Multi-Task Sparse Canonical Correlation Analysis with Application to Multi-Modal Brain Imaging Genetics

Lei Du⊙, Kefei Liu⊙, Xiaohui Yao, Shannon L. Risacher, Junwei Han⊙, Andrew J. Saykin, Lei Guo, and Li Shen⊙

Abstract—Brain imaging genetics studies the genetic basis of brain structures and functionalities via integrating genotypic data such as single nucleotide polymorphisms (SNPs) and imaging quantitative traits (QTs). In this area, both multi-task learning (MTL) and sparse canonical correlation analysis (SCCA) methods are widely used since they are superior to those independent and pairwise univariate analysis. MTL methods generally incorporate a few of QTs and could not select features from multiple QTs; while SCCA methods typically employ one modality of QTs to study its association with SNPs. Both MTL and SCCA are computational expensive as the number of SNPs increases. In this paper, we propose a novel multi-task SCCA (MTSCCA) method to identify bi-multivariate associations between SNPs and multi-modal imaging QTs. MTSCCA could make use of the complementary information carried by different imaging modalities. MTSCCA enforces sparsity at the group level via the $G_{2,1}$-norm, and jointly selects features across multiple tasks for SNPs and QTs via the $\ell_{2,1}$-norm. A fast optimization algorithm is proposed using the grouping information of SNPs. Compared with conventional SCCA methods, MTSCCA obtains better correlation coefficients and canonical weights patterns. In addition, MTSCCA runs very fast and easy-to-implement, indicating its potential power in genome-wide brain-wide imaging genetics.

Index Terms—Brain imaging genetics, sparse canonical correlation analysis, multi-task sparse canonical correlation analysis

✦

1 INTRODUCTION

IMAGING genetics is an emerging and important topic which integrates both the genetic factors and neuroimaging phenotypic measurements in brain science. This integration research of combining diverse genetic and genomic data is expected to uncover the genetic basis of brain structures and functionalities, and further offers new opportunities to interpret the causality of relationships between genetic variations and brain disorders such as the Alzheimer's disease (AD) [1], [2]. Modern neuroimaging techniques, such as magnetic resonance imaging (MRI) and positron-emission tomography (PET), image the morphometry and metabolic processes of the brain based on different techniques, and generate different imaging data describing the brain from different perspectives. These multi-modal imaging data provide complementary information and have been demonstrated to offer comprehensive understandings of the brain

structures, functionalities, and brain disorders [3]. Moreover, in biomedical studies, we usually face a huge size of genotyping biomarkers such as the single nucleotide polymorphisms (SNPs), which is a type of high-resolution markers in genome-wide association studies (GWAS). Therefore, developing the fast and efficient GWAS-oriented imaging genetics method which integrates multimodal imaging data simultaneously is of great importance and meaning.

The multivariate learning methods are very popular in brain imaging genetics since both imaging data and genetic data are multivariate. The multi-task learning (MTL) techniques are of this kind and widely used in brain imaging genetics [4], [5]. Generally, these methods choose a few important imaging QTs relevant to their aim as dependent variables and SNPs as independent variables. Then joint effect of multi-locus genotype on few phenotypes is studied. This paradigm can select SNPs that are simultaneously relevant to the candidate brain phenotypes. However, the brain is demonstrated to be comprised of multiple regions. Then using only a small proportion of them could be lack of power since they may lose important information carried by cerebral components which are not included.

Although a brain-wide MTL model can be used, they are still insufficient since they cannot select relevant brain phenotypes from multiple brain cerebral components. Therefore, bi-multivariate methods become more and more popular in brain imaging genetics recently. Sparse canonical correlation analysis (SCCA) is such a technique which usually identifies

- L. Du, J. Han, and L. Guo are with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China.
  E-mail: {dulei, jhan, lguo}@nwpu.edu.cn.
- K. Liu, X. Yao, and L. Shen are with the Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania's Perelman School of Medicine, Philadelphia, PA 19104 USA.
  E-mail: {kefei.liu, Xiaohui.Yao, Li.Shen}@pennmedicine.upenn.edu.
- S.L. Risacher and A.J. Saykin are with the Department of Radiology and Imaging Sciences, Indiana University's School of Medicine, Indianapolis, IN 46202 USA. E-mail: {srisache, asaykin}@iupui.edu.

the relationship between two views of data with sparse output induced by different regularization techniques [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]. These two-view SCCA methods have limited power since they only utilize one modal imaging QTs. Given multi-modal imaging data, incorporating them together could make use of the information carried by different modalities and would be beneficial to uncover interesting findings that using one modality cannot. Therefore, jointly analyzing the relationship between all the imaging phenotypes from different modalities and genetic factors via one single integrative SCCA model is desirable and of great interest. And this integrative model would be helpful to elucidate the shared mechanism of genetic factors on the brain.

One possible solution is the multi-view SCCA modelling which considers the pairwise relationship between all omics data involved. This multi-view SCCA is a naive extension to existing two-view SCCA models, and a three-view one has been introduced in [13]. It learns only one single canonical weight for genetic loci which is overstrict thus cannot make full use of the complementary information embedded in different modalities of imaging phenotypes.

Using brain-wide imaging QTs from multiple modalities, in this paper, we propose a Multi-Task learning based SCCA (MTSCCA) framework [16], [17] which can study bi-multivariate associations between these phenotypes and genotypes simultaneously. MTSCCA treats each SNP and QT as a feature, and then models the association between each imaging modality and SNPs as a learning task. Different from those conventional SCCA, including both two-view and three-view methods, MTSCCA learns one canonical weight matrix for SNPs, in which each column vector corresponds to one canonical weight of one SCCA task. In contrast, only one canonical weight vector is associated with each imaging modality. To make the model practical, we take into consideration the group structure such as the linkage disequilibrium (LD) [18] in human genome via the group $\ell_{2,1}$-norm ($\text{G}_{2,1}$-norm) [5] regularization. The joint individual feature selection for genetic and phenotypic markers is also taken into consideration via an $\ell_{2,1}$-norm constraint. In addition, we propose a fast and efficient optimization algorithm which is guaranteed to converge to a local optimum. We apply MTSCCA to a very large real neuroimaging genetic data set from the Alzheimer's disease neuroimaging initiative (ADNI) [19] cohort with all SNPs in the 19th chromosome and three different modalities of imaging QTs included. We intend to reveal the associations between these genetic markers and imaging phenotypes. Experimental results show that, compared with both two-view and multi-view SCCA methods, MTSCCA yields better canonical correlation coefficients and canonical weights. It also reports a compact set of SNPs and imaging QTs known to be associated with AD. Moreover, MTSCCA runs very fast and could be a powerful tool to genome-wide brain-wide bi-multivariate association analysis.

2 METHODOLOGY

We denote scalars as italic letters, column vectors as boldface lowercase letters, and matrices as boldface capitals. For $\mathbf{X} = (x_{ij})$, its ith row is denoted as $\mathbf{x}^i$ and jth column is $\mathbf{x}_j$, and $\mathbf{X}_i$ denotes the ith matrix. $\|\mathbf{x}\|_2$ denotes the euclidean norm, $\|\mathbf{X}\|_\text{F} = \sqrt{\sum_i \sum_j x_{ij}^2}$ denotes the Frobenius norm.

2.1 Background

Let $\mathbf{X}_i$ and $\mathbf{w}_i$, $i = 1, \ldots, I$, represent the data matrices and the corresponding canonical weights, respectively. Further, we use $\mathbf{X}_1$ load the SNP data, and those remaining $\mathbf{X}_k$'s ($k \neq 1$) load the imaging QT data of each imaging modality separately. Then the conventional SCCA is defined as

$$\min_{\mathbf{w}_1, \ldots, \mathbf{w}_I} \sum_{i < j} -\mathbf{w}_i^\top \mathbf{X}_i^\top \mathbf{X}_j \mathbf{w}_j$$
$$s.t. \quad \|\mathbf{X}_i \mathbf{w}_i\|_2^2 = 1, \Omega(\mathbf{w}_i) \leq b_i, \ i = 1, \ldots, I, \tag{1}$$

where $\Omega(\mathbf{w}_i)$ is the penalty function to accommodate sparsity and thus can select those features of interest. Many penalty functions have been studied in the literature such as Lasso (ℓ_1-norm function) [10], [13], [20], group Lasso [11] and graphical Lasso [9], [12]. Conventionally, we call it two-view SCCA (SCCA for short) when $I = 2$, and most existing studies fall into this category. Those methods using three or more sets of data ($I \geq 3$) are called multi-view or multi-set SCCA (mSCCA) [13]. The two-view SCCA only uses one modality of imaging QTs to study the genetic influence on brain functions or structures, and the mSCCA learns only one canonical weight for genetic data which must be correlated to all imaging QTs simultaneously.

2.2 MTSCCA

2.2.1 The MTSCCA Model

To distinguish from the notation in mSCCA, in this section, we use $\mathbf{X} \in \mathcal{R}^{n \times p}$ to represent the genetic data with n participants and p SNPs, and $\mathbf{Y}_j \in \mathcal{R}^{n \times q}(j = 1, \ldots, c)$ to represent the phenotype data with q imaging measurements, where c is the number of imaging modalities (tasks). Let $\mathbf{U} \in \mathcal{R}^{p \times c}$ be the canonical weight matrix associated with $\mathbf{X}$ and $\mathbf{V} \in \mathcal{R}^{q \times c}$ be that associated with imaging QTs with each $\mathbf{v}_j$ corresponding to $\mathbf{Y}_j$, we propose the novel multi-task based SCCA model as follows

$$\min_{\mathbf{u}_j, \mathbf{v}_j} \sum_j -\mathbf{u}_j^\top \mathbf{X}^\top \mathbf{Y}_j \mathbf{v}_j$$
$$s.t. \quad \|\mathbf{X}\mathbf{u}_j\|_2^2 = 1, \|\mathbf{Y}_j \mathbf{v}_j\|_2^2 = 1, \Omega(\mathbf{U}) \leq b_1, \Omega(\mathbf{V}) \leq b_2, \ \forall j. \tag{2}$$

Obviously, our model is distinct from those mCCA. First, MTSCCA employs the multi-task framework which learns a series of related SCCA tasks together. This simultaneous learning has been empirically [21], [22] and theoretically [21], [23] shown to improve performance dramatically compared with learning each task independently [24]. Second, our model learns a canonical weight matrix $\mathbf{U}$ for SNPs, in which each column $\mathbf{u}_j$ corresponds to an individual SCCA task. This is helpful since it does not require a unique canonical weight of SNPs to be associated with all modalities of imaging QTs at the same time. Third, MTSCCA learns one canonical weight corresponding to each imaging modality separately, indicating that we do not need to calculate multiple canonical weights for a specific imaging modality. This helps the model focus on the identification of markers from the genetic data, indicating it is quite suitable for imaging genetics analysis. Finally, our mode can be well scalable in terms of both modeling and computation. According to Eqs. (1) and (2), the number of tasks of MTSCCA equals the

number of imaging modalities which implies a linear relationship; while that of mCCA increases quadratically as the number of imaging modalities increase since it does a CCA task between every pair of data sets, including the pairwise SCCA among imaging modalities.

