

Published in final edited form as:

*Brain Inform Health (2015)*. 2015 ; 9250: 275–284.

# GN-SCCA: GraphNet based Sparse Canonical Correlation Analysis for Brain Imaging Genetics

**Lei Du**[1], **Jingwen Yan**[1], **Sungeun Kim**[1], **Shannon L. Risacher**[1], **Heng Huang**[2], **Mark Inlow**[3], **Jason H. Moore**[4], **Andrew J. Saykin**[1], and **Li Shen**[1,2,*] **for the Alzheimer's Disease Neuroimaging Initiative**[**]

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

[2]Computer Science and Engineering, University of Texas at Arlington, TX, USA

[3]Mathematics, Rose-Hulman Institute of Technology, IN, USA

[4]Biomedical Informatics, School of Medicine, University of Pennsylvania, PA, USA

## Abstract

Identifying associations between genetic variants and neuroimaging quantitative traits (QTs) is a popular research topic in brain imaging genetics. Sparse canonical correlation analysis (SCCA) has been widely used to reveal complex multi-SNP-multi-QT associations. Several SCCA methods explicitly incorporate prior knowledge into the model and intend to uncover the hidden structure informed by the prior knowledge. We propose a novel structured SCCA method using Graph constrained Elastic-Net (GraphNet) regularizer to not only discover important associations, but also induce smoothness between coefficients that are adjacent in the graph. In addition, the proposed method incorporates the covariance structure information usually ignored by most SCCA methods. Experiments on simulated and real imaging genetic data show that, the proposed method not only outperforms a widely used SCCA method but also yields an easy-to-interpret biological findings.

## 1 Introduction

Brain imaging genetics, which intends to discover the associations between genetic factors (e.g., the single nucleotide polymorphisms, SNPs) and quantitative traits (QTs, e.g., those extracted from neuroimaging data), is an emerging research topic. While single-SNP-single-QT association analyses have been widely performed [17], several studies have used regression techniques [9] to examine the joint effect of multiple SNPs on one or a few QTs. Recently, bi-multivariate analyses [6, 12, 7, 18], which aim to identify complex multi-SNP-multi-QT associations, have also received much attention.

*Correspondence to Li Shen (shenli@iu.edu).

**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/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Sparse canonical correlation analysis (SCCA) [14, 19], a type of bi-multivariate analysis, has been successfully used for analyzing imaging genetics data [12, 6], and other biology data [4, 5, 14, 19]. To simplify the problem, most existing SCCA methods assume that the covariance matrix of the data to be the identity matrix. Then the Lasso [14, 19] or group Lasso [6, 12] regularizer is often solved using the soft-thresholding method. Although this assumption usually leads to a reasonable result, it is worth pointing out that the relationship between those variables within either modality have been ignored. For neuroimaging genetic data, correlations usually exist among regions of interest (ROIs) in the brain and among linkage disequilibrium (LD) blocks in the genome. Therefore, simply treating the data covariance matrices as identity or diagonal ones will limit the performance of identifying meaningful structured imaging genetic associations.

Witten *et al.* [19, 20] proposed an SCCA method which employs penalized matrix decomposition (PMD) to yield two sparse canonical loadings. Lin *et al.* [12] extended Witten's SCCA model to incorporate non-overlapping group knowledge by imposing $l_{2;1}$-norm regularizer onto both canonical loadings. Chen *et al.* [3] proposed the ssCCA approach by imposing a smoothness penalty for one canonical loading of the taxa based on their relationship on the phylogenetic tree. Chen *et al.* [4, 5] treated the feature space as an undirected graph where each node corresponds to a variable and $r_{ij}$ is the edge weight between nodes $i$ and $j$. They proposed network based SCCA which penalizes the $l_1$ norm of $r_{ij}^2(u_i - sign(r_{ij})u_j)$ to encourage the weight values $u_i$ and $u_j$ to be similar if $r_{ij} > 0$, or dissimilar if $r_{ij} < 0$. A common limitation of these SCCA models is that they approximate $\mathbf{X}^\mathbf{T}\mathbf{X}$ by identity or diagonal matrix. Du *et al.* [7] proposed an S2CCA algorithm that overcomes this limitation, and requires users to explicitly specify non-overlapping group structures. Yan *et al.* [21] proposed KG-SCCA which uses $l_2$ norm of $r_{ij}^2(u_i - sign(r_{ij})u_j)$ to replace that in Chen's model [4, 5]. KG-SCCA also requires the structure information to be explicitly defined. Note that an inaccurate sign of $r_{ij}$ may introduce bias [10].

