Canonical Correlation Analysis of Imaging Genetics Data Based on Statistical Independence and Structural Sparsity

Yipu Zhang¹, Peng Peng¹, Yongfeng Ju¹, Gang Li¹, Vince D. Calhoun² [Fellow, IEEE], Yu-Ping Wang³,* [Senior Member, IEEE]
¹Yipu Zhang, Peng Peng, Yongfeng Ju and Gang Li are with the school of Electronics and Control Engineering, Chang’an University, Xi’an, Shaanxi, 710049, China.
²Vince D. Calhoun is with the Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS) (Georgia State University, Georgia Institute of Technology, Emory University), Atlanta, GA 30030, and also with the Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM 87131.
³Yu-Ping Wang is with Department of Biomedical Engineering, Tulane University, New Orleans, LA, 70118, USA.

Abstract

Current developments of neuroimaging and genetics promote an integrative and compressive study of schizophrenia. However, it is still difficult to explore how gene mutations are related to brain abnormalities due to the high dimension but low sample size of these data. Conventional approaches reduce the dimension of dataset separately and then calculate the correlation, but ignore the effects of the response variables and the structure of data. To improve the identification of risk genes and abnormal brain regions on schizophrenia, in this paper, we propose a novel method called Independence and Structural sparsity Canonical Correlation Analysis (ISCCA). ISCCA combines independent component analysis (ICA) and Canonical Correlation Analysis (CCA) to reduce the collinear effects, which also incorporate graph structure of the data into the model to improve the accuracy of feature selection. The results from simulation studies demonstrate its higher accuracy in discovering correlations compared with other competing methods. Moreover, applying ISCCA to a real imaging genetics dataset collected by Mind Clinical Imaging Consortium (MCIC), a set of distinct gene-ROI interactions are identified, which are verified to be both statistically and biologically significant.

Keywords

Canonical correlation analysis; Imaging Genetics; Statistical independence; Structural sparsity

I. INTRODUCTION

SCHIZOPHRENIA is a complex mental disease caused by both brain abnormalities and genetic mutations [1, 2]. Genetic data such as single nucleotide polymorphisms (SNPs) are
used to identify risk genes associated with schizophrenia, while brain imaging data such as functional magnetic resonance imaging (fMRI) are used to locate abnormal brain regions in patients with schizophrenia (SZ) [3–5]. Combining imaging with genetic data for association analysis fosters a new field of imaging genetics [6,7].

Canonical Correlation Analysis (CCA) [8] is one of the most popular algorithms in imaging genetics, which has been used to identify bi-multivariate associations between SNPs and fMRI in schizophrenia patient [9–11]. Due to the tremendous dimension of genomics and whole-brain imaging data, CCA methods often suffer from an overfitting issue. To address this issue, Witten et al. [12, 13] proposed the sparse canonical correlation analysis (SCCA), which integrated $L_1$ and $L_2$ regularization terms with CCA model. In recent study, some approaches utilized the prior information or incorporated the structure information of dataset to enhance the performance. For example, Lin et al. combined genetic and imaging group information with CCA and proposed group sparse CCA model [14] Some works combined genetic and imaging spatial structure information to design the structure sparse CCA [4, 15, 16]. However, these CCA-based methods only use second-order statistics (e.g., correlation and covariance) which will generate different latent variables (LVs). These LVs may be difficult to explain in practical applications [17].

Independent component analysis (ICA) [18, 19], also called blind source separation (BSS), are likely CCA and has become widely adopted tools for imaging genetics data analysis [20–23]. This approach is based on the notion that it is insufficient to consider only up to the second-order statistics for obtaining a unique LV model, especially the real neurophysiological data do not strictly follow multivariate Gaussian distributions. ICA attempts to find mutually statistically independent sources that are linearly superimposed to constitute multivariate data. Since independence is a much stronger condition than uncorrelatedness, ICA generally employs higher-order statistics to obtain unique solutions [24]. In recent studies, Calhoun et al. proposed several ICA-based methods, i.e., group ICA used the groups of brain regions as prior information to analysis brain fMRI data [20]. As a multimodal extension, joint ICA can analyze multitask fMRI data by the common sharing mixing matrix [21]. Liu et al. [22] developed parallel ICA (pICA) to identify intrinsic interrelationships between SNP and fMRI. On the basis of pICA, Vergara et al. [23] proposed three-way pICA to analyze fMRI, structural magnetic resonance imaging (sMRI) and SNP data. Lu et al. [25] proposed an automated ICA-based method to identify language network in brain tumor subjects. However, these ICA-based methods decompose datasets separately which ignore the effects of the response variables, and the independent components obtained by ICA are still poorly interpretable.

In this paper, inspired by both ICA and CCA, we proposed a new model that considers both correlation and independence, called Independence and Structural sparsity Canonical Correlation Analysis (ISCCA). ISCCA takes advantage of statistical independence and the structure of datasets, aiming to find the maximum correlation between risk genes and abnormal brain ROIs. ISCCA employs the independency constraint to measure the statistical independence among features, which can overcome the ambiguity of features selected by CCA. In addition, ISCCA imposes the structure information as the regularization term to improve the accuracy of feature selection.
To validate effectiveness and efficiency of our proposed model, we used both simulated data and real data to evaluate the performance of our method and compared it with several other methods, i.e., L₁-SCCA [12], FL-SCCA [13], AGN-SCCA [15] and Group SCCA [26]. Experimental results on the simulation data show that our proposed model yields improved performance in comparison to the competing methods under the metrics of correlation coefficient. Meanwhile, we applied ISCCA to the real SNP and fMRI data collected by the Mind Clinical Imaging Consortium and identified 11 risk genes and 15 brain ROIs having high correlations with schizophrenia.

The rest of the paper is organized as follows: In Section II, we reviewed the basic model and proposed our method. In Section III, we test ISCCA on both simulated and real datasets. Section IV summarizes the work of this paper.

