Bioinformatics, 36(21), 2020, 5214–5222
doi: 10.1093/bioinformatics/btaa650
Advance Access Publication Date: 19 July 2020
Original Paper

Genetics and population analysis

Co-sparse reduced-rank regression for association analysis between imaging phenotypes and genetic variants

Canhong Wen[1], Hailong Ba[1], Wenliang Pan[2] and Meiyan Huang [3,4,]*; for the Alzheimer's Disease Neuroimaging Initiative†

[1]International Institute of Finance, School of Management, University of Science and Technology of China, Hefei 230026, China, [2]Department of Statistical Science, School of Mathematics, Sun Yat-Sen University, Guangzhou 510275, China, [3]School of Biomedical Engineering and [4]Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou 510515, China

*To whom correspondence should be addressed.

†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.

Associate Editor: Bonnie Berger

Received on December 1, 2019; revised on May 22, 2020; editorial decision on July 11, 2020; accepted on July 14, 2020

Abstract

Motivation: The association analysis between genetic variants and imaging phenotypes must be carried out to understand the inherited neuropsychiatric disorders via imaging genetic studies. Given the high dimensionality in imaging and genetic data, traditional methods based on massive univariate regression entail large computational cost and disregard many-to-many correlations between phenotypes and genetic variants. Several multivariate imaging genetic methods have been proposed to alleviate the above problems. However, most of these methods are based on the l_1 penalty, which might cause the over-selection of variables and thus mislead scientists in analyzing data from the field of neuroimaging genetics.

Results: To address these challenges in both statistics and computation, we propose a novel co-sparse reduced-rank regression model that identifies complex correlations in a dimensional reduction manner. We developed an iterative algorithm based on a group primal dual-active set formulation to detect simultaneously important genetic variants and imaging phenotypes efficiently and precisely via non-convex penalty. The simulation studies showed that our method achieved accurate and stable performance in parameter estimation and variable selection. In real application, the proposed approach successfully detected several novel Alzheimer's disease-related genetic variants and regions of interest, which indicate that our method may be a valuable statistical toolbox for imaging genetic studies.

Availability and implementation: The R package **csrrr**, and the code for experiments in this article is available in Github: https://github.com/hailongba/csrrr.

Contact: huangmeiyan16@163.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

With the development of biomedicine and clinical science, the pathogenic mechanism of brain disorders must be studied from the perspective of internal genetics. Many research projects have been conducted around the world to generate massive high dimensional and complex brain imaging data and genome sequence data [e.g. the Alzheimer's Disease Neuroimaging Initiative (ADNI)] (Durston, 2010; Gilmore *et al.*, 2010). Therefore, brain imaging genetics has gained more attention in recent years. A major task of imaging genetics is to identify the relationships between the phenotypes extracted from imaging data and genotypes, such as single-nucleotide polymorphism (SNP); this process is expected to discover the genetic basis of brain structure and function, thereby further offering help for the prediction, diagnosis and treatment of various complex brain-related disorders, such as schizophrenia and Alzheimer's disease (AD) (Du *et al.*, 2018; Hashimoto *et al.*, 2015; Saykin *et al.*, 2015).

During the past decade, numerous studies on imaging genetics have been proposed, in which the most recently method used was genome-wide association study (GWAS). Traditional GWAS methods are based on the mass-univariate linear model. However, the association between imaging phenotypes and genetic variants is a complex many-to-many relationship. Thus, univariate analysis of traditional GWAS is insufficient to address the complex relationship between imaging and genetic data (Dudbridge, 2016; Li *et al.*, 2015). Moreover, running GWAS poses remarkable computational challenges given the high dimensionality of imaging (millions of brain locations) and genetic data ($\sim 10^6$ known variants) (Huang *et al.*, 2017). To accelerate the calculation of GWAS, Huang *et al.* (2015) proposed a fast voxel-wise genome-wide association analysis framework. By treating the imaging phenotypes as a whole, Wen *et al.* (2018) proposed the use of distance covariance to incorporate the correlations in imaging phenotypes for GWAS to alleviate the calculational cost. However, the above two methods are based on univariate regression analysis. Thus, the correlations between imaging phenotypes and genetic variants are ignored.

Several multivariate imaging genetic methods have been introduced to overcome shortcomings in univariate regression analysis. Among these methods, co-sparsity in imaging phenotypes and genetic variants is generally assumed. A sparse canonical correlation analysis (SCCA) was developed to explore the correlation between two high-dimensional datasets under the co-sparsity assumption (Witten *et al.*, 2009). However, SCCA only focuses on the correlation between the top two canonical correlation vectors and cannot find the exact mapping model between two datasets. To overcome this problem, Ma *et al.* (2014) proposed a thresholding singular value decomposition (TSVD) method by imposing sparse constraints on the left and right singular vectors of the coefficient matrix. In this way, they showed that the TSVD approach can realize co-sparsity in the responses and predictors and retain a low-rank structure. An alternative to achieve co-sparsity is the use of the sparse version of the reduced rank regression (RRR) model, which is a multivariate-to-multivariate model with a low-rank constraint on the coefficient matrix. To realize co-sparsity in RRR, Vounou *et al.* (2010) developed a sparse RRR method by simplifying the RRR model to rank-one models based on the assumption of diagonal covariance matrices in phenotypes and genetic variants. However, this assumption is often violated in practice (Kong *et al.*, 2020; Zhu *et al.*, 2014). To relieve this problem, Zhu *et al.* (2014) established a Bayesian framework to build a sparse version of RRR and applied it to an imaging genetics study. Kong *et al.* (2020) proposed a low-rank linear regression model to obtain a sparse estimate of the coefficient matrix based on trace norm regularization. However, most existing methods achieve co-sparsity via imposing l_1-type norms as regularizer, which might yield inadequate performance. The reason for the poor performance lies in the l_1 norm, which is a convex approximation of l_0 norm and introduces bias when the value is far from zero. Thus, methods based on the l_1 norm might cause bias in the estimation and thus lead to the over-selection problem.

To address the above challenges, we propose a new co-sparse RRR (CSRRR) model that simultaneously selects imaging phenotypes and genetic variants via the l_0 norm. However, variable selection via l_0 norm corresponds to the well-known best subset selection problem, which is thought to be NP-hard (Natarajan, 1995). Recently, a primal-dual active set (PDAS) formulation was presented for the best subset selection problem in linear, logistic and Cox regression model (Wen *et al.*, 2017). The best subset selection could be solved at considerably large problem sizes within seconds. Based on this idea, we extended the PDAS formulation to a group PDAS (GPDAS) formulation and developed an iterative algorithm to study the association between genetic variants and imaging phenotypes. This article offers four methodological contributions. The first one introduces the CSRRR model to directly solve high-dimensional imaging genetic data within a reasonable time frame. The second one uses the l_0 norm regularizer to jointly select the desired numbers of genetic variants and imaging phenotypes to realize the sparse structure. This norm penalty involves the best subset selection problem, which is hard to solve traditionally. The third one aims to develop an iterative algorithm based on GPDAS to solve the best subset selection problem and achieve an accurate solution efficiently and precisely. The fourth one reduces the computational spending and considers complex relationships between imaging phenotypes and genetic variants compared with traditional univariate linear methods. Based on the simulation studies, the proposed CSRRR can accurately estimate regression coefficients and jointly detect the causal SNPs and the affected regions of interest (ROIs) under a variety of simulation settings compared with several existing methods. We also demonstrated the effectiveness of the proposed method in an application of imaging genetic data analysis on ADNI data.

