HHS Public Access

Author manuscript

Med Image Anal. Author manuscript; available in PMC 2021 April 01.

Published in final edited form as:

Med Image Anal. 2020 April ; 61: 101656. doi:10.1016/j.media.2020.101656.

Detecting genetic associations with brain imaging phenotypes in Alzheimer's disease via a novel structured SCCA approach

Lei Du[a,*], **Kefei Liu**[b], **Xiaohui Yao**[b], **Shannon L. Risacher**[c], **Junwei Han**[a], **Andrew J. Saykin**[c], **Lei Guo**[a], **Li Shen**[b,*], **Alzheimer's Disease Neuroimaging Initiative**[**]

[a]School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

[b]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA

[c]Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN 46202, USA

Abstract

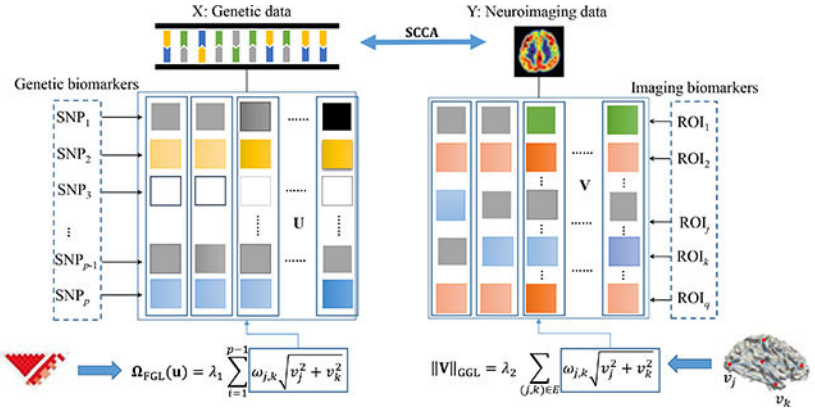
Brain imaging genetics becomes an important research topic since it can reveal complex associations between genetic factors and the structures or functions of the human brain. Sparse canonical correlation analysis (SCCA) is a popular bi-multivariate association identification method. To mine the complex genetic basis of brain imaging phenotypes, there arise many SCCA methods with a variety of norms for incorporating different structures of interest. They often use the group lasso penalty, the fused lasso or the graph/network guided fused lasso ones. However, the group lasso methods have limited capability because of the incomplete or unavailable prior knowledge in real applications. The fused lasso and graph/network guided methods are sensitive to the sign of the sample correlation which may be incorrectly estimated. In this paper, we introduce two new penalties to improve the fused lasso and the graph/network guided lasso penalties in structured sparse learning. We impose both penalties to the SCCA model and propose an optimization algorithm to solve it. The proposed SCCA method has a strong upper bound of grouping effect for both positively and negatively highly correlated variables. We show that, on both synthetic and real neuroimaging genetics data, the proposed SCCA method performs better than or equally to the conventional methods using fused lasso or graph/network guided fused lasso. In particular, the proposed method identifies higher canonical correlation coefficients and captures clearer canonical weight patterns, demonstrating its promising capability in revealing biologically meaningful imaging genetic associations.

[**]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.

[*]Corresponding to: Lei Du (dulei323@gmail.com) and Li Shen (Li.Shen@pennmedicine.upenn.edu).

Declarations of interest: none.

Graphical Abstract



Keywords

Brain imaging genetics; sparse canonical correlation analysis (SCCA); fused pairwise group Lasso; graph guided pairwise group Lasso

1. Introduction

Recently, brain imaging genetics becomes more and more popular in biomedical and bioinformatics studies. Brain imaging genetics aims to detect genetic associations with brain imaging phenotypes, and further to uncover how genetic factors influence the structure or function of the human brain using imaging measurements as the quantitative endophenotype (Potkin et al., 2009b; Vounou et al., 2010; Kim et al., 2013; Saykin et al., 2015). The genetic factors, such as the single nucleotide polymorphisms (SNPs), and imaging quantitative traits (QTs) are all multivariate. Therefore, identifying complex bi-multivariate associations that cannot be achieved by univariate methods is an important task in brain imaging genetics.

Sparse canonical correlation analysis (SCCA) gains wide attention in brain imaging genetics for its powerful capability in bi-multivariate association identification and feature selection. There are many SCCA methods depending on different type of sparsity-inducing techniques. The ℓ_1-norm penalty is among the most popular ones; however, it only pursuits individual feature level sparsity (Witten et al., 2009; Witten and Tibshirani, 2009; Parkhomenko et al., 2009; Hardoon and Shawe-Taylor, 2011; Chi et al., 2013). The biomarkers usually function jointly other than individually (Shen et al., 2010) in biomedical studies. For example, correlations usually exists between SNPs in a linkage disequilibrium (LD) block in the genome, and also among voxels in a region of interest (ROI) in the brain. Therefore, detecting the structural sparsity, such as the group level sparsity or the graph/network level sparsity, is of great interest and importance in brain imaging genetics (Shen et al., 2010, 2014).

To accommodate the structural sparsity, several structured SCCA methods have been proposed. They can be roughly classified into two kinds based on their different penalties (Du et al., 2016). The first kind of SCCA methods use the group lasso penalty, which is an

intra-group ℓ_2-norm and inter-group ℓ_1-norm (Silver et al., 2012; Chen et al., 2012; Chen and Liu, 2012; Chen et al., 2013; Lin et al., 2014; Du et al., 2014; Yan et al., 2014; Du et al., 2018, 2019). The group lasso tends to perform variable selection at the group level, and each group will be shrunk to zero or not as a whole (Yuan and Lin, 2006). To our knowledge, these SCCA methods require the group structure to be provided in advance, which limits their applications as the precise prior knowledge is hard to obtain in real biomedical studies (Du et al., 2016). The second kind of SCCA methods recover the structure information via the graph or network guided penalty (Du et al., 2016; Chen et al., 2012; Chen and Liu, 2012; Chen et al., 2013; Yan et al., 2014; Du et al., 2017). They are more flexible than the previous type since they can either use any available prior knowledge to recover the specific structure, or operate in a structure pursuing mode (Du et al., 2016; Chen et al., 2012). There are three types of graph guided penalties: (1) the graph guided fused lasso penalty and its variants (Du et al., 2016; Chen et al., 2012; Chen and Liu, 2012; Chen et al., 2013), (2) the sample correlation sign based graph guided fused ℓ_2-norm penalty (Chen and Liu, 2012; Yan et al., 2014), and (3) the graph guided absolute fused ℓ_2-norm penalty (Du et al., 2016). Du et al. (2016) has shown that the first two types of graph guided penalties can introduce estimation bias since the sign of the sample correlations may be incorrectly calculated. The reason could be that the sign can be easily swapped when removing a fraction of the data or perturbing the data as in bootstrap or sub-sampling. The third type of SCCA methods impose ℓ_2-norm on the structure penalty terms, which might not produce desirable structural sparsity (Du et al., 2017).

Inspired by the success of group lasso, we consider a case where each group consists of only two variables. Both variables will be simultaneously shrunk to zero or not with equal or similar weights. This motivates us to introduce two novel penalties, i.e. the fused pairwise group lasso (FGL) and the graph guided pairwise group lasso (GGL). The FGL imposes pairwise group lasso onto adjacent variables to introduce a chain of smoothness, and the GGL imposes pairwise group lasso terms guided by an undirect graph. The FGL encourages adjacent smoothness, and thus can identify successively highly correlated signals even though their signs are opposite. The GGL is more powerful than those conventional graphical lasso based methods as it is sample correlation sign independent too. Both FGL and GGL can be used in the data-driven mode where no prior knowledge is given, while FGL does assume that the genetic data has a sequential structure. Besides, GGL bridges the gap between the group lasso and graph guided penalties. As stated earlier, there usually exists a chain relationship across SNPs and a graphical relationship among voxels. To better solve these brain imaging genetic problems, we here propose a novel SCCA model (FGL-SCCA) which imposes the FGL penalty onto the genetic markers and GGL penalty onto the imaging measurements. FGL-SCCA intends to recover the adjacent and graphical smoothness and structure information automatically. It is sample correlation sign independent, which means it can assign equal or similar weights for those correlated variables no matter whether they are positively or negatively correlated. Thus FGL-SCCA is more robust than those existing SCCA methods using fused lasso and graph guided penalties. Meanwhile, we propose an efficient optimization algorithm to solve the FGL-SCCA problem. We also provide a quantitative upper bound for the grouping effect of FGL-SCCA to demonstrate its structure identification capability. Compared with three state-of-

the-art SCCA methods FL-SCCA (Witten and Tibshirani, 2009), NS-SCCA (Chen and Liu, 2012) and AGN-SCCA (Du et al., 2016), FGL-SCCA obtains higher or equal and more stable correlation coefficients on both synthetic data and real imaging genetic data from an Alzheimer's disease (AD) cohort. Besides, our method also identifies cleaner and sparser canonical weights than those benchmarks, and thus has the potential to provide an easier interpretation to guide subsequent analysis.

2. Methods

In this paper, we denote scalars as italic letters, column vectors as boldface lowercase letters, and matrices as boldface capitals. The Euclidean norm of a vector $\mathbf{u}$ is denoted as $\|\mathbf{u}\|$. $\mathbf{X} \in \mathscr{R}^{n \times p}$ is a matrix representing the SNP data, where n is the number of participants and p is the number of SNPs. $\mathbf{Y} \in \mathscr{R}^{n \times q}$ is the matrix of QT data with q being the number of imaging measurements.

2.1. The Fused Pairwise Group Lasso (FGL)

To recover the fused associations from the genetic data, we define the FGL penalty as follows

$$\Omega_{\mathrm{FGL}}(\mathbf{u}) = \lambda_1 \sum_{i=1}^{p-1} w_{i,i+1} \sqrt{u_i^2 + u_{i+1}^2}, \tag{1}$$

where $w_{i,i+1}$ is the weight for two adjacent variables, and λ_1 is a positive tuning parameter.

The FGL absorbs the advantages of both group lasso and fused lasso. Thus its merits are twofold. On one hand, the pairwise group lasso constraint introduces a chain of smoothness across all elements of the vector $\mathbf{u}$. This makes two adjacent as well as strongly associated variables being equal or similar with respect to their estimated weights. On the other hand, owing to the ℓ_2-norm, the FGL penalty is sample correlation sign independent. Therefore, it can extract signals that the fused lasso cannot, e.g. two adjacent features with negative correlation. We will demonstrate this later in Theorem 2.