II. METHODS

2.1 Materials

Canonical correlation analysis (CCA) is a classical method for multivariate statistical analysis proposed by Hotelling [8], which has been widely used in the field of pattern recognition and machine learning. In this paper, we use the random vectors \( x \in \mathbb{R}^{p \times 1} \) and \( y \in \mathbb{R}^{q \times 1} \) to represent the features of \( n \) subjects in two different datasets \( X \in \mathbb{R}^{n \times p} \) and \( Y \in \mathbb{R}^{n \times q} \), i.e., SNP and fMRI. The purpose of the CCA algorithm is to find two most correlated canonical vectors \( u \in \mathbb{R}^{p \times 1} \) and \( v \in \mathbb{R}^{q \times 1} \) by maximizing the correlation between linear combinations of features in \( u^T x \) and \( v^T y \). We describe the overall correlation coefficient (CC) by calculating the Pearson correlation coefficient as follows:

\[
\text{corr}(u^T x y^T v) = \frac{u^T x y^T v}{\sqrt{u^T x x^T u / v^T y y^T v}}
\]  

(1)

In imaging genetics study, the dimensions of genomic and imaging data are larger than the number of samples \( (p, q \gg n) \), that is, we have the high dimensional low sample size problem. L₁-norm regularization is commonly used to avoid overfitting, which assigns insignificant features to zero. To extract the significant features, Graph-constrained Elastic-Net (GraphNet) was proposed by employing a sparse graph in the L₂-norm penalty term [27]. Based on GraphNet, the regularization terms were proposed as follows:

\[
\begin{align*}
\|u\|_{GN} &= \lambda_1 |u^T L_1 u| + \zeta_1 \|u\|_1 \\
\|v\|_{GN} &= \lambda_2 |v^T L_2 v| + \zeta_2 \|v\|_1 
\end{align*}
\]  

(2)

where \( L_1 \) and \( L_2 \) are the Laplacian matrices of two different datasets, \( \lambda \) and \( \zeta \) are tuning parameters. According to Eq. (1) and Eq. (2), the objective function is then formulated as:

\[
\max_u u^T X Y^T v \\
\text{s.t.} \|X u\|_2^2 \leq 1, \|Y v\|_2^2 \leq 1, \|u\|_G \leq c_1, \|v\|_G \leq c_2
\]  

(3)

where \( c_1 \) and \( c_2 \) are the parameters to control sparsity.
### 2.2 The Proposed Method

Aiming to maximize the correlation while minimize the collinear effects of the features in one data, we hope to extract the unrelated LVs generated by independent sources. According to Ref. [28], these approximations are based on the maximum-entropy principle, which are formulated as:

\[
\begin{align*}
\max_u & \left[ E(G(u^T x)) - E(G(w_1)) \right]^2 \\
\text{s.t.} & \quad u^T x x^T u = 1
\end{align*}
\]  

(4)

and

\[
\begin{align*}
\max_v & \left[ E(G(v^T y)) - E(G(w_2)) \right]^2 \\
\text{s.t.} & \quad v^T y y^T v = 1
\end{align*}
\]  

(5)

where \(w_1\) and \(w_2\) are standard Gaussian variables. \(G(\cdot)\) is defined as a non-quadratic function:

\[
G(w) = \log \cosh(w)
\]  

(6)

To maximize the correlation between two datasets with the consideration of statistical independence, an intuitive way is to combine these three sub-objectives into one single objective function and achieve a good tradeoff between their respective goals. In addition, we employ the structure of data as prior in the penalty terms, and then obtain the following objective function:

\[
\begin{align*}
\max_{u,v} & \left[ a[u^T x y^T v] + \beta [E(G(u^T x)) - E(G(w_1))]^2 \\
& + \gamma [E(G(v^T y)) - E(G(w_2))]^2 \right] \\
\text{s.t.} & \quad u^T x x^T u = 1, \quad v^T y y^T v = 1, \quad \|u\|_{GN} \leq c_1, \quad \|v\|_{GN} \leq c_2
\end{align*}
\]  

(7)

where \(a, \beta\) and \(\gamma\) are the weights satisfying \(a + \beta + \gamma = 1\). For the input datasets \(X\) and \(Y\), Eq. (7) can be rewritten as:

\[
\begin{align*}
\min_{u,v} & \left[ a[u^T X Y^T v] - \beta \sum_{i=1}^{n} [G(u^T x_i) - G(w_i)]^2 \\
& - \gamma \sum_{i=1}^{n} [G(v^T y_i) - G(w_i)]^2 \right] \\
\text{s.t.} & \quad u^T x x^T u = 1, \quad v^T y y^T v = 1, \quad \|u\|_{GN} \leq c_1, \quad \|v\|_{GN} \leq c_2
\end{align*}
\]  

(8)

where \(x_i\) and \(y_i\) are the \(i\)-th row of matrices \(X\) and \(Y\), \(w_i\) is the \(i\)-th element of Gaussian random vector \(w \sim \mathcal{N}(0, 1)\).

### 2.3 Weight Adjustment

In Eq. (8), we combine the sub-objectives and optimize them simultaneously. The overall optimal solution depends on the relative values of the weights, i.e., \(a, \beta\) and \(\gamma\). To achieve
a good tradeoff between different sub-objectives, here we employ the following strategy to determine the weights.

