Published in final edited form as:

IEEE Trans Biomed Eng. 2018 February ; 65(2): 390–399. doi:10.1109/TBME.2017.2771483.

Adaptive sparse multiple canonical correlation analysis with application to imaging (epi)genomics study of schizophrenia

Wenxing Hu,

Biomedical Engineering Department, Tulane University, New Orleans, LA 70118, USA

Dongdong Lin,

Mind Research Network and Dept. of ECE, University of New Mexico, Albuquerque, NM, 87106

Shaolong Cao,

Department of Bioinformatics & Computational Biology, UT MD Anderson Cancer Center, Houston, TX

Jingyu Liu,

Mind Research Network and Dept. of ECE, University of New Mexico, Albuquerque, NM, 87106

Jiayu Chen,

Mind Research Network and Dept. of ECE, University of New Mexico, Albuquerque, NM, 87106

Vince D. Calhoun [Fellow, IEEE], and

Mind Research Network and Dept. of ECE, University of New Mexico, Albuquerque, NM, 87106

Yu-Ping Wang* [Senior Member, IEEE]

Biomedical Engineering Department, Tulane University, New Orleans, LA 70118, USA

Abstract

Finding correlations across multiple data sets in imaging and (epi)genomics is a common challenge. Sparse multiple canonical correlation analysis (SMCCA) is a multivariate model widely used to extract contributing features from each data while maximizing the cross-modality correlation. The model is achieved by using the combination of pairwise covariances between any two data sets. However, the scales of different pairwise covariances could be quite different and the direct combination of pairwise covariances in SMCCA is unfair. The problem of ‘unfair combination of pairwise covariances’ restricts the power of SMCCA for feature selection. In this paper, we propose a novel formulation of SMCCA, called adaptive SMCCA, to overcome the problem by introducing adaptive weights when combining pairwise covariances. Both simulation and real data analysis show the outperformance of adaptive SMCCA in terms of feature selection over conventional SMCCA and SMCCA with fixed weights. Large-scale numerical experiments show that adaptive SMCCA converges as fast as conventional SMCCA. When applying it to imaging (epi)genetics study of schizophrenia subjects, we can detect significant (epi)genetic variants and brain regions, which are consistent with other existing reports. In addition, several significant brain-development related pathways, e.g., neural tube development, are detected by our model, demonstrating imaging epigenetic association may be overlooked by conventional

SMCCA. All these results demonstrate that adaptive SMCCA are well-suited for detecting three-way or multi-way correlations and thus can find widespread applications in multiple omics and imaging data integration.

Index Terms

imaging genomics; genomic data analysis; canonical correlation analysis; multi-omics data integration; data fusion

I. Introduction

Multiple omics and imaging data integration has been a hot research field in the past decade. Many complex diseases, e.g., schizophrenia, and Alzheimer's disease, have been studied for hundreds of years but a clear disease mechanism is still unknown. Recently, techniques have emerged for characterizing biological mechanism from molecular level, e.g., single nucleotide polymorphisms (SNPs), DNA methylation, mRNA sequencing, to organ and tissue level changes (e.g., brain magnetic resonance imaging), which provide new perspectives on the study of complex human diseases. Given the huge amount of multiscale data generated by these techniques, efficient tools are needed to combine these heterogeneous datasets.

A large amount of works have been conducted to analyze these datasets separately, using simple statistical methods, such as univariate tests and regression models. Various biomarkers, such as specific genes[1] or brain regions[2], have been detected and then further verified by subsequent medical experiments. However, complex diseases are often associated with many different types of biomarkers from multiple biological measures. For instance, schizophrenia, a common mental disorder characterized by abnormal social behaviors, has been shown to be related to abnormal brain development during adolescence [3]. Abnormal brain development is usually reflected by abnormal gene expression [4]. Expression of genes is a complex process not only due to its complicated three-dimensional biological structures but also because of the complex regulation system involving diverse omics data such as DNA sequence, DNA methylation, RNA splicing, mRNA, micro-RNA, protein-protein-interaction, etc. All these factors can affect gene expression and have interactions with each other. Therefore, it is significant but challenging to integrate these heterogeneous omics data to enrich our understanding of biological mechanism underlying complex diseases.

Canonical correlation analysis (CCA) [5] is a popular method for data integration. CCA is a multivariate method that analyzes the correlation between two datasets collected from the same samples, by seeking two linear projections with the highest Pearson correlation. However, CCA suffers from severe overfitting problem when dealing with high dimensional but small sample sized biological data. Also, CCA's output, the canonical vector, is a combination of all variables, making it difficult to interpret the result. To address these issues, sparse CCA methods [6, 7] enforce sparsity regularizations on canonical vectors for biomarker selection. Some extensions of sparse CCA are subsequently proposed to take

advantage of the specific structure [8], group information [9] and phenotype class information that often exists in the data [10].

To integrate three or more datasets, sparse multiple CCA (SMCCA) [11] was proposed as an extension of sparse CCA. SMCCA has been used in many works for identifying biomarkers and for investigating disease mechanisms [12–15]. In addition, several variants of SMCCA, e.g., canonical variate regression [16], and collaborative regression [17], were developed to incorporate class label information. Collaborative regression is a combination of sparse regression and sparse CCA, and is considered as a special case of SMCCA when phenotype information serves as an additional dataset. It can discover a discriminant correlated subset of biomarkers which are relevant to the diseases, however, the prediction accuracy is not improved compared to conventional sparse regression, such as Lasso [18]. Canonical variate regression [16] improves collaborative regression by incorporating logistic regression into the model for the case of binary phenotype study.

In order to improve computational efficiency and to solve the optimization problem, both sparse CCA and SMCCA use a diagonal matrix approximation in the constraint conditions on the objective function. As a result of the diagonal matrix approximation, SMCCA's objective function, which originally is a combination of pairwise correlations, becomes a combination of covariances. The direct combination of pairwise covariances is however unfair in that pairwise covariances differ a lot in the scale. As a consequence, pairwise associations could be over-weighted when their covariances are large, while pairwise associations with small covariances could be overlooked even if their correlations are similar. The difference in the scale of covariance is partly due to the heterogeneous nature of multi-modal omics and imaging data.

In this paper, we propose an adaptively reweighted SMCCA to overcome the problem of 'unfair combination of pairwise covariances' in SMCCA. Inspired by IRLS (iteratively reweighted least squares) [19], adaptive SMCCA allocates weight coefficients iteratively during the optimization procedure. A convergence analysis about adaptive SMCCA is included in this paper. With application to simulated data generated using latent variable model, we show that adaptive SMCCA can significantly improve the performance of feature discovery. We also apply adaptive SMCCA to an imaging genetic and epigenetic study on schizophrenia, by integrating functional magnetic resonance imaging (fMRI), DNA methylation, and SNPs datasets. Real data analysis shows that adaptive SMCCA outperforms conventional SMCCA in that it can better identify important brain regions and genes closely related to schizophrenia.

