Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2018 December ; 2018: 356–361. doi:10.1109/BIBM.2018.8621298.

Fast Multi-Task SCCA Learning with Feature Selection for Multi-Modal Brain Imaging Genetics

Lei Du[1,*], **Kefei Liu**[2], **Xiaohui Yao**[2], **Shannon L. Risacher**[3], **Junwei Han**[1], **Lei Guo**[1], **Andrew J. Saykin**[3], **Li Shen**[2], and **for the Alzheimer's Disease Neuroimaging Initiative**

[1]School of Automation, Northwestern Polytechnical University

[2]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine

[3]Department of Radiology and Imaging Sciences, Indiana University School of Medicine

Abstract

Brain imaging genetics studies the genetic basis of brain structures and functions via integrating both genotypic data such as single nucleotide polymorphism (SNP) 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 analyses. MTL methods generally incorporate a few of QTs and are not designed for feature selection from a large number of QTs; while existing SCCA methods typically employ only one modality of QTs to study its association with SNPs. Both MTL and SCCA encounter computational challenges as the number of SNPs increases. In this paper, combining the merits of MTL and SCCA, we propose a novel multi-task SCCA (MTSCCA) learning framework 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. Using the $G_{2,1}$-norm regularization, MTSCCA treats all SNPs in the same group together to enforce sparsity at the group level. The $\ell_{2,1}$-norm penalty is used to jointly select features across multiple tasks for SNPs, and across multiple modalities for QTs. A fast optimization algorithm is proposed using the grouping information of SNPs. Compared with conventional SCCA methods, MTSCCA obtains improved performance regarding both correlation coefficients and canonical weights patterns. In addition, our method runs very fast and is easy-to-implement, and thus could provide a powerful tool for genome-wide brain-wide imaging genetic studies.

Keywords

Brain Imaging Genetics; Sparse Canonical Correlation Analysis; Multi-task SCCA

[*]Corresponding to: dulei@nwpu.edu.cn.

I. Introduction

In brain science, imaging genetics is an emerging and important topic which integrates both the genetic factors and neuroimaging phenotypic measurements. This integration strategy of combining diverse imaging and omics data is expected to uncover the genetic basis of brain structures and functions [1]–[3]. Modern neuroimaging techniques, such as magnetic resonance imaging (MRI) and positron-emission tomography (PET), image the structure and metabolic processes of the brain based on different techniques. These multi-modal imaging data provide complementary information for a more comprehensive understandings of brain structure, function and abnormality [4]. In biomedical studies, we usually face a large number of genotyping biomarkers such as the single nucleotide polymorphisms (SNPs). Therefore, developing fast and efficient imaging genetics methods which integrates multi-modal imaging data simultaneously is quite important.

The multivariate learning methods are very popular in brain imaging genetics since both imaging data and genetic data are multidimensional. The multi-task learning (MTL), especially MTL regression, are of this kind and widely used in brain imaging genetics [5], [6]. Generally, MTL methods treat a few important imaging QTs as dependent variables and SNPs as independent variables. Then joint effect of multi-locus genotype variables on a few phenotypes is studied. This paradigm can select SNPs that are simultaneously relevant to candidate brain phenotypes, but may ignore 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 phenotypes from multiple brain cerebral components. Therefore, bi-multivariate methods become more and more popular recently. Sparse canonical correlation analysis (SCCA) identifies the relationship between two views of data with sparse output induced by different penalties [7]–[12]. These SCCA methods have limited power since they only utilize QTs from one single imaging modality. Given multi-modal imaging data, incorporating them together 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 integral SCCA model is desirable and of great interest. This integration model would be helpful to elucidate the shared mechanism of genetic factors on the brain. Though the multi-view SCCA modelling could address this issue [12], it learns only one single canonical weight for genetic loci which is overstrict.