The rest of the article is organized as follows. Section 2 expounds our proposed methodology for co-sparsity best subset selection in the RRR model and presents an efficient iterative algorithm based on GPDAS for solving the CSRRR model with joint orthogonality constraints and non-convex sparsity. Section 3 demonstrates the competitive numerical performance of the proposed algorithm using simulation studies and applies our method to analyze large-scale imaging genetic data from the ADNI database. Section 4 provides the conclusions and discussions.

2 Method and algorithm

2.1 The CSRRR method

Suppose that we have n independent observations $(X, Y) = \{(x_i, y_i), i = 1, \ldots, n\}$, where $x_i \in \mathbb{R}^p$ represents the p-dimensional genetic variants, and $y_i \in \mathbb{R}^q$ denotes the q-dimensional imaging phenotypes. The RRR model with rank constraint r is defined by

$$Y = XC + E, \quad \mathrm{rank}(C) \le r, \tag{1}$$

where $C \in \mathbb{R}^{p \times q}$ is an unknown coefficient matrix with $\mathrm{rank}(C) \le r$, and $E = (\epsilon_1, \ldots, \epsilon_n)^\top \in \mathbb{R}^{n \times q}$ is a random error matrix with independent zero mean and finite variance entries. In addition to the centering assumption in responses and predictors, we further assumed that the predictors are normalized with unit variance.

In imaging genetic studies, we commonly assume that only a small subset of genetic variants contributes to any imaging phenotype, and that each selected genetic variant is associated with a small subset of phenotypes (Vounou *et al.*, 2010). Under this assumption, both imaging phenotypes and genetic variants are sparse. Thus, the regression coefficient C has sparsity in the row and column under the settings of the RRR model. Given the sparse levels k_x and k_y, we considered the following variable selection problem

$$\min_C ||Y - X\,C||_F^2, \quad \mathrm{s.t.rank(C)} \le \mathrm{r},\ ||\mathrm{C}||_{2,0} = \mathrm{k_x},\ ||\mathrm{C}^\top||_{2,0} = \mathrm{k_y}, \tag{2}$$

where $||\cdot||_F$ denotes the Frobenius norm, $\mathrm{rank}(\cdot)$ indicates the matrix rank and $||\mathrm{C}||_{2,0}$ counts the number of non-zero rows in C, $1 \le r \le \min(\mathrm{rank}(X), q, k_x, k_y)$, $1 \le k_x \le \min(p, n)$ and $1 \le k_y \le \min(q, n)$. For simplicity, we refer to problem (2) as the CSRRR problem henceforth.

2.2 Algorithm

The interplay between orthogonality constraints and non-convex sparsity in rank and sparsity creates substantial algorithmic challenges for solving the CSRRR problem in (2), for which numerous existing algorithms can become either inefficient or inapplicable.

To eliminate the rank constraint in (2), we expressed the coefficient matrix C as a product of two matrices, i.e. $C = BV^\top$, where $V \in \mathbb{R}^{q \times r}$ is an orthogonal matrix and $B \in \mathbb{R}^{p \times r}$. Then, the optimization problem in (2) can be rewritten as

$$\min_{B,V}||Y - XBV^{\top}||_F^2, \quad \mathrm{s.t.} \mathrm{V}^{\top}\mathrm{V} = \mathrm{I_r}, \ ||\mathrm{B}||_{2,0} = \mathrm{k_x}, \ ||\mathrm{V}||_{2,0} = \mathrm{k_y}. \tag{3}$$

The estimates of B and V obtained from problem (3) are not unique, i.e. B and V are not identifiable. Suppose that $(\hat{B}, \hat{V})$ denotes the solution of (3). Then, $(\tilde{B} = \hat{B}Q, \tilde{V} = \hat{V}Q)$ also solves problem (3), where Q is any orthogonal matrix. However, given the fact that $\tilde{C} = \tilde{B}\tilde{A}^{\top} = \hat{B}\hat{A}^{\top} = \hat{C}$, the estimation of C is unique, i.e. C is identifiable. This condition indicates that the division of C into a product of B and V causes no change in the estimation. Optimization was performed with respect to B and V, and it can be achieved in a block-wise iteration. To be precise, given the current estimate $(B^{(m)}, V^{(m)})$ for the mth iteration, the CSRRR optimization of problem (3) consists of the following iterations:

- B-step:

$$B^{(m+1)} = \mathrm{argmin}_B||Y - XB(V^{(m)})^{\top}||_F^2, \ \mathrm{s.t.}||\mathrm{B}||_{2,0} = \mathrm{k_x}.$$

- V-step:

$$V^{(m+1)} = \mathrm{argmin}_V||Y - XB^{(m+1)}V^{\top}||_F^2, \ \mathrm{s.t.}\mathrm{V}^{\top}\mathrm{V} = \mathrm{I_r}, \ ||\mathrm{V}||_{2,0} = \mathrm{k_y}.$$

B-step involves the best subset selection with each row being a group or group subset selection. Particularly, by the orthogonality constraint $(V^{(m)})^{\top}V^{(m)} = I_r$, the problem in B-step is equivalent to (omitting terms not involving B)

$$\min_{B\in\mathbb{R}^{p\times r}}||YV^{(m)} - XB||_F^2, \quad \mathrm{s.t.}||\mathrm{B}||_{2,0} = \mathrm{k_x}. \tag{4}$$

V-step involves orthogonality and sparsity constraints, which result in difficulty in directly addressing the optimization problem. Actually, it is a sparse version of the orthogonal Procrustes problem (Schönemann, 1966). To tackle the computational difficulty, we first identified the non-zero row of V, i.e. $\mathcal{A}_v$, by ignoring the orthogonal constraint and searched for an optimal orthogonal solution within a threshold parameter space $\mathcal{TS} = \{V \in R^{q\times r} : V_{(\mathcal{A}_v)^c,\cdot} = 0\}$. By ignoring the orthogonality constraint, V-step reduces the following orthogonality-removed problem

$$\min_{V\in\mathbb{R}^{q\times r}}||Y - XB^{(m+1)}V^{\top}||_F^2, \quad \mathrm{s.t.}||\mathrm{V}||_{2,0} = \mathrm{k_y}. \tag{5}$$

which is also a group subset selection problem. Within the threshold space $\mathcal{TS}$, V-step involves finding an optimal $V_{\mathcal{A}_v}$ that can be derived from the following Procrustes problem

$$\min_{V\in\mathbb{R}^{|\mathcal{A}_v|\times r}}||Y_{\mathcal{A}_v} - XBV^{\top}||_F^2, \quad \mathrm{s.t.}\mathrm{V}^{\top}\mathrm{V} = \mathrm{I_r}. \tag{6}$$

The above problem has a closed-form solution given by singular value decomposition (SVD), i.e. $V_{\mathcal{A}_v} = U_w V_w^{\top}$, where U_w and V_w are obtained from the SVD of $W = Y_{\mathcal{A}_v}^{\top} XB^{(m+1)}$, i.e. $W = U_w D_w V_w^{\top}$.

2.2.1 Group subset selection

B- and V-steps involve the group subset selection problem with elements in each row being a group [see problems in (4) and (5), respectively]. In this section, we develop a group-type generalization of the PDAS algorithm to solve the group subset selection problem. The PDAS algorithm was introduced to solve the best subset selection problem in parametric models with univariate response; it is computationally efficient in identifying the best sub-model as stated by Wen *et al.* (2019). The computational efficiency lies in utilizing an active set updating strategy and fitting the sub-models through the use of complementary primal and dual variables.