To avoid that one sub-objective may overwhelm the others, we take two aspects in consideration: the different scales and the different convergence speeds of the sub-objectives. The weights $\alpha$, $\beta$ and $\gamma$ are decomposed into three factors, $\alpha = \alpha_{\text{sig}} \cdot \alpha_{\text{scale}} \cdot \alpha_{\text{adj}}$, $\beta = \beta_{\text{sig}} \cdot \beta_{\text{scale}} \cdot \beta_{\text{adj}}$ and $\gamma = \gamma_{\text{sig}} \cdot \gamma_{\text{scale}} \cdot \gamma_{\text{adj}}$. Here, $\alpha_{\text{sig}}$, $\beta_{\text{sig}}$ and $\gamma_{\text{sig}}$ called significant factors indicating the relative significance of each sub-objective, which can be set according to the specific application. Here, $\alpha_{\text{scale}}$, $\beta_{\text{scale}}$ and $\gamma_{\text{scale}}$ are scale factors to unify the scale of the sub-objectives that are defined as follows.

$$
\alpha_{\text{scale}} = \frac{1}{\|H_1(u, v)\|_2}, \quad \beta_{\text{scale}} = \frac{1}{\|H_2(u)\|_2}, \quad \gamma_{\text{scale}} = \frac{1}{\|H_3(v)\|_2}
$$

where the functions in the denominators are corresponding to each sub-objective defined as:

$$
H_1(u, v) = \left(u^TXY^Tv\right)^2 + k_{11}(u^Tu - 1) + k_{12}(v^Tv - 1)
$$

$$
H_2(u) = \sum_{i=1}^{n} \left[G(u^Txi) - G(w_i)\right]^2 + k_2(u^Tu - 1)
$$

$$
H_3(v) = \sum_{i=1}^{n} \left[G(v^Ty_i) - G(w_i)\right]^2 + k_3(v^Tv - 1)
$$

where $k_{11}$, $k_{12}$, $k_2$ and $k_3$ are the Lagrange multipliers. $\alpha_{\text{adj}}$, $\beta_{\text{adj}}$ and $\gamma_{\text{adj}}$ are adjustable factors used to control different convergence rates and can be iteratively updated. Here we use the gradient function to estimate the convergence rates and define $\alpha_{\text{adj}}$, $\beta_{\text{adj}}$ and $\gamma_{\text{adj}}$ as:

$$
\alpha_{\text{adj}} = \frac{1}{\|\nabla_u H_1(u, v)\|_2 + \|\nabla_v H_1(u, v)\|_2}/2
$$

$$
\beta_{\text{adj}} = \frac{1}{\|\nabla_u H_2(u)\|_2}/2
$$

$$
\gamma_{\text{adj}} = \frac{1}{\|\nabla_v H_3(v)\|_2}
$$

where $\|\cdot\|_2$ in the denominator denotes $L_2$-norm regularization. The calculation of the first partial derivative in Eq. (9) are derived as follows.
\[ \nabla_u H_1(u, v) = 2(u^TXY^T)v - 2u^T \left( u^TXY^T \right) u \]
\[ \nabla_v H_1(u, v) = 2(v^TXu)Y^T - 2v^T \left( v^TXu \right) v \]
\[ \nabla_u H_2(u) = 2 \left( \sum_{i=1}^{n} \left[ (G(u^T x_i) - G(u_i))^2 \right] \right) u \]
\[ \nabla_v H_3(v) = 2 \left( \sum_{i=1}^{n} \left[ (G(v^T y_i) - G(v_i))^2 \right] \right) v \]

From Eq. (9), we find that the adjustable factor is inversely proportional to the gradient norm. That is, the gradients will be close to zero if the iteration results reach optimum. Thus, these three weights are set as:

\[
\begin{align*}
\alpha &= \alpha_{\text{sig}} \cdot \alpha_{\text{scale}} \cdot \alpha_{\text{adj}} \leq \tau \\
\beta &= \beta_{\text{sig}} \cdot \beta_{\text{scale}} \cdot \beta_{\text{adj}} \leq \tau \\
\gamma &= \gamma_{\text{sig}} \cdot \gamma_{\text{scale}} \cdot \gamma_{\text{adj}} \leq \tau \\
\alpha &= \alpha_{\text{sig}} \cdot \alpha_{\text{scale}} \cdot \alpha_{\text{adj}} > \tau \\
\beta &= \beta_{\text{sig}} \cdot \beta_{\text{scale}} \cdot \beta_{\text{adj}} > \tau \\
\gamma &= \gamma_{\text{sig}} \cdot \gamma_{\text{scale}} \cdot \gamma_{\text{adj}} > \tau
\end{align*}
\]

where \( \tau \) is a predefined positive threshold that constraint a small range around optimum.

When iterative searching route enters this range, the corresponding gradient approaches zero which means the weights are determined only by significant factors and scale factors.

### 2.4 Optimization Algorithm

To solve Eq. (8), we use Lagrange multiplier method and rewrite the objective function as:

\[
\mathcal{L}(u, v, \Psi) = -\alpha u^TXY^Tv - \beta \sum_{i=1}^{n} [G(u^T x_i) - G(u_i)]^2 + \frac{\lambda_1}{2} u^T L_1 u + \zeta_1 \|u\|_1 + \frac{\mu_1}{2} \|Xu\|_2^2 + \frac{\lambda_2}{2} v^T L_2 v + \zeta_2 \|v\|_1 + \frac{\mu_2}{2} \|Yv\|_2
\]

where \( \Psi = \{\lambda_1, \zeta_1, \mu_1, \lambda_2, \zeta_2, \mu_2\} \) denote control parameters of the respective regularizes.