2.2. The Graph Guided Pairwise Group Lasso (GGL)

Though the FGL could mine structure information, the smoothness is only imposed on adjacent variables. We sometimes are more interested in the network or graph structure hidden in the data. As discussed earlier, both functional and structural mechanisms of the human brain show a network structure rather than a group structure. Therefore, imposing the group-like constraint such as the group lasso or FGL might not be the best option. On this account, we extend the FGL to the graphical mode. Mapping the feature space in terms of $\mathbf{v}$ onto an undirected graph $\boldsymbol{G}$, we define the graph guided pairwise group lasso (GGL) as

$$\Omega_{\mathrm{GGL}}(\mathbf{v}) = \lambda_2 \sum_{(j,k) \in E} \omega_{j,k} \sqrt{v_j^2 + v_k^2}, \tag{2}$$

where E is the edge set guided by the graph G, and $w_{j,k}$ is the edge weight. λ_2 is a positive tuning parameter to control the amount of regularization.

The GGL penalty takes the advantage of both group lasso and graphical lasso. First, if there is no prior knowledge, every pairwise term will be included to encourage $|v_j| \approx |v_k|$ which is guaranteed by the pairwise group lasso. This holds for both positively and negatively correlated features, which is supported by Theorem 2. Second, if the connectivity information, e.g. such as the human connectome, is available, the constraint will be guided by the connectivity information. This will encourage $|v_j| = |v_k|$ if imaging markers j and k are in the same biological sub-network no matter whether they are positively or negatively correlated. Therefore, both imaging markers have a high probability to be selected.

2.3. The FGL-SCCA Model

Let both $\mathbf{X}$ and $\mathbf{Y}$ be centered and normalized, we impose FGL on the genetics data and GGL on the brain imaging data, and define the FGL-SCCA model as

$$\min_{\mathbf{u},\mathbf{v}} -\mathbf{u}^\top\mathbf{X}^\top\mathbf{Y}\mathbf{v} + \Omega_{\mathrm{FGL}}(\mathbf{u}) + \Omega_{\mathrm{GGL}}(\mathbf{v}) \; s.t. \; \left\|\mathbf{X}\mathbf{u}\right\|^2 \le 1, \left\|\mathbf{Y}\mathbf{v}\right\|^2 \le 1, \tag{3}$$

where $\mathbf{u}$ and $\mathbf{v}$ are called canonical loadings or canonical weights; $\Omega_{\mathrm{FGL}}(\mathbf{u})$ is the newly introduced FGL penalty to induce adjacent smoothness, and $\Omega_{\mathrm{GGL}}(\mathbf{v})$ is used to induce graphical smoothness. We do not artificially assume the in-set covariance matrices $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{Y}^\top\mathbf{Y}$ to be identity so that the auto-covariance information could be preserved in the proposed model. (Du et al., 2014).

The Lagrangian associated with the problem writes

$$\mathscr{L}(\mathbf{u},\mathbf{v}) = -\mathbf{u}^\top\mathbf{X}^\top\mathbf{Y}\mathbf{v} + \Omega_{\mathrm{FGL}}(\mathbf{u}) + \Omega_{\mathrm{GGL}}(\mathbf{v}) + \frac{\gamma_1}{2}(\|\mathbf{X}\mathbf{u}\|^2 - 1) + \frac{\gamma_2}{2}(\|\mathbf{Y}\mathbf{v}\|^2 - 1), \tag{4}$$

with γ_1 and γ_2 are positive tuning parameters. The solution is attained when the KKT conditions are satisfied. Thus the main difficulty in solving (4) becomes how to deduce the KKT conditions. This involves calculating the partial derivative of $\mathscr{L}(\mathbf{u},\mathbf{v})$ with respect to $\mathbf{u}$ and $\mathbf{v}$, especially the derivatives of the FGL and GGL penalty functions, which are very complicated.

2.4. Smoothing the penalties

Suppose $f(\mu,\upsilon) = \sqrt{\mu+\upsilon}$ with both μ and υ being non-negative variables, we have the following equation according to the Taylor's theorem

$$\begin{aligned} f(\mu,\upsilon) = \sqrt{\mu+\upsilon} = \sqrt{\mu^{(t)}+\upsilon^{(t)}} + f'_\mu(\mu^{(t)},\upsilon^{(t)})(\mu-\mu^{(t)}) \\ + f'_\upsilon(\mu^{(t)},\upsilon^{(t)})(\upsilon-\upsilon^{(t)}) + R(\mu) + R(\upsilon), \end{aligned} \tag{5}$$

where $(\mu^{(t)},\upsilon^{(t)})$ is the neighbour of (μ,υ), $f'_\mu(\mu^{(t)},\upsilon^{(t)}) = 1/(2\sqrt{\mu^{(t)}+\upsilon^{(t)}})$ is the gradient[1] of $f(\mu,\upsilon)$ with respect to μ at $(\mu^{(t)},\upsilon^{(t)})$, and $f'_\upsilon(\mu^{(t)},\upsilon^{(t)}) = 1/(2\sqrt{\mu^{(t)}+\upsilon^{(t)}})$ is that with respect to υ at $(\mu^{(t)},\upsilon^{(t)})$. $R(\mu)$ and $R(\upsilon)$ are the remainder terms. We then define the function $g(\mu,\upsilon)$ by dropping the remainders

$$\begin{aligned} g(\mu, \upsilon) &= \sqrt{\mu^{(t)} + \upsilon^{(t)}} + f'_{\mu}(\mu^{(t)}, \upsilon^{(t)})(\mu - \mu^{(t)}) + f'_{\upsilon}(\mu^{(t)}, \upsilon^{(t)})(\upsilon - \upsilon^{(t)}) \\ &= \sqrt{\mu^{(t)} + \upsilon^{(t)}} + \frac{1}{2\sqrt{\mu^{(t)} + \upsilon^{(t)}}}(\mu - \mu^{(t)}) + \frac{1}{2\sqrt{\mu^{(t)} + \upsilon^{(t)}}}(\upsilon - \upsilon^{(t)}) \\ &= \frac{\mu + \upsilon}{2\sqrt{\mu^{(t)} + \upsilon^{(t)}}} + \frac{\sqrt{\mu^{(t)} + \upsilon^{(t)}}}{2}. \end{aligned} \tag{6}$$

Obviously, $g(\mu,\upsilon)$ is an affine function of μ and υ. Thus it is continuous and differentiable, and we have the following proposition.

Proposition 1—*Given functions $f(\mu, \upsilon) = \sqrt{\mu + \upsilon}$, $g(\mu,\upsilon)$ with the form of Eq. (6), and $(\mu^{(t)},\upsilon^{(t)})$ is the neighbour of (μ,υ), then the following three rules hold.*

1. *$f(\mu,\upsilon)$ and $g(\mu,\upsilon)$ are equal at $(\mu^{(t)},\upsilon^{(t)})$, i.e. $f(\mu^{(t)},\upsilon^{(t)}) = g(\mu^{(t)},\upsilon^{(t)})$;*
2. *$f(\mu,\upsilon)$ and $g(\mu,\upsilon)$ have the same partial derivatives at $(\mu^{(t)},\upsilon^{(t)})$, i.e. $f'_{\mu}(\mu^{(t)}, \upsilon^{(t)}) = g'_{\mu}(\mu^{(t)}, \upsilon^{(t)})$ and $f'_{\upsilon}(\mu^{(t)}, \upsilon^{(t)}) = g'_{\upsilon}(\mu^{(t)}, \upsilon^{(t)})$;*
3. *$g(\mu,\upsilon)$ is an upper bound of $f(\mu,\upsilon)$, i.e. $f(\mu,\upsilon) \le g(\mu,\upsilon)$.*

Proof. The first and the second rules are obvious. Thus we put emphases on the third rule. Note that $f'_{\mu}(\mu^{(t)}, \upsilon^{(t)}) = f'_{\upsilon}(\mu^{(t)}, \upsilon^{(t)}) = 1/(2\sqrt{\mu^{(t)} + \upsilon^{(t)}})$, the difference between $f(\mu,\upsilon)$ and $g(\mu,\upsilon)$ is

$$\begin{aligned} g(\mu, \upsilon) - f(\mu, \upsilon) &= \frac{\mu + \upsilon}{2\sqrt{\mu^{(t)} + \upsilon^{(t)}}} + \frac{\sqrt{\mu^{(t)} + \upsilon^{(t)}}}{2} - \sqrt{\mu + \upsilon} \\ &= \frac{1}{2\sqrt{\mu^{(t)} + \upsilon^{(t)}}}\left(\sqrt{\mu^{(t)} + \upsilon^{(t)}} - \sqrt{\mu + \upsilon}\right)^2 \ge 0. \end{aligned} \tag{7}$$

This yields $f(\mu,\upsilon) \le g(\mu,\upsilon)$, which completes the proof. □

Substituting $\mu = u_i^2$ and $\upsilon = u_{i+1}^2$ into Eq. (6), we obtain

$$g(u_i^2, u_{i+1}^2) = \frac{u_i^2 + u_{i+1}^2}{2\sqrt{(u_i^{(t)})^2 + (u_{i+1}^{(t)})^2}} + \frac{\sqrt{(u_i^{(t)})^2 + (u_{i+1}^{(t)})^2}}{2} \tag{8}$$

where $u_i^{(t)}$ and $u_{i+1}^{(t)}$ are respectively the estimates of u_i and u_{i+1} in the t-th iteration in an optimizing procedure. Based on Poposition 1, $g(u_i^2, u_{i+1}^2)$ is a quadratic approximation to $f(u_i^2, u_{i+1}^2)$ at $(u_i^{(t)}, u_{i+1}^{(t)})$, and moreover it is an upper bound on $f(u_i^2, u_{i+1}^2)$. Thus embedding them into convex loss functions will lead to the same solution path. We then use $g(u_i^2, u_{i+1}^2)$

[1]Note that the gradient $f'_{\mu}(\mu^{(t)}, \upsilon^{(t)}) = 1/(2\sqrt{\mu^{(t)} + \upsilon^{(t)}})$ will not exist if $\sqrt{\mu^{(t)} + \upsilon^{(t)}} = 0$. We handle this issue by using $\sqrt{\mu^{(t)} + \upsilon^{(t)} + \zeta}$ for regularization, where ζ is a tiny positive number. It is easy to verify that $f'_{\mu}(\mu^{(t)}, \upsilon^{(t)}) = 1/(2\sqrt{\mu^{(t)} + \upsilon^{(t)} + \zeta})$ is the sub-gradient and thus inherits the same properties to the gradient in optimizing problems when $\zeta \to 0$.

as a surrogate of $f(u_i^2, u_{i+1}^2)$ in the remainder of this paper. Specifically, we have the surrogate function of the FGL penalty

$$\begin{aligned}\Omega_{\mathrm{FGL}}^{\mathrm{App}}(\mathbf{u}) &= \lambda_1 \sum_{i=1}^{p-1} w_{i,i+1} \cdot g(u_i^2, u_{i+1}^2) \\ &= \lambda_1 \sum_{i=1}^{p-1} w_{i,i+1} \cdot \left[\frac{u_i^2 + u_{i+1}^2}{2\sqrt{(u_i^{(t)})^2 + (u_{i+1}^{(t)})^2}} + \frac{\sqrt{(u_i^{(t)})^2 + (u_{i+1}^{(t)})^2}}{2} \right].\end{aligned} \tag{9}$$