For problem (4), let $B = (B_{1,\cdot}^{\top}, B_{2,\cdot}^{\top}, \ldots, B_{p,\cdot}^{\top})^{\top}$ be a coordinate-wise minimizer and $l_j(b) = ||YV^{(m)} - \sum_{i\neq j} X_{\cdot,i}B_{i,\cdot} - X_{\cdot,j}b^{\top}||_F^2$ with $b \in R^r$ be the partial loss, $j = 1, \ldots, p$. Minimizing the objective function $l_j(b)$ yields $b_j^{\top} = B_{j,\cdot} + \gamma_{j,\cdot}^B$, where $\gamma_{j,\cdot}^B = X_{\cdot,j}^{\top}(YV^{(m)} - XB)/n$.

The constraint $||B||_{2,0} = k_x$ indicates that $p - k_x$ rows would be forced to reach zero. To determine such $p - k_x$ rows, we considered the sacrifices of $l_j(b)$ which are given by $\Delta_j^B = ||B_{j,\cdot} + \gamma_{j,\cdot}^B||^2$, when $b_j^{\top}$ switches from $B_{j,\cdot} + \gamma_{j,\cdot}^B$ to $\mathbf{0}$. Among all the candidates, we may force these rows to reach zero if they contribute the least total sacrifices to the overall loss. To realize this condition, we let $\Delta_{[1]}^B \geq \ldots \geq \Delta_{[p]}^B$ denote the decreasing rearrangement of Δ_j^B's for $j = 1, \ldots, p$. Then, the ordered sacrifice vector at position k_x should be truncated. Combining the analytical result acquired before, we obtained

$$B_{j,\cdot} = \begin{cases} B_{j,\cdot} + \gamma_{j,\cdot}^B, & \text{if } \Delta_j^B \geq \Delta_{[k_x]}^B \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Algorithm 1 GPDAS algorithm for B-step

```
Input: Imaging phenotypes $Y$, genetic variants $X$, the desired levels of sparsity in genotypes ($k_x$) and the current estimate ($B, V$).
(1) Set $l=0$; Initialize $B^{(0)} = B$ and $\Gamma^{(0)} = X^{\top}(YV - XB^{(0)})/n$;
(2) While $B^{(l)}$ not converged do
    (2.a) Determine the sacrifies $\Delta_j^{(B,l)} = ||B_{j,\cdot} + \gamma_{j,\cdot}^B||^2$ for $j = 1, 2, \ldots, p$;
    (2.b) Determine the active and inactive sets for $B^{(l)}$ by
          $\mathcal{A}^{(B,l)} = \{j : \Delta_j^{(B,l)} \geq \Delta_{[k]}^{(B,l)}\}, \mathcal{I}^{(B,l)} = \mathcal{A}^{(B,l)c}$;
    (2.c) Update $B^{(l+1)}$ by
          $B_{\mathcal{A}^{(B,l)}}^{(l+1)} = (X_{\mathcal{A}^{(B,l)}}^{\top} X_{\mathcal{A}^{(B,l)}})^{-1} X_{\mathcal{A}^{(B,l)}}^{\top} YV$ and $B_{\mathcal{I}^{(B,l)}}^{(l+1)} = \mathbf{0}$;
    (2.d) Update $\Gamma^{(B,l+1)}$ by
          $\Gamma_{\mathcal{A}^{(B,l)}}^{(B,l+1)} = \mathbf{0}$ and $\Gamma_{\mathcal{I}^{(B,l)}}^{(B,l+1)} = \left(X_{\mathcal{I}^{(B,l)}}^{\top}(YV^{(m)} - XB^{(B,l)})\right)/n$;
    (2.e) $l = l + 1$.
Output: $\{\hat{B}, \hat{\mathcal{A}}^B\} = \{B^{(l+1)}, \mathcal{A}^{(B,l)}\}$.
```

In (7), $B = (B_{1,\cdot}^{\top}, \ldots, B_{p,\cdot}^{\top})^{\top}$ is a primal variable, $\Gamma^B = (\gamma_{1,\cdot}^{B\top}, \ldots, \gamma_{p,\cdot}^{B\top})^{\top}$ a dual variable and $\Delta^B = (\delta_{1,\cdot}^{B\top}, \ldots, \delta_{p,\cdot}^{B\top})^{\top}$ a reference sacrifice. With these quantities, we defined Equation (7) as the primal-dual condition and developed a group version of PDAS algorithm for solving problem (4) following:Similarly, we can derive the primal-dual condition for the coordinate-wise minimizer of the orthogonality-removed problem (5) as

$$V_{j,\cdot} = \begin{cases} V_{j,\cdot} + \gamma_{j,\cdot}^V, & \text{if } \Delta_j^V \geq \Delta_{[k_y]}^V \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where $\gamma_{j,\cdot}^V = Y_{\cdot,j}^{\top}(XB^{(m)} - Y)/n$ is the dual variable of $V_{j,\cdot}$, and $\Delta_j^V = ||V_{j,\cdot} + \gamma_{j,\cdot}^V||^2$ is the reference sacrifice, $j = 1, \ldots, q$. As mentioned by Wen *et al.* (2017), the PDAS algorithm for linear regression model usually converges in finite steps. Thus, rather than iteratively updating the estimate of V as in the B-step, we proposed to estimate V and the corresponding active set $\mathcal{A}_v$ within one iteration. In this regard, we could save computational time without substantial loss of accuracy. We summarize the algorithm for solving problems (5) and (6) as follows:

2.2.2 Algorithm for CSRRR

With the optimization of B-step and V-step, we can develop a block-wise iterative algorithm to solve the CSRRR problem (3). The algorithm is summarized as follows:To declare the convergence in Step 2 of Algorithm 3, we used the absolute tolerance method. In particular, the algorithm will stop when an absolute difference in the loss function (3) is lower than a pre-specified threshold τ, say $\tau = 0.001$.

The optimization problem (2) includes a considerable number of local optima because norm $||\cdot||_{2,0}$ is non-continuous and non-

```
Algorithm 2 One-step GPDAS algorithm for $V$-step

Input: Imaging phenotypes $Y$, genetic variants $X$, the desired levels of sparsity in phenotypes $(k_y)$ and the current estimate of $B$.

1. Compute $\gamma^V_{j,\cdot} = Y^\top_{\cdot,j}(XB - Y)/n$ and $\Delta^V_j = ||V_{j,\cdot} + \gamma^V_{j,\cdot}||^2$ for $j = 1, 2, \dots, q$;
2. Determine the active and inactive sets for $V^{(m+1)}$ by
       $\mathcal{A}^v = \{j : \Delta^V_j \geq \Delta^V_{[k_y]}\}, \quad \mathcal{I}^{(m+1)}_v = \{j : \Delta^V_j < \Delta^V_{[k_y]}\};$
3. Let $W = Y^\top_{\mathcal{A}_v,\cdot} XB$ and perform SVD of W, i.e. $W = U_w D_w V^\top_w$;
4. Determine $V$ by $V_{\mathcal{A}_v} = U_w V^\top_w$ and $V_{\mathcal{I}_v} = \mathbf{0}$

Output: $\{\hat{V}, \hat{\mathcal{A}}^V\} = \{V, \mathcal{A}^V\}$.
```

```
Algorithm 3 Algorithm for CSRRR problem

Input: Imaging phenotypes $Y$, genetic variants $X$, rank $r$, the desired levels of sparsity in genotypes $(k_x)$ and phenotypes $(k_y)$.

1. Initialize $m = 0$, and $V^{(0)}$ as the eigenvectors associated with the top $r$th eigenvalues of $Y^\top Y$.
2. While the value of objective function in (3) not converged do
     (1) Run Algorithm 1 with $k_x$ and $(B^{(m)}, V^{(m)})$;
   Output $\{B^{(m+1)}, \mathcal{A}^{(B,m+1)}\}$.
     (2) Run Algorithm 2 with $k_y$ and $B^{(m+1)}$;
   Output $\{V^{(m+1)}, \mathcal{A}^{(V,m+1)}\}$.
     (3) $m = m + 1$.
Output: $(\hat{C}, \hat{B}, \hat{V}, \hat{\mathcal{A}}^B, \hat{\mathcal{A}}^V)$
```