In this paper, we propose a Multi-Task learning based SCCA (MTSCCA) framework which can study bi-multivariate associations between phenotypes of multiple modalities and genotypes simultaneously. MTSCCA treats each SNP or QT as a feature, and then models the association between each imaging modality and SNPs as a learning task. Different from those conventional SCCA, 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. We take into consideration the group structure such as the linkage disequilibrium (LD) [13] in human genome via the group $\ell_{2,1}$-norm ($G_{2,1}$-norm) [6]. The jointly individual feature selection is also taken into consideration via a $\ell_{2,1}$-norm. In addition, we propose a fast and efficient

optimization algorithm to solve the MTSCCA problem. We apply MTSCCA to a large real neuroimaging genetic data set of the Alzheimer's disease neuroimaging initiative (ADNI) [14] cohort with SNPs from chromosome 19 and three different modalities of imaging QTs included. Experimental results show that, compared with conventional SCCA methods, MTSCCA yields both 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.

II. The Multi-Task SCCA Learning Method

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

A. The MTSCCA Method

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

$$\min_{\mathbf{u}_j, \mathbf{v}_j} \sum_j -\mathbf{u}_j^\top \mathbf{X}^\top \mathbf{Y}_j \mathbf{v}_j \quad (1)$$

$$s.t. \ \left\|\mathbf{X}\mathbf{u}_j\right\|_2^2 = 1, \left\|\mathbf{Y}_j \mathbf{v}_j\right\|_2^2 = 1, \Omega(\mathbf{U}) \le b_1, \Omega(\mathbf{V}) \le b_2, \forall j.$$

where $\mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2 \cdots \mathbf{u}_c]$, $\mathbf{V} = [\mathbf{v}_1 \ \mathbf{v}_2 \cdots \mathbf{v}_c]$.

Obviously, our model is distinct from those mCCA models. First, MTSCCA employs the multi-task framework which learns a series of related SCCA tasks together. This simultaneous learning has been shown to improve performance dramatically compared with learning each task independently [15], [16]. 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 model is well scalable in terms of both modeling and

computation. According to Eqs. (1), the number of tasks of MTSCCA is equal to the number of imaging modalities, which means the computation burden increases linearly.

1) Group-sparsity for Genetic Association and Joint Individual Feature Selection for SNPs: It is known that numerous SNPs inherently exhibit group structure in the genome. Thus we use the $G_{2,1}$-norm function [6] for regularization. Suppose the SNPs are partitioned into K non-overlapping groups $\mathcal{G} = \{g_k\}_{k=1}^{K}$, where m_g is the number of SNPs in group g, then the $G_{2,1}$-norm function is formulated as

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

$\mathbf{U}^k$ is a submatrix of $\mathbf{U}$ with rows in $\mathbf{U}$ indexed by g_k. This regularization penalizes the SNPs in the same group, i.e. $\{\mathbf{u}^i\}_{i \in g_k}$, as a whole and expects to estimate equal or similar coefficients for them.

Generally, within a specific group, an individual SNP could be relevant to the QTs and those remaining ones could be irrelevant. Therefore, we model this via the $\ell_{2,1}$-norm regularization which is usually used in multi-task models,

$$\|\mathbf{U}\|_{2,1} = \sum_{i=1}^{p} \|\mathbf{u}^i\|_2 = \sum_{i=1}^{p} \sqrt{\sum_{j=1}^{c} u_{ij}^2}. \quad (3)$$

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 features at the individual level across all SCCA tasks.

2) Joint Individual Feature Selection across Different Imaging Modalities: Identifying imaging biomarkers is also of great interest in out study. Since MTSCCA learns only one canonical weight for each imaging modality, the sparsity-inducing term $\ell_{2,1}$-norm is imposed across different imaging modalities, viz

$$\|\mathbf{V}\|_{2,1} = \sum_{i=1}^{q} \|\mathbf{v}^i\|_2 = \sum_{i=1}^{q} \sqrt{\sum_{j=1}^{c} v_{ij}^2}. \quad (4)$$

This motivation of using this penalty is as follows. Despite collected based on different imaging technologies, all modalities of imaging data are measured from the same brain space and have been map onto the same brain atlas via the segmentation and registration.

Therefore, it is reasonable to estimate equal or similar coefficients for those imaging features associated with the same brain area but attributed to different modalities.