Since GGL has a similar form to FGL, it is easy to obtain the surrogate function of the GGL penalty with respect to **v**, i.e.

$$\begin{aligned}\Omega_{\mathrm{GGL}}^{\mathrm{App}}(\mathbf{v}) &= \lambda_2 \sum_{(j,k)\in E} \omega_{j,k} \cdot g(v_j^2, v_k^2) \\ &= \lambda_2 \sum_{(j,k)\in E} \omega_{j,k} \left[\frac{v_j^2 + v_k^2}{2\sqrt{(v_j^{(t)})^2 + (v_k^{(t)})^2}} + \frac{\sqrt{(v_j^{(t)})^2 + (v_k^{(t)})^2}}{2} \right].\end{aligned} \tag{10}$$

2.5. The Surrogate Objective and Algorithm

Substituting $\Omega_{\mathrm{FGL}}(\mathbf{u})$ and $\Omega_{\mathrm{GGL}}(\mathbf{v})$ in Eq. (4) by $\Omega_{\mathrm{FGL}}^{\mathrm{App}}(\mathbf{u})$ and $\Omega_{\mathrm{GGL}}^{\mathrm{App}}(\mathbf{v})$ in Eqs. (9)–(10), respectively, we have the surrogate objective

$$\begin{aligned}\mathscr{L}(\mathbf{u},\mathbf{v}) = & -\mathbf{u}^\top \mathbf{X}^\top \mathbf{Y}\mathbf{v} + \Omega_{\mathrm{FGL}}^{\mathrm{App}}(\mathbf{u}) + \Omega_{\mathrm{GGL}}^{\mathrm{App}}(\mathbf{v}) + \frac{\gamma_1}{2}(\|\mathbf{X}\mathbf{u}\|^2 - 1) \\ & + \frac{\gamma_2}{2}(\|\mathbf{Y}\mathbf{v}\|^2 - 1).\end{aligned} \tag{11}$$

This objective is continuous, biconvex and differentiable with respect to **u** and **v**, and thus it is easy to solve. By respectively taking the partial derivatives of $\mathscr{L}(\mathbf{u},\mathbf{v})$ with respect to **u**, **v** and then setting the results to zero, we have the following equations.

$$\begin{aligned}\mathbf{0} = & -\mathbf{X}^\top \mathbf{Y}\mathbf{v} + (\lambda_1 \mathbf{D}_1 + \gamma_1 \mathbf{X}^\top \mathbf{X})\mathbf{u}, \ \|\mathbf{X}\mathbf{u}\|^2 - 1 = 0, \\ \mathbf{0} = & -\mathbf{Y}^\top \mathbf{X}\mathbf{u} + (\lambda_2 \mathbf{D}_2 + \gamma_2 \mathbf{Y}^\top \mathbf{Y})\mathbf{v}, \ \|\mathbf{Y}\mathbf{v}\|^2 - 1 = 0,\end{aligned} \tag{12}$$

where $\mathbf{D}_1$ is a diagonal matrix as follows

$$\mathbf{D}_1 = \begin{bmatrix} \frac{w_{1,2}}{\sqrt{(u_1^{(t)})^2 + (u_2^{(t)})^2}} & & & \\ & \ddots & & \\ & & \frac{w_{i-1,i}}{\sqrt{(u_{i-1}^{(t)})^2 + (u_i^{(t)})^2}} + \frac{w_{i,i+1}}{\sqrt{(u_i^{(t)})^2 + (u_{i+1}^{(t)})^2}} & \\ & & & \ddots \\ & & & \frac{w_{p-1,p}}{\sqrt{(u_{p-1}^{(t)})^2 + (u_p^{(t)})^2}} \end{bmatrix} \quad (13)$$

and

$$\mathbf{D}_2 = \begin{bmatrix} \sum_{k=1,(1,k)\in E}^{q} \frac{\omega_{1,k}}{\sqrt{(v_1^{(t)})^2 + (v_k^{(t)})^2}} & & \\ & \ddots & \\ & & \sum_{k=1,(j,k)\in E}^{q} \frac{\omega_{j,k}}{\sqrt{(v_j^{(t)})^2 + (v_k^{(t)})^2}} \\ & & & \ddots \\ & & & & \sum_{k=1,(q,k)\in E}^{q} \frac{\omega_{q,k}}{\sqrt{(v_q^{(t)})^2 + (v_k^{(t)})^2}} \end{bmatrix} \quad (14)$$

is also a diagonal matrix.

Now we have the closed-form updating rules

$$\mathbf{u}^{(t+1)} = (\lambda_1 \mathbf{D}_1^{(t)} + \gamma_1 \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \mathbf{v}^{(t)}, \ \mathbf{u}^{(t+1)} = \mathbf{u}^{(t+1)} / \left\| \mathbf{X}\mathbf{u}^{(t+1)} \right\|_2, \quad (15)$$

$$\mathbf{v}^{(t+1)} = (\lambda_2 \mathbf{D}_2^{(t)} + \gamma_2 \mathbf{Y}^\top \mathbf{Y})^{-1} \mathbf{Y}^\top \mathbf{X} \mathbf{u}^{(t+1)}, \ \mathbf{v}^{(t+1)} = \mathbf{v}^{(t+1)} / \left\| \mathbf{Y}\mathbf{v}^{(t+1)} \right\|_2, \quad (16)$$

where $\mathbf{D}_1^{(t)}$ and $\mathbf{D}_2^{(t)}$ denotes the t-th iteration of $\mathbf{D}_1$ and $\mathbf{D}_2$ respectively.

The procedure of the FGL-SCCA is shown in Algorithm 1. **u** and **v** are updated alternatively until the convergence criterion is met, such as the predefined termination condition or number of maximum iterations. In Algorithm 1, Steps 3–6 are repeated until convergence. In each iteration, Step 3 is easy to calculate as $\mathbf{D}_1$ can be computed via matrix computation 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 by solving a system of linear equations with approximative

quadratic complexity rather than computing the matrix inverse with cubic complexity. Thus the computation burden is dramatically reduced.

Algorithm 1

The FGL-SCCA Algorithm

Require:

$\mathbf{X} \in \mathscr{R}^{n \times p}$, $\mathbf{Y} \in \mathscr{R}^{n \times q}$, λ_1, λ_2, γ_1, γ_2

Ensure:

Canonical weights **u** and **v**.

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

while not converged **do**

  Update the diagonal matrix $\mathbf{D}_1$ according to Eq. (13);

  Solve **u** according to Eq. (15);

  Update the diagonal matrix $\mathbf{D}_2$ according to Eq. (14);

  Solve **v** according to Eq. (16);

end while

2.6. Convergence Analysis

We have the following theorem for Algorithm 1.

Theorem 1—*Solving the objective (11) is equivalent to solving the objective (4).*

Proof. The objective (11) and the objective (4) are only different in the penalties. According to Proposition 1 in Section 2.4, the three rules also hold for both objective (4) and (11): (1) $\mathscr{L}(\mathbf{u}^{(t)}, \mathbf{v}^{(t)}) = \mathscr{L}(\mathbf{u}^{(t)}, \mathbf{v}^{(t)})$, (2) $\mathscr{L}'_{\mathbf{u}}(\mathbf{u}^{(t)}, \mathbf{v}^{(t)}) = \mathscr{L}'_{\mathbf{u}}(\mathbf{u}^{(t)}, \mathbf{v}^{(t)})$ and $\mathscr{L}'_{\mathbf{v}}(\mathbf{u}^{(t)}, \mathbf{v}^{(t)}) = \mathscr{L}'_{\mathbf{v}}(\mathbf{u}^{(t)}, \mathbf{v}^{(t)})$, and (3) $\mathscr{L}(\mathbf{u}, \mathbf{v}) \leq \mathscr{L}(\mathbf{u}, \mathbf{v})$. Thus the objective (11) approximates to the objective (4) point-by-point during the iteration. Therefore, solving (11) is equivalent to solving (4). □

Actually, the Algorithm 1 is an alternating minimization method which will converge to the leading canonical pair (Golub and Zha, 1995). We further verify that **u** = {1, 0, …, 0} and **v** = {1, 0, …, 0} are a pair of feasible solution to the objective (11). This implies that the Slater's condition holds. Therefore, satisfying the KKT condition guarantees that Algorithm 1 will converge to one local optimum of objective (11), which is also the local optimum of the objective (4) as supported by Theorem 1. In the implementation, we solve a system of linear equations with quadratic complexity to update both **u** and **v**, without computing the inverse of the large covariance matrix with cubic complexity. Thus the whole algorithm is of desired efficiency.

2.7. The Grouping Effect

The grouping effect of FGL-SCCA refers to estimating equal or similar values for successive variables of **u** and for connected variables of **v.** This implies the simultaneous selection of adjacent genetic features and of correlated imaging features, which is guaranteed by the following theorem.

Theorem 2—*Given two views of data* $\mathbf{X}$ *and* $\mathbf{Y}$ *that have been centered and normalized, and the tuned parameters* (λ, γ). *Let* $\mathbf{u}^*$ *be the solution to the FGL-SCCA problem. For the sake of simplicity, we assume that only* $\mathbf{x}_i$ *and* $\mathbf{x}_{i+1}$ *are correlated. Let* $\rho_{i,i+1}$ *be their sample correlation. Then the optimal* u_i^* *and* u_{i+1}^* *associated with* $\mathbf{x}_i$ *and* x_{i+1} *satisfy,*

$$\left| \frac{|u_i^*| - |u_{i+1}^*|}{\sqrt{(u_i^*)^2 + (u_{i+1}^*)^2}} \right| \leq \frac{1+\gamma_1}{\lambda_1 w_{i,i+1}} \sqrt{2(1-|\rho_{i,i+1}|)}. \tag{17}$$

Proof. (1) We first prove the inequations when $\rho_{i,i+1} \geq 0$, indicating $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$ are positively correlated. Since $\mathbf{u}^*$ is the solution, we have $\frac{\partial \mathcal{L}}{\partial u_i}\Big|_{u_i^*} = 0$ and $\frac{\partial \mathcal{L}}{\partial u_{i+1}}\Big|_{u_{i+1}^*} = 0$, i.e.,

$$\lambda_1 \mathbf{D}_{1,i} u_i^* + \gamma_1 \mathbf{x}_i^\top \mathbf{X}\mathbf{u}^* = \mathbf{x}_i^\top \mathbf{Y}\mathbf{v}^*,\ \lambda_1 \mathbf{D}_{1,i+1} u_{i+1}^* + \gamma_1 \mathbf{x}_{i+1}^\top \mathbf{X}\mathbf{u}^* = \mathbf{x}_{i+1}^\top \mathbf{Y}\mathbf{v}^*. \tag{18}$$