Clearly, by taking the derivative of Eq. (12) with respect to \( u \) and \( v \), respectively. We obtain...
\[ \frac{\partial \mathcal{L}}{\partial u} = -\alpha X^T Y v \]
\[ -2\beta \sum_{i=1}^{n} \left[ (G(u^T x_i) - G(u_i)) \tanh(u^T x_i) x_i \right] + \left( \lambda_1 \hat{D}_1 + \zeta_1 D_1 + \mu_1 X^T X \right) u \]

\[ \frac{\partial \mathcal{L}}{\partial v} = -\alpha Y^T X u \]
\[ -2\gamma \sum_{i=1}^{n} \left[ (G(v^T y_i) - G(v_i)) \tanh(v^T y_i) y_i \right] + \left( \lambda_2 \hat{D}_2 + \zeta_2 D_2 + \mu_2 Y^T Y \right) v \]

where \( \hat{D}_1 \in \mathbb{R}^{p \times p} \) is a diagonal matrix with the \( r_1 \)-th element as \( \frac{1}{|L_1^r|} \), \( r_1 \in [1, p] \), \( L_1^r \) denotes the \( r_1 \)-th row of the Laplacian matrix \( L_1 \), and \( D_1 \in \mathbb{R}^{p \times p} \) is a diagonal matrix with the \( r_1 \)-the diagonal element as \( 1/(2|u_{r_1}|) \). Similarly, \( \hat{D}_2 \in \mathbb{R}^{q \times q} \) is a diagonal matrix with the \( r_2 \)-th element as \( \frac{1}{|L_2^r|} \), \( r_2 \in [1, q] \), \( L_2^r \) denotes the \( r_2 \)-th row of the Laplacian matrix \( L_2 \), and \( D_2 \in \mathbb{R}^{q \times q} \) is a diagonal matrix with the \( r_2 \)-the diagonal element as \( 1/(2|v_{r_2}|) \). To avoid \( |u_{r_1}| = 0 \), we normally utilize a small positive value \( \xi \) as \( |u_{r_1}| \approx 1/\sqrt{|u_{r_1}^2 + \xi|} \).

According to Eq. (13) and (14), we alternately update \( u \) and \( v \) by using the gradient descent method. Then we have the update equations

\[ u^{t+1} = u^t - \phi \frac{\partial \mathcal{L}}{\partial u} \]

\[ v^{t+1} = v^t - \phi \frac{\partial \mathcal{L}}{\partial v} \]

where \( t \) represents the \( t \)-th iteration and \( \phi = 0.1 \). 1 is the step size. In each iteration, we update \( u \) by fixing \( v \), and then update \( v \) by fixing \( u \). The iteration terminates when \( u \) and \( v \) satisfy \( |u^{t+1} - u^t| \leq 10^{-6} \) and \( |v^{t+1} - v^t| \leq 10^{-6} \).

**Algorithm 1.**

```
Input: \( X \in \mathbb{R}^{n \times p} \) and \( Y \in \mathbb{R}^{n \times q} \), parameters \( \Psi = \{ \lambda_1, \zeta_1, \mu_1, \lambda_2, \zeta_2, \mu_2 \} \).
Output: Canonical vectors \( u \) and \( v \).
1: Initialization: \( u, v, L_1 \) and \( L_2 \)
2: for \( t = 1 \) to \( \text{Max-Iteration} \) do
3: Update \( u \) by Eq. (15)
4: Update \( v \) by Eq. (16)
5: Until Convergence
6: end for
```
III. Results

3.1 Simulations

3.1.1 Data Generation—In order to validate the performance of our method, we generated four different datasets with different properties. As a simulation of the high dimensional low sample size problem, we set the number of samples be smaller than the number of features. Meanwhile, to ensure the diversity of datasets, the ground trues and the strengths of correlation coefficients within these data were distinct. We set the number of samples be $n = 80, 100, 100$ and $100$, the number of features in $X$ be $p = 100, 250, 500$ and $10000$, and the number of features in $Y$ be $q = 120, 600, 900$ and $12000$. Table I summarizes the dimensions of four simulated datasets.

The generation procedure was similar to that in Ref. [15, 16] with the following steps: 1) We generated $u$ and $v$ separately according to the prior information. 2) We randomly generated a latent vector $z \sim \mathcal{N}(0, I_{n \times n})$ and normalize it to unit length. 3) We generated $X$ with each sample $x_i \sim \mathcal{N}(z_i u, \Sigma_x)$, where $(\Sigma_x)_{ij} = \exp^{-|u_i - u_j|}$, and $Y$ with each sample $y_i \sim \mathcal{N}(z_i v, \Sigma_y)$, where $(\Sigma_y)_{ij} = \exp^{-|v_i - v_j|}$. In Specific, in the first simulated dataset, we set $u$ be a vector of length $p = 100$ with the elements from $0th$ to $40th$ be $-2$, from $80th$ to $90th$ be $1.5$ and the rest elements be $0$; we also generated $v$ with $q = 120$, with the elements of $v$ from $10th$ to $20th$ be $-4$, from $40th$ to $50th$ be $3$, from $110th$ to $120th$ be $1.5$ and the rest elements be $0$. We constructed other three datasets in the same way as follows: Let $u$ be a vector of length $p = 250, 500$ and $10000$, and $v$ be a vector of length $p = 600, 900$ and $12000$, respectively. For the second dataset, we set the elements of $u$ from $150th$ to $200th$ be $1$; the elements of $v$ from $100th$ to $200th$ be $-3$ and from $300th$ to $400th$ be $-2$. For the third dataset, the elements of $u$ from $300th$ to $400th$ are $1$; the elements of $v$ from $150th$ to $300th$ are $-3$ and from $450th$ to $600th$ are $-2$. For the fourth dataset, let the elements of $u$ from $1000th$ to $3000th$ be $1$, from $5000th$ to $6000th$ be $2$ and from $7000th$ to $9000th$ be $-2$; the elements of $v$ from $0th$ to $2000th$ be $-1$, from $3000th$ to $4000th$ be $2$ and from $8000th$ to $9000th$ be $-2$. The true correlation coefficients of the four datasets were $0.58, 0.62, 0.65$ and $0.88$, respectively. The ground truth of $u$ and $v$ in each simulated dataset can be observed in the first row of Fig. 1.