B. The Optimization Algorithm

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

$$\min_{\mathbf{u}_j,\mathbf{v}_j} \sum_{j=1}^{c} -\mathbf{u}_j^\top \mathbf{x}^\top \mathbf{Y}_j \mathbf{v}_j \quad (5)$$

$$s.t. \|\mathbf{X}\mathbf{u}_j\|_2^2 = 1, \|\mathbf{Y}_j\mathbf{v}_j\|_2^2 = 1, \|\mathbf{U}\|_{G_{2,1}} \le a, \|\mathbf{U}\|_{2,1} \le b_1, \|\mathbf{v}\|_{2,1} \le b_2, \forall j.$$

To solve Eq. (5), we first modify the loss function to

$$\min_{\mathbf{u}_j,\mathbf{v}_j} \sum_{j=1}^{c} \|\mathbf{X}\mathbf{u}_j - \mathbf{Y}_j\mathbf{v}_j\|_2^2 \quad (6)$$

$$s.t. \left\|\mathbf{X}\mathbf{u}_j\right\|_2^2 = 1, \left\|\mathbf{Y}_j\mathbf{v}_j\right\|_2^2 = 1, \left\|\mathbf{U}\right\|_{G_{2,1}} \le a, \left\|\mathbf{U}\right\|_{2,1} \le b_1, \left\|\mathbf{V}\right\|_{2,1} \le b_2, \forall j,$$

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

$$\mathscr{L}(\mathbf{U},\mathbf{V}) = \sum_{j=1}^{c} \left\|\mathbf{X}\mathbf{u}_j - \mathbf{Y}_j\mathbf{v}_j\right\|_2^2 + \beta\left(\left\|\mathbf{U}\right\|_{G_{2,1}} - a\right) + \lambda_1\left(\left\|\mathbf{U}\right\|_{2,1} - b_1\right) + \lambda_2\left(\left\|\mathbf{V}\right\|_{2,1} - b_2\right) \quad (7)$$
$$+ \gamma_1^*\left(\left\|\mathbf{X}\mathbf{u}\right\|_2^2 - 1\right) + \gamma_2^*\left(\left\|\mathbf{Y}\mathbf{v}\right\|_2^2 - 1\right),$$

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.

This problem is difficult to solve since it is non-convex in loss function and non-smooth in penalty functions. Fortunately, it is convex in **U** with **V** fixed. Moreover, this objective is convex in $\mathbf{v}_j$ with those remaining $\mathbf{v}_k$ $(k \neq j)$ and **U** fixed. On this account, we can solve this problem via the alternative update rule which is widely used in optimization community.

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

$$-\mathbf{X}^\top\left[\mathbf{Y}_1\mathbf{v}_1,\cdots,\mathbf{Y}_c\mathbf{v}_c\right]+\beta\widetilde{\mathbf{D}}\mathbf{U}+\lambda_1\mathbf{D}_1\mathbf{U}+\gamma_1\mathbf{X}^\top\mathbf{X}\mathbf{U}=\mathbf{0}, \quad (8)$$

where $2\widetilde{\mathrm{D}}\mathrm{U}$ is the subgradient of $\|\mathbf{U}\|_{G_{2,1}}$ and $2\mathbf{D}_1\mathbf{U}$ is that of $\|\mathbf{U}\|_{2,1}$; $\widetilde{\mathbf{D}}$ is a diagonal matrix with the k-th diagonal block 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 k-th group; $\mathbf{D}_1$ is also a diagonal matrix with the i-th entry being $\frac{1}{2\|\mathbf{u}^i\|_2}(i\in[1,p])$; and $\gamma_1=\gamma_1^*+1$. Then we can easily have

$$\mathbf{U}=\left(\beta\widetilde{\mathbf{D}}+\lambda_1\mathbf{D}_1+\gamma_1\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\left[\mathbf{Y}_1\mathbf{v}_1,\cdots,\mathbf{Y}_c\mathbf{v}_c\right]. \quad (9)$$

According to [6], this linear system in terms of $\mathbf{U}$ can be efficiently solved via an iterative algorithm by alternatively first updating $\widetilde{\mathbf{D}}$ 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. To accelerate the solution, we introduce the following theorem (proof is omitted and a similar proof can be found in [17]).