The rest of the paper is organized as follows. We first propose a novel SMCCA model, called adaptive SMCCA, in the method section. Several numerical experiments are conducted in the simulation section to compare the performance of our model with that of conventional SMCCA and weighted SMCCA. An application of adaptive SMCCA to combining fMRI imaging, genetic variates, and DNA methylation datasets for schizophrenia study is included in the application section. Finally, we draw conclusions about our model, with the discussion about its limitations.

II. Methods

A. Overview of sparse CCA and SMCCA methods

CCA is a statistical model that is widely used for correlation analysis. It seeks the most correlated canonical vectors, by solving an optimization problem. The canonical variable is obtained by maximizing the linear combination of original variables or features through a projection vector in each dataset, and can be derived via singular value decomposition (SVD) [20]. Specifically, given two datasets X_1 and X_2 , CCA seeks two projection vectors u_1 and u_2 by optimizing the following objective function:

$$(u_1, u_2) = \arg \max_{u_1, u_2} u_1^t X_1^t X_2 u_2, \text{ s. t. } u_1^t X_1^t X_1 u_1 \leq 1, u_2^t X_2^t X_2 u_2 \leq 1. \quad (1)$$

Sparse CCA method [6, 7, 21] was introduced recently to resolve the overfitting problem when dealing with high dimension but low sample size datasets. It enforces sparse penalty terms on the canonical vectors u_1 and u_2 to derive sparse vectors. The sparsity of the canonical vector equips sparse CCA with the ability of feature selection or dimension reduction, facilitating result interpretation. The formula of sparse CCA is given by

$$(u_1, u_2) = \arg \max_{u_1, u_2} u_1^t X_1^t X_2 u_2, \text{ s. t. } u_1^t u_1 \leq 1, u_2^t u_2 \leq 1, P(u_1) \leq \lambda_1, P(u_2) \leq \lambda_2. \quad (2)$$

where $P(*)$ is the sparse penalty term (e.g., L_1 norm), and λ_1, λ_2 are the tuning parameters controlling the sparsity level of u_1 and u_2 respectively. The inverse matrix of $X_1^t X_1$ is needed to solve the optimization problem (1). However, for high dimensional data, $(X_1^t X_1)^{-1}$ may not exist because $X_1^t X_1$ could be singular. To address this problem, with the assumption that the features are independent within each data set, the covariance matrix $X_1^t X_1$ is usually replaced with a diagonal or even an identity matrix, which will help yield a good result in practice and reduce computational cost [22]. As a result, constraints $u_i^t X_i^t X_i u_i \leq 1$ ($i=1, 2$) in Eqn. (1) become $u_i^t u_i \leq 1$ ($i=1, 2$) in Eqn. (2). However, without constraints $u_i^t X_i^t X_i u_i \leq 1$, the objective function in (2) becomes ‘to maximize the covariance $u_1^t X_1^t X_2 u_2$ ’ instead of ‘to maximize the correlation $u_1^t X_1^t X_2 u_2 / \|X_1 u_1\| \|X_2 u_2\|$ ’. This sacrifice is considered acceptable given that it can avoid computing $(X_1^t X_1)^{-1}$ and reduce computational expenses.

Sparse multiple CCA (SMCCA), proposed by D.M. Witten et al. [11], is an extension of two-way sparse CCA model when three or more datasets are considered for correlation analysis. SMCCA seeks optimal canonical loading vectors $\mu_1, \mu_2, \dots, \mu_m$ that maximize the linear combination of all the pairwise covariances, which is given by

$$(u_1, u_2, \dots, u_m) = \arg \max_{u_1, u_2, \dots, u_m} \sum_{i < j} u_i^t X_i^t X_j u_j \text{ s. t. } u_i^t u_i \leq 1, P_i(u_i) \leq \lambda_i, \forall i. \quad (3)$$

The objective function of SMCCA (Eqn. (3)) also maximizes pairwise covariance instead of pairwise correlation due to the loss of constraints $u_i^t X_i^t X_i u_i \leq 1, i=1, 2, \dots, m$. This sacrifice in objective function, which is acceptable in two-way sparse CCA, could lead to a severe problem in SMCCA. Unlike pairwise correlations, pairwise covariances differ a lot in terms of the scale of their values. Therefore, it is unfair to directly combine different pairwise covariances. This problem results from the fact that $corr_{ij} \approx corr_{i'j'}$ does not mean $cov_{ij} \approx cov_{i'j'}$. The $cov/corr$ ratio determines how unfair the direct combination of pairwise covariances is,

$$\frac{u_i^T X_i^T X_j u_j}{corr(X_i u_i, X_j u_j)} = \|X_i u_i\| \|X_j u_j\| \leq \|X_i\|_2 \|X_j\|_2, \text{ when } \|u_i\| \leq 1, \|u_j\| \leq 1, \quad (4)$$

where $\|*\|_2$ represents the spectral norm of matrices. According to Eqn. (4), the constraints of spectral norm cannot completely solve the ‘unfair combination’ problem. Similar to Frobenius norm, spectral norm is just another upper bound of the $cov/corr$ ratio. As shown in (4), even though the $cov/corr$ ratio has an upper bound $\|X_1\|_2 \|X_2\|_2$, its value still could be significantly influenced by the canonical loading vectors u_j . The $cov/corr$ ratio varies as canonical vectors change, making the $cov/corr$ ratio to be dynamical during the iterative computation. Therefore, the problem cannot be solved by only normalizing the raw data, or by assigning fixed weights to pairwise covariances. The effect of canonical vectors on the $cov/corr$ ratio was shown in Fig. 2.

As a consequence of the problem of ‘unfair combination of pairwise covariances’, some pairwise correlations are over-weighted when their cov_{ij} are large, while some are overlooked when their cov_{ij} are small even if the correlations are of similar importance. SMCCA is widely used for feature selection (biomarker detection) when integrating multiple datasets. The selected features correspond to the non-zero locations in sparse vector u_j . In consequence, features associated with the overlooked pairwise correlations tend to be neglected by SMCCA, whereas features associated with the over-weighted pairwise correlations tend to be over-selected as important biomarkers. This will substantially lead to false feature selection. To sum up, the unfair combination of pairwise covariances will reduce SMCCA’s power of feature selection (biomarker detection).