3.1.2 Parameters Selection—According to Eq. (16) and (17), we had six parameters $\lambda_1$, $\zeta_1$, $\mu_1$, $\lambda_2$, $\zeta_2$ and $\mu_2$ to be tuned. It is time consuming through a grid search, so we employed a strategy based on the observation of CCA. Since the major difference between CCA and our method is the penalty terms, using small parameters will yield the similar results for both ISCCA and CCA, and too large parameters may cause over-penalize for ISCCA. Here, we tuned these parameters within a neither too large nor too small range [15], i.e., $\Psi \in \{10^{-2}, 10^{-1}, 10^0, 10^1, 10^2\}$. All the parameters were tuned by a 5-fold cross-validation.
\[ CV(\Psi) = \frac{1}{5} \sum_{j=1}^{5} \text{corr}(X^\wedge u_-, Y^\wedge v_-) \]  

(20)

where \(X^\wedge\) and \(Y^\wedge\) are testing sets, and \(u_-\) and \(v_-\) are the canonical vectors obtained from training set. We selected the parameters which satisfy \(\text{argmax} \ CV(\lambda, \zeta, \mu)\) as the tuned optimal parameters. That is, the complete input data was randomly partitioned into five disjoint subsets of equal size; each subset was successively selected as the testing set, and the other four subsets were used for training.

3.1.3 Simulation Results—In our experiments, we compared the performance of the proposed ISCCA and four other competing methods: (1) \(L_1\)-SCCA (CCA with lasso); (2) FL-SCCA (CCA with fused lasso); (3) AGN-SCCA (sparse CCA with the absolute value based GraphNet) (4) Group SCCA (the group sparse CCA). We listed the parameters used for these methods in Appendix.

Table II shows the true correlation coefficients \((CC)\) and the estimated ones obtained by five methods from the testing set of four simulated datasets. As we can see from Table II, ISCCA consistently outperformed the other methods. We used the boldface to highlight the best value. Specifically, ISCCA achieved the equal value with the true \(CC\) in data 2, and obtained the best \(CCs\) of 0.56, 0.64 and 0.87 for data 1, data 3 and data 4, which were only slightly smaller than the true ones. For the compared methods, \(L_1\)-SCCA obtained the equal \(CC\) 0.64 with ISCCA in data 3, Group SCCA got the equal \(CC\) 0.87 with ISCCA for data 4. The 2\(^{nd}\)-best performance was obtained by AGN-SCCA for data 1 (0.55), and \(L_1\)-SCCA for data 2 (0.61) and data 3 (0.64). As bold italic shown in Table II, Group SCCA obtained the 3\(^{rd}\)-best performance for data 2 (0.64) and data 3 (0.70), even if it achieved the highest \(CCs\) (exceeding the true ones 0.02 and 0.05). It is obviously that ISCCA has the closest \(CC\) to the true ones for four datasets, i.e., 0.02, 0, 0.01 and 0.01, respectively.

Besides, we showed the estimated canonical vectors of five methods in Fig. 1. We can observe that \(L_1\)-SCCA, Group SCCA and ISCCA estimate a higher value of \(u\) and \(v\) (from \(-3\) to 3) than FL-SCCA (from \(-2\times10^{-5}\) to \(2\times10^{-5}\)) and AGN-SCCA (from \(-5\times10^{-5}\) to \(5\times10^{-5}\)). In comparison of \(L_1\)-SCCA and Group SCCA, ISCCA can accurately recover the true signals. Especially for the fourth dataset, our proposed method estimated a stronger \(u\) and \(v\) than \(L_1\)-SCCA and Group SCCA, the estimated values of the latter two methods were shown in color approaching zero. Moreover, comparing with \(L_1\)-SCCA, FL-SCCA and AGN-SCCA, ISCCA had a more stable results for the high dimensional low sample size problem when using 5-fold cross validation. Five testing results of ISCCA had some slight changes rather than generating only one result by AGN-SCCA. Clearly, Group SCCA has the most stable results in five methods, but ISCCA can estimate better values and had only small changes in each fold of cross validations, which showed a good trade-off between stability and accuracy.

In summary, our proposed method not only estimated the most accurate correlation coefficients, but also recovered the signals with the best accuracy in all simulations. The experiment results revealed that our method outperformed the competing methods, showing
that statistical independence and the structure-based regularization terms indeed improved the model performance.

### 3.2 Application to a Schizophrenia Data

#### 3.2.1 Data Preprocessing—
We applied our method to SNP and fMRI data collected by The Mind Clinical Imaging Consortium (MCIC). The data was from 208 subjects, among them 92 are schizophrenia patients (age: 34± 11, 22 females) and 116 healthy controls (age: 32±11, 44 females). We followed the same preprocessing procedures as in Ref. [1, 14], resulting in 41236 voxels and 777635 SNPs. In specific, fMRI data were collected during a sensorimotor task, a block-design motor response to auditory stimulation. The images were acquired on a Siemens3T Trio Scanner and 1.5 T Sonata with echo-planar imaging (EPI) sequences taking parameters (TR = 2000ms, TE = 30ms (3.0 T) /40ms (1.5 T), field of view = 22cm, slice thickness = 4mm, 1 mm skip, 27 slices, acquisition matrix = 64x64, flip angle = 90°). Data were pre-processed with SPM5 (http://www.fil.ion.ucl.ac.uk/spm) and were realigned, spatially normalized and resliced to 3×3×3mm, smoothed with a 10×10×10mm³ Gaussian kernel, and analyzed by multiple regression considering the stimulus and their temporal derivatives plus an intercept term as repressors. Finally, the stimulus-on versus stimulus-off contrast images were extracted with 53×63×46 voxels and all the voxels with missing measurements were excluded. 116 brain ROIs (regions of interest) were extracted based on the AAL brain atlas [29], which resulted in 41236 voxels left for analysis.

