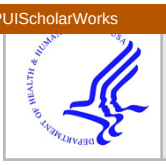

Published in final edited form as:
Inf Process Med Imaging. 2017 June ; 10265: 543–555. doi:10.1007/978-3-319-59050-9_43.

Identifying Associations Between Brain Imaging Phenotypes and Genetic Factors via A Novel Structured SCCA Approach

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

[1]School of Automation, Northwestern Polytechnical University, Xi'an China

[2]Radiology and Imaging Sciences, Indiana University School of Medicine, IN, USA

Abstract

Brain imaging genetics attracts more and more attention since it can reveal associations between genetic factors and the structures or functions of human brain. Sparse canonical correlation analysis (SCCA) is a powerful bi-multivariate association identification technique in imaging genetics. There have been many SCCA methods which could capture different types of structured imaging genetic relationships. These methods either use the group lasso to recover the group structure, or employ the graph/network guided fused lasso to find out the network structure. However, the group lasso methods have limitation in generalization because of the incomplete or unavailable prior knowledge in real world. The graph/network guided methods are sensitive to the sign of the sample correlation which may be incorrectly estimated. We introduce a new SCCA model using a novel graph guided pairwise group lasso penalty, and propose an efficient optimization algorithm. The proposed method has a strong upper bound for the grouping effect for both positively and negatively correlated variables. We show that our method performs better than or equally to two state-of-the-art SCCA methods on both synthetic and real neuroimaging genetics data. In particular, our method identifies stronger canonical correlations and captures better canonical loading profiles, showing its promise for revealing biologically meaningful imaging genetic associations.

1 Introduction

In recent years, brain imaging genetics becomes a popular research topic in biomedical and bioinformatics studies. Brain imaging genetics refers to the study of modeling and understanding how genetic factors influence the structure or function of human brain using the imaging measurements as the quantitative endophenotype [12, 11, 13]. Both the genetic factors, such as the single nucleotide polymorphisms (SNPs), and the imaging measurements

Correspondence to: Li Shen.

[**]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.

such as the imaging quantitative traits (QTs) are multivariate. Therefore, discovering meaningful bi-multivariate associations is an important task in brain imaging genetics.

Equipped with feature selection, sparse canonical correlation analysis (SC-CA) gains tremendous attention for its powerful ability in bi-multivariate association identification. There are many SCCA methods using different types of shrinkage techniques. The ℓ_1-norm penalty and its variants are widely used, but they only pursuit individual level sparsity [16, 8]. In biomedical studies, the genetic biomarkers usually function simultaneously other than individually [14]. This is also the case for the imaging measurements. Therefore, the structure level sparsity, such as the group structure or the graph/network structure, is of great interest and importance in brain imaging genetics [14, 15].

To capture the high-level structure information, several different structure-aware penalties have been proposed. There are roughly two kinds of structured SCCA methods according to their different penalties [4]. The first kind of SC-CA methods consider the group information using the group lasso regularizer, which is an intra-group ℓ_2-norm and inter-group ℓ_1-norm [1, 6]. The group lasso tends to assign equal weights for those variables in a same group, and each group will be selected or not as a whole [18]. To our knowledge, this type of SCCA methods require the priori knowledge to define the group structure. This limits their applications as it is hard to obtain precise priori knowledge in real biological studies [4]. The second kind of SCCA methods rebuild the structure information via the graph guided or network guided penalty [3, 6, 2, 5, 4]. These SCCA methods can capture the structure information using any available priori knowledge. Moreover, they can also recover the structure information based on the input data [4]. Three types of graph guided penalties have been used: (1) the graph guided fused lasso penalty and its variants [3, 1, 7], (2) the correlation sign based graph guided fused ℓ_2-norm penalty [2], and (3) the improved GraphNet based penalty [4]. Du *et al.* [4] has shown that the first two types of graph guided penalties could introduce estimation bias because of the sign of the correlations can be wrongly calculated. The reason could be that the sign of the correlations can be easily changed when removing a fraction of the data or perturbing the data as in bootstrap or sub-sampling. The improved GraphNet utilizes ℓ_2-norm with respect to the structure penalty terms, which may not produce desirable sparse results at structure level.

Inspired by the success of group lasso in group selection, we consider a case where each group is made up of only two variables. Both variables will be extracted together with similar or equal weights. Interestingly, this novel group lasso can be used in data-driven mode where no priori knowledge is provided. We call it graph guided pairwise group lasso (GGL) which bridges the gap between the group lasso and graph guided penalties. We then propose a new graph guided pairwise group lasso based sparse canonical correlation analysis model (GGL-SCCA) with intention to recover the structure information automatically. The proposed SCCA method is sample correlation sign independent and it is robust to those existing SCCA methods using graph guided penalty. We also propose an efficient optimization algorithm to solve the problem. Besides, we also provide a quantitative upper bound for the grouping effect of our method to demonstrate its structure identifying ability. Compared with the state-of-art SCCA methods such as NS-SCCA [2] and AGN-SCCA [4], GGL-SCCA can not only obtain higher or equal and more stable correlation coefficients

than the competing methods, but also find out cleaner canonical loading patterns on both synthetic data and real imaging genetic data.