It seems weighted SMCCA [17, 23] could relieve the problem of ‘unfair combination of pairwise covariances’ by allocating different weights to different pairwise covariances. However, the prior knowledge of ‘which pairwise covariances are overlooked’ is unknown. Therefore, some criteria are needed to select the weights by evaluating their performances. In practice the weights are usually determined using cross-validation algorithm with a criteria function to evaluate the performance of each candidate weight. For SMCCA, this

criterion function is defined as $\sum |u_i^t X_{i,test}^t X_{j,test} u_j|$, or $\sum ||u_i^t X_{i,test}^t X_{j,test} u_j| - |u_i^t X_{i,train}^t X_{j,train} u_j||$ [6, 21]. In order to get a higher correlation summation, cross validation tends to allocate larger weights to those larger pairwise covariances, making them even more dominant, as illustrated in the following analysis. Assuming $cov_1 = u_1^t X_1^t X_2 u_2$ and $cov_2 = u_1^t X_1^t X_3 u_3$ are two pairwise covariances, cov_1 is larger than cov_2 in absolute value, then

$$w_1 |cov_1| + w_2 |cov_2| > w_2 |cov_1| + w_1 |cov_2|$$

where w_1 and w_2 are the weights, and $w_1 > w_2$. As a result, weighted SMCCA might exacerbate the unfair combination of pairwise covariances.

Moreover, the iterative algorithm of SMCCA makes the ‘unfair combination’ problem even more complicated. According to Eqn. (4), the unfair combination changes dynamically as $u_1, u_2, \dots, u_m$ change during each iteration step. Therefore, weighted SMCCA cannot resolve the ‘unfair combination’ problem by just using fixed weights.

B. Adaptive SMCCA

To address the problem of ‘unfair combination of pairwise covariances’, raw data $X_1, X_2, \dots, X_m$ can be preprocessed so that they have unit spectral norms. However, as shown in (4), this technique only guarantee that they have the same upper bounds, i.e., 1, but the unfair combination problem still exists due to the influence of $u_1, u_2, \dots, u_m$. The effect of $u_1, u_2, \dots, u_m$ is illustrated in the simulation section.

In this paper, we develop a new form of SMCCA, called adaptive SMCCA, to resolve the problem of ‘unfair combination of pairwise covariances’. Adaptive SMCCA can overcome the unfair combination problem by introducing a coefficient to each pairwise covariance, which can adaptively fit the misestimated correlations. The objective function of adaptive SMCCA is given in the following

$$(u_1, u_2, \dots, u_m) = \arg \max_{u_1, u_2, \dots, u_m} \sum_{i < j} \Lambda_{ij} u_i^t X_i^t X_j u_j \text{ s. t. } u_i^t u_i \leq 1, P_i(u_i) \leq \lambda_i, \forall i, \quad (5)$$

where Λ_{ij} is a weight term used to adaptively fit the varying scales of different pairwise covariances during each iteration step. Λ_{ij} should be set properly so that the weighed objective function can reflect the original importance of each correlation, which is given by

$$\Lambda_{i1,j1} u_{i1}^t X_{i1}^t X_{j1} u_{j1} / \text{corr}(X_{i1} u_{i1}, X_{j1} u_{j1}) = \Lambda_{i2,j2} u_{i2}^t X_{i2}^t X_{j2} u_{j2} / \text{corr}(X_{i2} u_{i2}, X_{j2} u_{j2}), \forall i1, j1, i2, j2. \quad (6)$$

Solving (6), we have $\Lambda_{ij} = \|X_i u_i\|_2^{-1} \|X_j u_j\|_2^{-1}$.

An alternating optimization algorithm based iterative method is used to solve problem (5). Within each iteration step, we update $u_1, u_2, \dots, u_m$ alternatively; that is, $u_j (j \in [1, m], j \neq i)$ are fixed when updating u_i . Using the Lagrange multipliers method, the original objective function (5) can be re-written as

$$L(u_i, \Delta_i, \Delta_2) = -u_i^t \sum_{j \neq i} \Lambda_{ij} X_i^t X_j u_j + \Delta_1 \|u_i\|_2^2 + \Delta_2 \|u_i\|_1 \quad (7)$$

Its partial derivative w.r.t. u_i is

$$\frac{\partial L(u_i, \Delta_1, \Delta_2)}{\partial u_i} = -\sum_{j \neq i} \Lambda_{ij} X_i^t X_j u_j + 2\Delta_1 u_i + \Delta_2 \text{sgn}(u_i), \text{ if } u_i \neq 0. \quad (8)$$

According to the Karush-Kuhn-Tucker conditions [24], the optimal u_i satisfies $\|u_i\|_2^2 - 1 = 0$, $\|u_i\|_1 - \lambda_i = 0$, and $L(u_i, \Delta_1, \Delta_2) / \|u_i\|_1 = 0$. Then we have

$u_i = S \left(\sum_{j \neq i} \Lambda_{ij} X_i^t X_j u_j, c_i \right) / \|S \left(\sum_{j \neq i} \Lambda_{ij} X_i^t X_j u_j, c_i \right)\|_2$, where S denotes the soft-threshold operator, i.e., $S(x, c) := \text{sgn}(x)(|x| - c)_+$, and threshold c_i is set to satisfy $\|u_i\|_1 = \lambda_i$. A detailed algorithm for our model is provided in Table I. We use the first singular vectors as the initial input vectors, suggested by Zou et al. [25]. In Table I, $k_1, k_2, \dots, k_m$ are integers that control the sparsity level of $u_1, u_2, \dots, u_m$ respectively; and $u(k)$ denotes the k th largest element of u .

C. Parameter Selection

We use a 5-fold cross-validation approach to select the sparsity parameter. The total samples are divided into 5 subgroups, and during each step we pick up one subgroup as testing sample and use the rest 4 subgroups as training sample set. A fitting score is determined by the difference between the correlation of training sample and that of the test sample, which is used to evaluate the performance of selecting the sparsity parameters. It is very sensitive when using cross-validation to select sparsity parameter λ , which is directly used as the threshold value when updating loading vectors. The sensitivity is due to the non-uniform distribution of loading vector values. To overcome this problem, in this paper we use the sparsity numbers $k_1, k_2, \dots, k_m$ [26] rather than the conventional sparse parameter, to control the sparsity. After the sparsity parameters are fixed, a stability selection approach is used to select a stable subset [27]. We repeatedly perform bootstrapping, a resampling approach, on the whole sample set for B times, which produces B new sample sets. We implement adaptive SMCCA on each of the new sample sets and obtain B different canonical vectors. Candidate features can be selected based on their frequency of occurrence among these B canonical vectors using a frequency cutoff.