For SNP data, a blood sample was obtained for each participant and DNA was extracted. Genotyping for all participants was performed at the Mind Research Network using the Illumina Infinium HumanOmnil-Quad assay covering 1140419 SNP loci. Bead Studio was used to make the final genotype calls. PLINK software package (http://pngu.mgh.harvard.edu/~purcell/plink) was used to perform a series of standard quality control procedures, resulting in the final dataset spanning 777635 SNP loci. Each SNP was categorized into three clusters based on their genotype and was represented with discrete numbers: 0 for ‘BB’ (no minor allele), 1 for ‘AB’ (one minor allele) and 2 for ‘AA’ (two minor alleles). SNP with >20% missing data were deleted and missing data were further imputed. SNPs with minor allele frequency <1% were removed. To reflect the influence of genetic variation on brain behavior, SNPs included in top 75 schizophrenia genes listed on the SZ Gene database (http://www.szgene.org/) were selected for the analysis. This procedure yielded 3082 SNPs, which were annotated with 74 genes. There was no SNP found in the remaining one gene.

Then, to further reduce the data dimension, the voxels with the mean response less than 0.3 were removed while the SNPs included by the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway were selected [30], resulting in finally a dataset with 8891 voxels and 129145 SNPs.

#### 3.2.2 Statistical significance—
We applied our ISCCA method to analyze the correlation between SNP and fMRI datasets. In the real data experiment, 208 subjects were randomly divided into two subsets: 150 subjects for training and the remaining 58 subjects for testing. We used 5-fold cross validation to tune the optimal parameters on training
data, and the canonical vectors were estimated from the training data using the optimal parameters. Then these estimated vectors were applied to test set to select features. Here we performed random sampling from 208 subjects repeatedly for \( B = 50 \) times, selected the same proportion of subjects for training and test sets. Assume that the SNPs and voxels selected more frequently are more valuable for detecting the relationship between two datasets. We estimated \( \tilde{u}_b \) and \( \tilde{v}_b \) for each sample \( (b = 1, 2, \ldots, B) \) and measured the feature importance according to Ref. [14]:

\[
p_j = 1/B \sum_{b=1}^{B} I(\tilde{u}_j^b \neq 0)
\]

where \( \tilde{u}_j^b \) (or \( \tilde{v}_j^b \)) denotes the coefficient of the canonical vector \( u \) (or \( v \)) corresponding to \( j \)-th SNP (or voxel), and \( I(\cdot) \) is the indicator function. The features were then selected by a cut-off threshold 0.3. With the selected features, we measured the strength and significance of the correlation between gene and ROI by the hypothetical test. The null hypothesis of no correlation between a gene and a ROI is written as

\[
H_0 = \rho_{11} = \rho_{12} = \ldots = \rho_{ck} = \rho_{s_is_j} = 0
\]

versus the alternative hypothesis

\[
H_A: \exists c, k > 0, \rho_{ck} \neq 0
\]

where \( s_i \) and \( s_j \) are the number of SNPs and voxels in the \( i \)-th Gene and the \( j \)-th ROI, and \( \rho_{ck} \) is the correlation between the \( c \)-th SNP and the \( k \)-th voxel. To test the hypothesis, we first used the canonical vectors to calculate the pair-wise correlation between the SNPs from the \( i \)-th Gene and voxels from the \( j \)-th ROI, and averaged the correlation of each SNP-voxel pair to avoid the varying size of gene or ROI [14] as follows.

\[
\rho_{s_is_j} = \frac{1}{s_is_j s_j} \sum_{c=1}^{s_i} \sum_{k=1}^{s_j} |\rho_{ck}|
\]

We then evaluated the significance of the correlation via comparing \( \rho_{ij} \) with the null statistics \( \rho_{ij}^0 \) with 10000 times permutations of the samples.

\[
p_{\text{value}} = \frac{\{\rho_{ij}^0 \geq \rho_{ij}; b = 1, 2, \ldots, B\}}{10000}
\]

### 3.2.3 Experimental Results—
Based on the statistical measurement above, we finally summarized 15 SNPs from 11 genes linked with 533 voxels from 15 ROIs \((p\text{-value} < 0.05)\), the average correlation was \(0.1956\pm0.0569\) (mean+std). Table III lists the gene-ROI connections obtained by ISCCA (the index of ROI is defined by the AAL template).
Specially, BMP6, CDH19, CSMD1, CXCR4 and SYNE1 are correlated with ROI 7 with ROI 63, 13, 7, 51 and 89 ($\rho = 0.1352, 0.1105, 0.2125, 0.1421$ and $0.1725$), respectively. CD1D is associated with ROI 1, 7, 11, 62, 64, 89 and 90 ($\rho = 0.1734, 0.2099, 0.1541, 0.1893, 0.1912, 0.1410$ and $0.1843$). ETA1A is related to ROI 11, 57, 58, 61, 62, 63 and 64 ($\rho = 0.1734, 0.2099, 0.1541, 0.1893, 0.1912, 0.1410$ and $0.1843$). FRMD4A is correlated with ROI 7, 19, 58, 61, 64, 86 and 89 ($\rho = 0.2218, 0.2278, 0.2184, 0.1907, 0.2571, 0.2553$ and $0.1609$). FLJ46361 is correlated with ROI 1, 7, 11, 19, 57, 58, 61, 63, 64, 86, 89 and 90 ($\rho = 0.2819, 0.2281, 0.2056, 0.2841, 0.2162, 0.2637, 0.2663, 0.3222, 0.2885, 0.2552, 0.2681$ and $0.2760$). GRIK4 is linked to ROI 1, 7, 11, 63 and 64 ($\rho = 0.1101, 0.1752, 0.0607$ and $0.2047$). ZNF536 is connected with ROI 11, 51, 61 and 90 ($\rho = 0.1449, 0.1951, 0.1172$ and $0.2168$).