2 The Graph Guided Pairwise Group Lasso

Throughout this paper, we denote a vector as the boldface lowercase letter, and a matrix is denoted by a boldface uppercase one. The Euclidean norm of vector $\mathbf{u}$ is $\|\mathbf{u}\|$. Let $\mathbf{X} = [\mathbf{x}^1;\ldots;\mathbf{x}^n]^T \subseteq \mathbb{R}^p$ and $\mathbf{Y} = [\mathbf{y}^1;\ldots;\mathbf{y}^n]^T \subseteq \mathbb{R}^q$ be the SNP data and the QT data from the same participants.

We have known that the group lasso tends to extract a subset of the features. However, it depends on the priori knowledge and there is no overlap between groups. The graph guided fused lasso overcomes this limitation, but it requires the sign of the sample correlations to be defined in advance. This will introduce undesirable estimation bias [17]. In this paper, we introduce the graph guided pairwise group lasso penalty by taking advantage of both group lasso and graph guided fused lasso. The GGL penalty is defined as,

$$\Omega_{\mathrm{GGL}}(\mathbf{u}) = \sum_{(i,j)\in E} \sqrt{u_i^2 + u_j^2} \quad (1)$$

where E is the edge set of the graph where those highly correlated features are connected.

The GGL penalty has the following two merits. First, if there is no priori knowledge, every pairwise term will be included to encourage $|u_i| \approx |u_j|$ which is guaranteed by the pairwise ℓ_2-norm. This holds for both positively and negatively correlated features, which will be demonstrated later in Theorem 1. Second, if some priori knowledge such as the pathway information about genetic markers is provided, the whole penalty will be guided by the pathway information. This will encourage $|u_d| = |u_j|$ no matter whether they are positively or negatively correlated. Therefore, the two genetic markers have very high probability to be selected simultaneously. The same results hold for the imaging measurements if we have the brain connectivity pattern such as the human connectome.

3 Method

3.1 GGL-SCCA Model and Optimization

We then propose the GGL-SCCA model,

$$\min_{\mathbf{u},\mathbf{v}} -\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} \quad (2)$$

$$\text{s. t. } \|\mathbf{X}\mathbf{u}\|^2 \le 1, \|\mathbf{Y}\mathbf{v}\|^2 \le 1, \Omega_{\mathrm{GGL}}(\mathbf{u}) \le c_1, \Omega_{\mathrm{GGL}}(\mathbf{v}) \le c_2$$

where $\Omega_{\mathrm{GGL}}(\mathbf{u})$ and $\Omega_{\mathrm{GGL}}(\mathbf{v})$ are the GGL penalty to assure structure information. Of note, we use $\|\mathbf{Xu}\|^2 \leq 1$ instead of $\|\mathbf{u}\|^2 \leq 1$ to accommodate the in-set covariance $\mathbf{X}^T\mathbf{X}$ which can improve the model performance [6].

In order to solve this problem, we write the objective function of GGL-SCCA into matrix form using the Lagrange method,

$$\mathscr{L}(\mathbf{u},\mathbf{v}) = -\mathbf{u}^T\mathbf{X}^T\boldsymbol{Y}\mathbf{v} + \frac{\gamma_1}{2}\|\mathbf{Xu}\|^2 + \frac{\gamma_2}{2}\|\boldsymbol{Y}\mathbf{v}\|^2 + \lambda_1\Omega_{\mathrm{GGL}}(\mathbf{u}) + \lambda_2\Omega_{\mathrm{GGL}}(\mathbf{v}) \quad (3)$$

We approximate the objective function by a quadratic function. Obviously, the first term $\mathbf{u}^T\mathbf{X}^T\mathbf{Yv}$ is bilinear and biconvex in **u** and **v**. We then show the quadratic expression of the GGL term. Let $\mathbf{u}^t$ and $\mathbf{u}^{t+1}$ be the estimation at steps t and $t+1$ respectively, the first-order Taylor expansion of term $\sqrt{u_i^2+u_j^2}$ regarding $u_i^2+u_j^2$ is,

$$\begin{aligned}\sqrt{(u_i^{t+1})^2+(u_j^{t+1})^2} &\approx \sqrt{(u_i^t)^2+(u_j^t)^2} \\ &+\frac{1}{2\sqrt{(u_i^t)^2+(u_j^t)^2}}(((u_i^{t+1})^2+(u_j^{t+1})^2) - ((u_i^t)^2+(u_j^t)^2)) \\ &= \frac{(u_i^{t+1})^2}{2\sqrt{(u_i^t)^2+(u_j^t)^2}} + C\end{aligned} \quad (4)$$

where $C = \sqrt{(u_i^t)^2+(u_j^t)^2} + \frac{(u_j^{t+1})^2 - ((u_i^t)^2+(u_j^t)^2)}{2\sqrt{(u_i^t)^2+(u_j^t)^2}}$. From the point of view of optimization, the term C makes no contribution towards optimizing u_i.[3]

Then the GGL penalty can be simplified,

$$\Omega_{\mathrm{GGL}}(\mathbf{u}) \approx \sum_i\sum_j \frac{(u_i^{t+1})^2}{2\sqrt{(u_i^t)^2+(u_j^t)^2}} + C* \quad (5)$$

with C* being the sum of C across all pairwise penalty terms. Therefore, the GGL penalty is quadratically expressed.