In this paper, we impose the Graph-constrained Elastic Net (GraphNet) [8] into SCCA model and propose a new GraphNet constrained SCCA (GN-SCCA). Our contributions are twofold: (1) GN-SCCA estimates the covariance matrix directly instead of approximating it by the identity matrix **I**; (2) GN-SCCA employs a graph penalty using data-driven technique to induce smoothness by penalizing the pairwise differences between adjacent features. Thorough experiments on both simulation and real imaging genetic data show that our method outperforms a widely used SCCA implementation [19][5] by identifying stronger imaging genetic associations and more accurate canonical loading patterns.

## 2 Preliminaries

### 2.1 Sparse CCA

We use the boldface lowercase letter to denote the vector, and the boldface uppercase letter to denote the matrix. The $i$-th row and $j$-th column of $\mathbf{M} = (m_{ij})$ are represented as $\mathbf{m}^i$ and

[5]SCCA in the PMA software package is widely used as a benchmark algorithm. Here we simply use SCCA to denote the SCCA method in this software package. See http://cran.r-project.org/web/packages/PMA/ for details.

$\mathbf{m}_j$. Let $\mathbf{X} = \{\mathbf{x}^1; \ldots; \mathbf{x}^n\} \subseteq \mathbb{R}^p$ be the SNP data and $\mathbf{Y} = \{\mathbf{y}^1; \ldots; \mathbf{y}^n\} \subseteq \mathbb{R}^p$ be the QT data, where *n*, *p* and *q* are the subject number, SNP number and QT number respectively.

The SCCA model presented in [19, 20] is as follows:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{X}^T\mathbf{Y}\mathbf{v} \quad s.t.\ \|\mathbf{u}\|_2^2 \leq 1,\ \|\mathbf{v}\|_2^2 \leq 1,\ \|\mathbf{u}\|_1 \leq c_1, \|\mathbf{v}\|_1 \leq c_2, \quad (1)$$

where the two terms $\|\mathbf{u}\|_2^2 \leq 1$ and $\|\mathbf{v}\|_2^2 \leq 1$ originate from the equalities $\|\mathbf{u}\|_2^2=1$ and $\|\mathbf{v}\|_2^2=1$, where $\|\mathbf{u}\|_2^2=1$ and $\|\mathbf{v}\|_2^2=1$ approximate $\|\mathbf{X}\mathbf{u}\|_2^2=1$ and $\|\mathbf{Y}\mathbf{v}\|_2^2=1$ to simplify computation. This simplification approximates the covariance matrices $\mathbf{X^TX}$ and $\mathbf{Y^TY}$ by the identity matrix **I** (or sometimes a diagonal matrix), assuming that the features are independent. Most SCCA methods employ this simplification [3–5, 12, 19, 20]. Besides, $\|\mathbf{u}\|_1 \leq c_1$ and $\|\mathbf{v}\|_1 \leq c_2$ induce sparsity to control the sparsity of canonical loadings. In addition to the Lasso ($l_1$-norm), the fused Lasso can also be used [5, 14, 19, 20].

### 2.2 Graph Laplacian

The Graph Laplacian, also called the Laplacian matrix, has been widely used in the spectral clustering techniques and spectral graph theory [2], owing to its advantage in clustering those correlated features automatically. We denote a weighted undirected graph as *G* = (*V; E; W*), where *V* is the set of vertices corresponding to features of **X** or **Y**, *E* is the set of edges with $e_{i,j}$ indicating that two features $\mathbf{v}_i$ and $\mathbf{v}_j$ are connected, and $w_{i,j}$ is the weight of edge $e_{i,j}$. Here we consider *G* as a complete graph and thus every two vertices are connected.

Formally, the adjacency matrix of *G* is defined as:

$$A(i,j) = \begin{cases} w_{i,j}, & \text{if } i \neq j, \text{and} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Generally, $w_{i,j}$ is set to $|r_{i,j}|^d$, where $r_{i,j}$ is the sample correlation between the *i*-th and *j*-th variables. In this work, for simplicity, we set *d* = 2, i.e. $w_{i,j}=r_{i,j}^2$. It can also be decided by domain experts in other applications.

Let **D** be a diagonal degree matrix with the following diagonal entries: $D(i, i) = \Sigma_j A(i, j)$. Then the Laplacian matrix **L** is defined as **L** = **D** – **A** [8]. **L** has many merits such as the symmetry and the positive semi-definite structure. Most importantly, it can map a weighted graph onto a new space such that connected vertices stay as close as possible.