D. Convergence analysis

Adaptive SMCCA incorporates a set of coefficients into conventional SMCCA model, in order to achieve a fair estimation among pairwise correlations by adaptively adjusting the coefficients Λ_{ij} . The values of the coefficients reflect the intrinsic properties of those heterogeneous datasets. Each coefficient will gradually approach to an inherent value which fits the heterogeneous data. In consequence, the coefficient term converges very fast in practice.

A 20,000 times of numerical experiment on Gaussian distribution data shows that the average number of iteration steps for the convergence of our model is 15.84, whereas that of SMCCA is 14.70. The average time cost of our model is 0.88 *sec*, whereas that of SMCCA is 0.48 *sec*. This suggests adaptive SMCCA converges as fast as conventional SMCCA. The detailed convergence result is shown in Fig. 1.

III. Simulation

A. Simulation setup

To evaluate the performance of adaptive SMCCA, we simulated three datasets of different sizes: $X_1 \in \mathbb{R}^{n \times p_1}$, $X_2 \in \mathbb{R}^{n \times p_2}$, and $X_3 \in \mathbb{R}^{n \times p_3}$, where n is the sample size, p_1 , p_2 , and p_3 are the sizes of features of X_1 , X_2 , X_3 respectively. A latent variable model [28] as also used in [6, 9] was adopted to simulate the correlation among the three datasets. We create four independent latent variables: μ_1, μ_2, μ_3 , and $\mu_4 \in \mathbb{R}^{200 \times 1}$ (the sample size is set to be 200), which follow a Gaussian distribution. With a sparse loading vector $\beta \in \mathbb{R}^{1 \times p_i}$, each latent variable μ can be expanded to p_i features in each dataset.

$$X_1 = E_1 + \mu_1 \beta_1 + \mu_2 \beta_2 + \mu_4 \beta_3 \quad X_2 = E_2 + \mu_1 \beta_4 + \mu_3 \beta_5 + \mu_4 \beta_6 \quad X_3 = E_3 + \mu_2 \beta_7 + \mu_3 \beta_8 + \mu_4 \beta_9 \quad (12)$$

Each loading vector β has 100 non-zero entries $\beta_{non-zero} \sim U(0.4, 0.6)$. Without the loss of generality, the non-zero entries of $\beta_3, \beta_6, \beta_9$ (connected via μ_4) are located at the first 100 features; similarly, $\beta_1^{non-zero}$ and $\beta_4^{non-zero}$ (connected via μ_1) are located at the second 100 features; $\beta_2^{non-zero}$ and $\beta_7^{non-zero}$ (connected via μ_2) are located at the third 100 features; $\beta_5^{non-zero}$ and $\beta_8^{non-zero}$ (connected via μ_3) are located at the fourth 100 features. Thus, data X_1 and data X_2 have 100 correlated features generated by latent variable μ_1 . X_1 and X_3 have 100 correlated features generated by μ_2 . X_2 and X_3 have 100 correlated features generated by μ_3 . Also, X_1 , X_2 and X_3 have 100 correlated features generated by μ_4 . The goal of the SMCCA models is to detect the 100 three-way correlated features, which are generated by μ_4 . $E_1 \in \mathbb{R}^{200 \times 1000}$, $E_2 \in \mathbb{R}^{200 \times 5000}$, and $E_3 \in \mathbb{R}^{200 \times 20000}$ are noise signals, following a standard normal distribution. Normally we set the white noise signal to be stronger than the latent variable signal.

B. Pairwise covariance versus pairwise correlation

As mentioned in the previous sections, the problem of ‘unfair combination of covariances’ results from the difference of *cov/corr* ratios between different data-pairs. The *cov/corr* ratio

can be influenced by both the spectral norm, which determines the least upper bound of the ratio (Eqn. (4)), and the canonical loading vectors, which determine how far away the ratio can be from the upper bound. To illustrate the effect of canonical vectors and the scale difference of the *cov/corr* ratio, we analyze how *cov/corr* ratio changes with different canonical vector loadings, which is illustrated in Fig.2. During the simulation, all the spectral norms of X_j are set to be 1. However, as Fig.2 shows, *cov/corr* ratio vary greatly as canonical vectors change even though the least upper bounds are set to be equal (spectral norm = 1). With different canonical loadings, the *cov/corr* ratio, which determines how unfair the combination is, can change from around 0.05 to around 0.85, and a dominant pairwise covariance can even become a subsidiary one. In addition, the effect of canonical loading vectors is random and unpredictable. In practice, it can be influenced by some intrinsic factors, e.g. feature dimension, noise level, sparsity level, and data distribution. The *cov/corr* ratio varies as canonical vectors change, making the *cov/corr* ratio to be dynamical during the iterative computation. These simulation results demonstrate that the ‘unfair combination of pairwise covariances’ is a complicated and dynamic problem. Therefore, the problem cannot be solved by only normalizing the raw data, or by assigning fixed weights to pairwise covariances.

C. Performance comparison on simulated data

We compared the performance of our model with SMCCA and weighted SMCCA on the simulated data. The objective of the SMCCA models is to detect the three-way correlated features. For weighted SMCCA, we assigned 9 sets of different weight combination, and each weight coefficient was fixed as 0.1, 1, and 10, which are logarithmically equally spaced. We drew the average receiver operating characteristic (ROC) curves based on 200 replications in Fig. 3, for dataset X_1 , X_2 , and X_3 separately. The true positive rate (TPR) is defined as $\text{TPR} := \Sigma \text{ true positive} / \Sigma \text{ condition positive}$, and true positive means the correctly detected features. The false positive rate (FPR) is defined as $\text{FPR} := \Sigma \text{ false positive} / \Sigma \text{ condition negative}$; and false discovery rate (FDR) is defined as $\text{FDR} := \Sigma \text{ false positive} / \Sigma \text{ test outcome positive}$. As shown in Fig. 3, adaptive SMCCA performs much better than conventional SMCCA model, in that it provides a higher area under curve (AUC) score (SMCCA 0.819 vs. adaptive SMCCA 0.927). From Fig. 3, weighted SMCCA provides a worse feature detection when weights coefficients are set improperly. With proper weight coefficients, weighted SMCCA performs slightly better than conventional SMCCA, but still worse than the adaptive SMCCA. This is because adaptive SMCCA can ensure each pairwise covariance to be given a weight that can adjust adaptively according to the true pairwise correlation in each iteration step. As a result, adaptive SMCCA performs better since no data pair is neither over-weighted nor overlooked for association analysis. Similar results were obtained in noise analysis, as shown in Fig. 4. Adaptive SMCCA offers the best TPR and FDR almost under all noise levels. With a proper set of weight coefficients, weighted SMCCA can perform better than conventional SMCCA. An interesting observation is that the ROC of adaptive SMCCA seems to be the upper limit of the ROCs of weighted SMCCA equipped with different weights. This demonstrates that adaptive SMCCA can find the optimal weights combination that best reflects the original data importance. The difference of performance is more significant at lower noise levels compared to higher noise levels.