Now the objective function conveys to a quadratic function, and there exists a closed-form solution. Since GGL-SCCA is biconvex in **u** and **v**, we take the derivative with respect to them respectively. The solution to the Eq. (3) satisfies,

[3]Each u_i can be solved with u_j's ($j \neq i$) fixed (i.e., we use u_j^t to approximate u_j^{t+1} in C), thus u_j's do not contribute to the optimization of u_i [9].

$$\mathbf{0} \in -\mathbf{X}^T\boldsymbol{Y}\mathbf{v}+(\lambda_1\mathbf{D}_1+\gamma_1\mathbf{X}^T\mathbf{X})\mathbf{u}, \quad (6)$$

$$\mathbf{0} \in -\boldsymbol{Y}^T\mathbf{X}\mathbf{u}+(\lambda_2\mathbf{D}_2+\gamma_2\boldsymbol{Y}^T\boldsymbol{Y})\mathbf{v}, \quad (7)$$

Algorithm 1 The GGL-SCCA Algorithm

Require:

$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$, $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T$

Ensure:

Canonical loadings **u** and **v**.

Initialize $\mathbf{u} \in \mathbb{R}^{p\times 1}$, $\mathbf{v} \in \mathbb{R}^{q\times 1}$;

while not convergence **do**

  Update the diagonal matrix $\mathbf{D}_1$ by taking derivative of Eq. (5);

  Solve **u** according to Eq. (8);

  Update the diagonal matrix $\mathbf{D}_2$ by taking derivative of Eq. (5);

  Solve **v** according to Eq. (9);

end while

Scale **u** so that $\|\mathbf{Xu}\|_2^2 = 1$, and **v** so that $\|\mathbf{Yv}\|_2^2 = 1$

where $\mathbf{D}_1$ can be deduced from the previous step's value of **u** according to Eq. (5). $\mathbf{D}_2$ can be computed similarly. Therefore, $\mathbf{D}_1$ is a diagonal matrix with the k_1-th element being $\sum_{i,i\neq k_1}\frac{1}{\sqrt{u_{k_1}^2+u_i^2}}(k_1 \in [1,p])$, and $\mathbf{D}_2$ is a diagonal matrix with the k_2-th element being $\sum_{j,j\neq k_2}\frac{1}{\sqrt{v_{k_2}^2+v_j^2}}(k_2 \in [1,q])$.[4]

Therefore, **u** and **v** have the closed-form updating expressions,

$$\mathbf{u}^{t+1}=(\lambda_1\mathbf{D}_1^t+\gamma_1\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{Y}\mathbf{v}^t, \quad (8)$$

[4]Note that an element of diagonal matrix $\mathbf{D}_1$ will nonexist if $\sqrt{u_i^2+u_{k_1}^2}=0$. We handle this issue by regularizing it as $\sqrt{u_i^2+u_{k_1}^2+\zeta}$ with ζ being a tiny positive number. Then the objective function regarding **u** becomes $\tilde{\mathscr{L}}(\mathbf{u})=\sum_{i=1}^{p}(-\mathbf{u}_i\mathbf{x}_i^T\boldsymbol{Y}\mathbf{v}+\lambda_1\sum_{k_1}\sqrt{u_i^2+u_{k_1}^2+\zeta}+\frac{\gamma_1}{2}\|\mathbf{x}_iu_i\|_2^2)$. We can prove that $\tilde{\mathscr{L}}(\mathbf{u})$ will reduce to the original problem (3) when ζ approaching zero. Likewise, $\sqrt{v_j^2+v_{k_2}^2}=0$ can be regularized by the same method.

$$\mathbf{v}^{t+1}=(\lambda_2\mathbf{D}_2^t+\gamma_2\boldsymbol{Y}^T\boldsymbol{Y})^{-1}\boldsymbol{Y}^T\mathbf{X}\mathbf{u}^{t+1}. \quad (9)$$

We have known that GGL-SCCA model is biconvex with respect to **u** and **v** respectively. Then the Alternate Convex Search (ACS) method which is designed to solve the biconvex problem can be employed [10]. According to the ACS method, we address our SCCA model via alternative optimization by updating **u** and **v** alternatively. The procedure of the GGL-SCCA is shown in Algorithm 1. In every iteration, **u** and **v** are updated in turn till the algorithm converges or reaches a predefined stopping condition.

3.2 The Grouping Effect

In structured learning, a method that can estimate equal or similar values for a group of variables is more desirable [19, 4]. This is called grouping effect and of great importance. We have the following theorem with respect to the grouping effects of the GGL-SCCA method.