## 3 GraphNet based SCCA (GN-SCCA)

Inspired by the Graph Laplacian [11] and the GraphNet [8] technique, we define the penalty $P(\mathbf{u})$ and $P(\mathbf{v})$ as follows:

$$\begin{aligned} P(\mathbf{u}) &= \|\mathbf{u}\|_{GN} = \mathbf{u}^T\mathbf{L}_1\mathbf{u} \leq c_1, \\ P(\mathbf{v}) &= \|\mathbf{v}\|_{GN} = \mathbf{v}^T\mathbf{L}_2\mathbf{v} \leq c_2. \end{aligned} \quad (3)$$

where $\mathbf{L}_1$ and $\mathbf{L}_2$ are the Laplacian matrices of two complete undirect graphs defined by the sample correlation matrices of the SNP and QT training data, respectively. The terms $\mathbf{u}^T\mathbf{L}_1\mathbf{u}$ and $\mathbf{v}^T\mathbf{L}_2\mathbf{v}$ make each feature fair be penalized smoothly according to the correlation between the two features.

Applying the two penalties above, the GN-SCCA model takes the form:

$$\min_{\mathbf{u},\mathbf{v}} -\mathbf{u}^T\mathbf{X}^T\boldsymbol{Y}\mathbf{v} \quad (4)$$

$$s.t. \|\mathbf{X}\mathbf{u}\|_2^2 \leq 1, \|\boldsymbol{Y}\mathbf{v}\|_2^2 \leq 1, P(\mathbf{u}) \leq c_1, P(\mathbf{v}) \leq c_2, \|\mathbf{u}\|_1 \leq c_3, \|\mathbf{v}\|_1 \leq c_4,$$

where the terms $\|\mathbf{u}\|_1 \leq c_3$ and $\|\mathbf{v}\|_1 \leq c_4$ are used to induce sparsity; and the $P(\mathbf{u})$ and $P(\mathbf{v})$ are Graph Laplcaian based GraphNet constraints [8]. Note that we use $\|\mathbf{X}\mathbf{u}\|_2^2 \leq 1$ instead of $\|\mathbf{X}\|_2^2 \leq 1$, which is typically used in other models, and thus our model takes into consideration the full covariance information.

Using Lagrange multiplier and writing the penalties into the matrix form, the objective function of GN-SCCA is as follows:

$$\mathscr{L}(\mathbf{u},\mathbf{v}) = -\mathbf{u}^T\mathbf{X}^T\boldsymbol{Y}\mathbf{v} + \frac{\lambda_1}{2}\|\mathbf{u}\|_{GN} + \frac{\lambda_2}{2}\|\mathbf{v}\|_{GN} + \frac{\beta_1}{2}\|\mathbf{u}\|_1 + \frac{\beta_2}{2}\|\mathbf{v}\|_1 + \frac{\gamma_1}{2}\|\mathbf{X}\mathbf{u}\|_2^2 + \frac{\gamma_2}{2}\|\boldsymbol{Y}\mathbf{v}\|_2^2 \quad (5)$$

where ($\lambda_1$, $\lambda_2$, $\beta_1$, $\beta_2$) are tuning parameters, corresponding to ($c_1$,$c_2$,$c_3$,$c_4$). Take the derivative regarding **u** and **v** separately and let them be zero:

$$(\lambda_1\mathbf{L}_1 + \beta_1\mathbf{D}_1 + \gamma_1\mathbf{X}^T\mathbf{X})\mathbf{u} = \mathbf{X}^T\boldsymbol{Y}\mathbf{v}, \quad (6)$$

$$(\lambda_2\mathbf{L}_2 + \beta_2\mathbf{D}_2 + \gamma_2\boldsymbol{Y}^T\boldsymbol{Y})\mathbf{v} = \boldsymbol{Y}^T\mathbf{X}\mathbf{u}, \quad (7)$$

where $\mathbf{D}_1$ is a diagonal matrix with the $k_1$-th element as $\frac{1}{2\|u^{k_1}\|_1}(k_1 \in [1,p])$, and $\mathbf{D}_2$ is a diagonal matrix with the $k_2$-th element as $\frac{1}{2\|v^{k_2}\|_1}(k_2 \in [1,q])$[6].

Since $\mathbf{D}_1$ relies on **u** and $\mathbf{D}_2$ relies on **v**, we introduce an iterative procedure to solve this objective. In each iteration, we first fix **v** and solve for **u**, and then fix **u** and solve for **v**. The procedure stops until it satisfies a predefined stopping criterion. Algorithm 1 shows the pseudocode of the GN-SCCA algorithm.

### 3.1 Convergence Analysis of GN-SCCA

We first introduce Lemma 1 described in [13].

[6] If $\|u^{k_1}\|_1 = 0$ or $\|v^{k_2}\|_1 = 0$, we approximate it with $\sqrt{\|u^{k_1}\|_2^2 + \varsigma}$ or $\sqrt{\|v^{k_2}\|_2^2 + \varsigma}$, where $\zeta$ is a very small non-zero value. According to [13], this regularization will not affect the result when $\zeta \to 0$.