convex. To guarantee that the algorithm outputs the desired optimum, we adopted the warm start strategy for initialization. That is, the algorithm starts with $k_x = 0$ and $k_y = 0$ and computes the initial estimation $(\hat{B}^{(0)}, \hat{V}^{(0)})$. When running the algorithm with the increase in k_x, the results with $k_x - 1$ are fully utilized to derive the new initial estimators for the estimation with k_x, such that the active set is incrementally updated by adding the most promising index determined by Δ^B_j in (7). This warm start strategy is highly effective in finding a stationary solution. A similar phenomenon can be observed in the PDAS algorithm (Wen *et al.*, 2017).

Provided that the true relevant predictor is small and $p \leq O(e^n)$, the computational complexity of the PDAS algorithm for linear regression is $O(np)$, which is a linear time with respect to n and p (Wen *et al.*, 2017). Similarly, we can expect that the computational cost of GPDAS algorithm, that is, Algorithm 1 or Algorithm 2, is $O(npq)$. Through enormous experiments, Algorithm 3 can converge by several steps in most situations. The reason might be an unbiased estimation of B and V resulting from our GPDAS algorithm.

3 Experiments

3.1 Tuning parameter selection

The performance of Algorithm 3 depends on three tuning parameters in problem (2), i.e. the rank r, the sparsity k_x of genetic variants and the sparsity k_y of imaging phenotypes. Numerous well-established methods are available for the selection of tuning parameters. If an individual validation dataset is available, we can determine the optimal tuning parameters by selecting those with the smallest prediction error. However, real data are often scarce in practice. Thus, K-fold cross validation was used to estimate the prediction error and for comparison with different models. Here, the prediction error is defined by

$$PE(r, k_x, k_y) = \frac{1}{nq}||Y - X\hat{C}(r, k_x, k_y)||^2_F, \tag{9}$$

where $\hat{C}(r, k_x, k_y)$ is the CSRRR coefficient estimator with the combination of tuning parameters $\{r, k_x, k_y\}$.

The optimal tuning parameters were determined among a set of three-dimensional grid points. For each triple combination $\{r, k_x, k_y\}$, the CSRRR estimator was obtained via Algorithm 3, and the prediction error was calculated. We then selected the optimal model with the smallest prediction error. In simulations, we calculated the prediction error on a validation dataset to minimize the influence of tuning parameter selection methods on performance comparison. In our real data analysis, we used nested cross validation to compute the prediction error.

3.2 Simulation studies

In this section, we evaluated the performance of the proposed method on simulated data and compared the proposed method with four others: (i) threshold SVD method (TSVD, Ma *et al.*, 2014); (ii) sparse-reduced rank regression (sRRR, Mishra*et al.*, 2017); (iii) SCCA (Witten *et al.*, 2009) and (iv) fast voxel-wise genome-wide association analysis (FVGWAS, Huang *et al.*, 2015). The former three are the widely used multivariate methods in imaging genetics study, whereas the last one is a state-of-the-art univariate method.

3.2.1 Simulation settings

The SNP data were generated as follows. The linkage disequilibrium (LD) blocks between SNPs were simulated by the default method of Haplotview and PLINK (Barrett *et al.*, 2005; Gabriel *et al.*, 2002; Purcell *et al.*, 2007). We first simulated $n = 1000$ subjects by randomly combining the haplotypes of HapMap CEU subjects. We then used PLINK to determine the LD blocks based on these subjects. Finally, we randomly selected 150 blocks and combined the haplotypes of HapMap CEU subjects in each block to form genotype variables. In each block, we randomly drew 10 SNPs, achieving 1500 SNPs for each subject. To avoid collinearity in the model, we screened out the SNPs with the same values. After this quality control process, we attained $n = 1000$ observations and 999 SNPs in total.

For the imaging phenotypes, we generated ROI data from Model (1) with $C = BV^\top$. Matrix $B \in \mathbb{R}^{p \times r}$ corresponds to the coefficients of latent factors of SNPs, whereas matrix $V \in \mathbb{R}^{q \times r}$ corresponds to those of latent factors of ROIs. Therefore, the zero rows in B and V represent irrelevant ROI–SNP pairs. In our simulation, we pre-fixed the first k_y ROIs as the affected ROIs associated with the causal SNPs and regarded the first k_x SNPs as the causal SNP. This condition indicates that the first k_x rows of the low-rank matrix B and the first k_y rows of the low-rank matrix V were set to non-zero effect magnitude. To fully evaluate the stability among different associations between imaging phenotypes and genetic variants, we drew the non-zero elements from Uniform distribution $U(-1, 1)$ or standard Normal distribution $N(0, 1)$. To mimic the ADNI data, we fixed the dimension of ROIs as $q = 100$, the rank as $r = 3$ and the number of affected ROIs as $k_y = 10$. The number of causal SNPs ranged from $k_x = 30$ to $k_x = 150$ with a step size of 30 to explore the effect of k_x on the performance of different methods.

To fully evaluate the influence of noise E on the performance, we considered the following generation mechanisms:

Case 1: $E \sim \mathcal{N}(0, \Sigma)$, $\Sigma = \mathbf{I}$;

Case 2: $E \sim \mathcal{N}(0, \Sigma)$, $\Sigma_{ij} = 0.5^{|i-j|}$, $i, j = 1, \dots, q$;

Case 3: $E, E \sim \mathcal{N}(0, 4 \times \Sigma)$, $\Sigma_{ij} = 0.5^{|i-j|}$, $i, j = 1, \ldots, q$;

Case 4: $E \sim \chi^2(3)$.

Cases 1 and 2 investigate the finite performance with different correlation; Case 3 illustrates the influence by small signal-to-noise ratios. Case 4 considers the performance with the noise being drawn from non-Gaussian and skewness distribution.

Overall, 40 different scenarios of simulation settings and 100 replications were conducted for each setting. In each replication, we randomly divided the 1000 subjects into sets: a training dataset with a size of 700, a validation dataset with a size of 200 used for determining the optimal tuning parameters by the prediction error and a test dataset with a size of 100. For each coefficient matrix estimator $\hat{C}$, we measured the estimation and predictive accuracy in terms of the mean-squared error:

$$\text{Est} = ||C - \hat{C}||_F^2, \quad \text{Pred} = ||X_{test}C - X_{test}\hat{C}||_F^2/nq,$$

where X_{test} is obtained from the test dataset. We also reported the estimated number of causal SNPs $\hat{k}_x$ and the estimated number of affected ROIs $\hat{k}_y$. The performance of relevant variable detection was evaluated by sensitivity and specificity. For SNPs or predictors, the sensitivity and specificity are defined as

$$\text{Senx} = \frac{|\mathcal{A}_x \cap \hat{\mathcal{A}}_x|}{|\mathcal{A}_x|}, \quad \text{Spex} = \frac{|\mathcal{I}_x \cap \hat{\mathcal{I}}_x|}{|\mathcal{I}_x|};$$

where $\hat{\mathcal{A}}_x$ and $\hat{\mathcal{I}}_x$ are the selected casual SNPs set and non-casual SNPs set, respectively. $\mathcal{A}_x$ and $\mathcal{I}_x$ are the true casual SNPs set and non-casual SNPs set, respectively. Similarly, for ROIs or responses, the sensitivity and specificity are defined as

$$\text{Seny} = \frac{|\mathcal{A}_y \cap \hat{\mathcal{A}}_y|}{|\mathcal{A}_y|}, \quad \text{Spey} = \frac{|\mathcal{I}_y \cap \hat{\mathcal{I}}_y|}{|\mathcal{I}_y|},$$

where $\hat{\mathcal{A}}_y$ and $\hat{\mathcal{I}}_y$ are the selected affected and non-affected ROI sets, respectively. $\mathcal{A}_y$ and $\mathcal{I}_y$ are the true affected and non-affected ROI sets, respectively.