Theorem 1. Given two views of data **X** and **Y**, and the tuned parameters (λ, 7). Let **u*** be the solution to our SCCA problem. For the sake of simplicity, we assume there are only two features, e.g. u_i and u_j, are connected on the graph. Let ρ_{ij} be their sample correlation. Then the optimal **u*** satisfies,

$$\begin{aligned} |u_i^*-u_j^*| &\le \frac{(1+\gamma_1)\sqrt{u_i^{*2}+u_j^{*2}}}{\lambda_1}\sqrt{2\sqrt{(1-\rho_{ij})}}, \\ \text{if}\rho_{ij}\ge 0, \left|u_i^*-u_j^*\right| &\le \frac{(1+\gamma_1)\sqrt{u_i^{*2}+u_j^{*2}}}{\lambda_1}\sqrt{2\sqrt{(1-\rho_{ij})}}, \text{ if } \rho_{ij}<0. \end{aligned} \quad (10)$$

Proof. (1) We first prove the inequations when $\rho_{ij} \ge 0$ indicating features being positively correlated. We have the following two equations,

$$\frac{\partial\mathscr{L}}{\partial u_i}=\lambda_1 D_{1,i}u_i^*+\gamma_1\mathbf{x}_i^T\mathbf{X}\mathbf{u}=\mathbf{x}_i^T\boldsymbol{Y}\mathbf{v}, \frac{\partial\mathscr{L}}{\partial u_j}=\lambda_1 D_{1,i}u_i^*+\gamma_1\mathbf{x}_i^T\mathbf{X}\mathbf{u}=\mathbf{x}_i^T\boldsymbol{Y}\mathbf{v}. \quad (11)$$

Given u_i and u_j are the only connected features, we have $D_{1,i}=D_{1,j}=\frac{1}{\sqrt{u_i^2+u_j^2}}$. Then we arrive at

$$\frac{\lambda_1}{\sqrt{u_i^{*2}+u_j^{*2}}}u_i^*=\mathbf{x}_i^T\boldsymbol{Y}\mathbf{v}-\gamma_1\mathbf{x}_i^T\mathbf{X}\mathbf{u}, \frac{\lambda_1}{\sqrt{u_i^{*2}+u_j^{*2}}}u_j^*=\mathbf{x}_j^T\boldsymbol{Y}\mathbf{v}-\gamma_1\mathbf{x}_j^T\mathbf{X}\mathbf{u}. \quad (12)$$

Subtracting these two equations, we have

$$\frac{\lambda_1}{\sqrt{u_i^{*2}+u_j^{*2}}}(u_i^* - u_j^*)=(\mathbf{x}_i - \mathbf{x}_j)^T(\boldsymbol{Y}\mathbf{v} - \gamma_1\mathbf{X}\mathbf{u}) \quad (13)$$

Taking ℓ_2-norm on both sides, we arrive at

$$\frac{\lambda_1}{\sqrt{u_i^{*2}+u_j^{*2}}}|u_i^*-u_j^*|=\|\mathbf{x}_i-\mathbf{x}_j\|\|\boldsymbol{Y}\mathbf{v}-\gamma_1\mathbf{X}\mathbf{u}\|=\|\mathbf{x}_i-\mathbf{x}_j\|\sqrt{\|\boldsymbol{Y}\mathbf{v}\| - 2\gamma_1\mathbf{u}^\top\mathbf{X}^\top\boldsymbol{Y}\mathbf{v}+\gamma^2\|\mathbf{X}\mathbf{u}\|} \quad (14)$$

Using $\|\mathbf{x}_i - \mathbf{x}_j\| - \sqrt{2(1-\rho_{ij})}$, $\|\mathbf{X}\mathbf{u}\| \leq 1$, $\|\mathbf{Y}\mathbf{v}\| \leq 1$ and $-\mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} \leq 1$, we obtain the upper bound

$$|u_i^* - u_j^*| \leq \frac{(1+\gamma_1)\sqrt{u_i^{*2}+u_j^{*2}}}{\lambda_1}\sqrt{2(1-\rho_{ij})}. \quad (15)$$

(2) If $\rho_{ij} < 0$, it is clear that sign(u_i) = −sign(u_j). By adding both equations in Eq. (12) instead of subtracting them, we finally arrive at,

$$|u_i^*+u_j^*| \leq \frac{(1+\gamma_1)\sqrt{u_i^{*2}+u_j^{*2}}}{\lambda_1}\sqrt{2(1+\rho_{ij})}. \quad (16)$$

Note that GGL-SCCA model is symmetric about **u** and **v**, we can obtain the same upper bound of grouping effect for canonical weights **v**.

The Theorem 1 provides a qualitative theoretical description of the bound for both differences and sums of the coefficients. The bound between two coefficients directly depends on their correlation. If $\rho_{ij} \geq 0$, a higher correlation between two variables makes sure a smaller difference between their estimated coefficients. If $\rho_{ij} < 0$, a smaller value assures a smaller sum between their coefficients. This implies that the two coefficients will be approximate in amplitude. Therefore, the GGL-SCCA is capable of capture structure information no matter whether those features are positively or negatively correlated.