Theorem 1: If $\mathbf{X}^\top\mathbf{X}$ is a block diagonal matrix, Eq. (10) can be solved by

$$\mathbf{U}=\oplus_{k=1}^{K}\mathbf{U}^k=\begin{bmatrix}\mathbf{U}^1\\ \vdots\\ \mathbf{U}^K\end{bmatrix}, \quad (10)$$

$$\mathbf{U}^k=\left(\beta\underline{\widetilde{\mathbf{D}}}_{g_k}+\lambda_1\underline{\mathbf{D}_1}_{g_k}+\gamma_1\mathbf{X}_{g_k}^\top\mathbf{X}_{g_k}\right)^{-1}\mathbf{X}_{g_k}^\top\left[\mathbf{Y}_1\mathbf{v}_1,\cdots,\mathbf{Y}_c\mathbf{v}_c\right],$$

where $\underline{\widetilde{\mathrm{D}}}_{g_k}$ is the k-th diagonal block of $\widetilde{\mathbf{D}}$; $\underline{\mathbf{D}_1}_{g_k}$ is the k-th diagonal block of $\mathbf{D}_1$; and $\oplus$ denotes the concatenate operator for matrices along rows.

The advantages of this theorem are three folds. (1) The time complexity of Eq. (10) is $O\left(nm_k^2K\right)$ compared with that of Eq. (9) being $O(np^2)$, where m_k is the size of the k-th group, and $p=\sum_{k=1}^{K}m_k$. This is a significant improvement because that the LD block size is usually much smaller than the number of SNPs ($m_k\ll p$) in human genome [18]. (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 $\mathbf{x}_{g_k}^\top\mathbf{x}_{g_k}$. (3) Eq. (10) is easy to implement, demonstrating it is very promising in big imaging genetic analysis.

2) Updating $\mathbf{v}_j$: Note that each $\mathbf{v}_j$ is associated with each $\mathbf{Y}_j$ respectively. This means that these $\mathbf{v}_j$'s are not coupled (as compared to $\mathbf{u}_j$'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. (7), we take the derivative with respect to $\mathbf{v}_j$ and set it to zero

$$-\mathbf{Y}_j^\top \mathbf{X}\mathbf{u}_j + \lambda_2 \mathbf{D}_2 \mathbf{v}_j + \gamma_2 \mathbf{Y}_j^\top \mathbf{Y}_j \mathbf{v}_j = \mathbf{0}, \quad (11)$$

which yields

$$\mathbf{v}_j = \left(\lambda_2 \mathbf{D}_2 + \gamma_2 \mathbf{Y}_j^\top \mathbf{Y}_j\right)^{-1} \mathbf{Y}_j^\top \mathbf{X}\mathbf{u}_j, \quad (12)$$

where $\mathbf{D}_2$ is a diagonal matrix which is loaded by $\frac{1}{2\|\mathbf{v}^i\|_2}(i \in [1, q])$ on the diagonal; and $\gamma_2 = \gamma_2^* + 1$. Therefore, each $\mathbf{v}_j$ can also be solved alternatively through an iteration 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. (7)

Require:

$\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{Y}_j \in \mathbb{R}^{n \times q}$, $j \in [1, c]$, β, λ_1, λ_2, γ_1, γ_2

Ensure:

Canonical weights $\mathbf{U}$ and $\mathbf{V}$.

1: Initialize $\mathbf{U} \in \mathbb{R}^{p \times c}$, $\mathbf{V} \in \mathbb{R}^{q \times c}$;

2: **while** not convergence **do**

3: Update $\widetilde{\mathrm{D}}_{g_k}$ and $\mathbf{D}_{1_{g_k}}$;

4: Solve $\mathbf{U}$ according to Eq. (9), and normalize $\mathbf{u}_j$ to $\|\mathbf{X}\mathbf{u}_j\|_2^2 = 1$;

5: Update $\mathbf{D}_2$;

6: Solve $\mathbf{v}_j$ ($j = 1, \cdots, c$) in turn according to Eq. (12), and normalize $\mathbf{v}_j$ to $\|\mathbf{Y}_j \mathbf{v}_j\|_2^2 = 1$;

7: **end while**

III. Results

A. Experimental Setup

A nested 5-fold cross-validation strategy was used in this work. Specifically, those tuning parameters was determined in the inner loop where a group of them generating the highest

canonical coefficients will be chosen as the optimal parameters. Empirically, we fine tuned the β, λ_1 and λ_2 from {0.01, 0.1, 1, 10, 100} which usually yielded good results in this study. For the remaining γ_1 and γ_2, we simply set them to 1 as they have been shown to be insensitive to the learned results [8]. In the outer loop, the 5-fold training and testing results were calculated and presented.