2.2.2 Group-Sparsity and Joint Individual Feature Selection for SNPs

Since numerous SNPs inherently exhibit group structure in the genome, a realistic modeling method should take this information into consideration. In Eq. (2), a canonical weight matrix is associated with SNPs, and thus the conventional group Lasso which is used to penalize a vector cannot be employed directly. To tackle this issue, we use the $G_{2,1}$-norm function [5] which is formulated as

$$\|\mathbf{U}\|_{G_{2,1}} = \sum_{k=1}^{K} \|\mathbf{U}^k\|_F = \sum_{k=1}^{K} \sqrt{\sum_{i \in g_k} \sum_{j=1}^{c} u_{ij}^2}, \quad (3)$$

where the SNPs are partitioned into K groups $\mathcal{G} = \{g_k\}_{k=1}^{K}$. This regularization penalizes the SNPs in the same group as a whole and expects to estimate equal or similar coefficients for them. According to [5], this penalty has two major merits. First, it incorporates the group structural knowledge into the model via packaging all SNPs in the same group together. This makes the model practical because it is in accordance with the genetic mechanism. Second, it penalizes the canonical weight coefficients of a group of variables across all SCCA tasks jointly. This setup can mutually promote each individual task.

Although using $G_{2,1}$-norm regularization is meaningful, there is a lack of feature selection at individual level. For those disease related SNPs, they could hardly be located in the same group. Generally, within a specific group, an individual variable could be relevant to the QTs and those remaining ones could be irrelevant. Therefore, we also model this via the $\ell_{2,1}$-norm regularization which is the Lasso regularization adjusted for multi-task feature selection,

$$\|\mathbf{U}\|_{2,1} = \sum_{i=1}^{p} \|\mathbf{u}^i\|_2 = \sum_{i=1}^{p} \sqrt{\sum_{j=1}^{c} u_{ij}^2}. \quad (4)$$

Using both $G_{2,1}$-norm and $\ell_{2,1}$-norm regularization, MTSCCA can not only select features at the group level in accordance with the biological knowledge, but also jointly select feature at the individual level across all SCCA tasks.

2.2.3 Joint Individual Feature Selection Across Different Imaging Modalities

Apart from the identification of risk genetic factors, identifying the AD risk imaging biomarkers is also of great concern. In this study, in addition to the canonical weight matrix for SNPs, MTSCCA also learns one canonical weight for each imaging modality. For a larger number of imaging features, a non-sparse result without feature selection makes the model complex and hard to interpret. Therefore, sparsity-inducing regularization is necessary for those imaging biomarkers too.

In the MTSCCA model, we use the $\ell_{2,1}$-norm function on the imaging QTs, i.e.,

$$\|\mathbf{V}\|_{2,1} = \sum_{i=1}^{q} \|\mathbf{v}^i\|_2 = \sum_{i=1}^{q} \sqrt{\sum_{j=1}^{c} v_{ij}^2}. \quad (5)$$

At first glance this is similar to that used to jointly select individual features for SNPs, but it is employed here based on a different motivation. Although collected based on different imaging technologies, all modalities of imaging QTs are measured from the same brain geography and have been mapped onto the same brain atlas via the segmentation and registration. Thus it is reasonable to assume equal or similar weights for those imaging QTs associated with the same brain area but attributed to different modalities. Therefore, the ℓ_2-norm imposed on $\mathbf{v}^i$ penalizes those QTs from the same brain area but different modalities together, and then the ℓ_1-norm is utilized to select them jointly.

2.3 The Efficient Optimization Algorithm

Now we can write the MTSCCA with penalties explicitly exhibited, i.e.,

$$\min_{\mathbf{u}_j, \mathbf{v}_j} \sum_{j=1}^{c} -\mathbf{u}_j^\top \mathbf{X}^\top \mathbf{Y}_j \mathbf{v}_j$$
$$s.t. \quad \|\mathbf{X}\mathbf{u}_j\|_2^2 = 1, \|\mathbf{Y}_j\mathbf{v}_j\|_2^2 = 1,$$
$$\|\mathbf{U}\|_{G_{2,1}} \le a, \|\mathbf{U}\|_{2,1} \le b_1, \|\mathbf{V}\|_{2,1} \le b_2, \ \forall j. \quad (6)$$

In order to solve Eq. (6), we modify the loss function to

$$\min_{\mathbf{u}_j, \mathbf{v}_j} \sum_{j=1}^{c} \|\mathbf{X}\mathbf{u}_j - \mathbf{Y}_j \mathbf{v}_j\|_2^2$$
$$s.t. \quad \|\mathbf{X}\mathbf{u}_j\|_2^2 = 1, \|\mathbf{Y}_j\mathbf{v}_j\|_2^2 = 1,$$
$$\|\mathbf{U}\|_{G_{2,1}} \le a, \|\mathbf{U}\|_{2,1} \le b_1, \|\mathbf{V}\|_{2,1} \le b_2, \ \forall j, \quad (7)$$

which is equivalent to the original one since $\forall j, \|\mathbf{X}\mathbf{u}_j\|_2^2 = 1$ and $\|\mathbf{Y}_j\mathbf{v}_j\|_2^2 = 1$. Then we write its Lagrangian

$$\mathcal{L}(\mathbf{U}, \mathbf{V}) = \sum_{j=1}^{c} \left[\|\mathbf{X}\mathbf{u}_j - \mathbf{Y}_j \mathbf{v}_j\|_2^2 + \gamma_1(\|\mathbf{X}\mathbf{u}_j\|_2^2 - 1) + \gamma_2(\|\mathbf{Y}_j\mathbf{v}_j\|_2^2 - 1) \right]$$
$$+ \beta(\|\mathbf{U}\|_{G_{2,1}} - a) + \lambda_1(\|\mathbf{U}\|_{2,1} - b_1) + \lambda_2(\|\mathbf{V}\|_{2,1} - b_2), \quad (8)$$

where β, λ_1, λ_2, γ_1 and γ_2 are tuning parameters, and β, λ_1 and λ_2 are positive values which control the model sparsity. By dropping the constants, we further have

$$\mathcal{L}(\mathbf{U}, \mathbf{V}) = \sum_{j=1}^{c} \left[\|\mathbf{X}\mathbf{u}_j - \mathbf{Y}_j \mathbf{v}_j\|_2^2 + \gamma_1 \|\mathbf{X}\mathbf{u}_j\|_2^2 + \gamma_2 \|\mathbf{Y}_j\mathbf{v}_j\|_2^2 \right]$$
$$+ \beta \|\mathbf{U}\|_{G_{2,1}} + \lambda_1 \|\mathbf{U}\|_{2,1} + \lambda_2 \|\mathbf{V}\|_{2,1}, \quad (9)$$

from the point of view of optimization.

This equation is difficult to solve since it is non-convex in the loss function and non-smooth in penalty functions. Fortunately, it is convex in $\mathbf{U}$ with $\mathbf{V}$ fixed. Moreover, this objective is convex in $\mathbf{v}_j$ with those remaining $\mathbf{v}_k(k \neq j)$ and $\mathbf{U}$ fixed. On this account, we can solve this problem via the alternative update rule which is widely used in the optimizing community.
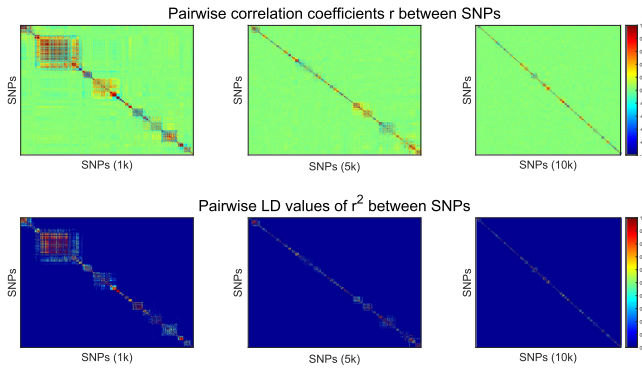
Fig. 1. Illustration of the pairwise correlation coefficients and LD values ($r^2 \geq 0.2$) of SNPs from Chromosome 19 of an ADNI database. (1) The three subfigures above show the correlation coefficients r among SNPs with a number of 1,000, 5,000, and 13,000. (2) The three subfigures below are the corresponding values of LD. All figures show that SNPs clearly form groups and the block diagonal structure always exists as the number of SNPs increases.

2.3.1 Updating $\mathbf{U}$

We first show solving $\mathbf{U}$ with $\mathbf{V}$ fixed. Since all $\mathbf{u}_j$'s are associated with $\mathbf{X}$, they can be jointly calculated via a multi-task framework. Taking the derivative of $\mathcal{L}(\mathbf{U}, \mathbf{V})$ with respect to $\mathbf{U}$ and letting it be 0, we arrive at

$$-\mathbf{X}^\top \mathbb{Y} + \beta \tilde{\mathbf{D}} \mathbf{U} + \lambda_1 \mathbf{D}_1 \mathbf{U} + \gamma'_1 \mathbf{X}^\top \mathbf{X} \mathbf{U} = \mathbf{0}, \qquad (10)$$

where $\mathbb{Y} = [\mathbf{Y}_1 \mathbf{v}_1 \quad \mathbf{Y}_2 \mathbf{v}_2 \quad \cdots \quad \mathbf{Y}_c \mathbf{v}_c]$, $2\tilde{\mathbf{D}} \mathbf{U}$ is the subgradient of $\|\mathbf{U}\|_{G_{2,1}}$ and $2\mathbf{D}_1 \mathbf{U}$ is that of $\|\mathbf{U}\|_{2,1}$; $\tilde{\mathbf{D}}$ is a block diagonal matrix with entries being $\frac{1}{2\|\mathbf{U}^k\|_F} \mathbf{I}_k (k \in [1, K])$, and $\mathbf{I}_k$ is an identity matrix of size equaling to the kth group; $\mathbf{D}_1$ is also a diagonal matrix with diagonal entries being $\frac{1}{2\|\mathbf{u}^i\|_2} (i \in [1, p])$; and $\gamma'_1 = \gamma_1 + 1$.