3.3 The Complexity Analysis

In Algorithm 1, Steps 2-7 are repeated until convergence. In each iteration, Step 3 is easy to calculate as $\mathbf{D}_1$ can be computed via matrix manipulation to avoid time consuming loop.

This is the same case for Step 5. Step 4 and Step 6 are the key steps, and we compute them via solving a system of linear equations with quadratic complexity instead of computing the matrix inverse with cubic complexity. This can reduce the computation burden greatly. Step 8 is a rescale steps and very simple to calculate. Therefore, the algorithm runs fast and efficiently.

In this study, we terminate Algorithm 1 when either of the two conditions satisfies, $\max\{|\delta| \mid \delta \in (\mathbf{u}_{t+1} - \mathbf{u}_t)\} \leq \epsilon$ and $\max\{|\delta| \mid \delta \in (\mathbf{v}_{t+1} - \mathbf{v}_t)\} \leq \epsilon$, where ϵ is a desirable estimation error. We chose $\epsilon = 10^{-5}$ empirically from experiments in this paper.

4 Experimental Study

4.1 Experimental Setup

We compare GGL-SCCA with two structure-aware SCCA methods. The first one is the network guided fused lasso based SCCA (NS-SCCA) which takes the sample correlation signs into consideration [2]. The second method is the AGN-SCCA which uses the absolute value based GraphNet to penalize those correlated variables [4]. These two methods are different in both modeling and optimizing techniques, and is deemed to be among the best structured SCCA methods by now.

We tune the parameters based on the following considerations to reduce time consumption. (1) According to Theorem 1, $\lambda_{i=1,2}$ and $\gamma_{i=1,2}$ contribute to the grouping effect oppositely. (2) The grouping effect is more sensitive to $\lambda_{i=1,2}$ than to $\gamma_{i=1,2}$. Therefore, we fix $\gamma_{i=1,2}$ to a moderate constant, and let $\gamma_{i=1,2} = 10$ in this paper. Finally, we have only two parameters $\lambda_{i=1,2}$ to be tuned and optimally tune them via a grid search from a moderate range 10^{-2} to 10^2 through nested five-fold cross-validation to make sure efficiency. The parameters that generate the highest correlation coefficients are used.

4.2 Results on Simulation Data

Four different data sets with different properties are generated in this study. We also set the number of observations be smaller than the number of features to simulate a large p small n problem. The details of the data sets are as follows. Firstly, $\mathbf{u}$ and $\mathbf{v}$ are generated according to the predefined structure. Secondly, a latent variable $\mathbf{z} \sim N(0, \mathbf{I}_{n\times n})$ is generated. And thirdly, $\mathbf{X}$ is created by $\mathbf{x}_i \sim N(Z_i\mathbf{u}, \Sigma_x)$, where $(\Sigma_x)_{jk} = \exp^{-|u_j - u_k|}$. Similarly, $\mathbf{Y}$ with the entry: $\mathbf{y}_i \sim N(Z_i\mathbf{v}, \Sigma_y)$, where $(\Sigma_y)_{jk} = \exp^{-|v_j - v_k|}$ is created. During this procedure, the true signals and the correlation strengths of the data are all distinct to assure diversity. This setup can make a thorough comparison.

We apply GGL-SCCA, NS-SCCA and AGN-SCCA to all four data sets. The true and estimated canonical loadings $\mathbf{u}$ and $\mathbf{v}$ are shown in Fig. 1. We observe that both GGL-SCCA and AGN-SCCA identify similar canonical loading profiles that are consistent to the ground truth across all data sets. NS-SCCA produces too many signals which are not so perfect to the ground truth. In addition, we also show the estimated correlation coefficients on both the training and testing sets calculated using the trained SCCA models in Table 1 (Left). The results show that GGL-SCCA obtains highest scores on both training and testing sets. Its testing result is only inferior to the NS-SCCA on the second data. The results implies that

GGL-SCCA has better training performance and generalization ability than those benchmarks. The area under ROC (AUC) shown in Table 1 (Right) indicates the sensitivity and specificity. It reveals that GGL-SCCA outperforms the competing methods as it holds the highest values for the most times. In summary, the simulation results demonstrate that GGL-SCCA could identify not only stronger testing associations but also more better signals on these diversified data sets.

4.3 Results on Real Neuroimaging Genetics Data

The real imaging genetics data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public private partnership, led by Principal Investigator Michael W. Weiner, MD. One primary goal of ADNI is 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 Alzheimers disease (AD). For up-to-date information, please refer to www.adni-info.org.

We use the genotyping and baseline amyloid imaging data (preprocessed [11C] Florbetapir PET scans) contributed by 567 non-Hispanic Caucasian participants. The amyloid imaging data used in this study are downloaded from LONI (adni.loni.usc.edu). Preprocessing is conducted to format this imaging data, and we finally generate 191 ROI level mean amyloid measurements in which the ROIs are defined by the MarsBaR AAL atlas [4]. The genotyping data includes 58 candidate SNP markers from the AD-related genes, such as the *APOE* gene. The aim is to evaluate the associations between the SNP data and the amyloid data, as well as which SNPs and amyloid measurements are correlated in this AD cohort.