Previous works reported these risk genes were correlated with schizophrenia. Lin et al., [31] discovered that BMP6 played a role in the selective impairments on sustained attention of schizophrenia. Ref. [32, 33] showed that CXCR4 may represent a common downstream mediator in the pathophysiology of schizophrenia and related mental conditions. CSMD1 was found closely related to schizophrenia by Ref. [34, 35]. Ref [36–38] reported that GRIK4 locus were strongly associated with both schizophrenia and bipolar disorder. Xu et al., [39] indicated CSMD1 and SYNE1 were associated with complex neuropsychiatric disorders. Ref [40, 41] revealed that TMEM38B and ZNF536 supported a central role of calcium-signaling in the pathogenesis of schizophrenia. In addition, CD1D [42], ETA1A [43], FRMD4A [44], CDH19 [45] and FLJ46361 [46] were also mentioned related to schizophrenia.

Fig. 2 displays the main brain regions related to genes. The number indicates the index of corresponding ROI in the AAL template, and the volume size of the sphere represents the number of voxels in each ROI. From Fig. 2, we can observe that the ROI 7, 57 and 58 are located at Middle frontal gyrus and Postcentral gyrus. These critical regions have been widely studied and are shown with significant evidence in schizophrenia [47]. ROI 1, 11, 13, 51, 86, 89 and 90 are located at Precental gyrus, Inferior frontal gyrus, Middle occipital gyrus and Inferior temporal gyrus which can be seen as the potential biomarkers for schizophrenia [48,49]. Moreover, ROI 19, 61, 62, 63 and 64 are located at Supplementary motor, Parietal, Inferior parietal and Supramarginal gyrus. Many studies showed these regions have potential associations with schizophrenia [50–52].

Finally, we utilized the selected SNPs for gene set over-representation analysis by ConsensusPathDB [53]. Table IV summarizes four main Gene ontology (GO) terms obtained by 11 genes CDH19, CSMD1, CXCR4, CRIK4, SYNE1, BMP6, CD1D, ETA1A, FLJ46361, FRMD4A and ZNF536. These four Go terms are related to neural activity enriched with p-value less than 0.01, including regulation of nervous system development, positive regulation of nervous system development, generation of neurons and neurogenesis.

### IV. Conclusion

In this paper, we proposed a novel CCA-based model integrating statistical independence and structural sparsity to explore the relationship between brain abnormalities and...
genetic mutations in schizophrenia. In our proposed model, we employed the statistical independence, which could overcome the ambiguity of features selected by CCA and reduce the collinear effects. Besides, we design the regularization terms incorporating graphical structure within the data to improve the accuracy of feature selection and overcome the overfitting issue. We validated the effectiveness of our proposed model on both simulated and real data. The experimental results of simulated data demonstrated that our method achieved superior performance in CC compared with other competing ones. Moreover, after applying ISCCA to MCIC data we discovered some abnormal interactions between 11 risk genes and 15 brain regions associated with schizophrenia.

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.

**Acknowledgments**

This work was supported in part by the National Institutes of Health under Grant R01GM109068, Grant R01MH104680, Grant R01MH107354, R01AR059781, R01EB006841, R01EB005846, and Grant P20GM103472, in part by the National Science Foundation (NSF), in part by the Fundamental Research Funds for the Central Universities, Chang' an University (CHD) NO. 300102329102, in part by the Natural Science Foundation of Shaanxi NO. 2019JM-536 and in part by the China Scholarship Council NO. 201806565009.

**References**


Fig. 1.
Canonical vectors estimated on four simulated datasets. The four columns correspond from Data 1 to Data 4. For each dataset, the estimated value of $u$ is shown on the left, and $v$ is on the right. Each row corresponds to: (1) Ground truth. (2) $L_1$-SCCA. (3) FL-SCCA. (4) AGN-SCCA. (5) Group SCCA. (6) ISCCA
Fig. 2.
The brain regions related to genes.
### TABLE I

DETAILS OF THE SIMULATED DATASETS

<table>
<thead>
<tr>
<th>Dataset</th>
<th>n</th>
<th>p</th>
<th>q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data1</td>
<td>80</td>
<td>100</td>
<td>120</td>
</tr>
<tr>
<td>Data2</td>
<td>100</td>
<td>250</td>
<td>600</td>
</tr>
<tr>
<td>Data3</td>
<td>100</td>
<td>500</td>
<td>900</td>
</tr>
<tr>
<td>Data4</td>
<td>100</td>
<td>10000</td>
<td>12000</td>
</tr>
<tr>
<td>Methods</td>
<td>Data 1 CC = 0.58 (mean±std)</td>
<td>Data 2 CC = 0.62 (mean±std)</td>
<td>Data 3 CC = 0.65 (mean±std)</td>
</tr>
<tr>
<td>-----------</td>
<td>----------------------------</td>
<td>----------------------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>$L_1$-SCCA</td>
<td>0.45±0.25</td>
<td>0.61±0.16</td>
<td>0.64±0.06</td>
</tr>
<tr>
<td>FL-SCCA</td>
<td>0.54±0.18</td>
<td>0.59±0.14</td>
<td>0.56±0.18</td>
</tr>
<tr>
<td>AGN-SCCA</td>
<td>0.55±0.10</td>
<td>0.59±0.17</td>
<td>0.53±0.12</td>
</tr>
<tr>
<td>Group SCCA</td>
<td>0.40±0.16</td>
<td>0.64±0.14</td>
<td>0.70±0.08</td>
</tr>
<tr>
<td>ISCCA</td>
<td>0.56±0.10</td>
<td>0.62±0.11</td>
<td>0.64±0.04</td>
</tr>
</tbody>
</table>