According to the definition of $\mathbf{D}_1$, we obtain

$$\begin{aligned} &\frac{\lambda_1 w_{i,i+1}}{\sqrt{(u_i^*)^2 + (u_{i+1}^*)^2}} u_i^* + \gamma_1 \mathbf{x}_i^\top \mathbf{X}\mathbf{u}^* = \mathbf{x}_i^\top \mathbf{Y}\mathbf{v}^*, \\ &\frac{\lambda_1 w_{i,i+1}}{\sqrt{(u_i^*)^2 + (u_{i+1}^*)^2}} u_{i+1}^* + \gamma_1 \mathbf{x}_{i+1}^\top \mathbf{X}\mathbf{u}^* = \mathbf{x}_{i+1}^\top \mathbf{Y}\mathbf{v}^*. \end{aligned} \tag{19}$$

Subtracting these two equations, we arrive at

$$\frac{\lambda_1 w_{i,i+1}}{\sqrt{(u_i^*)^2 + (u_{i+1}^*)^2}} (u_i^* - u_{i+1}^*) = (\mathbf{x}_i - \mathbf{x}_{i+1})^\top (\mathbf{Y}\mathbf{v}^* - \gamma_1 \mathbf{X}\mathbf{u}^*) \tag{20}$$

Taking ℓ_2-norm on both sides, we arrive at

$$\begin{aligned} &\frac{\lambda_1 w_{i,i+1}}{\sqrt{(u_i^*)^2 + (u_{i+1}^*)^2}} |u_i^* - u_{i+1}^*| \leq \|\mathbf{x}_i - \mathbf{x}_{i+1}\| \, \|\mathbf{Y}\mathbf{v}^* - \gamma_1 \mathbf{X}\mathbf{u}^*\| \\ &= \|\mathbf{x}_i - \mathbf{x}_{i+1}\| \sqrt{\|\mathbf{Y}\mathbf{v}^*\|^2 - 2\gamma_1 (\mathbf{u}^*)^\top \mathbf{X}^\top \mathbf{Y}\mathbf{v}^* + \gamma_1^2 \|\mathbf{X}\mathbf{u}^*\|^2} \end{aligned} \tag{21}$$

Since $\mathbf{X}$ and $\mathbf{Y}$ are centered and normalized, we have $\|\mathbf{x}_i - \mathbf{x}_{i+1}\| = \sqrt{2(1-\rho_{i,i+1})}$. Then using $\|\mathbf{X}\mathbf{u}^*\| = 1$, $\|\mathbf{Y}\mathbf{v}^*\| = 1$, $-(\mathbf{u}^*)^\top \mathbf{X}^\top \mathbf{Y}\mathbf{v}^* \leq 1$, we obtain the upper bound

$$\frac{|u_i^* - u_{i+1}^*|}{\sqrt{(u_i^*)^2 + (u_{i+1}^*)^2}} \leq \frac{1+\gamma_1}{\lambda_1 w_{i,i+1}} \sqrt{2(1-\rho_{i,i+1})}. \tag{22}$$

(2) If $\rho_{i,i+1} < 0$, it is clear that $\operatorname{sgn}(u_i^*) = -\operatorname{sgn}(u_j^*)$. By adding both equations in Eq. (19) instead of subtracting them, we finally arrive at,

$$\frac{\left|u_i^* + u_{i+1}^*\right|}{\sqrt{(u_i^*)^2 + (u_{i+1}^*)^2}} \leq \frac{1+\gamma_1}{\lambda_1 w_{i,i+1}} \sqrt{2(1+\rho_{i,i+1})}. \tag{23}$$

Combining Eqs. (22) and (23) together yields

$$\frac{\left||u_i^*| - |u_{i+1}^*|\right|}{\sqrt{(u_i^*)^2 + (u_{i+1}^*)^2}} \leq \frac{1+\gamma_1}{\lambda_1 w_{i,i+1}} \sqrt{2(1-|\rho_{i,i+1}|)}, \tag{24}$$

which completes the proof. □

The GGL has similar entries to the FGL by extending the adjacent smoothness to the graphical smoothness. Thus similar argument yields the upper bound of grouping effect in terms of canonical weight **v**, i.e.

$$\frac{\left||v_j^*| - |v_k^*|\right|}{\sqrt{(v_j^*)^2 + (v_k^*)^2}} \leq \frac{1+\gamma_2}{\lambda_2 \omega_{j,k}} \sqrt{2(1-|\rho_{j,k}|)}. \tag{25}$$

It is interesting that the Eqs. (24–25) give a normalized distance measurement for two variables. The range for this normalized distance varies from 0 to 1. This can clearly tell the similarity strength between two variables.

For the FGL penalizing canonical weight **u**, Theorem 2 provides a qualitative description of the bound accommodating the absolute value of differences between two successive variables. The bound directly depends on their sample correlation strength. If $\rho_{i,i+1} \geq 0$, a higher correlation between two variables pushes toward a smaller difference between their estimated coefficients. If $\rho_{i,i+1} < 0$, a smaller value promotes a smaller sum between their coefficients. This implies that the two coefficients will be equal or similar in amplitude. Therefore, the FGL-SCCA will strongly smooth between two highly correlated successive variables in terms of **u**. As for the GGL penalizing **v**, the same result exists between two connected variables which are not necessary to be neighbours.

3. Results

3.1. Benchmarks and Experimental Setup

One goal of this paper is to investigate the structure detection ability without requiring the prior knowledge. Three benchmark methods are used in this study for comparison. They are the FL-SCCA (fused lasso based SCCA) method which imposes the smoothness constraint between adjacent variables (Witten et al., 2009), the NS-SCCA (network structured SCCA) method whose penalty terms are network guided fused lasso (Chen and Liu, 2012), and the AGN-SCCA (absolute GraphNet SCCA) method whose penalty terms are also guided by graph but different to that of NS-SCCA (Du et al., 2016). The latter two methods are different in both modeling and optimizing techniques, and are deemed to be among the best structured SCCA methods by now. The group lasso based SCCA methods require prior

knowledge regarding the group information of variables, and hence we do not include them in the empirical study.

There are four parameters for all the SCCA methods, including the proposed FGL-SCCA. Blindly tuning them will incur heavy computation burden. For the efficiency purpose, we employ some heuristic strategy to lower down the computation burden regarding parameters tuning. Firstly, we observe that λ_i and γ_i ($i \in \{1, 2\}$) contribute oppositely to the grouping effect as shown in Theorem 2. Thus simultaneously increasing or decreasing both λ_i and γ_i ($i \in \{1, 2\}$) will lead to similar grouping results. Secondly, in this study, we prefer the structure pattern which is more sensitive to λ_i ($i \in \{1, 2\}$). Therefore, we fix γ_1 and γ_2, and only tune the remaining two parameters λ_1 and λ_2. Thirdly, an SCCA method and a conventional CCA will yield similar results if parameters of SCCA are too small. On the contrary, SCCA will over-penalize the result when the parameters are too large. So a neither too large nor too small parameter is more reasonable (Du et al., 2016). As a result, we optimally tune them via a grid search from 10^i ($i = -5, -4, \ldots, 0, \ldots, 4, 5$) through the nested five-fold cross-validation. Specifically, in the inner loop where the whole data are the training set of the external loop, we keep calculating $\mathrm{CV}(\lambda, \gamma) = \frac{1}{5}\sum_{j=1}^{5} Corr(\mathrm{X}_{-j}\mathbf{u}_j, \mathbf{Y}_{-j}\mathbf{v}_j)$ by changing only λ_1 or λ_2, where $\mathbf{X}_{-j}$ and $\mathbf{Y}_{-j}$ are the j-th subset of the inner testing set, and $\mathbf{u}_j$ and $\mathbf{v}_j$ are the canonical weights estimated from the inner training set. We choose parameters that generate the highest correlation coefficients (argmaxCV(λ, γ)) as the optimal parameters and use them in the external loop to generate the final results. All these methods utilize the same partition during cross-validation to make a fair comparison. Besides, we set the edge weight to be one i.e. $w_{i,i+1} = 1$ for FGL penalty and $w_{j,k} = 1$ for GGL penalty, and other type of weights can also be employed, e.g. $w_{i,i+1} = |\rho_{i,i+1}|^d$, where d is a positive integer to model the strength of the feature correlation. Finally, for each parameter setup we repeat the experiment 50 times and report the average results, which could further assure a stable performance.

For the proposed FGL-SCCA, we terminate the algorithm when both of the two conditions are satisfied, i.e. $\max_i \left|u_i^{(t+1)} - u_i^{(t)}\right| \leq \epsilon$ and $\max_j \left|v_j^{(t+1)} - v_j^{(t)}\right| \leq \epsilon$ where ϵ is the tolerable error. We empirically set $\epsilon = 10^{-5}$ from experiments in this paper. The implementation of the proposed method is available at github (https://github.com/dulei323/SCCA-FGL).

3.2. Simulation Study

We generate six data sets with different properties in this simulation study to assess the performance of FGL-SCCA in different situations. The properties, such as different ground truths, sparsity levels, and mix of positively/negatively cross-correlated features, of these six data sets are distinct to assure diversity, which could make a thorough comparison. The details of each data set are described as follows, and the true signal of every data set is also shown in Fig. 1 (top row).

- Data 1: This data set is generated with $n = 80$, $p = 120$, and $q = 100$. We first generate the vector $\mathbf{u} = (\underbrace{0, \ldots, 0}_{60}, \underbrace{2, \ldots, 2}_{40}, \underbrace{0, \ldots, 0}_{40})^\top$ and $\mathbf{v} = (\underbrace{0, \ldots, 0}_{25}, \underbrace{3, \ldots, 3}_{25}, \underbrace{0, \ldots, 0}_{50})^\top$. Then we create a latent variable $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_{n \times n})$. After that, $\mathbf{X}$ is created by

$\mathbf{x}^{\ell} \sim N(z_{\ell}\mathbf{u}^{\top}, 0.1\boldsymbol{\Sigma}_x)$, where $\mathbf{x}^{\ell}$ is the ℓth row of $\mathbf{X}$, and $(\boldsymbol{\Sigma}_x)_{i,i+1} = e^{-|u_i - u_i + 1|}$. Similarly, $\mathbf{Y}$ is created by $\mathbf{y}_{\ell} \sim N(z_{\ell}\mathbf{v}^{\top}, 0.1\boldsymbol{\Sigma}_y)$, where $(\boldsymbol{\Sigma}_y)_{jk} = e^{-|v_j - v_k|}$.