**Algorithm 1** GraphNet based Structure-aware SCCA (GN-SCCA)

**Require:**

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

**Ensure:**

Canonical vectors **u** and **v**.

1: Initialize $\mathbf{u} \in \mathbb{R}^{p \times 1}$, $\mathbf{v} \in \mathbb{R}^{q \times 1}$; $\mathbf{L}_1 = D_u - A_u$ and $\mathbf{L}_2 = D_v - A_v$ only from the training data;

2: **while** not converged **do**

3: **while** not converged regarding **u** **do**

4: Calculate the diagonal matrix $\mathbf{D}_1$, where the $k_1$-th element is $\frac{1}{2\|u^{k_1}\|_1}$;

5: Update $\mathbf{u} = (\lambda_1 \mathbf{L}_1 + \beta_1 \mathbf{D}_1 + \gamma_1 \mathbf{X}^T\mathbf{X})^{-1}\, \mathbf{X}^T\mathbf{Y}\mathbf{v}$;

6: **end while**

7: **while** not converged regarding **v** **do**

8: Calculate the diagonal matrix $\mathbf{D}_2$, where the $k_2$-th element is $\frac{1}{2\|v^{k_2}\|_1}$;

9: Update $\mathbf{v} = (\lambda_2 \mathbf{L}_2 + \beta_2 \mathbf{D}_2 + \gamma_2 \mathbf{Y}^T\mathbf{Y})^{-1}\, \mathbf{Y}^T\mathbf{X}\mathbf{u}$;

10: **end while**

11: **end while**

12: Scale **u** so that $\|\mathbf{X}\mathbf{u}\|_2 = 1$;

13: Scale **v** so that $\|\mathbf{Y}\mathbf{v}\|_2 = 1$.

**Lemma 1—**The following inequality holds for any two nonzero vectors $\tilde{\mathbf{u}}$, **u** with the same length,

$$\|\tilde{\mathbf{u}}\|_2 - \frac{\|\tilde{\mathbf{u}}\|_2^2}{2\|\mathbf{u}\|_2} \leq \|\mathbf{u}\|_2 - \frac{\|\mathbf{u}\|_2^2}{2\|\mathbf{u}\|_2}. \quad (8)$$

**Lemma 2—**For any real number $\tilde{u}$ and any nonzero real number u, we have

$$\|\tilde{u}\|_1 - \frac{\|\tilde{u}\|_1^2}{2\|u\|_1} \leq \|u\|_1 - \frac{\|u\|_1^2}{2\|u\|_1}. \quad (9)$$

**Proof:** The proof is obvious, given Lemma 1, $\|\tilde{u}\|_1 = \|\tilde{u}\|_2$ and $\|u\|_1 = \|u\|_2$.

**Theorem 1—**In each iteration, Algorithm 1 decreases the value of the objective function till the algorithm converges.

**Proof:** The proof consists of two phases. (1) Phase 1: For Steps 3–6, **u** is the only variable to estimate. The objective function Eq. (5) is equivalent to

$$\mathscr{L}(\mathbf{u},\mathbf{v}) = -\mathbf{u}^T\mathbf{X}^T\boldsymbol{Y}\mathbf{v} + \frac{\lambda_1}{2}\|\mathbf{u}\|_{GN} + \frac{\beta_1}{2}\|\mathbf{u}\|_1 + \frac{\gamma_1}{2}\|\mathbf{X}\mathbf{u}\|_2^2$$

From Step 5, we denote the updated value as $\tilde{\mathbf{u}}$. Then we have

$$\begin{aligned} &-\tilde{\mathbf{u}}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\boldsymbol{Y}\mathbf{v} + \lambda_1\tilde{\mathbf{u}}^{\mathbf{T}}\mathbf{L}_1\tilde{\mathbf{u}} + \beta_1\tilde{\mathbf{u}}^{\mathbf{T}}\mathbf{D}_1\tilde{\mathbf{u}} + \gamma_1\tilde{\mathbf{u}}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\tilde{\mathbf{u}} \\ &\leq -\mathbf{u}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\boldsymbol{Y}\mathbf{v} + \lambda_1\mathbf{u}^{\mathbf{T}}\mathbf{L}_1\mathbf{u} + \beta_1\mathbf{u}^{\mathbf{T}}\mathbf{D}_1\mathbf{u} + \gamma_1\mathbf{u}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{u} \end{aligned}$$