All three SCCA methods are performed on the real neuroimaging genetics data. Shown in Fig. 2 are the canonical loadings obtained from the training data, where the relevant imaging measurements and genetic markers are exhibited. It is clear that GGL-SCCA identifies two relevant ROIs and one SNPs for easy interpretation due to the novel GGL penalty. The two strongest imaging measurements come from the right frontal region, which are positively correlated with SNP rs429358, a confirmed AD related allele in *APOE* e4. The AGN-SCCA identifies similar results to our method, which however has many interfering signals for the genetic markers. The NS-SCCA finds out too many imaging signals that are very hard to interpret. To give a clear view, we map the canonical loadings regarding the imaging measurements of GGL-SCCA onto the brain. Fig. 3 clearly shows that our method only highlights a small region of the whole brain. Moreover, we present the training and testing correlations in Table 3. GGL-SCCA obtains the highest values on both training set and testing set. Although AGN-SCCA has the same *mean* on training data, its *standard deviation* is larger than GGL-SCCA. Moreover, GGL-SCCA obtains better testing results than both competing methods. This implies that GGL-SCCA is more stable and has better generalization ability than AGN-SCCA and NS-SCCA. The results on this real data demonstrate that GGL-SCCA has better bi-multivariate identification ability than the benchmark methods. The strong association between the frontal morphometry and the

APOE in AD cohort, indicating GGL-SCCA's promising and potential power in identifying biologically meaningful imaging genetic associations.

5 Conclusions

We have proposed a novel graph guided pairwise group lasso (GGL) based SC-CA method (GGL-SCCA) to identify associations between brain imaging measurements and genetic factors. The existing group lasso based methods [1, 6] were dependent on the priori knowledge which was not always available. The graph/netwrok guided fused lasso based approaches [3, 6, 2, 5, 4] only focus on the positively correlated variables, or depended on the signs of the sample correlation which were sensitive to the partition of the data. Our SCCA method combines the merits of group lasso and the graph/network guided fused lasso, which is independent to not only the signs of the sample correlation, but also the priori knowledge. Moreover, our method can also incorporate the priori knowledge to recover specific structures.

We have compared GGL-SCCA with two state-of-the-art structured SCCA methods on both synthetic data and real imaging genetic data. The results on the synthetic data show that GGL-SCCA performs better than both NS-SCCA and AGN-SCCA across all data sets. The results on real data show that, GGL-SCCA not only reports better canonical correlation values than the competing methods, but also obtains more accurate and cleaner canonical loading patterns. GGL-SCCA finds out a strong associations between the superior frontal morphometry and the *APOE* e4 SNP, revealing its power in brain imaging genetics. In this paper, we merely use the graph guided pairwise group lasso penalty to induce structured sparsity. In the future work, we will incorporate lasso into the model to assure additional sparsity, and incorporate the priori knowledge to evaluate the performance of GGL-SCCA.

Acknowledgments

This work was supported by NSFC under Grant 61602384, and the Fundamental Research Funds for the Central Universities under Grant 3102016OQD0065. This work was also supported by NIH R01 EB022574, R01 LM011360, U01 AG024904, P30 AG10133, R01 AG19771, UL1 TR001108, R01 AG 042437, R01 AG046171, and R01 AG040770, by DoD W81XWH-14-2-0151, W81XWH-13-1-0259, W81XWH-12-2-0012, and NCAA 14132004.

References

1. Chen J, Bushman FD, Lewis JD, Wu GD, Li H. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. Biostatistics. 2013; 14(2):244–258. [PubMed: 23074263]
2. Chen X, Liu H. An efficient optimization algorithm for structured sparse cca, with applications to eqtl mapping. Statistics in Biosciences. 2012; 4(1):3–26.
3. Chen X, Liu H, Carbonell JG. Structured sparse canonical correlation analysis. AISTATS. 2012
4. Du L, Huang H, Yan J, Kim S, Risacher SL, Inlow M, Moore JH, Saykin AJ, Shen L. Structured sparse canonical correlation analysis for brain imaging genetics: An improved GraphNet method. Bioinformatics. 2016; 32(10):1544–1551. [PubMed: 26801960]
5. Du L, Huang H, Yan J, Kim S, Risacher SL, Inlow M, Moore JH, Saykin AJ, Shen L. Structured sparse cca for brain imaging genetics via graph oscar. BMC Systems Biology. 2016:335–345.