To the best of our knowledge, this is the first multi-task SCCA method, and thus no previous work can be used to compare with. On this account, we choose the conventional SCCA, including both two-view SCCA and mSCCA [12] as benchmarks. This could help show the effectiveness of MTSCCA. Another issue is that these conventional methods suffer from heavy computational and memory requirement issues because they cannot handle the large covariance matrix calculation. To make the comparison feasible, based on Theorem 1, we implement the fast SCCA and the fast mSCCA. This yields the two benchmark methods in this study.

B. Data Sources

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

The neuroimaging data were from 755 non-Hispanic Caucasian subjects, including 281 AD, 292 MCI and 182 healthy control (HC) participants. The data contained multiple modalities including 18-F florbetapir PET (AV45) scans, fluorodeoxyglucose PET (FDG) scans, and structural MRI scans. These data were downloaded from the ADNI database (adni.loni.usc.edu). These multi-modality imaging data were aligned to each other for each participant. The structural MRI scans were processed with voxel-based morphometry (VBM) via SPM [19]. 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 as 2×2×2 mm^3 voxels, and had been smoothed with an 8mm 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. The studied measures include 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 for removing the effects of the baseline age, gender, education, and handedness.

The genotyping data of the same population were also downloaded from the ADNI website. The data were generated 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. Among all human chromosomes, chromosome 19 contains the largest

number of genes, in which the gene density is more than double the genome-wide average [20]. In addition, this chromosome also includes the well-known AD risk genes such as *APOE, APOC1* and *TOMM40*. 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. Among these SNPs, most of them might be irrelevant to AD, while only few of them could be relevant via influencing the intermediate brain imaging measurements. The aim of this study is to identify this small subset of SNPs in chromosome 19 that are related to brain imaging markers.

C. Experimental Results

We first use the canonical correlation coefficient (CCC) as an evaluation criteria. There will be three pairs of associations, and we denote them as SNP-AV45, SNP-FDG and SNP-VBM for the sake of description. For the three SCCA tasks, MTSCCA learns them together and generates a canonical weight matrix $\mathbf{U}$ for SNP data and one canonical weight vector $\mathbf{v}_j$ for AV45, FDG and VBM data. We then calculate CCCs in terms of SNP-AV45, SNP-FDG and SNP-VBM separately. The two-view SCCA naturally yields three CCCs for these three tasks. Although the mSCCA only learns one canonical weight vector for SNP data, we use it three times to generate three CCCs with respect to the three tasks.

Fig. 1 shows the CCCs of the SNP data with each type of imaging QT data, where CCCs of SNP-AV45, SNP-FDG and SNP-VBM are separately shown. In this figure, both the training CCCs and testing CCCs, as well as their standard deviations (SDs) are presented. By changing the number of selected features (10, 20, ⋯, 100 in this work) for both SNP and imaging QT data, 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 SNP-VBM. After investigation, this could be attributed to 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 because it seems counter-intuitive because more data (three different imaging modalities here) ought to provide more information. The reason might attribute to the modelling strategy of mSCCA. Demanding one set of features (SNPs) being associated with three sets of features (imaging QTs) simultaneously could be overstrict and thus harm the performance. In addition, we calculate the p-values between MTSCCA and two competing methods and show them in Table I. The p-values all reach the significance level which means that our method is significantly better than both competing methods. These results in terms of CCCs indicate 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.