The performance of adaptive SMCCA has also been tested under more complicated scenarios. Simulation experiments on new simulated data following a Gamma distribution (shape parameter 2, scale parameter 2) have been performed. The results were presented in Fig. 5 (a–c). According to Fig. 5(a–c), the performance gain of adaptive SMCCA is still significant when data follow a Gamma distribution. The AUCs (area under curve) were 0.899 (adaptive SMCCA) and 0.812 (SMCCA) respectively, demonstrating the performance gain of adaptive SMCCA was still significant when the data follow a Gamma distribution. Simulation experiments have been conducted under the scenario of different true pairwise correlations. The results were shown in Fig. 5 (d–f). The AUCs (area under curve) were 0.936 (adaptive SMCCA) and 0.828 (SMCCA) respectively, demonstrating the performance gain of adaptive SMCCA was still significant under the scenario of different real pairwise correlations. The value of the real/natural/true pairwise correlation does not affect the performance of adaptive SMCCA. The weights in adaptive SMCCA help re-scale the covariance to be close to its true correlation. Even when the true pairwise correlations are quite different, the adaptive weights will make the pairwise covariances to be close to the real correlation, rather than making the pairwise covariances to be equal to each other.

IV. Application to Schizophrenia Data

A. Data Collection and Preprocessing

Schizophrenia (SZ) is a complex mental disorder caused by both genetic risk [29] and environmental factors [30], and SZ patients usually show brain abnormalities [31]. Genome wide association study (GWAS) is commonly used to discover causal genetic variants of diseases, e.g., SNPs. fMRI is an imaging tool to investigate functional activity of the brain. DNA methylation is an epigenetic factor that regulates gene expression, and is influenced by environment factors. Our anticipation is that combining SNPs, fMRI, and DNA methylation will improve the study of the etiology and pathophysiology of schizophrenia. We applied adaptive SMCCA to a multisite schizophrenia study, namely the Mind Clinical Imaging Consortium (MCIC) [32], where three data modalities (SNPs, fMRI, and methylation) were collected from 183 subjects, including 79 SZ patients and 104 healthy controls.

The fMRI dataset was collected during a sensorimotor task. Datasets were pre-processed with SPM12 and were realigned and spatially normalized. After these procedures, the image data consisted of $53 \times 63 \times 46$ voxels. We extracted 41,236 voxels from 116 region of interest (ROI) for analysis, based on the AAL brain atlas [33]. A blood sample was obtained from each subject and DNA was extracted. Genotyping for all participants was performed at the Mind Research Network, covering 1,140,419 SNP loci, out of which 777,365 SNPs loci were retained after quality control. DNA methylation data was also acquired based on blood samples, assessed by the Illumina Infinium Methylation 27k Assay. 27,481 methylation CpG sites were retained after quality control [34], and 9,273 methylation sites were further selected after removing sites with variance less than 1×10^{-4} . In addition, we adjusted the value of these three datasets by removing the effects of age and gender; and the fMRI data were adjusted additionally by removing the sites effects.

B. Results

After data preprocessing, both conventional SMCCA and adaptive SMCCA were applied to the integration of SNPs, DNA methylation, and fMRI datasets. As mentioned in the method section, we first used the cross-validation to select the sparsity parameters, whose final values were 60 (SNPs), 80 (methylation), and 400 (fMRI) respectively; and then utilize the stability selection approach [27] to select the features with a higher frequency of occurrence. During stability selection procedure, both conventional SMCCA and adaptive SMCCA were applied to the MCIC data for 200 times, and each time a resampling was conducted to generate the training set. A cutoff of occurrence frequency is needed to identify the candidate voxels, SNP loci, and CpG sites. The frequency cutoff was set to be 0.6 in Meinshausen et al.'s work [27]. For our data, due to the ultra-high dimensionality and the small sample size, the consistency of biomarker occurrence is lower than that of Meinshausen et al.'s work. As a consequence, the frequency cutoff was set to be 0.2 in our work.

SMCCA detected 26 methylation CpG sites (corresponding to 26 genes) whereas adaptive SMCCA detected 31 CpG sites (corresponding to 30 genes), with 4 common CpG sites detected by both methods, as shown in S. Table I. As shown in Table II, 380 voxels from 7 ROIs (from fMRI data) are detected by conventional SMCCA, whereas 376 voxels from 7 ROIs are detected by adaptive SMCCA, with 4 ROIs in common. As listed in S. Table II, 40 SNPs from 37 genes are identified by SMCCA, whereas 21 SNPs from 20 genes are identified by adaptive SMCCA, with 10 SNPs in common. The slice view and render view of the selected voxels are shown in Fig. 6 and S. Fig 1 respectively. Furthermore, we plot the detailed pairwise correlations among the selected ROIs, SNPs (Genes), and CpG sites in Fig. 7. The pairwise correlations among features detected by adaptive SMCCA are 0.22 (SNP-methylation), 0.22 (SNP-fMRI), and 0.17 (methylation-fMRI). The pairwise correlations of the features detected by SMCCA are 0.20 (SNP-methylation), 0.27 (SNP-fMRI), and 0.11 (methylation-fMRI). From the heat-map (Fig. 7), the methylation-fMRI correlation of adaptive SMCCA (subfigure e.) shows a much warmer color than that of SMCCA (subfigure b).

The brain ROIs detected by the two methods were consistent (4 ROIs detected by both methods). According to a study of functional brain imaging of schizophrenia[35], the difference between schizophrenia patients and healthy controls were mostly located in hippocampus, frontal lobes, and temporal lobes. In particular, the hippocampus gyrus is important in the consolidation of information from short-term memory to long-term memory, and is one of the first brain regions to suffer damage in mental disorders. Both SMCCA and adaptive SMCCA detected some ROIs located within hippocampus, frontal lobes, and temporal lobes. The ROIs identified by adaptive SMCCA are more focused in the three lobes compared with that of conventional SMCCA in terms of both volume and percentage (adaptive SMCCA 6.09 cm^3 with 59.9% versus SMCCA 5.10 cm^3 with 49.7%). Moreover, conventional SMCCA identified cerebellum regions (Cerebellum_Crus2 and Vermis_8), which may be less relevant to the schizophrenia.