- Data 2: This data set is created similarly to the first data set, where $n = 50$, $p = 150$ and $q = 200$. $\mathbf{u} = (\underbrace{0, \ldots, 0}_{58}, 1, -1, 1, \underbrace{0, \ldots, 0}_{89})^{\top}$ and $\mathbf{v} = (\underbrace{0, \ldots, 0}_{40}, \underbrace{2, \ldots, 2}_{40}, \underbrace{0, \ldots, 0}_{40}, \underbrace{-3, \ldots, -3}_{40}, \underbrace{0, \ldots, 0}_{40})^{\top}$, $\mathbf{x}^{\ell} \sim N(z_{\ell}\mathbf{u}^{\top}, 0.2\boldsymbol{\Sigma}_x)$ with $(\boldsymbol{\Sigma}_x)_{i,i+1} = e^{-\sqrt{u_i^2 + u_i^2 + 1}}$ and $\mathbf{y}_{\ell} \sim N(z_{\ell}\mathbf{v}^{\top}, 0.2\boldsymbol{\Sigma}_y)$, where $(\boldsymbol{\Sigma}_y)_{jk} = e^{-|v_j - v_k|}$.
- Data 3: This data set is created by $n = 50$, $p = 150$ and $q = 200$, where $\mathbf{u} = (\underbrace{0, \ldots, 0}_{58}, 2, -2, \underbrace{0, \ldots, 0}_{90})^{\top}$, $\mathbf{v} = (\underbrace{0, \ldots, 0}_{40}, \underbrace{-1, 1, -1, 1, \ldots, -1, 1}_{40}, \underbrace{0, \ldots, 0}_{120})^{\top}$, $\mathbf{x}^{\ell} \sim N(z_{\ell}\mathbf{u}^{\top}, 0.2\boldsymbol{\Sigma}_x)$ with $(\boldsymbol{\Sigma}_x)_{i,i+1} = e^{-\sqrt{u_i^2 + u_i^2 + 1}}$ and $\mathbf{y}_{\ell} \sim N(z_{\ell}\mathbf{v}^{\top}, 0.2\boldsymbol{\Sigma}_y)$ with $(\boldsymbol{\Sigma}_y)_{jk} = e^{-|v_j - v_k|}$.
- Data 4: This data set is created by $n = 50$, $p = 150$ and $q = 200$, where $\mathbf{u} = (\underbrace{0, \ldots, 0}_{60}, \underbrace{-6, 6, -6, 6, \ldots, -6, 6}_{30}, \underbrace{0, \ldots, 0}_{60})^{\top}$, $\mathbf{v} = (\underbrace{0, \ldots, 0}_{40}, \underbrace{-2, \ldots, -2}_{20}, \underbrace{2, \ldots, 2}_{20}, \underbrace{0, \ldots, 0}_{120})^{\top}$, $\mathbf{x}^{\ell} \sim N(z_{\ell}\mathbf{u}^{\top}, 0.2\boldsymbol{\Sigma}_x)$ with $(\boldsymbol{\Sigma}_x)_{i,i+1} = e^{-\sqrt{u_i^2 + u_i^2 + 1}}$ and $\mathbf{y}_{\ell} \sim N(z_{\ell}\mathbf{v}^{\top}, 0.2\boldsymbol{\Sigma}_y)$ with $(\boldsymbol{\Sigma}_y)_{jk} = e^{-\sqrt{v_j^2 + v_k^2}}$.
- Data 5: This data set is created by $\boldsymbol{n} = 50$, $\boldsymbol{p} = 150$ and $\boldsymbol{q} = 200$, where $\mathbf{u} = (\underbrace{0, \ldots, 0}_{58}, 2, -2, -1, \underbrace{0, \ldots, 0}_{89})^{\top}$, $\mathbf{v} = (\underbrace{0, \ldots, 0}_{40}, \underbrace{-2, \ldots, -2}_{20}, \underbrace{2, \ldots, 2}_{20}, \underbrace{0, \ldots, 0}_{120})^{\top}$, $\mathbf{x}^{\ell} \sim N(z_{\ell}\mathbf{u}^{\top}, 0.2\boldsymbol{\Sigma}_x)$ with $(\boldsymbol{\Sigma}_x)_{i,i+1} = e^{-\sqrt{u_i^2 + u_i^2 + 1}}$ and $\mathbf{y}_{\ell} \sim N(z_{\ell}\mathbf{v}^{\top}, 0.2\boldsymbol{\Sigma}_y)$ with $(\boldsymbol{\Sigma}_y)_{jk} = e^{-\sqrt{v_j^2 + v_k^2}}$.
- Data 6: This data set is created by $\boldsymbol{n} = 50$, $\boldsymbol{p} = 150$ and $\boldsymbol{q} = 200$, where $\mathbf{u} = (\underbrace{0, \ldots, 0}_{58}, 1, -1, 1, \underbrace{0, \ldots, 0}_{89})^{\top}$, $\mathbf{v} = (\underbrace{0, \ldots, 0}_{40}, \underbrace{-2, 2, -2, 2, \ldots, -2, 2}_{40}, \underbrace{0, \ldots, 0}_{120})^{\top}$, $\mathbf{x}^{\ell} \sim N(z_{\ell}\mathbf{u}^{\top}, 0.1\boldsymbol{\Sigma}_x)$ with $(\boldsymbol{\Sigma}_x)_{i,i+1} = e^{-\sqrt{u_i^2 + u_i^2 + 1}}$ and $\mathrm{y}_{\ell} \sim N(z_{\ell}\mathbf{v}^{\top}, 0.1\boldsymbol{\Sigma}_y)$ with $(\boldsymbol{\Sigma}_y)_{jk} = e^{-\sqrt{v_j^2 + v_k^2}}$.

The ground truth and estimated canonical weights **u** and **v** of each method are presented in Fig. 1. In each subfigure, the vertical axis represents the indices of each **u** (left panel) or **v** (right panel), and the horizontal axis represents 250 runs of experiments (50 times of 5-fold cross-validation). Our FGL-SCCA identifies similar canonical weights that are consistent with the ground truth across all six data sets. Interestingly, when the true signals have group structures of both **X** and **Y**, i.e. data 1 and data 4, almost every method can find the true signals correctly. This demonstrates the group information identification ability of these SCCA methods which have been analyzed in their respective papers (Witten et al., 2009; Chen and Liu, 2012; Du et al., 2016). However, when the true signals of **X** involve only two

successive negatively correlated variables, i.e. data 3, the FGL-SCCA still correctly find out them with a clear pattern. Those competing methods, on the contrary, report too many nonzero signals which cannot easily help find out true signals. For the remaining three data sets, i.e. data 2, data 5 and data 6, there are three successive nonzero variables with negative relationship on **X,** while different group structures on **Y**. We observe that FGL-SCCA can also accurately find out the true signals, and those benchmarks cannot.

We consider a feature as relevant if its estimated weight $\hat{u}_i$ (or $\hat{v}_j$) is larger in absolute value than a predefined threshold. The larger the $|\hat{u}_i|$ (or $|\hat{v}_j|$) is, the more contribution the *i*-th genetic feature (or the *j*-th imaging feature) makes to the canonical correlation. We then sort the features in descending order of their $|\hat{u}_i|$ (or $|\hat{v}_j|$), and vary the threshold to obtain a sequence of true positive rate (TPR) - false positive rate (FPR) pairs and to calculate the area under the ROC curve (AUC). Table 1 shows the area under ROC (AUC) which stands for the sensitivity and specificity in terms of canonical weights. We observe that the proposed FGL-SCCA obtains the highest value on all six data sets in terms of both **u** and **v**. NS-SCCA and AGN-SCCA are suboptimal and FL-SCCA performs the worst in terms of these evaluation criteria. This means that FGL-SCCA could be the best choice in structure information extraction followed by NS-SCCA and AGN-SCCA. In addition, we also show the training and testing correlation coefficients calculated from the trained SCCA models in Table 2. It is clear that all methods obtain good results on all training sets. Interestingly, the proposed FGL-SCCA outperforms those competing methods on the testing sets. This indicates that FGL-SCCA possesses better generalization ability than fused lasso and graphical penalty based benchmarks. To summarize, results on these six diverse data sets demonstrate that FGL-SCCA can not only identify similar or higher training and testing bi-multivariate associations, but also better canonical weights profiles.

3.3. Real Neuroimaging Genetics Study

We also compared the proposed structure-aware SCCA method with benchmarks using real neuroimaging and genetics data. 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. The 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). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. For up-to-date information, see www.adni-info.org.

The brain imaging measurements data (i.e., amyloid imaging data) of 567 non-Hispanic Caucasian participants at the ADNI-GO/2 baseline were downloaded from the LONI website (adni.loni.usc.edu). Shown in Table 3 are the characteristics of these subjects, including 196 healthy control (HC), 343 MCI and 28 AD participants. The [11C] Florbetapir PET scans were averaged, aligned to a standard space, resampled to a standard image and

voxel size, smoothed to a uniform resolution and normalized to a cerebellar gray matter reference region resulting in standardized uptake value ratio (SUVR) images as previously described (Jagust et al., 2010). After this, the images were aligned to each participant's same visit MRI scan and normalized to MNI space as $2\times2\times2$ m^3 voxels using parameters from the MRI segmentation. We further extracted region of interest (ROI) level amyloid measurements, and generated 191 mean amyloid measurements spanning all brain ROI level based on the MarsBaR AAL atlas.

The single nucleotide polymorphism (SNP) data were also downloaded from the ADNI website. They were genotyped using the Human 610-Quad or OmniExpress Array (Illumina, Inc., San Diego, CA), 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. As the second pre-processing step, the quality-controlled SNPs were imputed using the MaCH software (Li et al., 2010) to estimate the missing genotypes. The genotyping data here includes 1,000 SNP markers from chromosome 19 near the *APOE* gene. The aim is to detect the associations between SNPs and amyloid measurements, as well as which SNPs and amyloid markers are simultaneously correlated with diagnostic status.

All four SCCA methods were applied to this real neuroimaging genetics data. Fig. 2 presents the canonical weights estimated from the training set by each method, showing those relevant SNPs and imaging measurements. In each subfigure, the horizontal axis represents the reference number of each individual SNP (left panel) or imaging ROI (right panel), and the vertical axis represents every run and there are 250 runs in total (50 times of 5-fold cross-validation). We can clearly observe that FGL-SCCA identifies two relevant groups of successive SNPs and a very small proportion of ROIs for easy interpretation due to the novel FGL and GGL penalties. The peak signal on the genetic data originates from rs429358, which codes for the *APOE* ε4 allele. This locus has been confirmed to be associated with AD previously (Ramanan et al., 2014). The locus rs56131196 with the second highest weight comes from the *APOC1* gene, which is recently identified to be correlated with both Type 2 Diabetes (T2D) and AD (Gao et al., 2016). The two strongest imaging ROIs are from the frontal brain area. They are the right superior frontal gyrus and the left middle frontal gyrus, which have been demonstrated to be correlated with AD. The non-zero signal with the third largest weight is from the caudate nucleus, which has been reported as an AD related brain area (Jiji et al., 2013). Those competing methods, such as the FL-SCCA, NS-SCCA and the AGN-SCCA, find out many interfering signals for both imaging markers across the brain and genetic markers of chromosome 19. FL-SCCA reports the most non-zero signals for both imaging and genetics markers, followed by NS-SCCA and AGN-SCCA. In biomedical studies, results with many non-zero signals are very hard to interpret since they cannot imply a clear clue for further investigation. We also show the training and testing correlation coefficients in Table 4. In the table, both *mean* and *std* are contained, and the *p*-values which are calculated between each benchmark and FGL-SCCA are also shown. The proposed FGL-SCCA obtains better canonical correlation coefficients on both training set and testing set. Moreover, all *p*-values are significant (< 0.05) indicating that FGL-SCCA

outperform those benchmarks on this real imaging genetic data set. Table 5 shows the runtime results on this real data, in which we could observe that FL-SCCA, NS-SCCA and FGL-SCCA run much faster than AGN-SCCA. In summary, the results on this real data reveal that FGL-SCCA has better bi-multivariate identification capability than both fused lasso and graphical lasso based SCCA methods in this ADNI cohort study.