According to the definition of $\mathbf{D_1}$, we obtain

$$\begin{aligned} &-\tilde{\mathbf{u}}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\boldsymbol{Y}\mathbf{v} + \lambda_1\tilde{\mathbf{u}}^{\mathbf{T}}\mathbf{L}_1\tilde{\mathbf{u}} + \beta_1\sum_{k_1}\frac{\|\tilde{u}^{k_1}\|_1^2}{2\|u^{k_1}\|_1} + \gamma_1\tilde{\mathbf{u}}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\tilde{\mathbf{u}} \\ &\leq -\mathbf{u}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\boldsymbol{Y}\mathbf{v} + \lambda_1\mathbf{u}^{\mathbf{T}}\mathbf{L}_1\mathbf{u} + \beta_1\sum_{k_1}\frac{\|u^{k_1}\|_1^2}{2\|u^{k_1}\|_1} + \gamma_1\mathbf{u}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{u} \end{aligned} \quad (10)$$

Then summing Eq. (9) and Eq. (10) on both sides, we obtain

$$-\tilde{\mathbf{u}}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\boldsymbol{Y}\mathbf{v} + \lambda_1\tilde{\mathbf{u}}^{\mathbf{T}}\mathbf{L}_1\tilde{\mathbf{u}} + \beta_1\|\tilde{\mathbf{u}}\|_1 + \gamma_1\|\mathbf{X}\tilde{\mathbf{u}}\|_2^2 \leq -\mathbf{u}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\boldsymbol{Y}\mathbf{v} + \lambda_1\mathbf{u}^{\mathbf{T}}\mathbf{L}_1\mathbf{u} + \beta_1\|\mathbf{u}\|_1 + \gamma_1\|\mathbf{X}\mathbf{u}\|_2^2$$

Let $\lambda_1^*=2\lambda_1$, $\gamma_1^*=2\gamma_1$, $\beta_1^*=2\beta_1$, we arrive at

$$-\tilde{\mathbf{u}}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\boldsymbol{Y}\mathbf{v} + \frac{\lambda_1^*}{2}\|\tilde{\mathbf{u}}\|_{GN} + \frac{\beta_1^*}{2}\|\tilde{\mathbf{u}}\|_1 + \frac{\gamma_1^*}{2}\|\mathbf{X}\tilde{\mathbf{u}}\|_2^2 \leq -\mathbf{u}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\boldsymbol{Y}\mathbf{v} + \frac{\lambda_1^*}{2}\|\mathbf{u}\|_{GN} + \frac{\beta_1^*}{2}\|\mathbf{u}\|_1 + \frac{\gamma_1^*}{2}\|\mathbf{X}\mathbf{u}\|_2^2. \quad (11)$$

Thus, the objective value decreases during Phase 1: $\mathscr{L}(\tilde{\mathbf{u}}, \mathbf{v}) \leq \mathscr{L}(\mathbf{u}, \mathbf{v})$.

(2) Phase 2: For Steps 7–10, $\mathbf{v}$ is the variable to estimate. Similarly, we have

$$-\tilde{\mathbf{u}}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\boldsymbol{Y}\tilde{\mathbf{v}} + \frac{\lambda_2^*}{2}\|\tilde{\mathbf{v}}\|_{GN} + \frac{\beta_2^*}{2}\|\tilde{\mathbf{v}}\|_1 + \frac{\gamma_2^*}{2}\|\boldsymbol{Y}\tilde{\mathbf{v}}\|_2^2 \leq -\tilde{\mathbf{u}}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\boldsymbol{Y}\mathbf{v} + \frac{\lambda_2^*}{2}\|\mathbf{v}\|_{GN} + \frac{\beta_2^*}{2}\|\mathbf{v}\|_1 + \frac{\gamma_2^*}{2}\|\boldsymbol{Y}\mathbf{v}\|_2^2 \quad (12)$$

Thus, the objective value decreases during Phase 2: $\mathscr{L}(\tilde{\mathbf{u}}, \tilde{\mathbf{v}}) \leq \mathscr{L}(\tilde{\mathbf{u}}, \mathbf{v})$.

Applying the transitive property of inequalities, we obtain $\mathscr{L}(\tilde{\mathbf{u}}, \tilde{\mathbf{v}}) \leq \mathscr{L}(\mathbf{u}, \mathbf{v})$. Therefore, Algorithm 1 decreases the objective function in each iteration.

We set the stopping criterion of Algorithm 1 as $\max\{|\delta| \mid \delta \in (\mathbf{u}_{t+1}-\mathbf{u}_t)\} \leq \tau$ and $\max\{|\delta| \mid \delta \in (\mathbf{v}_{t+1}-\mathbf{v}_t)\} \leq \tau$, where $\tau$ is a desirable estimate error. In this paper, $\tau = 10^{-5}$ is empirically chosen in the experiments.