Then we can easily have

$$(\beta \tilde{\mathbf{D}} + \lambda_1 \mathbf{D}_1 + \gamma'_1 \mathbf{X}^\top \mathbf{X}) \mathbf{U} = \mathbf{X}^\top \mathbb{Y}, \qquad (11)$$

and further

$$\mathbf{U} = (\beta \tilde{\mathbf{D}} + \lambda_1 \mathbf{D}_1 + \gamma'_1 \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{Y}. \qquad (12)$$

According to [5], this linear system in terms of $\mathbf{U}$ can be efficiently solved via an iterative algorithm by alternatively first updating $\tilde{\mathbf{D}}$ and $\mathbf{D}_1$ and then $\mathbf{U}$. However, if the number of SNPs becomes larger and larger, this iterative algorithm is still computationally expensive.

A Fast Implementation. The primary difficulty of the $\mathbf{U}$-update is the calculation of the covariance matrix $\mathbf{X}^\top \mathbf{X}$ when there are a large number of features of $\mathbf{X}$. In this paper, we use an approximation method to assure a fast implementation of $\mathbf{X}^\top \mathbf{X}$ via making use of the priori knowledge, i.e., the inherent structure of the SNPs within the genome. Fig. 1 is the illustration of the pairwise correlation coefficients and LD values in r^2 among a segment of SNPs at different loci from chromosome 19. The SNPs naturally form block structure along the diagonal, indicating a clear pattern of intra-block high correlation and inter-block low correlation. Since $\mathbf{X}$ is centered and normalized, $\mathbf{X}^\top \mathbf{X}$ is the same as the pairwise correlation coefficients as shown in Fig. 1. This indicates that $\mathbf{X}^\top \mathbf{X}$ holds block diagonal structure too, and its off-block-diagonal elements are nearly zero, i.e., $\mathbf{X}_{g_k}^\top \mathbf{X}_{g_t} \approx \mathbf{0}$ ($k \neq t$). In a word, the
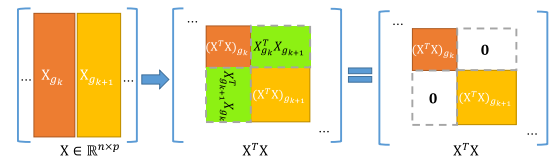

Fig. 2. Illustration of the simplified covariance matrix $\mathbf{X}^\top \mathbf{X}$, where $\mathbf{X}_{g_k}$ and $\mathbf{X}_{g_{k+1}}$ are two LD blocks, and $\mathbf{X}_{g_k}^\top \mathbf{X}_{g_k}$ is abbreviated as $(\mathbf{X}^\top \mathbf{X})_{g_k}$. Since the correlation between the two blocks are very low ($\mathbf{X}_{g_k}^\top \mathbf{X}_{g_{k+1}} \approx \mathbf{0}$ and $\mathbf{X}_{g_{k+1}}^\top \mathbf{X}_{g_k} \approx \mathbf{0}$), their covariance can be ignored.

information of the covariance matrix are mainly carried by a series of block matrices along the diagonal. Most importantly, the size of these blocks are quite small compared with the original covariance matrix attributing to the fact that the LD block is usually much smaller than the number of SNPs ($p_k \ll p$) in human genome [25].

This structure has been widely used to guide the recovery of group relationship among SNPs via the group Lasso [26] or $G_{2,1}$-norm [5]. However, they suffer from heavy computational issues caused by the enormous SNPs, and only when artificially assuming that $\mathbf{X}^\top \mathbf{X}$ is an identity matrix could alleviate this issue [6], [7], [13]. From Fig. 1, we know that the identity assumption will inevitably lose information carried by those blocks along the diagonal [11]. In this study, we not only make use of this grouping information to identify relationships among SNPs, but also explore a fast and easy-to-implement method to handle the computational issues.

Based on the analysis above, we propose that $\mathbf{X}^\top \mathbf{X}$ can be computationally simplified by a series of $(\mathbf{X}^\top \mathbf{X})_{g_k}$ (abbreviated from $\mathbf{X}_{g_k}^\top \mathbf{X}_{g_k}$) along the diagonal. We only omit those off-block-diagonal elements which has little influence on the performance. Fig. 2 is the illustration of the approximation where the off-block-diagonal elements are replaced by zero. It is clear that the primary information of $\mathbf{X}^\top \mathbf{X}$ are well preserved since we take into consideration the LD structure. Therefore, compared with those methods using identity assumption, our method preserves more information of the data, and could be useful in identification of important genetic markers [11]. Most importantly, other than those methods calculating $\mathbf{X}^\top \mathbf{X}$ via the brute force [11], we have a very fast implementation which is supported by the following theorem.

Theorem 1. *If $\mathbf{X}^\top \mathbf{X}$ is a block diagonal matrix, Eq. (11) can be solved by*

$$\mathbf{U} = \oplus_{k=1}^{K} \mathbf{U}^k = \begin{bmatrix} \mathbf{U}^1 \\ \vdots \\ \mathbf{U}^K \end{bmatrix},$$

$$\mathbf{U}^k = (\beta \underline{\tilde{\mathbf{D}}}_{g_k} + \lambda_1 \underline{\mathbf{D}_1}_{g_k} + \gamma'_1 (\mathbf{X}^\top \mathbf{X})_{g_k})^{-1} \mathbf{X}_{g_k}^\top [\mathbf{Y}_1 \mathbf{v}_1, \ldots, \mathbf{Y}_c \mathbf{v}_c],$$

$$(13)$$

where $\underline{\tilde{\mathbf{D}}}_{g_k}$ is the kth block of the diagonal matrix $\tilde{\mathbf{D}}$; $\underline{\mathbf{D}_1}_{g_k}$ is the kth block of $\mathbf{D}_1$; and $\oplus$ denotes the operation that concatenates matrices vertically.

Proof. Since SNPs exhibit group structures, we denote $\mathbf{X} = (\ldots, \mathbf{X}_{g_k}, \ldots)$ with k being the index of the kth group. Then the covariance matrix $\mathbf{X}^\top \mathbf{X}$ can be represented as

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} \ddots & & \\ & (\mathbf{x}^\top \mathbf{x})_{g_k} & \\ & & \ddots \end{bmatrix}.$$

We have known that $\tilde{\mathbf{D}}$ and $\mathbf{D}_1$ are diagonal matrices, indicating they both are diagonally separable. Then according to Eq. (12), we have

$$
\mathbf{U} = \left(\beta \begin{bmatrix} \ddots & & \\ & \tilde{\mathbf{D}}_{g_k} & \\ & & \ddots \end{bmatrix} + \lambda_1 \begin{bmatrix} \ddots & & \\ & \underline{\mathbf{D}_1}_{g_k} & \\ & & \ddots \end{bmatrix} \right.
$$
$$
\left. + \gamma_1' \begin{bmatrix} \ddots & & \\ & \mathbf{X}_{g_k}^\top \mathbf{X}_{g_k} & \\ & & \ddots \end{bmatrix} \right)^{-1} \begin{bmatrix} \vdots \\ \mathbf{X}_{g_k}^\top \\ \vdots \end{bmatrix} \mathbb{Y}
$$
$$
= \begin{bmatrix} \vdots \\ (\beta \tilde{\mathbf{D}}_{g_k} + \lambda_1 \underline{\mathbf{D}_1}_{g_k} + \gamma_1' (\mathbf{X}^\top \mathbf{X})_{g_k})^{-1} \mathbf{X}_{g_k}^\top \mathbb{Y} \\ \vdots \end{bmatrix}
$$
$$
= \oplus_{k=1}^K \mathbf{U}^k. \qquad \square
$$

The advantages of this theorem are threefold. (1) The time complexity of Eq. (13) is $O(np_k^2 K)$ compared with that of Eq. (12) being $O(np^2)$, where p_k is the size of the kth group, and $p = \sum_{k=1}^K p_k$. This is a significant improvement because that the LD block size is usually quite small, i.e., $p_k \ll p$. (2) Benefiting from the computation effort reduction, the memory requirement is also saved a lot because storing $\mathbf{X}^\top \mathbf{X}$ is very memory expensive than storing several $(\mathbf{x}^\top \mathbf{x})_{g_k}$. (3) According to the proof, Eq. (13) is quite easy to implement, demonstrating it is very promising in big imaging genetic analysis. This is one of the contributions of this study and might provide a powerful tool for genome-wide and brain-wide bi-multivariate analysis.

2.3.2 Updating $\mathbf{v}_j$

Note that each $\mathbf{v}_j$ is associated with each $\mathbf{Y}_j$ separately. This means that these $\mathbf{v}_j$'s are not closely coupled such as $\mathbf{u}_j$'s and should be tackled with separately. Next we will show how to solve $\mathbf{v}_j$ with $\mathbf{v}_k (k \neq j)$ and $\mathbf{U}$ being fixed. Based on Eq. (9), we take the derivative with respect to $\mathbf{v}_j$ and set it to zero

$$-\mathbf{Y}_j^\top \mathbf{X} \mathbf{u}_j + \lambda_2 \mathbf{D}_2 \mathbf{v}_j + \gamma_2' \mathbf{Y}_j^\top \mathbf{Y}_j \mathbf{v}_j = \mathbf{0}, \qquad (14)$$

which can be rewritten as

$$(\lambda_2 \mathbf{D}_2 + \gamma_2' \mathbf{Y}_j^\top \mathbf{Y}_j) \mathbf{v}_j = \mathbf{Y}_j^\top \mathbf{X} \mathbf{u}_j, \qquad (15)$$

i.e.,

$$\mathbf{v}_j = (\lambda_2 \mathbf{D}_2 + \gamma_2' \mathbf{Y}_j^\top \mathbf{Y}_j)^{-1} \mathbf{Y}_j^\top \mathbf{X} \mathbf{u}_j, \qquad (16)$$

where $\mathbf{D}_2$ is a diagonal matrix with its ith entry being $\frac{1}{2\|\mathbf{v}^i\|_2} (i \in [1, q])$ on the diagonal; and $\gamma_2' = \gamma_2 + 1$. Therefore, each $\mathbf{v}_j$ can also be solved alternatively through an iterative algorithm.

Now that the building blocks regarding updating $\mathbf{U}$ and each individual $\mathbf{v}_j$ are created, we present the pseudocode in Algorithm 1.

Algorithm 1. Algorithm to Solve Eq. (9)

Require:
   $\mathbf{X} \in \mathcal{R}^{n \times p}$, $\mathbf{Y}_j \in \mathcal{R}^{n \times q}$, $j = 1, \ldots, c$, β, λ_1, λ_2, γ_1, γ_2
Ensure:
   Canonical weights $\mathbf{U}$ and $\mathbf{V}$.
1: Initialize $\mathbf{U} \in \mathcal{R}^{p \times c}$, $\mathbf{V} \in \mathcal{R}^{q \times c}$;
2: **while** not convergence **do**
3:    Update $\tilde{\mathbf{D}}_{g_k}$ and $\underline{\mathbf{D}_1}_{g_k}$;
4:    Solve $\mathbf{U}$ according to Eq. (13), and normalize $\mathbf{u}_j$ to $\|\mathbf{X} \mathbf{u}_j\|_2^2 = 1$;
5:    Update $\mathbf{D}_2$;
6:    Solve $\mathbf{v}_j (j = 1, \ldots, c)$ according to Eq. (16), and normalize $\mathbf{v}_j$ to $\|\mathbf{Y}_j \mathbf{v}_j\|_2^2 = 1$;
7: **end while**

2.4 Convergence Analysis

We have the following theorem for Algorithm 1.