6. Du L, Yan J, Kim S, Risacher SL, Huang H, Inlow M, Moore JH, Saykin AJ, Shen L. A novel structure-aware sparse learning algorithm for brain imaging genetics. MICCAI. 2014:329–336. [PubMed: 25320816]
7. Du, L., Yan, J., Kim, S., Risacher, SL., Huang, H., Inlow, M., Moore, JH., Saykin, AJ., Shen, L., et al. BIH. Springer; 2015. GN-SCCA: GraphNet based sparse canonical correlation analysis for brain imaging genetics; p. 275-284.
8. Du, L., Zhang, T., Liu, K., Yao, X., Yan, J., Risacher, SL., Guo, L., Saykin, AJ., Shen, L. BIBM. IEEE Computer Society; 2016. Sparse canonical correlation analysis via truncated ℓ_1-norm with application to brain imaging genetics; p. 707-711.
9. Friedman JH, Hastie T, Hofling H, Tibshirani R. Pathwise coordinate optimization. The Annals of Applied Statistics. 2007; 1(2):302–332.
10. Gorski J, Pfeuffer F, Klamroth K. Biconvex sets and optimization with biconvex functions: a survey and extensions. Mathematical Methods of Operations Research. 2007; 66(3):373–407.
11. Kim S, Swaminathan S, Inlow M, Risacher SL, Nho K, Shen L, Foroud TM, Petersen RC, Aisen PS, Soares H, et al. Influence of genetic variation on plasma protein levels in older adults using a multi-analyte panel. PLoS One. 2013; 8(7):e70269. [PubMed: 23894628]
12. Potkin SG, Turner JA, Guffanti G, Lakatos A, Torri F, Keator DB, Macciardi F. Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: methodological considerations. Cognitive neuropsychiatry. 2009; 14(4-5):391–418. [PubMed: 19634037]
13. Saykin AJ, Shen L, Yao X, Kim S, Nho K, Risacher SL, Ramanan VK, Foroud TM, Faber KM, Sarwar N, et al. Genetic studies of quantitative mci and ad phenotypes in adni: Progress, opportunities, and plans. Alzheimer's & Dementia. 2015; 11(7):792–814.
14. 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. 2010; 53(3):1051–63. [PubMed: 20100581]
15. Shen L, Thompson PM, Potkin SG, Bertram L, Farrer LA, Foroud TM, Green RC, Hu X, Huentelman MJ, Kim S, et al. Genetic analysis of quantitative phenotypes in ad and mci: imaging, cognition and biomarkers. Brain imaging and behavior. 2014; 8(2):183–207. [PubMed: 24092460]
16. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009; 10(3):515–34. [PubMed: 19377034]
17. Yang, S., Yuan, L., Lai, YC., Shen, X., Wonka, P., Ye, J. KDD. ACM; 2012. Feature grouping and selection over an undirected graph; p. 922-930.
18. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006; 68(1):49–67.
19. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005; 67(2):301–320.

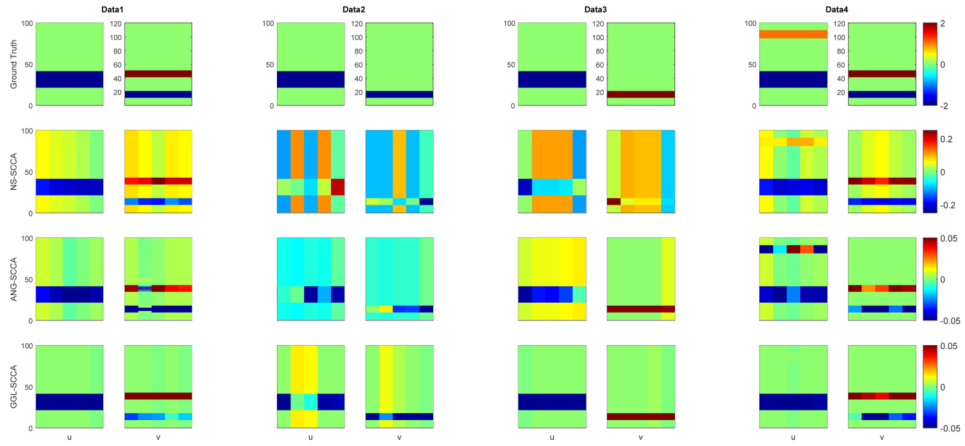

Fig. 1.

Canonical loadings estimated on synthetic data. The first row is the ground truth, and each remaining row corresponds to a SCCA method: (1) NS-SCCA, (2) AGN-SCCA, and (3) GGL-SCCA. For each method, the estimated weights of **u** are shown on the left panel, and those of **v** are shown on the right.

