variation, the 1000 genomes project consortium," *Nature*, vol. 526, no. 7571, pp. 68–74, 2015.
- [14] L. Shen, S. Kim, S. L. Risacher, et al., "Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort," *NeuroImage*, vol. 53, no. 3, pp. 1051–1063, 2010.
- [15] J. L. Stein, X. Hua, S. Lee, et al., "Voxelwise genome-wide association study (vGWAS)," *NeuroImage*, vol. 53, no. 3, pp. 1160–1174, 2010.
- [16] V. Anttila, B. S. Winsvold, P. Gormley, et al., "Genome-wide meta-analysis identifies new susceptibility loci for migraine," *Nature Genetics*, vol. 45, no. 8, pp. 912–917, 2013.
- [17] T. W. Muhleisen, M. Leber, T. G. Schulze, et al., "Genome-wide association study reveals two new risk loci for bipolar disorder," *Nature Commun.*, vol. 5, no. 3, 2014, Art. no. 3339.
- [18] M. A. Nalls, N. Pankratz, C. M. Lill, et al., "Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease," *Nature Genetics*, vol. 46, no. 9, pp. 989–993, 2014.
- [19] D. P. Hibar, O. Kohannim, J. L. Stein, et al., "Multilocus genetic analysis of brain images," *Frontiers Genetics*, vol. 2, pp. 73–84, 2011.
- [20] H. Zhu, S. Zhang, and Q. Sha, "A novel method to test associations between a weighted combination of phenotypes and genetic variants," *Plos One*, vol. 13, no. 1, 2018, Art. no. e0190788.
- [21] M. Vounou, T. E. Nichols, and G. Montana, "Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach," *NeuroImage*, vol. 53, no. 3, pp. 1147–1159, 2010.
- [22] J. Wan, S. Kim, M. Inlow, et al., "Hippocampal surface mapping of genetic risk factors in AD via sparse learning models," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2011, pp. 376–383.
- [23] E. C. Chi, G. I. Allen, H. Zhou, et al., "Imaging genetics via sparse canonical correlation analysis," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2013, pp. 740–743.
- [24] D. Lin, V. D. Calhoun, and Y. Wang, "Correspondence between fMRI and SNP data by group sparse canonical correlation analysis," *Med. Image Anal.*, vol. 18, no. 6, pp. 891–902, 2014.
- [25] J. Yan, L. Du, S. Kim, et al., "Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm," *Bioinf.*, vol. 30, no. 17, pp. 564–571, 2014.
- [26] P. Zille, V. D. Calhoun, and Y. Wang, "Enforcing co-expression within a brain-imaging genomics regression framework," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2561–2571, Dec. 2018.
- [27] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