Theorem 2. *Algorithm 1 decreases the objective value of Eq. (9) in each iteration.*

Proof. In order to prove this theorem, we need two essential conclusions: (1) Eq. (12) decreases the objective Eq. (9) in each iteration; and (2) Eq. (16) decreases the objective Eq. (9) in each iteration.

We first prove the conclusion (1). According to Eq. (12), we have

$$
\sum_{j=1}^c \left\| \mathbf{X} \mathbf{u}_j^{(t+1)} - \mathbf{Y}_j \mathbf{v}_j^{(t)} \right\|_2^2 + \gamma_1' Tr(\mathbf{U}^{(t+1)\top} \mathbf{X}^\top \mathbf{X} \mathbf{U}^{(t+1)})
$$
$$
+ \beta Tr(\mathbf{U}^{(t+1)\top} \tilde{\mathbf{D}} \mathbf{U}^{(t+1)}) + \lambda_1 Tr(\mathbf{U}^{(t+1)\top} \mathbf{D}_1 \mathbf{U}^{(t+1)})
$$
$$
\leq \sum_{j=1}^c \left\| \mathbf{X} \mathbf{u}_j^{(t)} - \mathbf{Y}_j \mathbf{v}_j^{(t)} \right\|_2^2 + \gamma_1' Tr(\mathbf{U}^{(t)\top} \mathbf{X}^\top \mathbf{X} \mathbf{U}^{(t)})
$$
$$
+ \beta Tr(\mathbf{U}^{(t)\top} \tilde{\mathbf{D}} \mathbf{U}^{(t)}) + \lambda_1 Tr(\mathbf{U}^{(t)\top} \mathbf{D}_1 \mathbf{U}^{(t)})
$$
$$
\Rightarrow \sum_{j=1}^c \left[\left\| \mathbf{X} \mathbf{u}_j^{(t+1)} - \mathbf{Y}_j \mathbf{v}_j^{(t)} \right\|_2^2 + \gamma_1' \left\| \mathbf{X} \mathbf{u}_j^{(t+1)} \right\|_2^2 \right]
$$
$$
+ \beta \sum_{k=1}^K \frac{\left\| \mathbf{U}^{k(t+1)} \right\|_F^2}{2 \left\| \mathbf{U}^{k(t)} \right\|_F} + \lambda_1 \sum_{i=1}^p \frac{\left\| \mathbf{u}^{i(t+1)} \right\|_2^2}{2 \left\| \mathbf{u}^{i(t)} \right\|_2}
$$
$$
\leq \sum_{j=1}^c \left[\left\| \mathbf{X} \mathbf{u}_j^{(t)} - \mathbf{Y}_j \mathbf{v}_j^{(t)} \right\|_2^2 + \gamma_1' \left\| \mathbf{X} \mathbf{u}_j^{(t)} \right\|_2^2 \right]
$$
$$
+ \beta \sum_{k=1}^K \frac{\left\| \mathbf{U}^{k(t)} \right\|_F^2}{2 \left\| \mathbf{U}^{k(t)} \right\|_F} + \lambda_1 \sum_{i=1}^p \frac{\left\| \mathbf{u}^{i(t)} \right\|_2^2}{2 \left\| \mathbf{u}^{i(t)} \right\|_2}.
$$

(17)

According to Lemma 1 in [5], $\left\| \mathbf{U}^{k(t+1)} \right\|_F - \frac{\left\| \mathbf{U}^{k(t+1)} \right\|_F^2}{2 \left\| \mathbf{U}^{k(t)} \right\|_F} \leq \left\| \mathbf{U}^{k(t)} \right\|_F - \frac{\left\| \mathbf{U}^{k(t)} \right\|_F^2}{2 \left\| \mathbf{U}^{k(t)} \right\|_F}$, and $\left\| \mathbf{u}^{i(t+1)} \right\|_2 - \frac{\left\| \mathbf{u}^{i(t+1)} \right\|_2^2}{2 \left\| \mathbf{u}^{i(t)} \right\|_2} \leq \left\| \mathbf{u}^{i(t)} \right\|_2 - \frac{\left\| \mathbf{u}^{i(t)} \right\|_2^2}{2 \left\| \mathbf{u}^{i(t)} \right\|_2}$. Then applying two inequations to Eq. (17) with respect to each group and individual feature, we have

$$\sum_{j=1}^{c}\left[\left\|\mathbf{X}\mathbf{u}_j^{(t+1)} - \mathbf{Y}_j\mathbf{v}_j^{(t)}\right\|_2^2 + \gamma_1'\left\|\mathbf{X}\mathbf{u}_j^{(t+1)}\right\|_2^2\right]$$
$$+ \beta\sum_{k=1}^{K}\left\|\mathbf{U}^{k(t+1)}\right\|_{\mathrm{F}} + \lambda_1\sum_{i=1}^{p}\left\|\mathbf{u}^{i(t+1)}\right\|_2$$
$$\leq \sum_{j=1}^{c}\left[\left\|\mathbf{X}\mathbf{u}_j^{(t)} - \mathbf{Y}_j\mathbf{v}_j^{(t)}\right\|_2^2 + \gamma_1'\left\|\mathbf{X}\mathbf{u}_j^{(t)}\right\|_2^2\right]$$
$$+ \beta\sum_{k=1}^{K}\left\|\mathbf{U}^{k(t)}\right\|_{\mathrm{F}} + \lambda_1\sum_{i=1}^{p}\left\|\mathbf{u}^{i(t)}\right\|_2, \tag{18}$$

which can be rewritten in matrix form as

$$\sum_{j=1}^{c}\left[\left\|\mathbf{X}\mathbf{u}_j^{(t+1)} - \mathbf{Y}_j\mathbf{v}_j^{(t)}\right\|_2^2 + \gamma_1\left\|\mathbf{X}\mathbf{u}_j^{(t+1)}\right\|_2^2\right]$$
$$+ \beta\left\|\mathbf{U}^{(t+1)}\right\|_{\mathrm{G}_{2,1}} + \lambda_1\left\|\mathbf{U}^{(t+1)}\right\|_{2,1}$$
$$\leq \sum_{j=1}^{c}\left[\left\|\mathbf{X}\mathbf{u}_j^{(t)} - \mathbf{Y}_j\mathbf{v}_j^{(t)}\right\|_2^2 + \gamma_1\left\|\mathbf{X}\mathbf{u}_j^{(t)}\right\|_2^2\right]$$
$$+ \beta\left\|\mathbf{U}^{(t)}\right\|_{\mathrm{G}_{2,1}} + \lambda_1\left\|\mathbf{U}^{(t)}\right\|_{2,1}. \tag{19}$$

γ_1' is replaced by γ_1 since $\left\|\mathbf{X}\mathbf{u}_j^{(t+1)}\right\|_2^2$ and $\left\|\mathbf{X}\mathbf{u}_j^{(t)}\right\|_2^2$ have been normalized to 1. Thus the objective value is decreased in each iteration regarding updating $\mathbf{U}$.

Similarly, we have the following inequality.

$$\sum_{j=1}^{c}\left[\left\|\mathbf{X}\mathbf{u}_j^{(t+1)} - \mathbf{Y}_j\mathbf{v}_j^{(t+1)}\right\|_2^2 + \gamma_2\left\|\mathbf{Y}_j\mathbf{v}_j^{(t+1)}\right\|_2^2\right] + \lambda_2\left\|\mathbf{V}^{(t+1)}\right\|_{2,1}$$
$$\leq \sum_{j=1}^{c}\left[\left\|\mathbf{X}\mathbf{u}_j^{(t+1)} - \mathbf{Y}_j\mathbf{v}_j^{(t)}\right\|_2^2 + \gamma_2\left\|\mathbf{Y}_j\mathbf{v}_j^{(t)}\right\|_2^2\right] + \lambda_2\left\|\mathbf{V}^{(t)}\right\|_{2,1}. \tag{20}$$

Now based on Eqs. (19) and (20), we have $\mathcal{L}(\mathbf{U}^{(t+1)}, \mathbf{V}^{(t+1)}) \leq \mathcal{L}(\mathbf{U}^{(t+1)}, \mathbf{V}^{(t)}) \leq \mathcal{L}(\mathbf{U}^{(t)}, \mathbf{V}^{(t)})$, which completes the proof. □

According to Eq. (9), the objective is lower bounded by 0, and thus iteratively decreasing the objective value will converge to a local optimum. The proposed algorithm runs very fast owning to (1) its closed-form solution for each update; and (2) the divide-and-conquer strategy which is supported by Theorem 1.

3 RESULTS

3.1 Benchmarks and Experimental Setup

In order to evaluate the performance of the proposed multi-task SCCA method, we choose the closely related mSCCA [13] and the conventional two-view SCCA as benchmark methods. A common problem of the two-view SCCA and mSCCA is that they suffer from heavy computational and memory requirement issues because they cannot handle the large covariance matrix calculation. To make the comparison available, based on Theorem 1, we implement the fast two-view SCCA and the fast mSCCA. This yields the three benchmark methods in this study, confirming another contribution of this study.