4. Discussion

To further investigate the performance of our FGL-SCCA method, we average the canonical weights across five folds and select the top ten SNPs and top ten ROI measurements and show them in Tables 6 and 7.

4.1. Top Selected Genetic Markers

In Table 6, the first column shows the reference number of each identified SNP, the second one shows the gene name, the third column is the averaged weight across 250 runs (50 times of 5-fold cross-validation), and the fourth column is the percent showing that each SNP is selected as top ten markers in 250 runs. The last column is the p-value of the main effect of each SNP on the diagnosis. There are three groups of loci associated with the top ten SNPs. The first group are rs429358, rs769449, rs769450, rs1081105, and they all locate in the *APOE* gene which is related to AD. Interestingly, the sign of SNP rs769450 is different from its neighbouring SNPs (rs769449, rs429358 and rs1081105). This demonstrates that the FGL-SCCA can perform feature grouping as long as two adjacent variables exhibit high similarity in absolute values. Moreover, although rs769450 and rs1081105 are non-significant in this data, they both are jointly selected by the newly introduced FGL penalty. The second group of loci are all from *APOC1* gene. They are rs12721051, rs56131196 and rs4420638, and are recently identified to be shared genetic factors between T2D and AD (Gao et al., 2016). The third group of loci are rs10414043, rs7256200 and rs483082 located between the *APOE* and *APOC1* gene, and they also have been identified to show association with the longevity in humans (Zeng et al., 2016).

4.2. Top Selected Brain Imaging ROIs

Table 7 presents the top ten brain imaging ROIs identified by the averaged canonical weights. In this table, the first column exhibits the name of the brain region, the second one shows the averaged weight across 250 runs, and the third column is the percent showing that each ROI is selected as top ten risk markers in 250 runs. The p-value of the diagnostic effect measured by ANOVA was shown for each imaging ROI in the last column. In our experimental setting, there might be more than one variable associated with the same label in the automated anatomical labeling (AAL) atlas because we have 191 brain regions corresponding to 116 AAL regions. Thus the first and the fifth imaging measures are from the same AAL region. The p-values of all the ten markers are relatively small indicating that they are significantly correlated with diagnostic status. At the same time, a literature search also shows that all these ten imaging markers have been reported to be more or less associated with AD, such as the frontal gyrus (Hirono et al., 1998; Bi et al., 2018) and the caudate nucleus (Jiji et al., 2013). The first two markers, i.e., the left middle frontal gyrus and the right superior frontal gyrus, have similar estimated weights owing to the newly

introduced graphical GGL penalty. We further find that the correlation between them is 0.8827 which is a very high value in this data set. This demonstrates that FGL-SCCA could group a pair of highly correlated variables if they both are associated with diagnostic status. We note that the highest correlation value exists between the orbital part of left middle frontal gyrus and the right one after looking into the pairwise correlation matrix. It looks strange that FGL-SCCA does not estimate similar weights for both of them. The reason might boil down to three aspects: (1) both variables should be correlated with each other; (2) both variables should be correlated with those SNPs identified by our algorithm; and (3) a brain ROI is connected to more than one ROI according to the GGL penalty. Thus the final weight of an ROI will be affected by the combination of several grouping effects. This also explains why all the top ten brain imaging measurements and top ten SNPs hold very high correlations (0.485) in this study, which dominates the relationship between this leading pair of canonical weights. To give a clear spatial view, we map the averaged canonical weights regarding these imaging measurements of FGL-SCCA onto the brain atlas. Fig. 3 shows that our method only highlights a small region of the whole brain. This is quite meaningful since it provides a clear and clean clue for further targeted analysis.

As a structured method, it is very important to verify the performance on the identified graph structure. In this study, the estimated canonical weights **v** imply the identified graph structure of the brain ROIs. If two ROIs have the same or similar weight values, they will be in the same subgraph according to Theorem 2. Based on this, we could obtain the graph with two considerations. First, both v_i and v_j, i.e. the weight values of ROI_i and ROI_j, should be important, which means their weight values are larger than a threshold τ_1, i.e. $|v_i| > \tau_1$ and $|v_j| > \tau_1$. Second, v_i and v_j should be equal or similar, indicating that their difference is small enough, e.g. $\| |v_i| - |v_j| \| / |v_j| < \tau_2$, where τ_2 is the maximum tolerance difference. Suppose $\tau_1 = 0.0001$ and $\tau_2 = 0.1$ (both thresholds could be changed accordingly), the identified graph structure is shown in Fig. 4. We clearly observe that there are three subgraphs identified by FGL-SCCA. Interestingly, all the nodes (ROIs) in these three subgraphs have been verified to be correlated to AD. This demonstrates the effectiveness of our method in identifying meaningful subgraphs in this ADNI study, which verifies the correctness of our model design.

4.3. Refined Analysis

Based on the top ten selected SNPs and brain ROIs, Fig. 5 shows the heat map of pairwise correlations of every brain ROI-SNP pair. As expected, most ROI-SNP pairs have considerable correlation values. We observe that rs769450 from the *APOE* gene has the negative correlation with all these ten ROIs. In order to further understand this, we choose to use rs429358 as the comparison based on the following considerations. First, rs429358 has been confirmed to be the top risk factor for late onset AD via affecting multiple brain structures (Potkin et al., 2009a). Second, both rs429358 and rs769450 are from the *APOE* gene and they are jointly selected by the FGL-SCCA method. Besides, they hold opposite coefficient signs in our model. The frontal lobe region is a well-known AD related brain area, and the clumping together of beta-amyloid proteins could be a major AD hallmark. Therefore, we use the beta-amyloid deposition measurement in the frontal lobe as the target imaging marker.

Using the amyloid accumulation in the left middle frontal gyrus as the response, we conducted the two-way ANOVA to show that the main effects of rs769450 genotype, diagnosis and their SNP-by-diagnosis interaction effect. As shown in Fig. 6(a), the main effects of rs769450 genotype ($p < 0.01$), diagnosis ($p < 0.01$) and their SNP-by-diagnosis interaction effect ($p < 0.01$) all reached the significant level when age, gender, education and handedness were included as covariates. The pairwise comparison results showed that the amyloid accumulation in AD participants was significantly higher than that of both MCI and HC groups (all $p < 0.01$). In addition, MCI participants also showed a significantly increased amyloid deposition than HCs ($p < 0.01$). In order to investigate the genotype effect within each baseline diagnosis group separately, we conducted pairwise comparisons among the heterozygotes AG, homozygous AA and GG within ADs, MCIs and HCs respectively. The results showed that within ADs, those patients holding the homozygous AA have lower beta-amyloid deposition measurements compared with those holding GG and AG. This pattern can also be observed within the MCI participants but not in the HCs. By contrast, in Fig. 6(b), the two-way ANOVA results from rs429358 showed that within both ADs and MCIs, participants holding the homozygous CC have higher beta-amyloid deposition compared with those having TT. It is easy to observe that the genotype polymorphisms of rs769450 and rs429358 have opposite effects on the beta-amyloid deposition in the group of ADs and MCIs. Specifically, the major homozygote of rs769450 in AD patients were vulnerable to increase beta-amyloid deposition in left middle frontal gyrus. On the contrary, AD patients with the minor homozygote of rs429358 were vulnerable to have higher beta-amyloid deposition measurement in this ADNI cohort.

5. Conclusions

We have introduced two novel penalties such as the fused pairwise group lasso (FGL) and graph guided pairwise group lasso (GGL). We proposed a novel structure-aware sparse canonical correlation analysis (SCCA) method using FGL and GGL as constraints to identify associations between brain imaging measurements and genetic factors. The existing group lasso based methods (Chen et al., 2013; Du et al., 2014) were dependent on the prior knowledge which usually was not always available. The graph/network guided fused lasso based methods (Chen et al., 2012; Du et al., 2014; Chen and Liu, 2012; Du et al., 2016) focus on the positively correlated variables, or depended on the signs of the sample correlation which were sensitive to the partition of data. The proposed SCCA method combines the advantages of both group lasso and graphical fused lasso, which is independent of the sign of the sample correlation. Moreover, FGL-SCCA can be used in data-driven mode which means it does not require the prior knowledge, and can incorporate the prior knowledge to recover specific structures, too. FGL-SCCA recovers a chain of smoothness on the genetic factors and graphical smoothness on the brain imaging measurements.

We have compared FGL-SCCA with three state-of-the-art structure-aware SCCA methods on both synthetic data and real imaging genetic data. The results on the synthetic data show that FGL-SCCA performs similarly or better than all three benchmarks. The results on real data show that, FGL-SCCA not only estimates better canonical correlation coefficients than the competing methods, but also obtains more clear, cleaner and sparser canonical weights.

FGL-SCCA detects a strong association between a few group of loci (from *APOE* and *APOC1*) and frontal and caudate morphometries. All three group of loci, including the SNP rs429358 etc., and the imaging measurements such as the frontal gyrus have been identified to be highly associated with AD, demonstrating FGL-SCCA's power in brain imaging genetics. Since the GGL penalty becomes quite complicated as the number of variables increases, one interesting future direction is to improve the efficiency and scalability for FGL-SCCA in more realistic settings. Moreover, given the prominent role of the *APOE* signal in this application, it is also of great importance to identify other AD-relevant SNPs. Therefore, another future direction is to use our method to identify the new AD-relevant SNPs in addition to those from *APOE.*

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12–2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Funding: This research was supported by the National Natural Science Foundation of China [61973255, 61602384]; the Natural Science Basic Research Plan in Shaanxi Province of China [2017JQ6001]; the China Postdoctoral Science Foundation [2017M613202]; Science and Technology Foundation for Selected Overseas Chinese Scholar, Department of Human Resources and Social Security in Shaanxi Province [2017022]; the Postdoctoral Science Foundation of Shaanxi Province; and the Fundamental Research Funds for the Central Universities at Northwestern Polytechnical University.

This work was also supported by the National Institutes of Health [R01 EB022574, RF1 AG063481, U01 AG024904, P30 AG10133, R01 AG19771] at University of Pennsylvania and Indiana University.

References

Bi X.-a., Jiang Q, Sun Q, Shu Q, Liu Y, 2018 Analysis of alzheimer's disease based on the random neural network cluster in fmri. Frontiers in Neuroinformatics 12, 60. [PubMed: 30245623]

Chen J, Bushman FD, Lewis JD, Wu GD, Li H, 2013 Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. Biostatistics 14 (2), 244–258. [PubMed: 23074263]

Chen X, Han L, Carbonell J, 2012 Structured sparse canonical correlation analysis. In: Artificial Intelligence and Statistics. pp. 199–207.