3.2.2 Simulation results

The average results are reported over 100 repetitions in Supplementary Tables. Supplementary Tables S1–S8 summarize the results of all scenarios. Given that SCCA and FVGWAS offer no coefficient matrix as output, the values of Pred and Est are unavailable for these specific methods. In addition, FVGWAS based on hypothesis test method needs to pre-specify the selected number of causal $\hat{k}_x$ and affected ROIs $\hat{k}_y$. Thus, for each scenario, we set the true k_x and k_y as corresponding estimated $\hat{k}_x$ and $\hat{k}_y$, respectively, and opted not to report the results.

The values of Pred, Est and bias in $\hat{k}_x$ increased as k_x increased for all methods except SCCA and FVGWAS, where the values were unavailable, and the Senx value from FVGWAS decreased significantly. These findings suggest that a high k_x value or number of causal SNPs leads to increased challenges in effective detection and parameter estimation. However, in all cases, our CSRRR method consistently yielded better performance than the other methods in terms of prediction and estimation error with better Pred values and smallest Est values. For the variable selection performance, the estimated number of causal SNPs $\hat{k}_x$ and the estimated number of affected ROIs $\hat{k}_y$ from our method are always close to the true values. This condition suggests that our CSRRR approach is less influenced by the number of true causal SNPs than the other existing methods.

Given the influence of different correlations within the noise matrix E, the results in Supplementary Tables S1–S4 remained almost the same for all methods. In Case 3, when the variance of noise was doubled, the measurements for all methods worsened compared with those in Case 2. Still, among these methods, our CSRRR consistently yielded a model with higher sensitivity and specificity. Moreover, compared with sRRR and TSVD, CSRRR had the smallest reduction in terms of prediction and estimation accuracy. In Case 4, in which a non-Gaussian noise was considered, both sRRR and TSVD showed poor performance in selecting related ROI–SNP pairs and estimating the regression coefficient matrix. By contrast, CSRRR yielded the smallest prediction error and estimation error and excellently identified the true causal SNPs and affected ROIs.

The above analysis reflects that the generation mechanism of E considerably influences the performance, whereas CSRRR is competitive for its anti-interference capability. We also examined the effect of different generation mechanisms of non-zero elements in C by comparing the results in Supplementary Tables S1–S8. The results from the uniformly distributed coefficients are expectedly better than those from normally distributed coefficients because the variance in non-zero elements in C is notably smaller than that in uniform distribution. In summary, the proposed CSRRR method has superior accuracy and stability in prediction, estimation accuracy and selection of relevant ROI–SNP pairs. The results suggest that our CSRRR approach is suitable for use in real data analysis.

3.2.3 Computation time

In this section, we present three numerical experiments to show how the proposed method scales with the number of subjects n, dimension of phenotypes q and the number of SNPs p. Following the settings in Section 3.2.1, we generated simple simulated data with rank $r = 3$, number of causal SNPs $k_x = 100$ and number of affected ROIs $k_y = 50$. The first k_x rows and k_y columns of the coefficient matrix $C \in \mathbb{R}^{p \times q}$ and noise matrix $E \in \mathbb{R}^{n \times q}$ were drawn from the standard Normal distribution. The baseline parameters for the sample size n, the number of SNPs p and the number of ROI q is $(n, p, q) = (500, 1000, 100)$. In each experiment, we varied one parameter while fixing the other two at the baseline values. Specifically, we allowed p to increase from 6000 to 15 000 by a step size of 1000 in Experiment 1, let q increase from 500 to 5000 by a step size of 500 in Experiment 2 and enabled n to increase from 1000 to 10 000 by a step size of 1000 in Experiment 3.

Figure 1 summarizes the results. The computation time increased at a linear rate of n, p and q, coinciding with the analysis of computational complexity in Section 2.2.2. Furthermore, our proposed yield is a sparse estimator obtained in several seconds with p in 10 000 s, and it shows the possibility of our proposal in handling high-dimensional data.

3.3 ADNI data analysis

3.3.1 Data processing

To illustrate the usefulness of the proposed method, we considered the genetic data and structural magnetic resonance imaging (MRI) scans provided by ADNI database (http://adni.loni.usc.edu/). ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and non-profit organizations as a $60 million and 5-year public–private partnership. The primary goal of ADNI was to test whether serial MRI, PET and other biological markers are useful in clinical trials of mild cognitive impairment (MCI) and early AD. The 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 and estimate the time and cost of clinical trials. ADNI subjects aged 55–90 years old and from over 50 sites across the USA and Canada participated in the research; more detailed information is available at (www.adni-info.org).

Data from 708 (421 men and 287 women, aged 75.42 ± 6.76 years) subjects with the structural MRI data of 164 AD, 346 MCI and 198 normal control provided by the ADNI dataset were used. All MRI data were processed under the following steps to extract ROI data from the MRI data: (i) non-parametric non-uniform bias correction for image intensity inhomogeneity correction (Sled *et al.*, 1998); (ii) skull stripping (Wang *et al.*, 2014) and warping a labeled template to each skull-stripped image for the removal of the cerebellum (aBEAT in version 1.0, http://www.nitrc.org/projects/abeat); (iii) registration of all images to a common template using

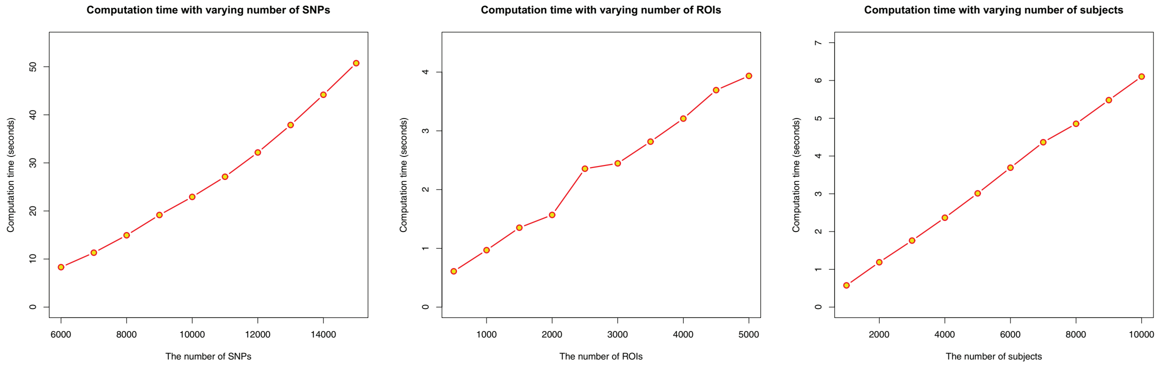

Fig. 1. Computation time with varying number of SNPs p (left panel), the number of ROIs q (middle panel) and the number of subjects n (right panel)

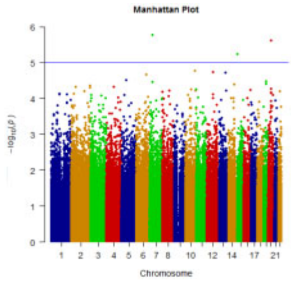

Fig. 2. Manhattan plot of ADNI ROI volume GWAS

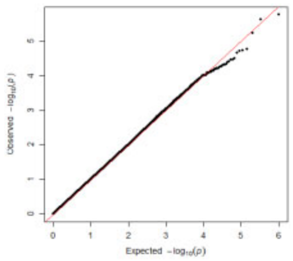

Fig. 3. QQ plot of ADNI ROI volume GWAS

the 4 D-HAMMER method (Shen and Davatzikos, 2004); (iv) automatic labeling of 93 ROIs on the template (Tzourio-Mazoyer *et al.*, 2002); (v) ROI label projection from the template image to each MRI image and (vi) calculation of the volume of each ROI in the labeled image.