All the methods contain parameters that should be fine tuned before running experiments. We apply the nested

5-fold cross-validation in this work. Specifically, those tuning parameters were determined in the inner loop where a group of them generating the highest mean correlation coefficients, i.e., $\mathrm{CV}(\lambda, \beta, \gamma) = \frac{1}{5}\sum_{j=1}^{5}Corr(\mathbf{X}_{-j}\mathbf{u}_j, \mathbf{Y}_{-j}\mathbf{v}_j)$, will be chosen as the optimal parameters, where $\mathbf{X}_{-j}$ and $\mathbf{Y}_{-j}$ are the jth subset of the inner testing set, and $\mathbf{u}_j$ and $\mathbf{v}_j$ are the canonical weights estimated from the inner training set. Once determined, these parameters are used in the external loop to generate the final results. Before tuning parameters, we use some heuristic strategy to reduce the computation burden since blindly tuning them by grid search is computational intensive. For all methods, γ_1 and γ_2 are used to address the scaling issue when calculating the correlation coefficient. On this account, fixing the denominator to be 1 or other integers will just affect the magnitude of $\mathbf{U}$ and $\mathbf{V}$, and the relative relationship among each element remains the same. For example, suppose $u_{1,1} = 5$, $u_{1,2} = 1$ and $\|\mathbf{X}\mathbf{u}_1\|_2^2 = 20$ ($\|\mathbf{u}_1\|_2^2 = 20$ for two-view SCCA and mSCCA), tuning $\|\mathbf{X}\mathbf{u}_1\|_2^2 = 1$ will lead to $u_{1,2} = 0.25$ and $u_{1,2} = 0.05$; while tuning $\|\mathbf{X}\mathbf{u}_1\|_2^2 = 10$ will lead to $u_{1,1} = 2.5$ and $u_{1,2} = 0.5$. This will not affect the feature selection as $u_{1,1}$ will always be selected with higher priority than $u_{1,2}$. Therefore, we set $\gamma_1 = \gamma_2 = 1$ in this paper. Generally, too large parameters yield over-penalized results while too small ones yield less-penalized results. To avoid this issue, we tune the remaining parameters λ_1, λ_2, β from a moderate range 10^i ($i = -5, -4, \ldots, 0, \ldots, 4, 5$) via the grid search strategy. Finally, in order to make the results stable, we repeat each experiment 100 times and show the average results. In the experiments, all methods are stopped when both $\max_i\left|u_i^{(t+1)} - u_i^{(t)}\right| \leq \epsilon$ ($\forall u_i$) and $\max_j\left|v_j^{(t+1)} - v_j^{(t)}\right| \leq \epsilon$ ($\forall v_j$) are satisfied, where ϵ is the tolerable error. We empirically set $\epsilon = 10^{-5}$ from experiments in this paper.

3.2 Simulation Study

This section present the comparison results on the synthetic data. We generate four data sets with different number of samples and features, sparsity levels and noise levels to assure a thorough comparison. The first three data sets are generated using the same ground truth but with different noise strengthes. The $\mathbf{X}$, $\mathbf{Y}_j$ ($j = 2$) and $\mathbf{z}$ of them are all with $n = 80$, $p = 120$, $q_1 = 100$ and $q_2 = 100$. This could help show the performance when treating with different noises. The fourth data set is created to access the performance under high-dimensional situation, and $n = 500$, $p = 2,000$, $q_1 = 1,000$, $q_2 = 1,000$ respectively. The details of each data set are described as follows.

- Data 1: We first set $\mathbf{u} = (\underbrace{0, \ldots, 0}_{60}, \underbrace{1, \ldots, 1}_{20}, \underbrace{0, \ldots, 0}_{40})^{\top}$, $\mathbf{v}_1 = (\underbrace{0, \ldots, 0}_{25}, \underbrace{2, \ldots, 2}_{25}, \underbrace{0, \ldots, 0}_{50})^{\top}$ and $\mathbf{v}_2 = (\underbrace{0, \ldots, 0}_{25}, \underbrace{3, \ldots, 3}_{25}, \underbrace{0, \ldots, 0}_{50})^{\top}$. Then we generate a random latent vector μ of length n and normalize it to unit length. The data matrix $\mathbf{X}$ is created by $x_{\ell,i} \sim N(\mu_\ell u_i, \sigma_x)$, where $\sigma_x = 5$ denotes the noise strength. Similarly, $\mathbf{Y}_j$ is created by $(y_{\ell,i})_j \sim N(\mu_\ell v_{i,j}, \sigma_{y_j})$ with $\sigma_{y_1} = 5$ and $\sigma_{y_2} = 5$.

TABLE 1
Performance Comparison on Synthetic Data

	Training Results			Testing Results		
	SCCA	mSCCA	MTSCCA	SCCA	mSCCA	MTSCCA
Data 1	0.28 ± 0.06	0.40 ± 0.10	$\mathbf{0.99 \pm 0.00}$	0.25 ± 0.14	0.16 ± 0.10	0.23 ± 0.16
Data 2	0.59 ± 0.06	0.49 ± 0.06	$\mathbf{0.63 \pm 0.06}$	0.31 ± 0.15	0.25 ± 0.19	0.41 ± 0.18
Data 3	0.95 ± 0.01	0.95 ± 0.01	$\mathbf{0.96 \pm 0.01}$	0.91 ± 0.04	0.91 ± 0.05	$\mathbf{0.95 \pm 0.03}$
Data 4	0.89 ± 0.02	0.89 ± 0.02	$\mathbf{0.99 \pm 0.00}$	0.85 ± 0.05	0.85 ± 0.06	$\mathbf{0.97 \pm 0.01}$

Training and testing canonical correlation coefficients (mean ± std) of 5-fold cross-validation are shown for SCCA, mSCCA, and MTSCCA. The best values are shown in boldface.

- Data 2 - Data 3: These two data sets are created with the same ground truth as the first one with different noises, i.e., $\sigma_x = \sigma_{y_1} = \sigma_{y_2} = 1$ for Data 2 and $\sigma_x = \sigma_{y_1} = \sigma_{y_2} = 0.1$ for Data 3. Therefore, the correlation coefficients of these three data sets are different, and that of the first data set is the smallest and that of the third one is the highest.

- Data 4: In this data set, $\mathbf{u} = (\underbrace{0,\ldots,0}_{500}, \underbrace{1,\ldots,1}_{100}, \underbrace{0,\ldots,0}_{400},$ $\underbrace{2,\ldots,2}_{100}, \underbrace{0,\ldots,0}_{900})^\top$, $\mathbf{v}_1 = (\underbrace{0,\ldots,0}_{200}, \underbrace{1.5,\ldots,1.5}_{300}, \underbrace{0,\ldots,0}_{500})^\top$, $\mathbf{v}_2 = (\underbrace{0,\ldots,0}_{200}, \underbrace{2.5,\ldots,2.5}_{300}, \underbrace{0,\ldots,0}_{500})^\top$, and $\sigma_x = \sigma_{y_1} = \sigma_{y_2} = 0.1$. The data matrix $\mathbf{X}$ is created by $x_{\ell,i} \sim N(\mu_\ell u_i, \sigma_x)$, and $\mathbf{Y}_j$ is generated by $(y_{\ell,i})_j \sim N(\mu_\ell v_{i,j}, \sigma_{y_j})$, with the random latent vector μ of length n.

We first show the training and testing canonical correlation coefficients (CCCs) in Table 1. On the first three data sets, all methods obtain a good score when the true CC is high, while perform poorly (overfitting) when the true CC is excessively low due to the high percentage of noise. MTSCCA identifies the highest training CCCs among all three unsupervised methods, including both two-view SCCA and mSCCA. This demonstrates that MTSCCA performs better than the two single-task based SCCA methods. In Data 4, we observe that MTSCCA obtains higher training and testing CCCs than two-view SCCA and mSCCA in this high-dimensional data set. This indicates that, owing to the multi-task modeling strategy, the ability of identifying bi-multivariate association can be improved.

In addition, the feature selection ability is also of great interest and importance. In Fig. 3, we show the scatter of the estimated $\mathbf{u}$ and $\mathbf{v}_j$'s. For the two-view SCCA, each $\mathbf{u}_j$ is calculated independently from each single-task SCCA, and $\mathbf{u}$ is obtained by averaging $\mathbf{u}_j$'s. The $\mathbf{u}$ of MTSCCA is also obtained by averaging $\mathbf{u}_j$'s associating with three SCCA tasks. There are two estimated $\mathbf{v}_j$'s for all methods and we show them separately in Fig. 4. In order to show the performance clearly, the ground truthes are also presented in the figure (first row). Within each subfigure, the horizontal axis represents the indices and the vertical axis represents the weight values. A feature with a larger canonical weight (in absolute value) contributes more to the bi-multivariate correlation. We observe that all methods cannot find out the correct locations of signals in the first data owing to the low signal-to-noise ratio. Combining the results of the first three data sets together, all their performances improve from the

first data set to the third one. MTSCCA holds the best canonical profiles being consistent with the ground truthes, showing its better performance in feature selection than the two-view and multiple-view SCCA. In Data 4 where the feature dimensionality is high, MTSCCA always identifies correct signal locations. To make the comparison more formal, Tables 2 and 3 show the sensitivity and specificity in terms of canonical weights $\mathbf{u}$ and $\mathbf{v}_j$'s. Both metrics are calculated as follows. Features are selected based on their absolute weight values, and the larger the $|u_i|$ (or $|v_i|$) is, the more relevant to the canonical correlation. Generally, given a predefined threshold, those features with larger-than-threshold values are selected. However, it is hard to predefine an appropriate threshold. To overcome this issue, in this paper, the sensitivity is calculated via $\frac{\text{\#true positive in the top } K \text{ selected features}}{K}$ where K is the number of non-zero features of the ground truth. Similarly, the specificity is calculated by $\frac{K}{\text{\#selected features required to cover the ground truth}}$. The results show that all methods obtain good sensitivity and specificity across these simulated data sets. MTSCCA performs slightly better than those single-task based SCCA methods owing to the multi-task modeling strategy. It is worth noting that in the original implementations, both two-view SCCA and multi-view SCCA fail since they cannot treat the large matrix calculation on the same platform as MTSCCA does. By incorporating Theorem 1, the two methods become feasible to high-dimensional data sets. The runtime of each method is shown in Table 4, and there is no significant difference between these methods based on Theorem 1. This again demonstrates the effectiveness and practice of our fast implementation strategy.
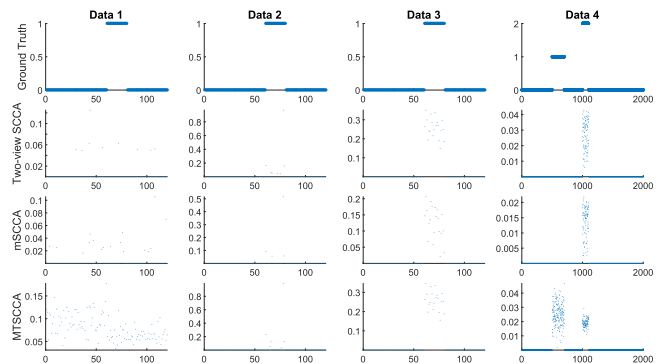