The SNPs detected by the two methods have a considerable overlap (10 SNPs in common). Both SMCCA and adaptive SMCCA identified SNPs related to brain development, brain

neuronal activity, and DNA methylation. The SNPs and the corresponding supporting literatures are listed in S. Table II. Among the 10 common detected SNPs, gene *BSX*, *PFTK1*, *AMIGO2* have been reported to have brain development or neural signaling associations. *BSX*, a vital gene for early stage brain structure development, is only expressed in the embryonic period. *BSX* has influence on methylation level, and on some transcription factor, such as *REST* [36]. *PFTK1* transcription is regulated by gene *TRH* [37], which is relevant to brain degenerative diseases. *TRH* can be used as a therapy for mental depression, and cerebellar ataxia. *AMIGO2* protects neuron by producing a protein which is associated with calcium-dependent survival signaling and axon tract development. The expression level of *AMIGO2* is prominent in dorsal and ventral hippocampus and other brain sub-regions [38]. *DCC* is required for mesocorticolimbic dopamine development, and reduced transcription of *DCC* is related to behavioral abnormalities, e.g., schizophrenia [39]. According to a study in the Ashkenazi-Jewish population and the Chinese population, *SMPD1* mutations play a role in some mental disorders, such as Parkinson's disease [40]. *ABCB1* has been used as a treatment of depression. Some studies [41, 42] show that *ABCB1*'s expression at the blood-brain barrier affects brain concentration of many antidepressants, and its genetic variations account for the difference of antidepressant responses in patients. Gene family *CNTNs* have been suggested to be related to neuron development disorders, e.g., bipolar disorder and schizophrenia, etc. [43]. As a gene belonging to *CNTN* family, *CNTN3* promotes the outgrowth of neuron, and has an effect on neuronal networks in the broad brain areas, such as dentate gyrus, which is a part of the hippocampus and hippocampal formation [44].

The detected CpG sites from DNA methylation data have a relatively smaller overlap (4 CpG sites in common). For the CpG sites detected from methylation data (see S. Table I), we performed a pathway enrichment analysis using ConsensusPathDB [45]. 6 pathways and gene ontology (GO) terms were enriched from SMCCA; whereas 6 pathways and GO terms were enriched from adaptive SMCCA, which are listed in Table III. From Table III, adaptive SMCCA detected pathways with more significant p-values and q-values compared to SMCCA. In addition, the pathways detected by adaptive SMCCA had strong brain development or activity associations while that of conventional SMCCA did not. For instance, 'neural tube development' is relevant to brain development and functional communication. 'Tube morphogenesis' is associated with the development and shaping of both blood vascular and neurons [46], which is in accordance with the bold signal in fMRI. The strong relationship between methylation data and the development of nervous systems shows that DNA methylation is relevant to brain disorders, and has the ability to facilitate disease study by conducting epigenetics imaging study.

Compared with conventional SMCCA, adaptive SMCCA provides a more significant set of selected features with a stronger inter-omics association, especially for methylation-fMRI association. Both adaptive SMCCA and conventional SMCCA detected many mental disorders related SNP loci, which reflected the SNP-fMRI interactions. However, adaptive SMCCA detected several significant 'neuron development' related pathways, linked with gene ontology (GO) terms from methylation data (see Table III) whereas conventional SMCCA did not. In addition, the brain ROIs identified by adaptive SMCCA were more focused in significant schizophrenia related lobes compared with that identified by

conventional SMCCA. This outperformance of adaptive SMCCA indicates that the fMRI-methylation association might be overlooked in conventional SMCCA. This overlook may be a consequence of the ‘unfair combination of covariances’ problem, that is, fMRI-methylation covariance is relatively small compared to SNP-methylation and SNP-fMRI covariances. The result shows that adaptive SMCCA is a better method for multiple omics data integration than conventional SMCCA because it can detect those inter-omics relationships overlooked by SMCCA. Since the unfair combination of covariances is very common in practice due to the heterogeneity of multiple omics data, adaptive SMCCA model can find more applications.

V. Discussion and Conclusion

For multiple data integration, the unfair combination of pairwise covariances limits the power of conventional SMCCA for feature selection, resulting from the relaxation of constraint conditions and the heterogeneity among multiple omics. This problem may lead to an oversight of those omics data with smaller pairwise covariances, and/or an over-consideration of those omics data with larger covariances. This unfair combination problem limits the performance of SMCCA for biomarker detection when integrating multiple omics and imaging modalities.

It seems weighted SMCCA [23] might provides a possible way to relieve the unfair combination by introducing weights coefficients. However, according to both theoretical analysis and simulation results, weighted SMCCA tends to perform even worse than conventional SMCCA in terms of feature selection based criteria, e.g., TPR, FDR, and AUC. When equipped with a proper set of weight coefficients, weighted SMCCA performs slightly better than conventional SMCCA. However, the weight tuning coefficients are usually chosen based on the cross-validation method. As discussed in the method section, cross-validation will cause a more severe unbalance since it tends to allocate larger weights to larger pairwise covariances. In real data analysis, the ground truth information is usually unknown and the sample size is limited. Therefore, it is difficult to use training sample to select proper weight coefficients. Moreover, the unfair combination problem is dynamical due to the updating of canonical vectors so that the unfair combination problem cannot be solved by just using fixed weights.

In this paper, the proposed adaptive SMCCA can overcome above problem by adaptively adjusting the weight coefficients on those pairwise covariances, which can improve the power of identifying correlated features among multiple heterogeneous datasets. As a matter of fact, our simulation shows that adaptive SMCCA performs much better than conventional SMCCA and weighted SMCCA in terms of feature selection. In addition, we further applied adaptive SMCCA model to real imaging (epi)genetic datasets (SNPs, fMRI and DNA methylation). Compared with SMCCA, adaptive SMCCA found a more significant subset of SNPs, fMRI and DNA methylation data, with stronger inter-associations that have been reported by previous studies. In particular, our model detected a more significant subset of methylation CpG sites, with a stronger relationship with brain functional development. Further gene enrichment on the CpG sites identified by adaptive SMCCA found some pathways and gene ontology terms, e.g., neural tube development, which were relevant to

the brain. In comparison, no brain related pathway can be enriched based on those methylation sites identified by SMCCA. This indicates that the relation between DNA methylation and fMRI was overlooked by SMCCA. In conclusion, adaptive SMCCA is more powerful for detecting subtle correlations and hence can find widespread applications for multiple omics integration and disease mechanisms discovery.

We conducted imaging (epi)genetics study on schizophrenia subjects as an example of application of adaptive SMCCA. Several schizophrenia related brain regions, e.g., Hippocampus and fusiform, were detected to have relationship with some (epi)genetic biomarkers, which have been verified by previous studies and by further gene enrichment analysis. Several significant brain development related pathways and Gene Ontology terms, e.g., neural tube development, are enriched based on the selected (epi)genetic biomarkers. In conclusion, both simulation and real data experiments show that adaptive SMCCA has a better performance on finding correlations across multiple omics datasets. Moreover, adaptive SMCCA can be extended to accommodate the integration of other types of datasets for disease mechanisms study and disease diagnosis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research is supported by NSF and NIH.