We considered the 818 subjects' genotype variables, which includes 620 901 SNPs, acquired using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA) in the ADNI database. We performed quality control and SNP screening on the SNP data as introduced in a previous study (Huang *et al.*, 2015). Moreover, we imputed the remaining missing genotype variables as the modal value. After these procedures, we retained 708 subjects, and each subject had 501 584 SNPs during the subsequent analysis.

3.3.2 Data analysis

We considered the volumes of 93 ROIs as multivariate phenotypic responses to carry out ADNI data analysis via CSRRR. Given that only a small set of SNPs is effective and causal, instead of using the whole set of 501 584 SNPs, we focused on important SNPs that have significant associations with the volumes of 93 ROIs in this study. To detect these important SNPs, we first pre-screened the whole set of 501 584 SNPs based on the P-values obtained from multivariate analysis of variance (MANOVA). We used Pillai's trace test in MANOVA because it is often considered a robust and powerful test statistic (Tabachnick *et al.*, 2007). Specifically, we regressed all volumes of 93 ROIs to each SNP and selected SNPs corresponding to the first 1000 minimum P-values as the predictors of our algorithm. Moreover, to eliminate the influence of covariates (an intercept, gender, age, whole brain volume and the top five principal component scores in SNPs), we regressed ROIs to these covariates and regarded the resulting residuals as the responses of our algorithm.

With the pre-screening procedure, the P-values of the whole set of 501 584 SNPs were calculated. Figures 2 and 3 show the Manhattan and QQ plots of the GWAS results of all volumes of 93 ROIs, respectively. The Manhattan plot is a plot of the $-\log_{10}(P)$-values of the association statistic on the y-axis versus the chromosomal position of SNPs on the x-axis. SNPs with high $-\log_{10}(P)$-values represent high associations with ROIs. In Figure 2, three SNPs are associated with the 93 ROIs in the pre-screening procedure at the 10^{-5} significant level. The QQ plot shows the observed association $-\log_{10}(P)$-values for all SNPs on the y-axis versus the expected uniform distribution of $-\log_{10}(P)$-values under the null hypothesis stating the lack of association on the x-axis. As shown in Figure 3, the observed association $-\log_{10}(P)$-values fit well with the expected $-\log_{10}(P)$-values. The $-\log_{10}(P)$-values in the upper-right tail of the distribution showed a significant deviation, suggesting strong associations between these SNPs and the 93 ROIs.

In this study, rank r, the number of causal SNPs k_x and the number of affected ROIs k_y should be determined in our optimization problem. We set the maximum range $(r_{max}, k_{x_{max}}, k_{y_{max}})$ of these parameters to 10, 30 and 30. Thus, 9000 candidate parameter combinations (r, k_x, k_y) were obtained in total. We applied two nested cross-validation loops to evaluate the proposed method and select the optimal parameters, where 3- and 5-fold cross validations were used for the external and inner loops, respectively. For the external 3-fold cross validation, all 708 samples were divided into three subsets with the same proportion of each class label to retain the consistency of data distribution with the whole 708 samples. For each run, we successively selected one of the three parts as the testing set to calculate the prediction error, whereas the remaining samples in the other two subsets were combined and used as the training set for model fitting. Moreover, parameter tuning was evaluated with the inner 5-fold cross validation of the training set. Specifically, the training set can be further split into a training and validation parts. By varying the values of the different parameters, the proposed model was fitted, and the coefficient matrix was estimated using the samples in the training part. The prediction errors were obtained during validation. This process was repeated five times. Therefore, for each parameter combination, we can achieve an error by averaging the five prediction error values. Finally, we used the training set and the parameter combination with the smallest average error to fit the model again and calculated the prediction error in the testing set. Then, we performed the external 3-fold cross validation to obtain three prediction errors and selected the parameters with the minimum prediction error. With these procedures, we finally

Table 1. Six selected SNPs, and '–' in the table indicates the item was not found to correspond to genes

CHR	SNP	BP	Gene	min *P*-value
2	rs10174624	152 599 705	–	1.252e–04
4	rs2555646	175 437 613	HPGD	1.440e–04
8	rs1866698	137 522 988	–	3.805e–04
11	rs835989	44 949 238	TSPAN18	2.915e–05
12	rs7297570	95 471 330	FGD6	1.688e–04
19	rs2075650	45 395 619	TOMM40	2.419e–09

Table 2. *P*-values between SNP rs2075650 and 13 selected ROIs

	ROI	*P*-value		ROI	*P*-value
1	Amygdala right	2.419e–09	2	Hippocampal formation left	4.532e–08
3	Hippocampal formation right	1.141e–07	4	Entorhinal cortex right	2.747e–06
5	Amygdala left	6.370e–06	6	Uncus left	9.752e–06
7	Superior temporal gyrus right	8.832e–05	8	Insula left	1.214e–03
9	Middle temporal gyrus right	1.350e–03	10	Uncus right	1.584e–03
11	Middle temporal gyrus left	1.739e–03	12	Superior temporal gyrus left	3.661e–03
13	Medial occipitotemporal gyrus right	3.175e–01			

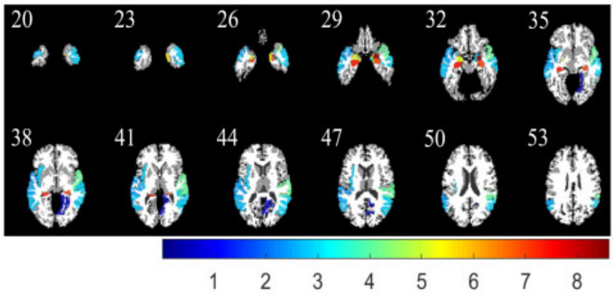

Fig. 4. Selected slices of $-\log_{10}(P)$-values for the selected ROIs corresponding to the SNP rs2075650

obtained the three optimal parameters, whose values were $\hat{r} = 2$, $\hat{k}_x = 6$ and $\hat{k}_y = 13$, respectively.

The six selected SNPs are listed in Table 1, where the minimum *P*-value of each SNP is the smallest *P*-value by regressing the SNP to each selected ROI. From Table 1, the minimum *P*-value was achieved by SNP rs2075650 in translocase of outer mitochondrial membrane 40 homolog (TOMM40). Thus, we regressed each selected ROI to SNP rs2075650 to calculate the *P*-values between this SNP and 13 selected ROIs in Table 2. As shown from the *P*-values in Table 2, the TOMM40 gene is significantly related to amygdala and hippocampus, which play primary roles in the processing of memory and cognition. Therefore, the TOMM40 gene, amygdala and hippocampus are highly related to AD. Figure 4 shows the maps of several selected slices of $-\log_{10}(P)$-values for the selected ROIs corresponding to the SNP rs2075650. In the figure, symmetric clustering across the left and right hemispheres was inspected in selected 13 ROIs. For most brain regions, the brain structures are highly symmetric between hemispheres. Therefore, symmetric associations of SNPs and ROIs can be biologically observed.