Fig. 3. Canonical weights $\mathbf{u}$ (mean value) estimated on synthetic data. The first row is the ground truth, and each remaining row corresponds to an SCCA method: (1) Two-view SCCA, (2) mSCCA (Multi-view SCCA), and (3) MTSCCA (Multi-task SCCA). In each subfigure, the horizontal axis represents the indices of each $\mathbf{u}$, and the vertical axis represents the estimated weight value.
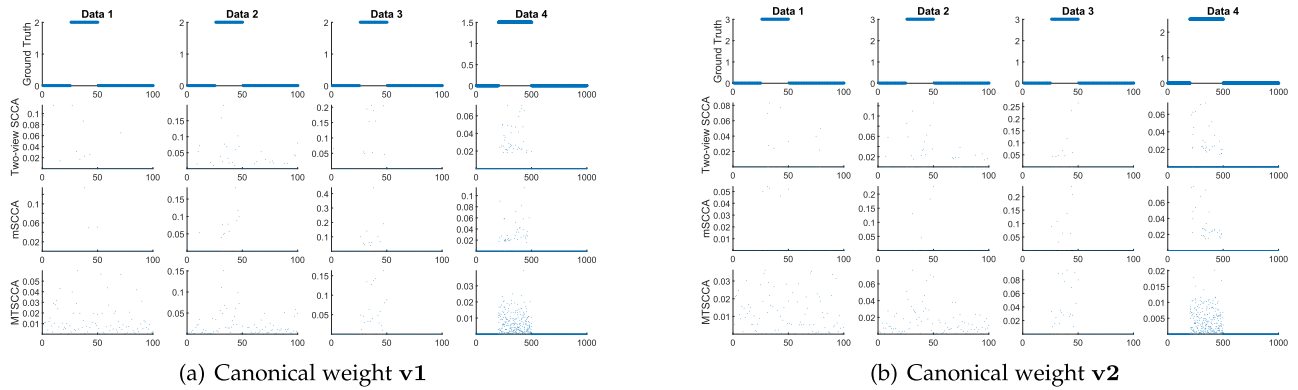
Fig. 4. Canonical weights **V** (mean value) estimated on synthetic data. The first row is the ground truth, and each remaining row corresponds to an SCCA method. (1) Two-view SCCA. (2) mSCCA (Multi-view SCCA). (3) MTSCCA (Multi-task SCCA). In each subfigure, the horizontal axis represents the indices of $\mathbf{v}_j$ $(j = 1, 2)$, and the vertical axis represents the estimated weight value.

TABLE 2
Comparison of the Sensitivity of Canonical Weights on Synthetic Data

	u				v1				v2			
	Data 1	Data 2	Data 3	Data 4	Data 1	Data 2	Data 3	Data 4	Data 1	Data 2	Data 3	Data 4
SCCA	0.25	0.45	1.00	1.00	0.88	1.00	0.85	0.99	0.44	0.56	0.84	1.00
mSCCA	0.20	0.45	1.00	1.00	0.60	0.84	0.96	1.00	0.76	0.92	0.96	1.00
MTSCCA	0.05	0.55	1.00	1.00	0.32	0.56	1.00	1.00	0.32	0.64	1.00	1.00

TABLE 3
Comparison of the Specificity of Canonical Weights on Synthetic Data

	u				v1				v2			
	Data 1	Data 2	Data 3	Data 4	Data 1	Data 2	Data 3	Data 4	Data 1	Data 2	Data 3	Data 4
SCCA	0.85	0.89	1.00	1.00	0.77	0.83	0.96	1.00	0.81	0.85	0.95	1.00
mSCCA	0.84	0.89	1.00	1.00	0.87	0.95	0.99	1.00	0.92	0.97	0.99	1.00
MTSCCA	0.81	0.91	1.00	1.00	0.77	0.85	1.00	1.00	0.77	0.88	1.00	1.00

In summary, this simulation study using data sets with diverse characteristics demonstrates that MTSCCA is effective in bi-multivariate association identification with multiple data modalities. Moreover, MTSCCA identifies the best canonical loading profiles which is consistent with the ground truth compared to the single-task SCCA methods. In addition, it also reveals that the group structure can not only help prompt the identification performance, but also help reduce the time effort in high-dimensional scenario in multi-modal bi-multivariate association analysis.

3.3 Real Neuroimaging Genetics Study

The genotying and brain imaging data used in this article were obtained from the Alzheimer's Disease Neuroimaging Initiative database (adni.loni.usc.edu). One primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease. For up-to-date information, see www.adni-info.org.

The neuroimaging data were from 755 non-Hispanic Caucasian participants, including 281 AD, 292 MCI and 182

healthy control (HC). They were 18-F florbetapir PET (AV-45) scans, fluorodeoxyglucose positron emission tomography (FDG) scans, and structural MRI scans which were downloaded from the ADNI database (adni.loni.usc.edu). Details of this data set are exhibited in Table 5. The multimodality imaging data were aligned to each participant's same visit. The structural MRI scans were processed with voxel-based morphometry (VBM) via SPM [27]. Generally, all scans had been aligned to a T1-weighted template image, segmented into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) maps, normalized to the standard Montreal Neurological Institute (MNI) space

TABLE 4
Runtime Comparison of Synthetic Data

	Runtime		
	SCCA	mSCCA	MTSCCA
Data 1	0.19 ± 0.24	0.19 ± 0.24	0.19 ± 0.23
Data 2	0.15 ± 0.16	0.16 ± 0.18	0.18 ± 0.22
Data 3	0.11 ± 0.18	0.17 ± 0.18	0.13 ± 0.15
Data 4	1.49 ± 5.58	2.59 ± 5.52	2.59 ± 5.86

TABLE 5
Participant Characteristics

	HC	MCI	AD
Num	182	292	281
Gender(M/F, %)	47.16/52.84	54.52/45.48	47.37/52.63
Handedness(R/L, %)	90.91/9.09	87.35/12.65	91.50/8.50
Age (mean±std)	72.97 ± 6.00	71.81 ± 7.62	72.38 ± 7.31
Education (mean±std)	16.52 ± 2.58	15.97 ± 2.78	16.14 ± 2.78

TABLE 6
The p-Values of t-Tests for CCCs Comparison between MTSCCA and Two-View SCCA and mCCA

	SNPs-AV45	SNPs-FDG	SNPs-VBM
		Training	
Two-view SCCA	5.46E-24	3.39E-25	6.00E-15 (-)
mSCCA	7.98E-27	1.51E-27	4.77E-18
		Testing	
Two-view SCCA	1.46E-23	8.60E-43	4.99E-24
mSCCA	3.71E-27	3.91E-31	4.80E-22

The '-' in parenthesis means that MTSCCA loses on this trial.

as $2 \times 2 \times 2$ mm^3 voxels, and had been smoothed with an 8 mm FWHM kernel. The FDG-PET and AV45-PET scans were also registered into the same MNI space by SPM. We then subsampled the whole brain and generated 116 regions of interest (ROI) level measurements based on the MarsBaR automated anatomical labeling (AAL) atlas. They were the mean gray matter densities for structural MRI, amyloid values for AV45 scans and glucose utilization for FDG scans. Using the regression weights derived from the healthy control participants, these imaging measures were pre-adjusted to remove the effects of the baseline age, gender, education, and handedness.

The genotyping data of the same population were downloaded from the LONI website. They were genotyped using the Human 610-Quad or OmniExpress Array (Illumina, Inc., San Diego, CA, USA), and preprocessed using the standard quality control (QC) and imputation steps. The QC criteria for the SNP data include (1) call rate check per subject and per SNP marker, (2) gender check, (3) sibling pair identification, (4) the Hardy-Weinberg equilibrium test, (5) marker removal by the minor allele frequency and (6) population stratification. In second pre-processing step, following the quality-controlled SNPs, those missing genotypes were imputed using the MaCH software [28]. Among all human chromosomes, the chromosome 19 sequence contains the most number of genes, in which the gene density is more than double the genome-wide average [29], [30]. In addition, this chromosome also includes the well-known AD risk genes such as *APOE*, *TOMM40* and *ABCA7*. Therefore, a bi-multivariate association study between this chromosome and whole brain imaging markers could be of great interest, and has potential to yield interesting AD risk factors. As a result, all the SNPs from chromosome 19 were included, i.e., 152,787 SNPs were involved in this study.
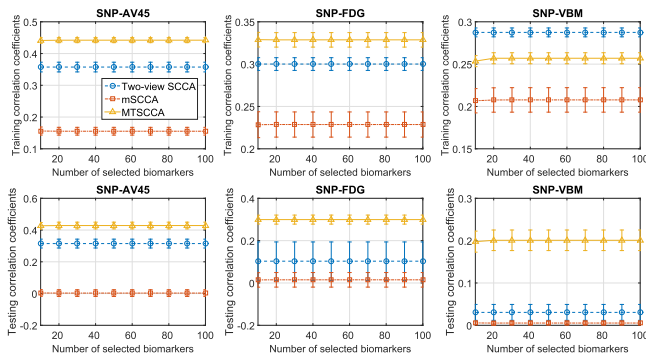

Fig. 5. Performance comparison: The mean and standard deviation (SD) of the canonical correlation coefficients (CCCs) obtained from 5-fold cross-validation trials are plotted, where each error bar indicates ±0.5SD. The subtitle SNPs-AV45 means the CCCs are calculated between the SNPs data and the AV45-PET data.

Among these enormous SNPs, most of them might be irrelevant to AD, while only a few of them could be relevant via influencing the intermediate brain imaging measurements. The aim is to identify this small subset of SNPs in chromosome 19 correlating to imaging markers and AD.

3.4 Improved Bi-Multivariate Association

In this subsection we evaluate the proposed method in identifying the bi-multivariate associations between one genetic data and three sets of imaging phenotypes. Thus there will be three pairs of associations, and we denote them as SNPs-AV45, SNPs-FDG and SNPs-VBM for the sake of description. For the three SCCA tasks, the proposed MTSCCA learns them together and generate a canonical weights matrix $\mathbf{U}$ for SNPs and one canonical weight vector $\mathbf{v}_j$ for AV45, FDG and VBM. We then calculate three canonical correlation coefficients in terms of SNPs-AV45, SNPs-FDG and SNPs-VBM separately. The two-view SCCA naturally yields three CCCs for these three tasks. Though the mSCCA only learns one canonical weight vector for SNPs, we use it three times to generate three CCCs with respect to the three tasks.

Fig. 5 shows the CCCs of the SNPs data with each imaging QTs data, where CCCs estimated from SNPs-AV45, SNPs-FDG and SNPs-VBM are separately shown. In this figure, both the training CCCs and testing CCCs, as well as their standard deviations (SD) are presented. By changing the number of selected features $(10, 20, \ldots, 100$ in this work$)$ for both SNPs and imaging QTs, the CCCs can be generated and then these curves are plotted. It is clear that the proposed MTSCCA obtains higher CCCs on both training and testing sets across all imaging modalities except for training results of SNPs-VBM. After investigation, this could be that the two-view SCCA runs into overfitting since it holds high training CCCs and quite low CCCs simultaneously. We also observe that mSCCA always obtains the lowest CCCs on both training and testing sets across three tasks in this data. This is very interesting as it seems to violate the truth because more data (three different imaging QTs here) ought to provide more information. The reason might attribute to its modelling strategy. Demanding one set of features (SNPs) being associated with three sets of features (imaging QTs) simultaneously could be overstrict and thus harm the performance. This is also the reason that two-SCCA generally holds better CCCs than mSCCA.