Apart from the CCCs, the selected features in terms of SNPs are a major concern. We show the top ten selected SNPs according to the canonical weight values of each individual method in Table II. 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. Owning to the joint learning paradigm, 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. rs429358 (*APOE*), rs56131196 (*APOC1*), rs12721051 (*APOC1*), rs4420638 (*APOC1*), rs111789331 (4.5 kb of *APOC1*), rs66626994 (5.6 kb of *APOC1*), rs146275714 (*PVRL2*), rs147711004 (71 kb of *APOE*) and rs10119 (*TOM-M40*), have been reported to show increasing risk of AD in previous studies. This indicates the ability of MTSCCA in identifying meaningful SNPs from massive genetic markers. The two-view SCCA identifies rs429358 and five other AD related SNPs (rs10414043, rs147711004, rs7256200, rs73052335 and rs66626994). But it identifies four SNPs that are not reported by now and thus further investigation is warranted. The mSCCA performs inadequately since it does not find out the most important locus rs429358. Moreover, except the marker rs623264, the remaining nine identified SNPs of mSCCA have not been reported yet. This reveals that MTSCCA could be a suitable tool in discovering meaningful genetic markers in a very large scenario.

Fig. 2 presents the canonical weights on each imaging modality (AV45, FDG and VBM) across the five trials. We observe that those imaging markers with nonzero coefficients generated by MTSCCA are all associated with AD. We also show the top ten selected QTs of each imaging modal data of MTSCCA in Table III. 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. Most importantly, these markers are all have been documented to be related to AD or MCI [21]–[25]. These results indicate that MTSCCA could identify meaningful imaging markers that are associated with the status of dementia. The mSCCA also identifies a few of AD related markers such as the hippocampus. The two-view method is rambling and thus is not a good option in this scenario. To summarize, the top ten selected SNPs and imaging QTs are highly correlated with each other, and with AD, demonstrating that MTSCCA could be a very promising method in brain imaging genetics.

IV. Conclusion

In this paper, we have proposed a novel multi-task based SCCA (MTSCCA) method and applied it to imaging genetic problem with multi-modal brain imaging QTs. Different from existing SCCA methods, MTSCCA incorporates multiple sets of imaging modalities into a single integral model. MTSCCA has better modeling capability than both conventional SCCA and MTL regression. A fast optimization algorithm is proposed which avoids calculating the large covariance during solution.

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 succeeded in identifying a small set of SNPs from a large number of genetic markers from chromosome 19. It is worth noting that all top ten selected SNPs of MTSCCA were known AD risk factors. In addition, the canonical weight patterns of imaging QTs were also of great

importance. The identified imaging QTs were highly correlated to AD or MCI. These results demonstrated that the proposed multi-task SCCA could be a powerful tool in big data mining in brain imaging genetics. We plan to extend MTSCCA to genome-wide brain-wide imaging analysis in the future work.

Acknowledgments

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. This work was supported by NSFC [61602384]; Natural Science Basic Research Plan in Shaanxi Province of China [2017JQ6001]; China Postdoctoral Science Foundation [2017M613202]; Science and Technology Foundation for Selected Overseas Chinese Scholar [2017022]; Postdoctoral Science Foundation of Shaanxi Province [2017BSHEDZZ81]; Fundamental Research Funds for Central Universities at Northwestern Polytechnical University. This work was also supported by the National Institutes of Health [R01 EB022574, R01 LM011360, U01 AG024904, P30 AG10133, R01 AG19771, R01 AG 042437, R01 AG046171, R01 AG040770] at University of Pennsylvania and Indiana University.

References

[1]. Potkin SG, Turner J, Fallon J, Lakatos A, Keator D, Guffanti G, and Macciardi F, "Gene discovery through imaging genetics: identification of two novel genes associated with schizophrenia," Molecular psychiatry, vol. 14, no. 4, p. 416, 2009. [PubMed: 19065146]

[2]. Saykin AJ, Shen L, Yao X, Kim S, Nho K 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.

[3]. Shen L, Kim S, Risacher SL, Nho K, Swaminathan S, West JD, Foroud T, Pankratz N, Moore JH, Sloan CD 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. 1051–63, 2010. [PubMed: 20100581]

[4]. Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, Trojanowski JQ, Toga AW, and Beckett L, "The alzheimer's disease neuroimaging initiative," Neuroimaging Clinics of North America, vol. 15, no. 4, pp. 869–877, 2005. [PubMed: 16443497]

[5]. Lee S, Zhu J, and Xing EP, "Adaptive multi-task lasso: with application to eqtl detection," in International Conference on Neural Information Processing Systems, 2010, pp. 1306–1314.