Chen X, Liu H, 2012 An efficient optimization algorithm for structured sparse cca, with applications to eqtl mapping. Statistics in Biosciences 4 (1), 3–26.

Chi EC, Allen GI, Zhou H, Kohannim O, Lange K, Thompson PM, 2013 Imaging genetics via sparse canonical correlation analysis. In: ISBI. pp. 740–743.

Du L, Huang H, Yan J, Kim S, Risacher SL, Inlow M, Moore JH, Saykin AJ, Shen L, 2016 Structured sparse canonical correlation analysis for brain imaging genetics: An improved graphnet method. Bioinformatics 32 (10), 1544–1551. [PubMed: 26801960]

Du L, Liu K, Zhang T, Yao X, Yan J, Risacher SL, Han J, Guo L, Saykin AJ, Shen L, 2018 A novel SCCA approach via truncated ℓ_1-norm and truncated group lasso for brain imaging genetics. Bioinformatics 34 (2), 278–285. [PubMed: 28968815]

Du L, Liu K, Zhu L, Yao X, Risacher SL, Guo L, Saykin AJ, Shen L, 2019 Identifying progressive imaging genetic patterns via multitask sparse canonical correlation analysis: a longitudinal study of the adni cohort. Bioinformatics 35 (14), i474–i483. [PubMed: 31510645]

Du L, Yan J, Kim S, Risacher SL, Huang H, Inlow M, Moore JH, Saykin AJ, Shen L, 2014 A novel structure-aware sparse learning algorithm for brain imaging genetics. In: MICCAI. pp. 329–336. [PubMed: 25320816]

Du L, Zhang T, Liu K, Yan J, Yao X, Risacher SL, Saykin AJ, Han J, Guo L, Shen L, 2017 Identifying associations between brain imaging phenotypes and genetic factors via a novel structured scca approach. In: International Conference on Information Processing in Medical Imaging Springer, pp. 543–555.

Gao L, Cui Z, Shen L, Ji H-F, 2016 Shared genetic etiology between type 2 diabetes and alzheimers disease identified by bioinformatics analysis. Journal of Alzheimer's Disease 50 (1), 13–17.

Golub GH, Zha H, 1995 The canonical correlations of matrix pairs and their numerical computation. Linear Algebra for Signal Processing 69, 27–49.

Hardoon DR, Shawe-Taylor J, 2011 Sparse canonical correlation analysis. Machine Learning 83 (3), 331–353.

Hirono N, Mori E, Ishii K, Ikejiri Y, Imamura T, Shimomura T, Hashimoto M, Yamashita H, Sasaki M, 1998 Frontal lobe hypometabolism and depression in alzheimer's disease. Neurology 50 (2), 380–383. [PubMed: 9484357]

Jagust WJ, Bandy D, Chen K, Foster NL, Landau SM, Mathis CA, Price JC, Reiman EM, Skovronsky D, Koeppe RA, et al., 2010 The alzheimer's disease neuroimaging initiative positron emission tomography core. Alzheimer's & Dementia 6 (3), 221–229.

Jiji S, Smitha KA, Gupta AK, Pillai VPM, Jayasree RS, 2013 Segmentation and volumetric analysis of the caudate nucleus in alzheimer's disease. European journal of radiology 82 (9), 1525–1530. [PubMed: 23664648]

Kim S, Swaminathan S, Inlow M, Risacher SL, Nho K, Shen L, Foroud TM, Petersen RC, Aisen PS, Soares H, Toledo JB, Shaw LM, Trojanowski JQ, Weiner MW, McDonald BC, Farlow MR, Ghetti B, Saykin AJ, 2013 Influence of genetic variation on plasma protein levels in older adults using a multi-analyte panel. PLoS One 8 (7), e70269. [PubMed: 23894628]

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR, 2010 MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34 (8), 816–34. [PubMed: 21058334]

Lin D, Calhoun VD, Wang Y-P, 2014 Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. Medical image analysis 18 (6), 891–902. [PubMed: 24247004]

Parkhomenko E, Tritchler D, Beyene J, 2009 Sparse canonical correlation analysis with application to genomic data integration. Statistical applications in genetics and molecular biology 8 (1), 1–34.

Potkin SG, Guffanti G, Lakatos A, Turner JA, Kruggel F, Fallon JH, Saykin AJ, Orro A, Lupoli S, Salvi E, et al., 2009a Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for alzheimer's disease. PloS one 4 (8), e6501. [PubMed: 19668339]

Potkin SG, Turner JA, Guffanti G, Lakatos A, Torri F, Keator DB, Macciardi F, 2009b Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: methodological considerations. Cognitive neuropsychiatry 14 (4–5), 391–418. [PubMed: 19634037]

Ramanan VK, Risacher SL, Nho K, Kim S, Swaminathan S, et al., 2014 Apoe and bche as modulators of cerebral amyloid deposition: a florbetapir pet genome-wide association study. Molecular psychiatry 19 (3), 351–357. [PubMed: 23419831]

Saykin AJ, Shen L, Yao X, Kim S, Nho K, Risacher SL, Ramanan VK, Foroud TM, Faber KM, Sarwar N, et al., 2015 Genetic studies of quantitative mci and ad phenotypes in adni: Progress, opportunities, and plans. Alzheimer's & Dementia 11 (7), 792–814.

Shen L, Kim S, Risacher SL, Nho K, Swaminathan S, West JD, Foroud T, Pankratz N, Moore JH, Sloan CD, Huentelman MJ, Craig DW, DeChairo BM, Potkin SG, Jack CR, Weiner MW, Saykin

AJ, 2010 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 53 (3), 1051–63. [PubMed: 20100581]

Shen L, Thompson PM, Potkin SG, Bertram L, Farrer LA, Foroud TM, Green RC, Hu X, Huentelman MJ, Kim S, Kauwe JSK, Li Q, Liu E, Macciardi F, Moore JH, Munsie L, Nho K, Ramanan VK, Risacher SL, Stone DJ, Swaminathan S, Toga AW, Weiner MW, Saykin AJ, 2014 Genetic analysis of quantitative phenotypes in ad and mci: imaging, cognition and biomarkers. Brain imaging and behavior 8 (2), 183–207. [PubMed: 24092460]

Silver MJ, Montana G, Initiative ADN, 2012 Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. Statistical Applications in Genetics and Molecular Biology 11 (1), 1–43.

Vounou M, Nichols TE, Montana G, 2010 Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. Neuroimage 53 (3), 1147–1159. [PubMed: 20624472]

Witten DM, Tibshirani R, Hastie T, 2009 A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10 (3), 515–34. [PubMed: 19377034]

Witten DM, Tibshirani RJ, 2009 Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical applications in genetics and molecular biology 8 (1), 1–27.

Yan J, Du L, Kim S, Risacher SL, Huang H, Moore JH, Saykin AJ, Shen L, 2014 Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. Bioinformatics 30 (17), i564–i571. [PubMed: 25161248]

Yuan M, Lin Y, 2006 Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (1), 49–67.

Zeng Y, Nie C, Min J, Liu X, Li M, Chen H, Xu H, Wang M, Ni T, Li Y, 2016 Novel loci and pathways significantly associated with longevity. Scientific Reports 6, 21243. [PubMed: 26912274]

Highlights

- We present a novel fused penalty and a new graph-guided penalty.
- A novel structured SCCA model and optimization algorithm are proposed.
- Our method has a qualitative upper bound for the grouping effect.
- Our SCCA improves state-of-the-art methods with more reasonable canonical weights.

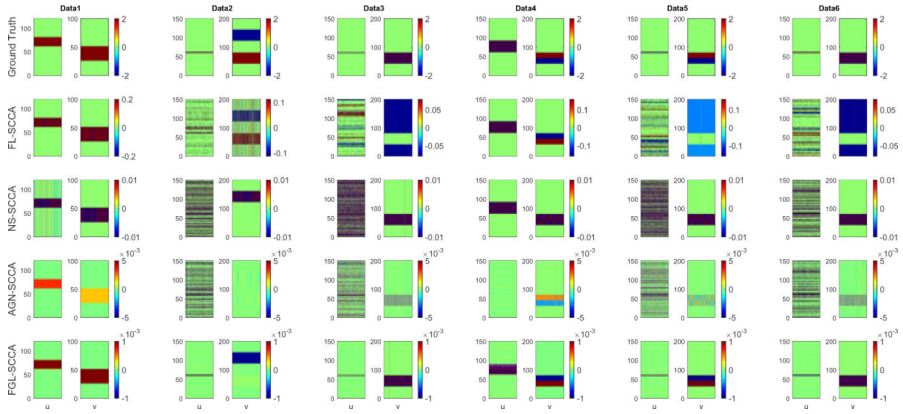


Figure 1:

Canonical weights estimated on synthetic data. The first row is the ground truth, and each remaining row corresponds to an SCCA method: (1) FL-SCCA, (2) NS-SCCA, (3) AGN-SCCA, and (4) FGL-SCCA. For each method, the estimated weights of **u** are shown on the left panel, and those of **v** are shown on the right. In each subfigure, the vertical axis represents the indices of each **u** (left panel) or **v** (right panel), and the horizontal axis represents 250 runs of experiments (50 times of 5-fold cross-validation).

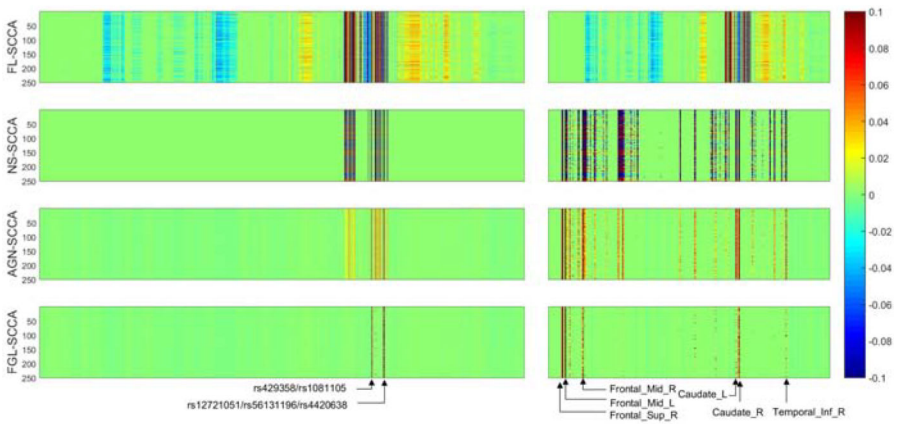


Figure 2:

Canonical weights estimated on real imaging genetics data set. Each row corresponds to a method: (1) FL-SCCA, (2) NS-SCCA, (3) AGN-SCCA, and (4) FGL-SCCA. For each method, the estimated weights of **u** are shown on the left panel, and those of **v** are shown on the right. In each subfigure, the horizontal axis represents the reference number of each individual SNP (left panel) or imaging ROI (right panel), and the vertical axis represents every run and there are 250 runs in total (50 times of 5-fold cross-validation).