The relationship between TOMM40 gene and AD has been discovered and confirmed in several biomedical and clinical studies (Huang *et al.*, 2015). Figures 2 and 3 indicate that several SNPs that were considered significant. This strong interaction weakened the effect of other SNPs on ROI. Thus, the effects of other SNPs on AD are not conducive to identification. To identify more meaningful SNPs and ROIs related to AD, we set a suitable rank $r = 3$ and directly performed Algorithm 3 on 30 SNPs and their corresponding 20 ROIs, i.e. $k_x = 30$ and $k_y = 20$.

Table 3 lists the selected 30 important SNPs associated with 20 ROIs, including the corresponding SNPs, CHR IDs, base pair values, minimum *P*-values and genes. Among the 30 SNPs, the following genes were detected: TOMM40 in CHR 19 is related to AD; PPM1H in CHR 12 is linked with increased AD risk (Badhwar *et al.*, 2017); variation close to MTNR1A (CHR 4) is a shared genetic risk factor for AD in old age (Sulkava *et al.*, 2018); ZNF827 in CHR 4 is associated with APoE4 non-carriers of AD (Jiang *et al.*, 2015); RPA1 in CHR 17 is identified to late-onset AD (Cong *et al.*, 2017). Moreover, we observed several SNPs with potential risks and whose influence on AD has not been revealed in literature. For example, SLC2A1 is related to brain development and function (Gao *et al.*, 2012); SGCZ and TSPAN18 are reported to be associated with schizophrenia and bipolar disorder (Chen *et al.*, 2017); LRRTM4 is linked to cognitive impairment (Chen *et al.*, 2019). Therefore, further investigation should be conducted on these genes in the progression of AD in the future. Such findings might be beneficial to the discovery of new AD-related genetic variations and the early prediction and treatment of this disease.

Figure 5 shows the $-\log_{10}(P)$-values of significant SNP–ROI pairs. As shown in Figure 5, symmetric ROIs across the left and right hemispheres were inspected in 20 ROIs. Among these ROIs, hippocampus, amygdala, parahippocampal gyrus, entorhinal cortex and perirhinal cortex are related to memory; superior temporal gyrus is associated with auditory processing, social cognition processes and function of language; middle temporal gyrus is linked to language processes; inferior temporal gyrus is one of the higher levels of the ventral stream of visual processing. These findings are consistent with those reported in AD prediction and AD imaging genetics studies (Huang *et al.*, 2015; Ning *et al.*, 2018; Zhou *et al.*, 2019). Therefore, the detected ROIs in our study are considered trustworthy.

4 Discussion

In this article, we developed a novel co-sparse best subset selection procedure in multivariate RRR model for the efficient association analysis between genetic variants and imaging phenotypes. Our CSRRR approach accurately identifies the causal SNPs and affected the ROIs simultaneously via l_0 norm constraints. The simulation studies demonstrated that our method achieved competitive accuracy and superior stability in estimating regression coefficients and detected significant ROI-SNP pairs compared with the existing methods. Finally, we applied CSRRR to the association analysis in

Table 3. Thirty selected SNPs associated with 20 selected ROIs

CHR	SNP	BP	min *P*-value	Gene	CHR	SNP	BP	min *P*-value	Gene
19	rs2075650	45 395 619	2.419e–09	TOMM40	4	rs2165666	187 461 706	8.679e–06	MTNR1A
20	rs6014689	54 907 329	1.635e–05	–	11	rs835989	44 949 238	2.915e–05	TSPAN18
4	rs2555646	17 543 7613	1.440e–04	HPGD	9	rs10812321	2 598 844	2.005e–04	VLDLR-AS1
12	rs10506138	40 312 496	2.352e–04	SLC2A13	1	rs12723891	217 143 384	2.459e–04	ESRRG
8	rs7812836	30 352 258	2.483e–04	RBPMS	4	rs1472866	146 849 715	3.112e–04	ZNF827
1	rs12756859	69 009 194	3.565e–04	–	8	rs1866698	137 522 988	3.805e–04	–
8	rs2294066	2 046 700	3.992e–04	MYOM2	3	rs4245878	32 264 265	5.426e–04	–
15	rs4886417	75 330 159	7.874e–04	PPCDC	5	rs10805539	29 931 737	8.277e–04	–
8	rs6995750	14 900 324	1.222e–03	SGCZ	21	rs2298694	47 650 362	1.287e–03	MCM3AP-AS1
11	rs478693	117 263 103	1.583e–03	CEP164	3	rs13075467	135 609 551	1.909e–03	–
14	rs393170	60 618 604	2.297e–03	DHRS7	2	rs10520177	77 228 622	2.668e–03	LRRTM4
10	rs1904592	72 121 024	2.904e–03	LRRC20	19	rs10420030	55 064 414	3.449e–03	–
17	rs8077346	1 786 133	4.265e–03	RPA1	12	rs771972	63 275 017	5.751e–03	PPM1H
1	rs17408450	57 057 316	1.177e–02	–	7	rs1014136	38 229 107	1.994e–02	STARD3NL
1	rs2982285	22 506 253	2.592e–02	–	2	rs4853285	77 188 061	3.393e–02	LRRTM4

Note: '–' in the table indicates the item was not found to correspond to genes.

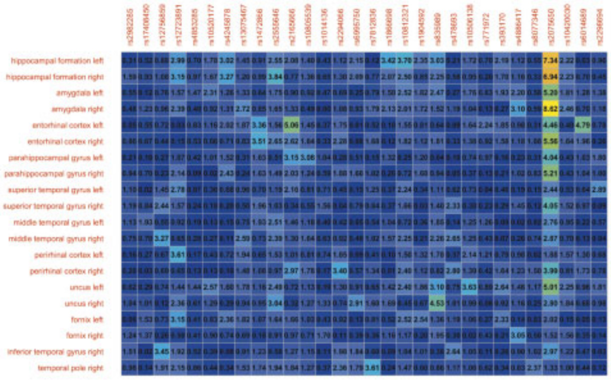

Fig. 5. $-\log_{10}(P)$-values of 30 selected SNPs associated with 20 selected ROIs

imaging genetic data from the ADNI study and obtained meaningful results. Therefore, the CSRRR approach is a valuable statistical toolbox for high-dimensional imaging genetic analysis.

Several potential issues should be addressed in future research. First, the brain structures are highly symmetric between hemispheres for most brain regions. Our proposal detected symmetric association between SNP rs2075650 and ROIs across the left and right hemispheres (Fig. 4) although we did not incorporate this symmetric information into Procedures (2) or (3). We can also consider this information by representing the symmetric structure as a network graph $G = (V, E)$, where V denotes the ROIs, and E denotes the set of undirected edges. Specifically, an edge exists between two ROIs if they are symmetric. To characterize the network G, we defined the adjacency matrix of G by $A = (a_{ij})_{q\times q}$, where $a_{ij} = 1$ if ROI i and j are linked, and $a_{ij} = 0$ otherwise. Then, the Laplacian of network G was defined by $L = D - A$, where D is the diagonal matrix with diagonal element $d_i = \sum_{j=1}^{q} a_{ij}$. In this manner, we can extend the proposed method to incorporate the symmetric information across the left and right hemispheres following

$$\begin{aligned}&\min_C ||Y - XC||_F^2 + \lambda tr(CLC^\top)\\&\text{s.t. } \operatorname{rank}(C) \le r,\ ||C||_{2,0} = k_x,\ ||C^\top||_{2,0} = k_y,\end{aligned}$$

where $\lambda(> 0)$ is a hyperparameter to control the amount of regularization for the symmetric structure, and $tr(M)$ denotes the trace of matrix M. We expect that a similar algorithm, similar to the one presented in this article, will be applied for solving the above problem.