[6]. Wang H, Nie F, Huang H, Kim S, Nho K, Risacher SL, Saykin AJ, Shen L et al., "Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort," Bioinformatics, vol. 28, no. 2, pp. 229–237, 2012. [PubMed: 22155867]

[7]. Chen J, Bushman FD et al., "Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis," Biostatistics, vol. 14, no. 2, pp. 244–258, 2013. [PubMed: 23074263]

[8]. Chen X and Liu H, "An efficient optimization algorithm for structured sparse cca, with applications to eqtl mapping," Statistics in Biosciences, vol. 4, no. 1, pp. 3–26, 2012.

[9]. Du L, Huang H, Yan J, Kim S, Risacher SL et al., "Structured sparse canonical correlation analysis for brain imaging genetics: An improved graphnet method," Bioinformatics, vol. 32, no. 10, pp. 1544–1551, 2016. [PubMed: 26801960]

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

[11]. Lin D, Calhoun VD, and Wang YP, "Correspondence between fMRI and SNP data by group sparse canonical correlation analysis," Medical Image Analysis, 2013.

[12]. Witten DM and Tibshirani RJ, “Extensions of sparse canonical correlation analysis with applications to genomic data,” Statistical applications in genetics and molecular biology, vol. 8, no. 1, pp. 1–27, 2009.

[13]. Pritchard JK and Przeworski M, “Linkage disequilibrium in humans: Models and data,” American Journal of Human Genetics, vol. 69, no. 1, pp. 1–14, 2001. [PubMed: 11410837]

[14]. Weiner MW, Aisen PS, Jack CR, Jagust WJ, Trojanowski JQ, Shaw L, Saykin AJ, Morris JC, Cairns N, Beckett LA et al., “The alzheimer’s disease neuroimaging initiative: progress report and future plans,” Alzheimer’s & Dementia, vol. 6, no. 3, pp. 202–211, 2010.

[15]. Ando RK and Zhang T, “A framework for learning predictive structures from multiple tasks and unlabeled data,” Journal of Machine Learning Research, vol. 6, pp. 1817–1853, 2005.

[16]. Argyriou A, Evgeniou T, and Pontil M, “Multi-task feature learning.” Advances in Neural Information Processing Systems, vol. 73, no. 3, pp. 41–48, 2006.

[17]. Huang Y, Du L, Liu K, Yao X, Risacher SL, Guo L, Saykin AJ, and Shen L, “A fast scca algorithm for big data analysis in brain imaging genetics,” in Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics. Cham: Springer International Publishing, 2017, pp. 210–219.

[18]. Rosenfeld JA, Mason CE, and Smith TM, “Limitations of the human reference genome for personalized genomics.” PLOS ONE, vol. 7, no. 7, 2012.

[19]. Ashburner J and Friston KJ, “Voxel-based morphometry-the methods,” Neuroimage, vol. 11, no. 6, pp. 805–21, 2000. [PubMed: 10860804]

[20]. Grimwood J, Gordon LA, Olsen A, Terry A, Schmutz J, Lamerdin J, Hellsten U, Goodstein D, Couronne O, Tran-Gyamfi M et al., “The dna sequence and biology of human chromosome 19,” Nature, vol. 428, no. 6982, p. 529, 2004. [PubMed: 15057824]

[21]. Hunt A, Schönknecht P, Henze M, Seidl U, Haberkorn U, and Schröder J, “Reduced cerebral glucose metabolism in patients at risk for alzheimer’s disease,” Psychiatry Research Neuroimaging, vol. 155, no. 2, pp. 147–154, 2007.