The authors wish to thank the National Institutes of Health (R01GM109068, R01MH104680, R01MH107354, and P20GM103472) and the NSF EPSCoR program (#1539067) for the partial support.

Conflict of Interest: none declared.

References

1. Cao S, Qin H, Gossmann A, Deng HW, Wang YP. Unified tests for fine-scale mapping and identifying sparse high-dimensional sequence associations. *Bioinformatics*. 2015; 32:330–337. [PubMed: 26458888]
2. Deng S-P, Hu W, Calhoun VD, Wang Y-P. Integrating Imaging Genomic Data in the Quest for Biomarkers for Schizophrenia Disease. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2017
3. Toro C, Deakin J. Adult neurogenesis and schizophrenia: a window on abnormal early brain development? *Schizophrenia research*. 2007; 90:1–14. [PubMed: 17123784]
4. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*. 2007; 445:168–176. [PubMed: 17151600]
5. Hotelling H. Relations between two sets of variates. *Biometrika*. 1936; 28:321–377.
6. Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*. 2009; 8:1–34.
7. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*. 2009:kxp008.
8. Du L, Huang H, Yan J, Kim S, Risacher SL, Inlow M, et al. Structured sparse canonical correlation analysis for brain imaging genetics: an improved GraphNet method. *Bioinformatics*. 2016:btw033.
9. Lin D, Calhoun VD, Wang YP. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. *Medical image analysis*. 2014; 18:891–902. [PubMed: 24247004]

10. Fang J, Lin D, Schulz SC, Xu Z, Calhoun VD, Wang YP. Joint sparse canonical correlation analysis for detecting differential imaging genetics modules. *Bioinformatics*. 2016; 32:3480–3488. [PubMed: 27466625]
11. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*. 2009; 8:1–27.
12. Huopaniemi I, Suviataival T, Nikkilä J, Orešič M, Kaski S. Multivariate multi-way analysis of multi-source data. *Bioinformatics*. 2010; 26:i391–i398. [PubMed: 20529933]
13. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The annals of applied statistics*. 2013; 7:523. [PubMed: 23745156]
14. Shen R, Wang S, Mo Q. Sparse integrative clustering of multiple omics data sets. *The annals of applied statistics*. 2013; 7:269. [PubMed: 24587839]
15. Hu, W., Lin, D., Calhoun, VD., Wang, Y-p. Integration of SNPs-FMRI-methylation data with sparse multi-CCA for schizophrenia study. *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*; 2016. p. 3310-3313.
16. Luo C, Liu J, Dey DK, Chen K. Canonical variate regression. *Biostatistics*. 2016:kxw001.
17. Gross SM, Tibshirani R. Collaborative regression. *Biostatistics*. 2015; 16:326–338. [PubMed: 25406332]
18. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1996:267–288.
19. Holland PW, Welsch RE. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods*. 1977; 6:813–827.
20. Golub GH, Reinsch C. Singular value decomposition and least squares solutions. *Numerische mathematik*. 1970; 14:403–420.
21. Waaijenborg S, Verselwe de Witt Hamer PC, Zwinderman AH. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical applications in genetics and molecular biology*. 2008; 7
22. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*. 2003:104–117.
23. Chalise P, Batzler A, Abo R, Wang L, Fridley BL. Simultaneous analysis of multiple data types in pharmacogenomic studies using weighted sparse canonical correlation analysis. *Omics: a journal of integrative biology*. 2012; 16:363–373. [PubMed: 22734853]
24. Kuhn, HWT., AW. *Nonlinear Programming*. University of California Press; Berkeley, Calif: 1951.
25. Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. *Journal of computational and graphical statistics*. 2006; 15:265–286.
26. Xu Z, Chang X, Xu F, Zhang H. $L_{1/2}$ regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on neural networks and learning systems*. 2012; 23:1013–1027. [PubMed: 24807129]
27. Meinshausen N, Bühlmann P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2010; 72:417–473.
28. Jöreskog KG, Goldberger AS. Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*. 1975; 70:631–639.
29. Harrison, PJ., Weinberger, DR. Nature Publishing Group. Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. 2005.
30. Morgan C, Fisher H. Environment and schizophrenia: environmental factors in schizophrenia: childhood trauma—a critical review. *Schizophrenia bulletin*. 2007; 33:3–10. [PubMed: 17105965]
31. Ho BC, Andreasen NC, Nopoulos P, Arndt S, Magnotta V, Flaum M. Progressive structural brain abnormalities and their relationship to clinical outcome: a longitudinal magnetic resonance imaging study early in schizophrenia. *Archives of general Psychiatry*. 2003; 60:585–594. [PubMed: 12796222]
32. Gollub RL, Shoemaker JM, King MD, White T, Ehrlich S, Sponheim SR, et al. The MCIC collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics*. 2013; 11:367–388. [PubMed: 23760817]

33. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*. 2002; 15:273–289. [PubMed: 11771995]
34. Liu J, Chen J, Ehrlich S, Walton E, White T, Perrone-Bizzozero N, et al. Methylation patterns in whole blood correlate with symptoms in schizophrenia patients. *Schizophrenia bulletin*. 2014; 40:769–776. [PubMed: 23734059]
35. Kircher TT, Thienel R. Functional brain imaging of symptoms and cognition in schizophrenia. *Progress in brain research*. 2005; 150:299–604. [PubMed: 16186032]
36. Park SY, Kim JB, Han YM. REST is a key regulator in brain-specific homeobox gene expression during neuronal differentiation. *Journal of neurochemistry*. 2007; 103:2565–2574. [PubMed: 17944879]
37. Shibusawa N, Hashimoto K, Yamada M. Thyrotropin-releasing hormone (TRH) in the cerebellum. *The Cerebellum*. 2008; 7:84. [PubMed: 18418668]
38. Laeremans A, Nys J, Luyten W, D’Hooge R, Paulussen M, Arckens L. AMIGO2 mRNA expression in hippocampal CA2 and CA3a. *Brain Structure and Function*. 2013; 218:123–130. [PubMed: 22314660]
39. Grant A, Fathalli F, Rouleau G, Joobor R, Flores C. Association between schizophrenia and genetic variation in DCC: a case–control study. *Schizophrenia research*. 2012; 137:26–31. [PubMed: 22418395]
40. Gan-Or Z, Orr-Urtreger A, Alcalay RN, Bressman S, Giladi N, Rouleau GA. The emerging role of SMPD1 mutations in Parkinson’s disease: Implications for future studies. *Parkinsonism & related disorders*. 2015; 21:1294–1295. [PubMed: 26320887]
41. Brückl TM, Uhr M. ABCB1 genotyping in the treatment of depression. *Pharmacogenomics*. 2016; 17:2039–2069. [PubMed: 27918249]
42. Xie WW, Zhang L, Wu RR, Yu Y, Zhao JP, Li LH. case-control association study of aBcB1 gene and major depressive disorder in a local chinese han population. *Neuropsychiatric disease and treatment*. 2015; 11:1967. [PubMed: 26347319]
43. Hu J, Liao J, Sathanoori M, Kochmar S, Sebastian J, Yatsenko SA, et al. CNTN6 copy number variations in 14 patients: a possible candidate gene for neurodevelopmental and neuropsychiatric disorders. *Journal of neurodevelopmental disorders*. 2015; 7:26. [PubMed: 26257835]
44. Akane H, Saito F, Shiraki A, Imatanaka N, Akahori Y, Itahashi M, et al. Gene expression profile of brain regions reflecting aberrations in nervous system development targeting the process of neurite extension of rat offspring exposed developmentally to glycidol. *Journal of Applied Toxicology*. 2014; 34:1389–1399. [PubMed: 24395379]
45. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. *Nucleic acids research*. 2013; 41:D793–D800. [PubMed: 23143270]
46. Lubarsky B, Krasnow MA. Tube morphogenesis: making and shaping biological tubes. *Cell*. 2003; 112:19–28. [PubMed: 12526790]