## 4 Experimental Results

### 4.1 Results on Simulation Data

We used four simulated data sets to compare the performances of GN-SCCA and a widely used SCCA implementation [19]. We applied two different methods to generate these data

with distinct structures to assure diversity. The first two data sets (both with $n = 1000$ and $p = q = 50$, but with different built-in correlations) were generated as follows: 1) We created a random positive definite group structured covariance matrix **M**. 2) Data set **Y** with covariance structure **M** was calculated by Cholesky decomposition. 3) Data set **X** was created similarly. 4) Canonical loadings **u** and **v** were created so that the variables in one group share the same weights based on the group structures of **X** and **Y** respectively. 5) The portion of the specified group in **Y** were replaced based on the **u**, **v**, **X** and the assigned correlation. The last two data sets (with different $n$, $p$, $q$ and built-in correlations) were created using the simulation procedure described in [5]: 1) Predefined structure information was used to create **u** and **v**. 2) Latent vector $z$ was generated from $N(\mathbf{0}, \mathbf{I}_{n\times n})$. 3) **X** was created with each $\mathbf{x}_i \sim N(z_i\mathbf{u}, \mathbf{I}_{p\times p})$ and **Y** with each $\mathbf{y}_i \sim N(z_i\mathbf{v}, \Sigma_y)$ where

$$\left(\sum\nolimits_y\right)_{jk} = exp^{-|v_j - v_k|}.$$

According to Eqs. (6–7), six parameters need to be decided for GN-SCCA. Here we choose the value of tuning range based on two considerations: 1) Chen and Liu [4] showed that the results were insensitive to $\gamma_1$ and $\gamma_2$ in a similar study; 2) The major difference between traditional CCA and SCCA is the penalty terms. Thus their results will be the same if small parameters are used. With this observation, we tune $\gamma_1$ and $\gamma_2$ from small range of [1,10,100], and tune the remaining ones from $10^{-1}$ to $10^3$ through **nested** 5-fold cross-validation.

The true signals and estimated **u** and **v** are shown in Fig. 1. The estimated canonical loadings **u** and **v** of GN-SCCA were consistent with the ground truth on all simulated data sets, while SCCA only found an incomplete portion of the true signals. Shown in Table 1 are the cross-validation performances of the two methods. The left part of the table shows that GN-SCCA outperformed SCCA consistently and significantly, and it has better test accuracy than SCCA on testing data. The right part of Table 1 presents the area under ROC (AUC), where GN-SCCA also significantly outperformed SCCA on all data sets. These results demonstrated that GN-SCCA identifies the correlations and signal locations more accurately and more stably than SCCA.

### 4.2 Results on Real Neuroimaging Genetics Data

We used the real neuroimaging and SNP data downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database to assess the performances of GN-SCCA and SCCA. One goal of ADNI is to test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. Please see www.adni-info.org for more details.

Both the SNP and MRI data were downloaded from the LONI website (adni.loni.usc.edu). There are 204 healthy control (HC), 363 MCI and 176 AD participants. The structural MRI scans were processed with voxel-based morphometry (VBM) in SPM8 [1, 15]. Briefly, scans were aligned to a T1-weighted template image, segmented into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) maps, normalized to MNI space, and

smoothed with an 8mm FWHM kernel. We subsampled the whole brain and yielded 465 voxels spanning all brain ROIs. These VBM measures were pre-adjusted for removing the effects of the baseline age, gender, education, and handedness by the regression weights derived from HC participants. We investigated SNPs from the top 5 AD risk genes [16] and APOE e4. In total we have 379 SNPs in this study. Our task was to examine correlations between the voxels (GM density measures) and genetic biomarker SNPs.

Shown in Table 2 are the 5-fold cross-validation results of GN-SCCA and SCCA. GN-SCCA significantly and consistently outperformed SCCA in terms of identifying stronger correlations from the training data. For the testing performance, SCCA did not do well possibly due to over-fitting, while GN-SCCA consistently outperformed SCCA. Fig. 2 shows the heat maps of the trained canonical loadings learned from cross-validation. We could observe that both weights, i.e. **u** and **v**, estimated by GN-SCCA were quite sparse and presented a clear pattern which could be easier to interpret. However, SCCA identified many signals which could be harder to explain. The strongest genetic signal, identified by GN-SCCA, was the APOE e4 SNP rs429358; and the strongest imaging signals came from the hippocampus. They were negatively correlated with each other. This reassures that our method identified a well-known correlation between APOE and hippocampal morphometry in an AD cohort. These results show the capability of GN-SCCA to identify biologically meaningful imaging genetic associations.

## 5 Conclusions

We proposed a GraphNet constrained SCCA (GN-SCCA) to mine imaging genetic associations, and incorporated the covariance information ignored by many existing SCCA methods. The GraphNet term induces smoothness by penalizing the pairwise differences between adjacent features in a complete graph or an user-given graph (correlation matrix used in this study). Our experimental study showed that GN-SCCA accurately discovered the true signals from the simulation data and obtained improved performance and biologically meaningful findings from real data. In this work, we only did comparative study between GN-SCCA and a widely-used SCCA method [19]. We have observed many recent developments in structured SCCA models. Some (e.g., [6, 18, 19, 14, 12]) ignored the covariance structure information of the input data, which was usually helpful to imaging genetics applications. A few other models (e.g., [7, 21]) overcome this limitation but impose different sparsity structures. Work is in progress to compare the proposed GN-SCCA with these structured SCCA models. Given the mathematically simple formulation of GN-SCCA, we feel it is a valuable addition which is complementary to the existing SCCA models.