[22]. Nakao T, Radua J, Rubia K, and Mataix-Cols D, “Gray matter volume abnormalities in adhd: voxel-based meta-analysis exploring the effects of age and stimulant medication.” Am J Psychiatry, vol. 168, no. 11, pp. 1154–1163, 2011. [PubMed: 21865529]

[23]. Frisoni GB, Ganzola R, Canu E, Rub U, Pizzini FB, Alessandrini F, Zoccatelli G, Beltramello A, Caltagirone C, and Thompson PM, “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. [PubMed: 18988639]

[24]. Sjobeck M and Englund E, “Alzheimer’s disease and the cerebellum: a morphologic study on neuronal and glial changes,” Dement Geriatr Cogn Disord, vol. 12, no. 3, pp. 211–218, 2001. [PubMed: 11244215]

[25]. Vasavada MM, Wang J, Eslinger PJ, Gill DJ, Sun X, Karunanayaka P, and Yang QX, “Olfactory cortex degeneration in alzheimer’s disease and mild cognitive impairment,” Journal of Alzheimers Disease Jad, vol. 45, no. 3, pp. 947–58, 2015. [PubMed: 25633674]

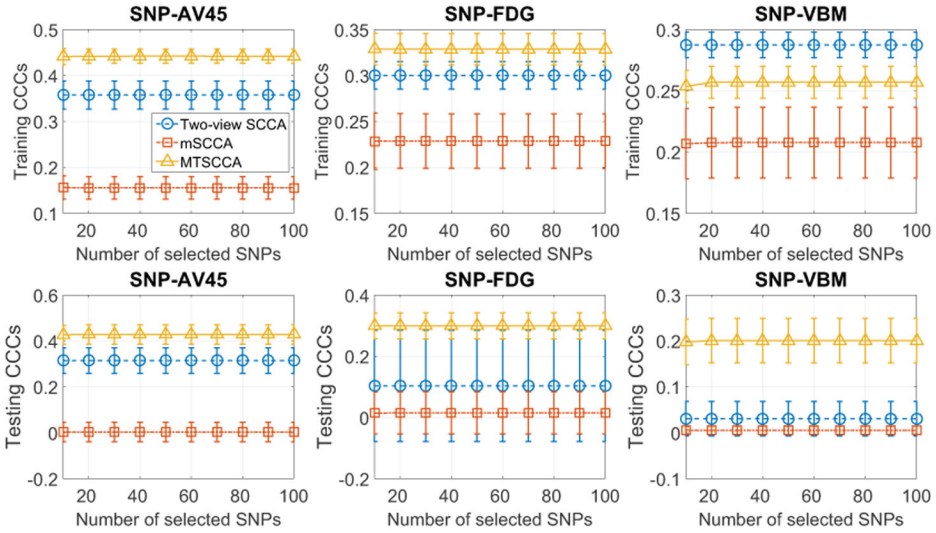

Fig. 1.

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

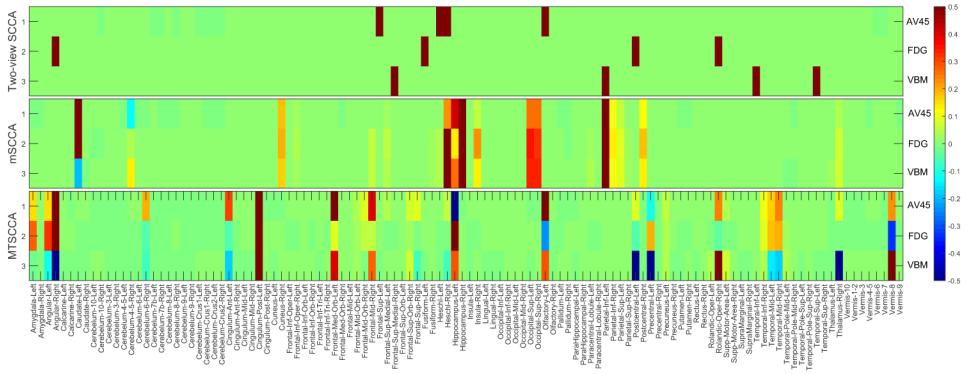

Fig. 2.

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.

TABLE I

The *p*-values of *t*-tests for CCCs comparison between MTSCCA and benchmarks. The '-' in parenthesis means that MTSCCA loses on this trial.

	SNP-AV45	**SNP-FDG**	**SNP-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

TABLE II

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 III

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