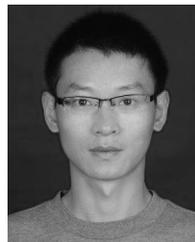
- [28] W. Hu, D. Lin, S. Cao, et al., "Adaptive sparse multiple canonical correlation analysis with application to imaging (Epi)genomics study of schizophrenia," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 2, pp. 390–399, Feb. 2018.
- [29] X. Hao, C. Li, J. Yan, et al., "Identification of associations between genotypes and longitudinal phenotypes via temporally-constrained group sparse canonical correlation analysis," *Bioinf.*, vol. 33, no. 14, pp. i341–i349, 2017.
- [30] J. Fang, D. Lin, S. C. Schulz, et al., "Joint sparse canonical correlation analysis for detecting differential imaging genetics modules," *Bioinf.*, vol. 32, no. 22, pp. 3480–3488, 2016.
- [31] L. Du, H. Huang, J. Yan, et al., "Structured sparse canonical correlation analysis for brain imaging genetics: An improved graphnet method," *Bioinf.*, vol. 32, no. 10, pp. 1544–1551, 2016.
- [32] L. Du, J. Yan, S. Kim, et al., "GN-SCCA: Graphnet based sparse canonical correlation analysis for brain imaging genetics," *Brain Inform. Health*, vol. 9250, pp. 275–284, 2015.
- [33] W. Shao, J. Cheng, L. Sun, Z. Han, Q. Feng, D. Zhang, and K. Huang, "Ordinal multi-modal feature selection for survival analysis of early-stage renal cancer," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2018, pp. 648–656.
- [34] T. Sun and S. Chen, "Locality preserving CCA with applications to data visualization and pose estimation," *Image Vis. Comput.*, vol. 25, no. 5, pp. 531–543, 2007.
- [35] L. Qiao, S. Chen, and X. Tan, "Sparsity preserving projections with applications to face recognition," *Pattern Recognit.*, vol. 43, no. 1, pp. 331–341, 2010.
- [36] J. Liu, S. Ji, and J. Ye, "SLEP: Sparse learning with efficient projections," Arizona State University. 2009. [Online]. Available: <http://www.public.asu.edu/~jye02/Software/SLEP>.
- [37] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: A survey and extensions," *Math. Methods Operations Res.*, vol. 66, no. 3, pp. 373–407, 2007.
- [38] X. Hao, C. Li, X. Yao, et al., "Mining outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease," *Sci. Rep.*, vol. 7, 2017, Art. no. 44272.
- [39] L. Du, J. Yan, S. Kim, et al., "A novel structure-aware sparse learning algorithm for brain imaging genetics," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2014, pp. 329–336.
- [40] J. Chen, F. D. Bushman, J. D. Lewis, et al., "Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis," *Biostatistics*, vol. 14, no. 2, pp. 244–258, 2013.
- [41] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, et al., "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002.
- [42] V. K. Ramanan, S. L. Risacher, K. Nho, et al., "APOE and BCHE as modulators of cerebral amyloid deposition: A florbetapir PET genome-wide association study," *Mol. Psychiatry*, vol. 19, no. 3, pp. 351–357, 2014.
- [43] M. de Leon, A. Convit, S. DeSanti, et al., "The hippocampus in aging and Alzheimer's disease," *Neuroimaging Clinics North Amer.*, vol. 5, no. 1, pp. 1–17, 1995.
- [44] J. C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, et al., "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease," *Nature Genetics*, vol. 45, pp. 1452–1458, 2013.
- [45] M. Xia, J. Wang, and Y. He, "BrainNet viewer: A network visualization tool for human brain connectomics," *Plos One*, vol. 8, no. 7, 2013, Art. no. e68910.
- [46] H. A. Wishart, A. J. Saykin, T. W. McAllister, et al., "Regional brain atrophy in cognitively intact adults with a single APOE epsilon4 allele," *Neurology*, vol. 67, no. 7, pp. 1221–1224, 2006.
- [47] Y. Liu, J. Yu, H. Wang, et al., "APOE genotype and neuroimaging markers of Alzheimer's disease: Systematic review and meta-analysis," *J. Neurology Neurosurgery Psychiatry*, vol. 86, no. 2, pp. 127–134, 2015.
- [48] E. M. Reiman, R. J. Caselli, L. S. Yun, et al., "Preclinical evidence of Alzheimer's disease in persons homozygous for the epsilon 4 allele for apolipoprotein E," *New England J. Med.*, vol. 334, no. 12, pp. 752–758, 1996.
- [49] T. Zhang and C. Davatzikos, "Optimally-discriminative voxel-based morphometry significantly increases the ability to detect group differences in schizophrenia, mild cognitive impairment, and Alzheimer's disease," *NeuroImage*, vol. 79, pp. 94–110, 2013.



**Meiling Wang** received the MS degree in information and communication engineering from the Nanjing University of Information Science and Technology, China, in 2016. She is currently working toward the PhD degree in the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. Her research interests include machine learning and brain imaging genetics.



**Wei Shao** received the PhD degrees from the Nanjing University of Aeronautics and Astronautics, in 2018. His current research interests include machine learning and bioinformatics.



**Xiaoke Hao** received the BS and MS degrees from the Nanjing University of Information Science and Technology, Nanjing, China, in 2009 and 2012, respectively, and the PhD degree in computer science and technology from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2017. He is currently an assistant professor with the School of Artificial Intelligence of the Hebei University of Technology. His research interests include machine learning and imaging genetics. He is a member of the IEEE.



**Li Shen** received the BS degree from Xian Jiao Tong University, the MS degree from Shanghai Jiao Tong University, and the PhD degree from Dartmouth College, all in computer science. He is a professor of the Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine School of Medicine. His research interests include medical image computing, bioinformatics, data mining, network science, systems biology, brain imaging genomics, and brain connectomics.



**Daoqiang Zhang** received the BS and PhD degrees in computer science from the Nanjing University of Aeronautics and Astronautics (NUAA), China, in 1999, and 2004, respectively. He joined the Department of Computer Science and Engineering of NUAA as a lecturer, in 2004, and is a professor at present. His research interests include machine learning, pattern recognition, data mining, and medical image analysis. In these areas, he has published more than 150 scientific articles in refereed international journals such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *IEEE Transactions on Medical Imaging*, the *NeuroImage*, *Human Brain Mapping*, *Medical Image Analysis*; and conference proceedings such as IJCAI, AAAI, NIPS, CVPR, and MICCAI, with more than 8,000 citations by Google Scholar. He was nominated for the National Excellent Doctoral Dissertation Award of China, in 2006, won the best paper award or best student award of several international conferences such as PRICAI'06, STMI'12, and BICS'16, etc. He has served as a program committee member for several international and native conferences such as IJCAI, AAAI, NIPS, SDM, PRICAI, and ACML, etc. He is a member of the Machine Learning Society of the Chinese Association of Artificial Intelligence (CAAI), and the Artificial Intelligence & Pattern Recognition Society of the China Computer Federation (CCF).

► For more information on this or any other computing topic, please visit our Digital Library at [www.computer.org/csdl](http://www.computer.org/csdl).