In addition, we calculate the p-values between our method and two competing methods and show them in Table 6, where the '-' in parenthesis indicates that MTSCCA fails. The p-values are all reach the significance level which means that our method is significantly better than both competing methods. These results in terms of CCCs indicate

TABLE 7
Runtime Comparison with the mean±SD
being Presented

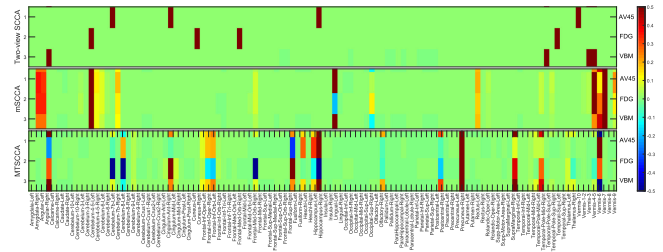| Runtime (seconds)		
Two-view SCCA	mSCCA	MTSCCA
342 ± 0.37	114 ± 0.30	361 ± 0.93

Fig. 6. Comparison of canonical weights in terms of each imaging modality across five trials. Each row corresponds to a SCCA method. (1) Two-view SCCA. (2) mSCCA. (3) MTSCCA. Within each panel, there are three rows corresponding to three type of imaging QTs, i.e., AV45, FDG, and VBM.

that the proposed joint bi-multivariate learning method indeed has better association identification capability than those SCCA methods, including both two-view and multiple-view ones. Table 7 shows the runtime in seconds of each method, where that of the two-view SCCA are summation of three two-view tasks. The runtime results indicate that all three methods run fast on this large data set. This attributes to the grouping strategy used in the implementation according to Theorem 1. In contrast, both competing methods are incapacitated in their original implementations since they cannot manipulate a big matrix with hundreds of thousands of features included. This again assures our contribution to accelerate both our method and conventional methods via making use of the grouping structures.

3.5 Genetic Marker Selection

Apart from the CCCs, the selected features in terms of SNPs are a major concern. This can help reveal those SNPs being highly related to imaging QTs and AD status at the same time. We show the top ten selected SNPs according to the canonical weight values of each individual method. In order to make the selection results stable, we average the canonical weight matrix into a vector and then choose the top ten SNPs based on their absolute values for MTSCCA. The top ten markers of two-view SCCA method are calculated via averaging the three separate canonical weights. Those of mSCCA are obtained by its canonical weight vector. The results of selected SNPs are shown in Table 8. Owning to the jointly learning paradigm, the proposed MTSCCA yields a surprisingly meaningful result with respect to selected features (SNPs). As expected, the notable AD risk markers rs429358 gains the highest weight value, and all of the remaining nine SNPs of MTSCCA, i.e., rs56131196 (*APOC1*), rs12721051 (*APOC1*), rs4420638 (*APOC1*), rs111789331 (4.5 kb of *APOC1*), rs66626994 (5.6 kb of *APOC1*), rs146275714 (*PVRL2*), rs41289512, rs147711004 (71 kb of *APOE*) and rs10119 (*TOMM40*), have been reported to show increasing risk of

AD in previous studies [31], [32], [33]. This indicates the ability of MTSCCA in identifying meaningful SNPs from massive genetic markers. The two-view SCCA also identifies the rs429358 as its first important SNPs, and five other AD related SNPs (rs10414043, rs147711004, rs7256200, rs73052335 and rs66626994) have been reported previously. But it identifies four SNPs that are not reported by now and thus further investigation should be taken place. The mSCCA performs unacceptably in this comparison since it does not find out rs429358. Moreover, except for the marker rs623264, all identified SNPs of mSCCA have not been reported yet in the current stage. In summary, the results in terms of selected SNPs show that MTSCCA performs better than both competing methods. This reveals that MTSCCA could be a suitable tool and very helpful in discovering meaningful genetic markers in a very large scenario.

3.6 Brain Imaging Marker Selection

Besides the genetic markers, as a bi-multivariate method, MTSCCA also selects features from the multiple imaging QTs. Fig. 6 presents the canonical weights of every method on each imaging modality (AV45, FDG and VBM) across the five trials. We observe that all those imaging markers with nonzero coefficients have been shown to be associated with the progression of AD. To make it clear, we show the top ten selected QTs of each imaging modal data of MTSCCA in Table 9. There are five markers (the right angular gyrus, the left posterior cingulum cortex, the left hippocampus, the left olfactory cortex and the vermis 8) reported in all three modalities owning to the joint feature selection via the $\ell_{2,1}$-norm regularization. Most importantly, these

TABLE 8
Top Ten SNPs Selected by Integrated Canonical Weights

Two-view SCCA	mSCCA	MTSCCA
rs429358	rs138339429	rs429358
rs10414043	rs141300647	rs56131196
rs147711004	rs58501143	rs12721051
rs146291812	rs17363184	rs4420638
rs623264	rs623264	rs111789331
rs7256200	rs11881833	rs66626994
rs186235601	rs7253576	rs146275714
rs73052335	rs1749316	rs41289512
rs66626994	rs139402102	rs147711004
rs415966	rs4605289	rs10119

TABLE 9
Top Ten Imaging QTs Selected by Canonical Weights
of Each Imaging Modality of MTSCCA

AV45	FDG	VBM
Frontal_Med_Orb_Left	Cingulum_Post_Left	Postcentral_Left
Angular_Right	Angular_Right	Precentral_Left
Cingulum_Post_Left	Hippocampus_Left	Angular_Right
Hippocampus_Left	Vermis_8	Cingulum_Post_Left
Olfactory_Left	Angular_Left	Vermis_8
Frontal_Mid_Right	Amygdala_Left	Thalamus_Right
Cingulum_Ant_Left	Olfactory_Left	Rolandic_Oper_Right
Rolandic_Oper_Right	Temporal_Mid_Right	Frontal_Med_Orb_Left
Temporal_Mid_Right	Precentral_Left	Hippocampus_Left
Vermis_8	Temporal_Mid_Left	Olfactory_Left

markers are have all been documented to be related to AD in the literature independently. For example, the significant reduction of glucose metabolism in the right angular gyrus has been observed in aging-associated cognitive decline (AACD) patients [34]. The declined metabolism in the left posterior cingulum cortex is an early sign of Alzheimer's disease [35]. This brain tract is also connected to the hippocampus which is a notable sign of AD and MCI [36], [37]. The remaining left olfactory cortex [38] and vermis 8 [39], have been separately validated to be a reflection of AD or MCI. These results indicate that MTSCCA could find out meaningful imaging QTs markers that are associated with the status of dementia. The mSCCA also identifies a few of AD related markers such as the hippocampus. The results of the two-view method are rambling and thus lack of biological meanings. To summarize, the proposed MTSCCA can not only obtain higher CCCs than conventional SCCA methods, but also yield better canonical weights for both SNPs and imaging QTs. The top ten selected SNPs and imaging QTs are highly correlated with each other, as well as AD status, which demonstrates that MTSCCA could be very promising in brain imaging genetics.

4 CONCLUSION

High-throughput genotyping technique and neuroimaging techniques provide us a large amount of biomedical data, and finding out their bi-multivariate associations is important. In this paper, we have proposed a novel multi-task sparse canonical correlation analysis framework and apply it to imaging genetics with multi-modal brain imaging QTs. Different from existing SCCA, MTSCCA can incorporate multiple sets of imaging modalities data into a single integrative model. Furthermore, MTSCCA is a multiple bi-multivariate method and thus has better modeling capability than both SCCA and MTL regression. A fast optimization algorithm is proposed which avoids calculating the large covariance and its inverse. The algorithm is guaranteed to converge to a local optimum, and runs very fast with hundreds of thousands of features involved.

We compared MTSCCA with the conventional two-view and multi-view SCCA on an ADNI cohort. Our method obtained better performance than the benchmarks with higher correlation coefficients and clearer canonical weight patterns. MTSCCA succeeds in identifying a small set of SNPs from enormous genetic markers from the 19th chromosome. It is worth noting that all top ten selected SNPs of MTSCCA are AD risk factors. In addition, the canonical weight patterns of imaging QTs were also of great success. The identified imaging QTs were highly correlated to AD or MCI. These promising results demonstrated that the proposed multi-task SCCA framework could be a powerful tool in big brain imaging genetics. Since a GWAS based bi-multivariate analysis is of much concern, in the future work, we will keep looking into the merit of MTSCCA and use it to genemo-wide brain-wide imaging analysis.

ACKNOWLEDGMENTS

REFERENCES

[1] A. J. Saykin, L. Shen, X. Yao, S. Kim, K. Nho, et al., "Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans," *Alzheimer's Dementia*, vol. 11, no. 7, pp. 792–814, 2015.

[2] L. Shen, P. M. Thompson, S. G. Potkin, L. Bertram, L. A. Farrer, et al., "Genetic analysis of quantitative phenotypes in AD and MCI: Imaging, cognition and biomarkers," *Brain Imag. Behav.*, vol. 8, no. 2, pp. 183–207, 2014.

[3] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "The alzheimer's disease neuroimaging initiative," *Neuroimaging Clinics North Amer.*, vol. 15, no. 4, pp. 869–877, 2005.

[4] S. Lee, J. Zhu, and E. P. Xing, "Adaptive multi-task Lasso: With application to eQTL detection," in *Proc. 23rd Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 1306–1314.

[5] H. Wang, F. Nie, H. Huang, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, and L. Shen, "Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the ADNI cohort," *Bioinf.*, vol. 28, no. 2, pp. 229–237, 2012.

[6] J. Chen, F. D. Bushman, J. D. Lewis, G. D. Wu, and H. Li, "Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis," *Biostatistics*, vol. 14, no. 2, pp. 244–258, 2013.

[7] X. Chen and H. Liu, "An efficient optimization algorithm for structured sparse CCA, with applications to eQTL mapping," *Statist. Biosciences*, vol. 4, no. 1, pp. 3–26, 2012.

[8] D. Lin, V. D. Calhoun, and Y. P. Wang, "Correspondence between fMRI and SNP data by group sparse canonical correlation analysis," *Med. Image Anal.*, vol. 18, pp. 891–902, 2014.

[9] L. Du, H. Huang, J. Yan, S. Kim, S. L. Risacher, M. Inlow, J. H. Moore, A. J. Saykin, and L. Shen, "Structured sparse canonical correlation analysis for brain imaging genetics: An improved GraphNet method," *Bioinf.*, vol. 32, no. 10, pp. 1544–1551, 2016.