*std denotes the standard deviation.*
## Table III

**The gene-ROI interactions discovered by ISCCA**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Brain region</th>
<th>The ROI's index</th>
<th>$\rho$ (correlation value)</th>
<th>The number of SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMP6</td>
<td>SupraMarginal_L</td>
<td>63</td>
<td>0.1352</td>
<td>2</td>
</tr>
<tr>
<td>CDH19</td>
<td>Frontal_Inf_Tri_L</td>
<td>13</td>
<td>0.1105</td>
<td>2</td>
</tr>
<tr>
<td>CSMD1</td>
<td>Frontal_Mid_L</td>
<td>7</td>
<td>0.2125</td>
<td>1</td>
</tr>
<tr>
<td>CXCR4</td>
<td>Occipital_Mid_L</td>
<td>51</td>
<td>0.1421</td>
<td>1</td>
</tr>
<tr>
<td>SYNE1</td>
<td>Temporal_Inf_L</td>
<td>89</td>
<td>0.1725</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Precentral_L</td>
<td>1</td>
<td>0.1734</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Frontal_Mid_L</td>
<td>7</td>
<td>0.2099</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Frontal_Inf_Oper_L</td>
<td>11</td>
<td>0.1541</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parietal_Inf_L</td>
<td>62</td>
<td>0.1893</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>SupraMarginal_R</td>
<td>64</td>
<td>0.1912</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Temporal_Inf_L</td>
<td>89</td>
<td>0.1410</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Temporal_Inf_R</td>
<td>90</td>
<td>0.1843</td>
<td></td>
</tr>
<tr>
<td>CD1D</td>
<td>Frontal_Inf_Oper_L</td>
<td>11</td>
<td>0.0968</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Postcentral_L</td>
<td>57</td>
<td>0.1510</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Postcentral_R</td>
<td>58</td>
<td>0.1673</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parietal_Inf_L</td>
<td>61</td>
<td>0.1472</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parietal_Inf_R</td>
<td>62</td>
<td>0.1704</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SupraMarginal_L</td>
<td>63</td>
<td>0.1721</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SupraMarginal_R</td>
<td>64</td>
<td>0.1772</td>
<td></td>
</tr>
<tr>
<td>ETAA1</td>
<td>Frontal_Inf_Oper_L</td>
<td>11</td>
<td>0.0968</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Postcentral_L</td>
<td>57</td>
<td>0.1510</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Postcentral_R</td>
<td>58</td>
<td>0.1673</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parietal_Inf_L</td>
<td>61</td>
<td>0.1472</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parietal_Inf_R</td>
<td>62</td>
<td>0.1704</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SupraMarginal_L</td>
<td>63</td>
<td>0.1721</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SupraMarginal_R</td>
<td>64</td>
<td>0.1772</td>
<td></td>
</tr>
<tr>
<td>FLJ46361</td>
<td>Precentral_L</td>
<td>1</td>
<td>0.2819</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Frontal_Mid_L</td>
<td>7</td>
<td>0.2281</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Frontal_Inf_Tri_L</td>
<td>13</td>
<td>0.2056</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Supp_Motor_Area_L</td>
<td>19</td>
<td>0.2841</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Postcentral_L</td>
<td>57</td>
<td>0.2162</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Postcentral_R</td>
<td>58</td>
<td>0.2637</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Parietal_Inf_L</td>
<td>61</td>
<td>0.2663</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SupraMarginal_L</td>
<td>63</td>
<td>0.3222</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SupraMarginal_R</td>
<td>64</td>
<td>0.2885</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Temporal_Mid_R</td>
<td>86</td>
<td>0.2552</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Temporal_Inf_L</td>
<td>89</td>
<td>0.2681</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Temporal_Inf_R</td>
<td>90</td>
<td>0.2760</td>
<td></td>
</tr>
<tr>
<td>FRMD4A</td>
<td>Frontal_Mid_L</td>
<td>7</td>
<td>0.2218</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Supp_Motor_Area_L</td>
<td>19</td>
<td>0.2278</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Postcentral_R</td>
<td>58</td>
<td>0.2184</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parietal_Inf_L</td>
<td>61</td>
<td>0.1907</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SupraMarginal_R</td>
<td>64</td>
<td>0.2571</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Temporal_Mid_R</td>
<td>86</td>
<td>0.2553</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Brain region</td>
<td>The ROI’s index</td>
<td>$\rho$ (correlation value)</td>
<td>The number of SNPs</td>
</tr>
<tr>
<td>--------</td>
<td>--------------------</td>
<td>-----------------</td>
<td>----------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td></td>
<td>Temporal_Inf_L</td>
<td>89</td>
<td>0.1609</td>
<td></td>
</tr>
<tr>
<td>GRIK4</td>
<td>Precentral_L</td>
<td>1</td>
<td>0.1101</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Frontal_Mid_L</td>
<td>7</td>
<td>0.1752</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Frontal_Inf_Oper_L</td>
<td>11</td>
<td>0.0607</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>SupraMarginal_L</td>
<td>65</td>
<td>0.2047</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SupraMarginal_R</td>
<td>64</td>
<td>0.1826</td>
<td></td>
</tr>
<tr>
<td>ZNFS6</td>
<td>Frontal_Inf_Oper_L</td>
<td>11</td>
<td>0.1449</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Occipital_Mid_L</td>
<td>51</td>
<td>0.1951</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parietal_Inf_L</td>
<td>61</td>
<td>0.1172</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Temporal_Inf_R</td>
<td>90</td>
<td>0.2168</td>
<td></td>
</tr>
</tbody>
</table>
Table IV
The enriched gene ontology terms that are related to the neural activity

<table>
<thead>
<tr>
<th>Go term</th>
<th>Gene</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>regulation of nervous system development</td>
<td>CDH19, CSMD1,</td>
<td>1.16e-3</td>
</tr>
<tr>
<td>positive regulation of nervous system develop</td>
<td>CXCR4, GRIK4, SYNE1, BMP6,</td>
<td>2.73e-3</td>
</tr>
<tr>
<td>generation of neurons</td>
<td>CD1D, ETAA1, FLJ46361,</td>
<td>7.39e-3</td>
</tr>
<tr>
<td>neurogenesis</td>
<td>FRMD4A, ZNF536</td>
<td>9.25e-3</td>
</tr>
</tbody>
</table>