Second, in ADNI data analysis, we removed the exact collinearity before analyzing the data to guarantee the identification of coefficient matrix C. Instead, we dealt with the collinearity problem where the pair-wise correlations between covariates are extremely high by reformulating problem (2) such as that in elastic net (Zou and Hastie, 2005). A regularization term $||C||_F^2 = \lambda tr(C^\top C)$ was added to the objective function in problem (2). With this reformulation, we can solve not only the collinearity problem but also select groups of correlated variables together, such as the SNPs in the same genes or genes in the same biological 'pathway'. Here, we attempted to provide a general and basic framework for co-sparse subset selection. Our future investigation will focus on the collinearity problem.

Funding

This work was supported by the National Natural Science Foundation of China [11801540 to C.H.W.]; the Natural Science Foundation of Guangdong [2017A030310572 to C.H.W.]; the Fundamental Research Funds for the Central Universities [WK2040170015, WK2040000016 to C.H.W.]; the National Natural Science Foundation of China [81601562 to M.Y.H.]; and Science and Technology Planning Project of GuangZhou [201904010417 to M.Y.H.]; the National Natural Science Foundation of China [11701590 to W.L.P.]; the Natural Science Foundation of Guangdong Province of China [2017A030310053 to W.L.P.]; the Young teacher program/Fundamental Research Funds for the Central Universities [17lgpy14 to W.L.P]. Data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012).

Conflict of Interest: none declared.

References

Badhwar,A. *et al.* (2017) Proteomic differences in brain vessels of Alzheimer's disease mice: normalization by PPAR agonist pioglitazone. *J. Cerebral Blood Flow Metab.*, **37**, 1120–1136.

Barrett,J.C. *et al.* (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Chen,X. *et al.* (2017) A novel relationship for schizophrenia, bipolar, and major depressive disorder. Part 8: a hint from chromosome 8 high density association screen. *Mol. Neurobiol.*, **54**, 5868–5882.

Chen,Y.-C. *et al.* (2019) LRRTM4 and PCSK5 genetic polymorphisms as markers for cognitive impairment in a hypotensive aging population: a genome-wide association study in Taiwan. *J. Clin. Med.*, **8**, 1124.

Cong,W. *et al.*; for the Alzheimer's Disease Neuroimaging Initiative. (2017) Genome-wide network-based pathway analysis of CSF t-tau/1-42 ratio in the ADNI cohort. *BMC Genomics*, **18**.

Du,L. *et al.*; for the Alzheimer's Disease Neuroimaging Initiative. (2018) A novel SCCA approach via truncated ℓ1-norm and truncated group lasso for brain imaging genetics. *Bioinformatics*, **34**, 278–285.

Dudbridge,F. (2016) Polygenic epidemiology. *Genet. Epidemiol.*, **40**, 268–272.

Durston,S. (2010) Imaging genetics in ADHD. *Neuroimage*, **53**, 832–838.

Gabriel,S.B. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Gao,J. *et al.* (2012) An exploratory analysis on gene–environment interactions for Parkinson disease. *Neurobiol. Aging*, **33**, 2528.e1–2528.e6.

Gilmore,J.H. *et al.* (2010) Genetic and environmental contributions to neonatal brain structure: a twin study. *Hum. Brain Mapping*, **31**, 1174–1182.

Hashimoto,R. *et al.* (2015) Imaging genetics and psychiatric disorders. *Curr. Mol. Med.*, **15**, 168–175.

Huang,C. *et al.* (2017) FGWAS: functional genome wide association analysis. *Neuroimage*, **159**, 107–121.

Huang,M. *et al.* (2015) FVGWAS: fast voxelwise genome wide association analysis of large-scale imaging genetic data. *Neuroimage*, **118**, 613–627.

Jiang,S. *et al.* (2015) Identification of novel quantitative traits-associated susceptibility loci for APOE epsilon 4 non-carriers of Alzheimer's disease. *Curr. Alzheimer Res.*, **12**, 218–227.

Kong,D. *et al.* (2020) L2RM: low-rank linear regression models for high-dimensional matrix responses. *J. Am. Stat. Assoc.*, **115**, 403–447.

Li,P. *et al.* (2015) An overview of SNP interactions in genome-wide association studies. *Brief. Funct. Genomics*, **14**, 143–155.

Ma,X. *et al.* (2014) Learning regulatory programs by threshold SVD regression. *Proc. Natl. Acad. Sci. USA*, **111**, 15675–15680.

Mishra,A. *et al.* (2017) Sequential co-sparse factor regression. *J. Comput. Graph. Stat.*, **26, 814–825.**

Natarajan,B.K. (1995) Sparse approximate solutions to linear systems. *SIAM J. Comput.*, **24**, 227–234.

Ning,K. *et al.* (2018) Classifying Alzheimer's disease with brain imaging and genetic data using a neural network framework. *Neurobiol. Aging*, **68**, 151–158.

Purcell,S. *et al.* (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Saykin,A.J. *et al.*; Alzheimer's Disease Neuroimaging Initiative. (2015) Genetic studies of quantitative MCI and AD phenotypes in ADNI: progress, opportunities, and plans. *Alzheimer's Dement.*, **11**, 792–814.

Schönemann,P.H. (1966) A generalized solution of the orthogonal procrustes problem. *Psychometrika*, **31**, 1–10.

Shen,D. and Davatzikos,C. (2004) Measuring temporal morphological changes robustly in brain MR images via 4-dimensional template warping. *Neuroimage*, **21**, 1508–1517.

Sled,J.G. *et al.* (1998) A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans. Med. Imaging*, **17**, 87–97.

Sulkava,S. *et al.* (2018) Melatonin receptor type 1A gene linked to Alzheimer's disease in old age. *Sleep*, **41, zsy103**.

Tabachnick,B.G. *et al.* (2007) *Using Multivariate Statistics*, **Vol. 5**. Pearson, Boston, MA.

Tzourio-Mazoyer,N. *et al.* (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, **15**, 273–289.

Vounou,M. *et al.* (2010) Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. *Neuroimage*, **53**, 1147–1159.

Wang,Y. *et al.*; for the Alzheimer's Disease Neuroimaging Initiative. (2014) Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates. *PLoS One*, **9**, e77810.

Wen,C. *et al.* (2019) Best subset selection in reduced rank regression. *arXiv preprint arXiv : 1912.06590*.

Wen,C. *et al.* (2018) Whole genome association study of brain-wide imaging phenotypes: a study of the ping cohort. *Genet. Epidemiol.*, **42**, 265–275.

Wen,C. *et al.* (2017) Bess: an r package for best subset selection in linear, logistic and CoxPH models. *J. Stat. Softw.*, *arXiv preprint arXiv : 1709.06254*.

Witten,D.M. *et al.* (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.

Zhou,T. *et al.* (2019) Brain-wide genome-wide association study for Alzheimer's disease via joint projection learning and sparse regression model. *IEEE Trans. Biomed. Eng.*, **66**, 165–175.

Zhu,H. *et al.* (2014) Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association*, **109**, 977–990.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **67**, 301–320.