## Acknowledgments

This work was supported by NIH R01 LM011360, U01 AG024904 (details available at http://adni.loni.usc.edu), RC2 AG036535, R01 AG19771, P30 AG10133, and NSF IIS-1117335 at IU, by NSF CCF-0830780, CCF-0917274, DMS-0915228, and IIS-1117965 at UTA, and by NIH R01 LM011360, R01 LM009012, and R01 LM010098 at Dartmouth.

## References


1. Ashburner J, Friston KJ. Voxel-based morphometry–the methods. Neuroimage. 2000; 11(6):805–21. [PubMed: 10860804]
2. Belkin, M.; Niyogi, P. Proceedings of the 18th annual conference on Learning Theory. Springer-Verlag; 2005. Towards a theoretical foundation for laplacian-based manifold methods; p. 486-500.
3. Chen J, Bushman FD, et al. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. Biostatistics. 2013; 14(2):244–258. [PubMed: 23074263]
4. 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.
5. Chen X, Liu H, Carbonell JG. Structured sparse canonical correlation analysis. International Conference on Artificial Intelligence and Statistics. 2012
6. Chi E, Allen G, et al. Imaging genetics via sparse canonical correlation analysis. Biomedical Imaging (ISBI), 2013 IEEE 10th Int Sym on. 2013:740–743.
7. Du L, et al. A novel structure-aware sparse learning algorithm for brain imaging genetics. International Conference on Medical Image Computing and Computer Assisted Intervention. 2014:329–336.
8. Grosenick L, et al. Interpretable whole-brain prediction analysis with graphnet. NeuroImage. 2013; 72:304–321. [PubMed: 23298747]
9. Hibar DP, Kohannim O, et al. Multilocus genetic analysis of brain images. Front Genet. 2011; 2:73. [PubMed: 22303368]
10. Kim S, Xing EP. Statistical estimation of correlated genome associations to a quantitative trait network. PLoS genetics. 2009; 5(8)
11. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008; 24(9):1175–1182. [PubMed: 18310618]
12. Lin D, Calhoun VD, Wang YP. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. Med Image Anal. 2013
13. Nie F, Huang H, Cai X, Ding CH. Efficient and robust feature selection via joint 2, 1-norms minimization. Advances in Neural Information Processing Systems. 2010:1813–1821.
14. Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. Statistical Applications in Genetics and Molecular Biology. 2009; 8(1): 1–34.
15. Risacher SL, Saykin AJ, et al. Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. Curr Alzheimer Res. 2009; 6(4):347–61. [PubMed: 19689234]
16. Shah RD, Samworth RJ. Variable selection with error control: another look at stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2013; 75(1):55–80.
17. Shen L, Kim S, 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]
18. Vounou M, Nichols TE, Montana G. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. NeuroImage. 2010; 53(3): 1147–59. [PubMed: 20624472]
19. 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]
20. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical applications in genetics and molecular biology. 2009; 8(1):1–27.
21. Yan J, Du L, et al. Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. Bioinformatics. 2014; 30(17):i564–i571. [PubMed: 25161248]

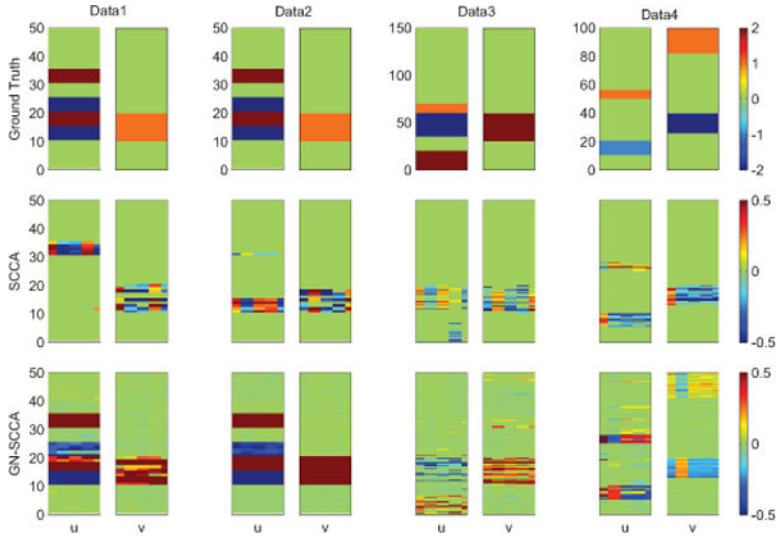


**Fig. 1.**

Comparisons on estimated canonical loadings using 5-fold cross-validation on synthetic data. The ground truth (the top row), SCCA results (the middle row) and GN-SCCA results (the bottom row) are all shown. For each panel pair, the 5 estimated **u**'s are shown on the left panel, and the 5 estimated **v**'s are shown on the right.