[10] L. Du, K. Liu, T. Zhang, X. Yao, J. Yan, S. L. Risacher, J. Han, L. Guo, A. J. Saykin, and L. Shen, "A novel SCCA approach via truncated ℓ_1-norm and truncated group Lasso for brain imaging genetics," *Bioinf.*, vol. 34, no. 2, pp. 278–285, 2018.

[11] L. Du, J. Yan, S. Kim, S. L. Risacher, H. Huang, M. Inlow, J. H. Moore, A. J. Saykin, and L. Shen, "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.

[12] L. Du, T. Zhang, K. Liu, J. Yan, X. Yao, S. L. Risacher, A. J. Saykin, J. Han, L. Guo, and L. Shen, "Identifying associations between brain imaging phenotypes and genetic factors via a novel structured SCCA approach," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2017, pp. 543–555.

[13] D. M. Witten and R. J. Tibshirani, "Extensions of sparse canonical correlation analysis with applications to genomic data," *Statistical Appl. Genetics Mol. Biol.*, vol. 8, no. 1, pp. 1–27, 2009.

[14] I. Wilms and C. Croux, "Sparse canonical correlation analysis from a predictive point of view," *Biometrical J.*, vol. 57, no. 5, pp. 834–851, 2015.

[15] Q. Mai and X. Zhang, "An iterative penalized least squares approach to sparse canonical correlation analysis," *Biometrical J.*, vol. 57, no. 5, pp. 834–851, 2019.

[16] L. Du, K. Liu, X. Yao, S. L. Risacher, J. Han, L. Guo, A. J. Saykin, and L. Shen, "Fast multi-task SCCA learning with feature selection for multi-modal brain imaging genetics," in *Proc. IEEE Int. Conf. Bioinf. Biomedicine*, 2018, pp. 356–361.

[17] L. Du, K. Liu, L. Zhu, X. Yao, S. L. Risacher, L. Guo, A. J. Saykin, and L. Shen, "Identifying progressive imaging genetic patterns via multi-task sparse canonical correlation analysis: A longitudinal study of the adni cohort," *Bioinf.*, vol. 35, no. 14, pp. i474–i483, 2019.

[18] J. K. Pritchard and M. Przeworski, "Linkage disequilibrium in humans: Models and data," *Amer. J. Human Genetics*, vol. 69, no. 1, pp. 1–14, 2001.

[19] M. W. Weiner, P. S. Aisen, C. R. Jack, W. J. Jagust, J. Q. Trojanowski, L. Shaw, A. J. Saykin, J. C. Morris, N. Cairns, L. A. Beckett, A. W. Toga, R. C. Green, S. Walter, H. Soares, P. J. Snyder, E. Siemers, W. Z. Potter, P. E. Cole, and M. E. Schmidt, "The alzheimer's disease neuroimaging initiative: Progress report and future plans," *Alzheimer's Dementia*, vol. 6, no. 3, pp. 202–211, 2010.

[20] E. Parkhomenko, D. Tritchler, and J. Beyene, "Sparse canonical correlation analysis with application to genomic data integration," *Statistical Appl. Genetics Mol. Biol.*, vol. 8, no. 1, pp. 1–34, 2009.

[21] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learn. Res.*, vol. 6, pp. 1817–1853, 2005.

[22] B. Bakker and T. Heskes, "Task clustering and gating for Bayesian multitask learning," *J. Mach. Learn. Res.*, vol. 4, pp. 83–99, 2003.

[23] S. Bendavid and R. Schuller, "Exploiting task relatedness for multiple task learning," in *Proc. Learn. Theory Kernel Machines*, 2003, pp. 567–580.

[24] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. 19th Int. Conf. Neural Inf. Process. Syst.*, 2006, vol. 73, pp. 41–48.

[25] J. A. Rosenfeld, C. E. Mason, and T. M. Smith, "Limitations of the human reference genome for personalized genomics," *PLoS One*, vol. 7, no. 7, 2012, Art. no. e40294.

[26] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statistical Soc.: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[27] J. Ashburner and K. J. Friston, "Voxel-based morphometry–the methods," *NeuroImage*, vol. 11, no. 6, pp. 805–21, 2000.

[28] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis, "MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes," *Genetic Epidemiology*, vol. 34, no. 8, pp. 816–834, 2010.

[29] J. Grimwood, L. A. Gordon, A. Olsen, A. Terry, J. Schmutz, J. Lamerdin, U. Hellsten, D. Goodstein, O. Couronne, M. Tran-Gyamfi, et al., "The dna sequence and biology of human chromosome 19," *Nature*, vol. 428, no. 6982, 2004, Art. no. 529.

[30] J. C. Venter, "The sequence of the human genome," *Sci.*, vol. 292, no. 5523, pp. 1838–1838, 2001.

[31] L. Gao, Z. Cui, L. Shen, and H.-F. Ji, "Shared genetic etiology between type 2 diabetes and alzheimer's disease identified by bioinformatics analysis," *J. Alzheimer's Disease*, vol. 50, no. 1, pp. 13–17, 2016.

[32] X. Zhou, Y. Chen, K. Y. Mok, T. C. Kwok, V. C. Mok, Q. Guo, F. C. Ip, Y. Chen, N. Mullapudi, P. Giusti-Rodríguez, et al., "Noncoding variability at the APOE locus contributes to the alzheimer's risk," *Nature Commun.*, vol. 10, no. 1, 2019, Art. no. 3310.

[33] G. Davies, N. Armstrong, J. C. Bis, J. Bressler, V. Chouraki, S. Giddaluru, E. Hofer, C. A. Ibrahim-Verbaas, M. Kirin, J. Lahti, et al., "Genetic contributions to variation in general cognitive function: A meta-analysis of genome-wide association studies in the charge consortium (n= 53 949)," *Mol. Psychiatry*, vol. 20, no. 2, 2015, Art. no. 183.

[34] A. Hunt, P. Schönknechta, M. Henze, U. Seidl, U. Haberkorn, and J. Schrödera, "Reduced cerebral glucose metabolism in patients at risk for alzheimer's disease," *Psychiatry Res. Neuroimaging*, vol. 155, no. 2, pp. 147–154, 2007.

[35] T. Nakao, J. Radua, K. Rubia, and D. Mataix-Cols, "Gray matter volume abnormalities in ADHD: Voxel-based meta-analysis exploring the effects of age and stimulant medication," *Amer. J. Psychiatry*, vol. 168, no. 11, pp. 1154–1163, 2011.

[36] L. Delano-Wood, N. H. Stricker, S. F. Sorg, D. A. Nation, A. J. Jak, S. P. Woods, D. J. Libon, D. C. Delis, L. R. Frank, and M. W. Bondi, "Posterior cingulum white matter disruption and its associations with verbal memory and stroke risk in mild cognitive impairment," *J. Alzheimer's Disease*, vol. 29, no. 3, pp. 589–603, 2012.

[37] G. B. Frisoni, et al., "Mapping local hippocampal changes in alzheimer's disease and normal ageing with MRI at 3 tesla," *Brain*, vol. 131, no. 12, pp. 3266–3276, 2008.

[38] M. M. Vasavada, J. Wang, P. J. Eslinger, D. J. Gill, X. Sun, P. Karunanayaka, and Q. X. Yang, "Olfactory cortex degeneration in alzheimer's disease and mild cognitive impairment," *J. Alzheimer's Disease*, vol. 45, no. 3, pp. 947–958, 2015.

[39] M. Sjobeck and E. Englund, "Alzheimer's disease and the cerebellum: A morphologic study on neuronal and glial changes," *Dementia Geriatric Cogn. Disorders*, vol. 12, no. 3, pp. 211–218, 2001.
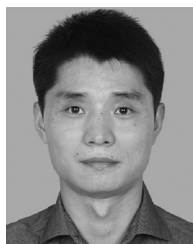
Lei Du received the PhD degree in computer science from the School of the Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2013. Currently, he is an assistant professor in the School of Automation, Northwestern Polytechnical University, Xi'an, China. His research interests include brain imaging genetics, bioinformatics, machine learning, and big data mining. He is a member of the IEEE.

Kefei Liu received the BSc degree in mathematics from Wuhan University and the PhD degree in electronic engineering from the City University of Hong Kong, in 2006 and 2013 respectively. He is a postdoctoral research associate in the Department of Biostatistics, Epidemiology and Informatics at the University of Pennsylvania. He is interested in developing machine learning and statistical methods for the analysis of large-scale heterogeneous biological data.

Xiaohui Yao received the BS degree in computer science and technology from Qing Dao University, in 2009, the MS degree in computer software and theory from the University of Science and Technology of China, in 2012, and the PhD degree in bioinformatics from Indiana University, in 2018. She is a postdoctoral fellow with the Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania. Her research interests include imaging genetics, multidimensional data mining, systems biology, and information visualization.

Shannon L. Risacher received the BS degree in psychology from Indiana University-Purdue University Indianapolis, in 2006, and the PhD degree in medical neuroscience from Indiana University's School of Medicine, in 2011, where she is currently an assistant professor of radiology and imaging sciences. Her main research interest is in identifying biomarkers for early detection of Alzheimer's disease pathology before clinical symptoms.

Junwei Han received the PhD degrees in pattern recognition and intelligent systems from the School of Automation, Northwestern Polytechnical University, Xi'an, China, in 2003. He is currently a professor in the School of Automation, Northwestern Polytechnical University. His research interests include computer vision and multi-media processing. He is a senior member of the IEEE.

Andrew J. Saykin received the BA degree in psychology from the University of Massachusetts Amherst, in 1977, the MS degree in clinical psychology, in 1977, and the PsyD degree in clinical neuropsychology from Hahnemann Medical College, in 1982. He is the Raymond C. Beeler professor of radiology and professor of medical and molecular genetics at Indiana University's School of Medicine. His expertise is in the areas of multi-modal neuroimaging research, human genetics, and neuropsychology/cognitive neuroscience. He has a longstanding interest in the structural, functional, and molecular substrates of cognitive deficits in Alzheimer's disease, cancer, brain injury, schizophrenia, and other neurological and neuropsychiatric disorders.

Lei Guo received the BS, MS, and PhD degrees in 1982, 1986, and 1993, respectively. He is currently a professor of pattern recognition at Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, image processing, image segmentation, object detection, and tracking.

Li Shen received the BS degree from Xi'an Jiao Tong University, in 1993, the MS degree from Shanghai Jiao Tong University, in 1996, and the PhD degree from Dartmouth College, all in computer science, in 2004. He is a professor of informatics at the University of Pennsylvania's Perelman School of Medicine. He is an elected fellow of the American Institute for Medical and Biological Engineering (AIMBE). His research interests include medical image computing, bioinformatics, machine learning, brain imaging genomics, and big data science in biomedicine. He is a senior member of the IEEE.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.