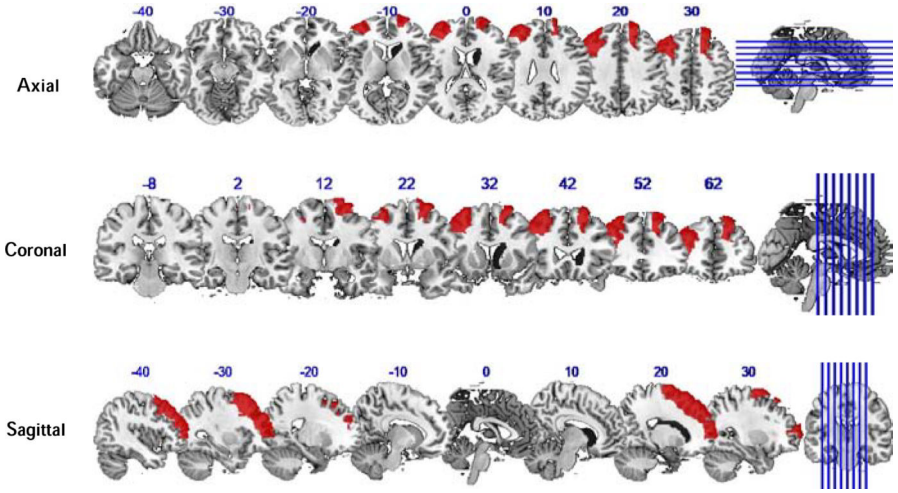


Figure 3:
Mapping averaged canonical weights **v** of FGL-SCCA onto the brain.

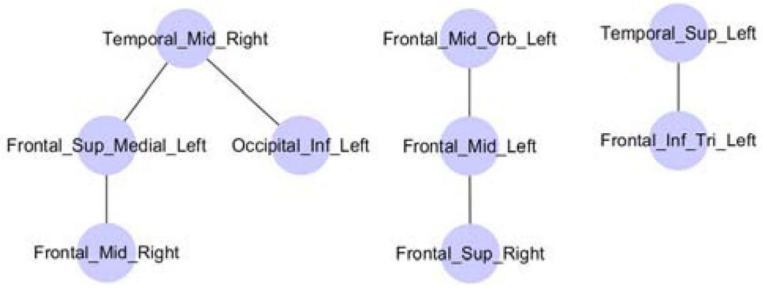


Figure 4:
Heat map of brain ROI-SNP associations of top selected markers.

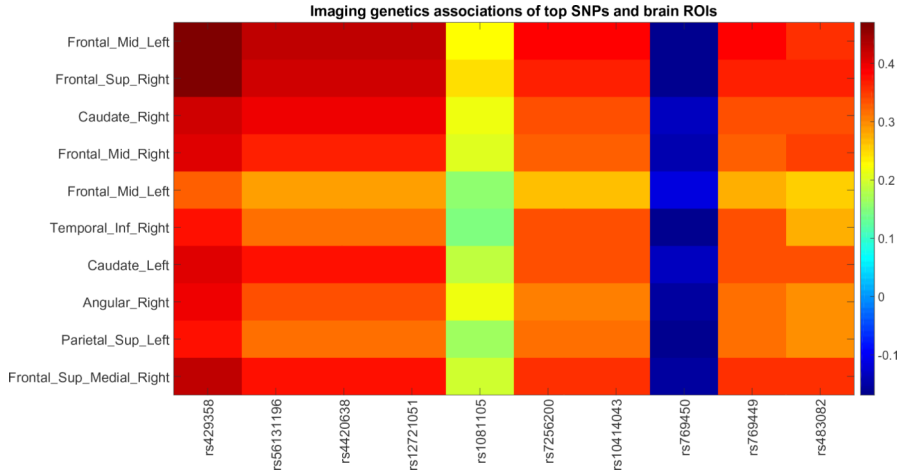
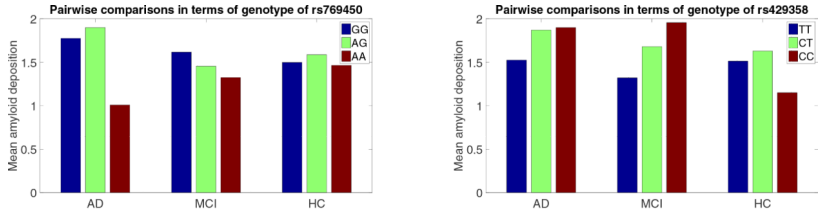


Figure 5:
Heat map of brain ROI-SNP associations of top selected markers.



(a) (b)

Figure 6:

Pairwise comparisons in terms of genotype of rs769450 and rs429358 within ADs, MCIs and HCs respectively. Two-way ANOVA was applied to examine the effects of rs769450 and baseline diagnosis on left middle frontal gyrus (a). Age, gender, education, handedness were included as covariates. The results of rs429358 were also shown for comparison (b).

Table 1:

The AUC (area under ROC curve) values (mean±std) of canonical weights are also shown.

	Area under ROC Curve (AUC): u			
	FL-SCCA	**NS-SCCA**	**AGN-SCCA**	**FGL-SCCA**
Data 1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
Data 2	0.86±0.26	1.00±0.00	0.99±0.02	1.00±0.00
Data 3	0.37±0.04	1.00±0.04	0.99±0.02	1.00±0.00
Data 4	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
Data 5	0.44±0.11	1.00±0.00	0.99±0.01	1.00±0.00
Data 6	0.96±0.04	1.00±0.00	0.98±0.03	1.00±0.00
Area under ROC Curve (AUC): v				
Data 1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
Data 2	0.78±0.41	0.75±0.00	1.00±0.00	1.00±0.00
Data 3	0.00±0.00	1.00±0.06	1.00±0.04	1.00±0.06
Data 4	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
Data 5	0.00±0.06	1.00±0.00	1.00±0.00	1.00±0.00
Data 6	0.00±0.00	1.00±0.00	0.97±0.10	1.00±0.00

Table 2:

Performance comparison on synthetic data. Training and testing correlation coefficients (mean±std.) of 5-fold cross-validation are shown for FL-SCCA, NS-SCCA, AGN-SCCA and FGL-SCCA. The best testing correlation coefficients with the smallest std. value are shown in boldface.

	Training Resuilts			
	FL-SCCA	**NS-SCCA**	**AGN-SCCA**	**FGL-SCCA**
Data 1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
Data 2	0.91±0.05	0.95±0.01	0.87±0.19	0.93±0.01
Data 3	0.76±0.03	0.96±0.01	0.89±0.19	0.95±0.01
Data 4	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
Data 5	0.76±0.04	0.96±0.01	0.92±0.14	0.95±0.01
Data 6	0.85±0.03	0.93±0.01	0.81±0.23	0.86±0.01
	Testing Results			
Data 1	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
Data 2	0.77±0.28	0.89±0.06	0.74±0.20	**0.92±0.04**
Data 3	0.34±0.19	0.91±0.06	0.80±0.20	**0.95±0.07**
Data 4	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
Data 5	0.23±0.18	0.94±0.04	0.86±0.15	**0.94±0.03**
Data 6	0.38±0.22	0.82±0.10	0.59±0.23	**0.86±0.06**

Table 3:

Participant characteristics.

	HC	MCI	AD
Num	196	343	28
Gender(M/F, %)	52.04/47.96	59.18/40.82	64.29/35.71
Handedness(R/L, %)	90.82/9.18	90.09/9.91	82.14/17.86
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 4:

Performance comparison on real data. Averaged training and testing correlation coefficients by 50 times 5-fold cross-validation are shown for FL-SCCA, NS-SCCA, AGN-SCCA and FGL-SCCA (mean±std). The best mean±std is shown in boldface. The *p*-values of FL-SCCA, NS-SCCA and AGN-SCCA compared with FGL-SCCA via Student's *t*-tests are also shown.

Method	Training Results	*p*-value	Testing Results	*p*-value
FL-SCCA	0.41±0.02	4.89E-201	0.35±0.08	1.25E-75
NS-SCCA	0.44±0.02	8.03E-164	0.42±0.07	1.39E-37
AGN-SCCA	0.47±0.02	2.37E-105	0.43±0.07	3.43E-21
FGL-SCCA	**0.49±0.02**	-	**0.45±0.07**	-

Table 5:

Runtime comparison on real data.

Method	FL-SCCA	NS-SCCA	AGN-SCCA	FGL-SCCA
time (sec.)	1.84	2.70	61.80	3.63

Table 6:

Top ten SNPs selected by averaged canonical weights. The *p*-value of ANOVA results were shown to indicate the statistical significance of the relevance of each SNP to diagnostic status, where age, gender, education, handedness were included as covariates.

RS_NO	Gene	Weight	Percent	*p*-value
rs429358	*APOE*	5.45E-01	100%	2.30E-06
rs56131196	*APOC1*	1.65E-01	100%	1.14E-03
rs4420638	*APOC1*	1.24E-01	100%	1.14E-03
rs12721051	*APOC1*	1.24E-01	100%	1.14E-03
rs1081105	*APOE*	6.55E-02	96.40%	1.73E-01
rs7256200	*APOE* (dist=3285), *APOC1* (dist=1642)	8.71E-03	76.40%	3.35E-05
rs10414043	*APOE* (dist=3061), *APOC1* (dist=2208)	7.35E-03	72.40%	3.35E-05
rs769450	*APOE*	−7.27E-03	63.20%	5.10E-02
rs769449	*APOE*	5.76E-03	64.80%	1.53E-05
rs483082	*APOE* (dist=3526), *APOC1* (dist=1743)	4.67E-03	66.40%	7.98E-05

Table 7:

Top ten brain imaging markers selected by averaged canonical weights. The *p*-value of ANOVA results were shown to indicate the statistical significance of the relevance of each brain imaging marker to diagnostic status, where age, gender, education, handedness were included as covariates.

Brain Region	Weight	Percent	*p*-value
Left middle frontal gyrus	4.07E-01	100.00%	6.23E-07
Right superior frontal gyrus	3.96E-01	100.00%	8.65E-07
Right caudate nucleus	1.05E-01	98.80%	4.34E-06
Right middle frontal gyrus	7.43E-02	98.00%	1.49E-04
Left middle frontal gyrus	3.38E-02	78.00%	8.26E-03
Right inferior temporal gyrus	2.23E-02	69.60%	6.44E-05
Left caudate nucleus	1.46E-02	86.40%	6.01E-08
Right angular gyrus	9.03E-03	52.00%	9.01E-06
Superior parietal lobule	7.52E-03	41.60%	8.86E-04
Right frontal superior medial gyrus	4.47E-03	59.20%	2.36E-09