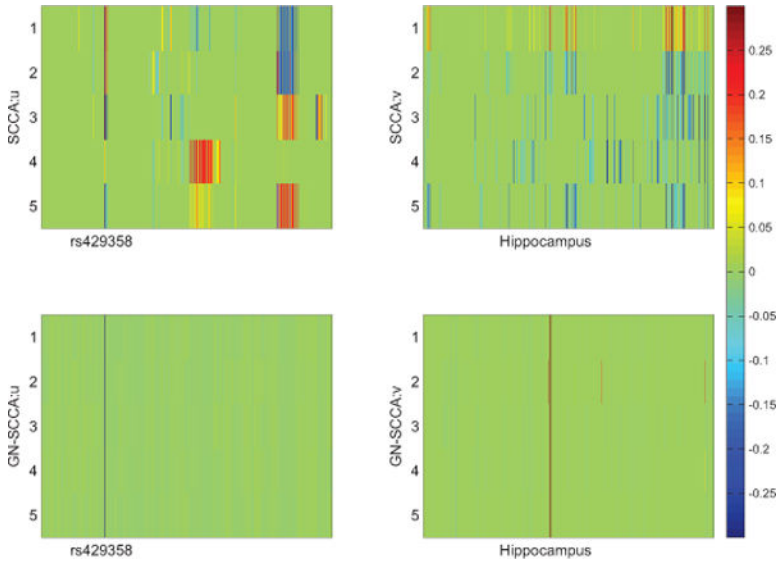


**Fig. 2.**

Comparisons on estimated canonical loadings using 5-fold cross-validation on real data. The SCCA results (the top row) and GN-SCCA results (the bottom row) are shown. For each panel pair, the 5 estimated **u**'s are shown on the left panel, and the 5 estimated **v**'s are shown on the right.

**Table 1**

5-fold nested cross-validation results on synthetic data: Mean±std. is shown for estimated correlation coefficients and AUC regarding the estimated canonical loadings. *p*-values of paired t-test between GN-SCCA and SCCA are also shown.

| | Correlation Coefficient (CC) | | | Area under ROC (AUC) | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| True CC | SCCA | GN-SCCA | *p* | SCCA:u | GN-SCCA:u | *p* | SCCA:v | GN-SCCA:v | *p* |
| Data1(0.80) | 0.48±0.03 | 0.80±0.01 | 1.52E-05 | 0.65±0.02 | 1.00±0.00 | 1.44E-06 | 0.81±0.04 | 1.00±0.00 | 2.65E-04 |
| Data2(0.90) | 0.56±0.04 | 0.90±0.01 | 8.38E-06 | 0.66±0.01 | 1.00±0.00 | 3.15E-07 | 0.79±0.04 | 1.00±0.00 | 1.79E-04 |
| Data3(0.92) | 0.55±0.15 | 0.89±0.06 | 1.54E-03 | 0.67±0.01 | 0.89±0.04 | 2.49E-04 | 0.81±0.04 | 1.00±0.00 | 2.23E-04 |
| Data4(0.98) | 0.97±0.01 | 0.98±0.01 | 6.82E-02 | 0.89±0.05 | 0.98±0.03 | 3.45E-03 | 0.69±0.01 | 1.00±0.00 | 1.88E-07 |

**Table 2**

5-fold nested cross-validation results on real data: The models learned from training data were used to estimate the correlation coefficients for both training and testing cases. *p*-values of paired t-tests between GN-SCCA and SCCA are shown.

| | SCCA | | | | | | GN-SCCA | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Correlation coefficients** | **F1** | **F2** | **F3** | **F4** | **F5** | **mean±std.** | **F1** | **F2** | **F3** | **F4** | **F5** | **mean±std.** | ***P*** |
| Training | 0.22 | 0.23 | 0.24 | 0.20 | 0.21 | 0.22± 0.02 | 0.28 | 0.27 | 0.28 | 0.26 | 0.27 | 0.27± 0.01 | 2.25E-4 |
| Testing | 0.07 | 0.04 | 0.09 | 0.05 | 0.16 | 0.07± 0.03 | 0.21 | 0.28 | 0.24 | 0.31 | 0.27 | 0.26± 0.04 | 9.14E-4 |