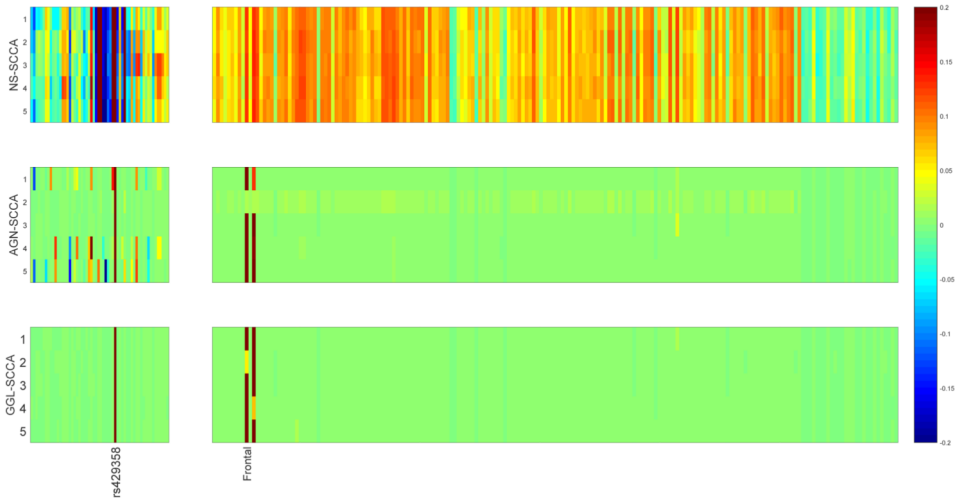

Fig. 2.

Canonical loadings estimated on real imaging genetics data set. Each row corresponds to a SCCA method: (1) NS-SCCA, (2) AGN-SCCA, and (3) GGL-SCCA. For each method, the estimated weights of **u** are shown on the left panel, and those of **v** are shown on the right.

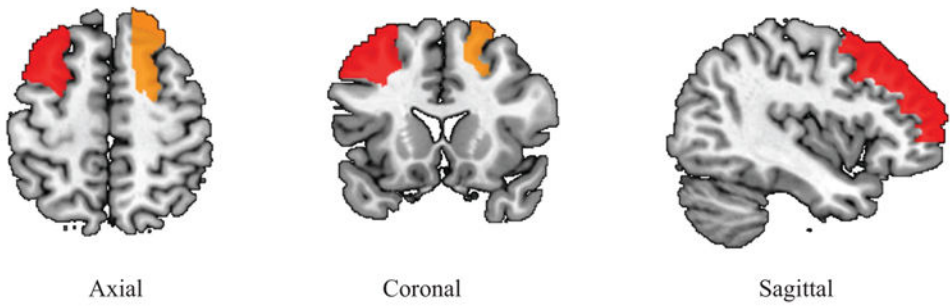

Fig. 3.
Mapping averaged canonical loading **v** of GGL-SCCA onto the brain.

Table 1

Performance comparison on synthetic data. Training and testing correlation coefficients (mean±std) of 5-fold cross-validation are shown for NS-SCCA, AGN-SCCA and GGL-SCCA. The best value are shown in boldface. The AUC (area under the curve) values (mean±std) of canonical loadings are also shown.

	Training Correlation Coefficients			**Area under ROC (AUC): u**		
Data set	**NS-SCCA**	**AGN-SCCA**	**GGL-SCCA**	**NS-SCCA**	**AGN-SCCA**	**GGL-SCCA**
data1	0.39±0.07	0.53±0.10	**0.60±0.07**	1.00±0.00	1.00±0.00	1.00±0.00
data2	0.31±0.08	0.35±0.08	**0.48±0.08**	0.20±0.45	0.60±0.55	0.60±0.55
data3	0.20±0.07	0.29±0.07	**0.40±0.07**	0.20±0.45	0.80±0.45	1.00±0.00
data4	0.44±0.08	0.44±0.07	**0.50±0.05**	1.00±0.00	1.00±0.00	0.93±0.15
	Testing Correlation Coefficients			Area under ROC (AUC): v		
data1	0.42±0.10	0.60±0.10	**0.62±0.23**	1.00±0.00	0.96±0.09	1.00±0.00
data2	**0.25±0.18**	0.21±0.14	0.22±0.08	0.20±0.45	0.80±0.45	1.00±0.00
data3	0.28±0.19	0.33±0.24	**0.43±0.21**	0.20±0.45	1.00±0.00	1.00±0.00
data4	0.25±0.10	0.32±0.24	**0.44±0.24**	1.00±0.00	1.00±0.00	1.00±0.00

Table 2

Participant characteristics.

	HC	MCI	AD
Num	196	343	28
Gender(M/F)	102/94	203/140	18/10
Handedness(R/L)	178/18	309/34	23/5
Age (mean±std)	74.77±5.39	71.92±7.47	75.23±10.66
Education (mean±std)	15.61±2.74	15.99±2.75	15.61±2.74

Table 3

Performance comparison on real data. Training and testing correlation coefficients (each fold and mean±std) of 5-fold cross-validation are shown for NS-SCCA, AGN-SCCA and GGL-SCCA. The best mean±std is shown in boldface.

Method	Training Results					mean±std	Testing Results					mean±std
NS-SCCA	0.41	0.40	0.43	0.39	0.41	0.41±0.01	0.37	0.41	0.23	0.43	0.37	0.36±0.08
AGN-SCCA	0.49	0.43	0.52	0.49	0.51	0.49±0.03	0.48	0.46	0.33	0.55	0.43	0.45±0.08
GGL-SCCA	0.48	0.48	0.52	0.46	0.49	**0.49±0.02**	0.51	0.45	0.34	0.55	0.47	**0.46±0.08**