Biographies

WenXing Hu received B.Sc. degree in Applied Mathematics from Xi'an Jiaotong University, China, 2011. Now, he is a Ph.D student in Biomedical Engineering, Tulane University, USA. His research interests include machine learning and statistical learning, dimension reduction, correlation analysis, and multi-omics data integration.

Dr. Dongdong Lin received the bachelor's degree in biomedical engineering from Chongqing University in China in 2007 and the PhD degree in biomedical engineering from the Tulane University, in 2015. He is a postdoctoral researcher in the Mind Research Network institute. His primary research is on mathematical models for integrative analysis of high-dimensional biological data and multi-modal medical imaging to deepen the insights of genetic mechanisms of brain activity changes as well as the pathophysiology underlying complex disorders. He is the coauthor of 18 journal papers and more than 20 conference proceedings.

Dr. Shaolong Cao received B.Sc. degree in Applied Mathematics from Xi'an Jiaotong University, China, 2011, and the Ph.D degree in Biomedical Engineering from Tulane University, USA. Now, he is a postdoctoral fellow in the department of Bioinformatics and Computational Biology at University of Texas, MD Anderson Cancer Center, USA. His research interests include High-throughput genomic data analysis, High-dimensional genetic data inference, generalized linear mixed model and genetic network based approach.

Dr. Jingyu Liu received the Master degree in electrical engineering from the Northern Jiaotong University, Beijing, China, in 1997, and the Ph.D. degree in electrical engineering from the University of New Mexico (UNM), USA, in 2004.

She is currently Associate Professor in Translational Neuroscience at Mind Research Network, Albuquerque, NM, and Research Assistant Professor in the Department of Electrical and Computer Engineering, UNM.

Her research interests include design of multi-model data mining algorithms applied to biomedical data, such as brain imaging and signals, genetic and epigenetic data, clinical and neuropsychological assessments, etc. She has been focusing on the interdisciplinary research field to bridge engineering with neuroscience, and has published more than 50 scientific journal papers.

Dr. Jiayu Chen received the Bachelor degree in electrical engineering from the Shanghai Jiao Tong University, Shanghai, China, in 2002, and the Ph.D. degree in electrical engineering from the University of New Mexico, USA, in 2013. She is currently a research scientist at The Mind Research Network, Albuquerque, NM, USA.

Her research is focused on developing computational algorithms for integrating multi-modal imaging, genetic and phenotypic data to better understand psychiatric disorders including but not limited to schizophrenia, bipolar disorder and attention deficit/hyperactivity disorder. She has published more than 30 peer-reviewed papers in this interdisciplinary field involving engineering, neuroscience, genetics and psychiatry.

Dr. Vince D. Calhoun is currently President of the Mind Research Network and a Distinguished Professor in the Department of Electrical and Computer Engineering at the University of New Mexico. He is the author of more than 550 full journal articles and over 650 technical reports, abstracts and conference proceedings. His work includes the development of flexible methods to analyze functional magnetic resonance imaging such as independent component analysis (ICA), data fusion of multimodal imaging and genetics data, neuroinformatics, and the identification of biomarkers for disease. Among other things, he leads an NIH P20 COBRE center grant on multimodal imaging of schizophrenia, bipolar disorder, and major depression as well as an NSF EPSCoR grant focused on brain imaging and epigenetics of adolescent development. Dr. Calhoun is a fellow of the Institute of Electrical and Electronic Engineers, The American Association for the Advancement of Science, The American Institute of Biomedical and Medical Engineers, The American College of Neuropsychopharmacology, and the International Society of Magnetic Resonance in Medicine.

Dr. Yu-Ping Wang received the BS degree in applied mathematics from Tianjin University, China, in 1990, and the MS degree in computational mathematics and the PhD degree in communications and electronic systems from Xi'an Jiaotong University, China, in 1993 and 1996, respectively. After his graduation, he had visiting positions at the Center for Wavelets, Approximation and Information Processing of the National University of Singapore and Washington University Medical School in St. Louis. From 2000 to 2003, he worked as a senior research engineer at Perceptive Scientific Instruments, Inc., and then Advanced Digital Imaging Research, LLC, Houston, Texas. In the fall of 2003, he returned to academia as an assistant professor of computer science and electrical engineering at the University of Missouri-Kansas City. He is currently a Professor of Biomedical Engineering and Biostatistics & Bioinformatics at Tulane University School of Science and Engineering & School of Public Health and Tropical Medicine. He is also a member of Tulane Center of Bioinformatics and Genomics, Tulane Cancer Center and Tulane Neuroscience Program. His research interests have been computer vision, signal processing and machine learning with applications to biomedical imaging and bioinformatics, where he has about 200 peer reviewed publications. He has served on numerous program committees and NSF/NIH review panels, and served as editors for several journals such as *Neuroscience Methods*.

Fig. 1.

The convergence analysis of SMCCA and adaptive SMCCA for Gaussian distributed data. Both SMCCA and adaptive SMCCA were applied and the number of iteration steps for convergence was recorded. This procedure was replicated 20,000 times to produce the boxplot.

Fig 2.

The figure showing the effect of canonical loading vectors on cov/corr ratio. X-axis represents 10 groups of canonical vectors randomly generated and normalized. Y-axis represents the corresponding cov/corr ratios.

Fig. 3.

The comparison of adaptive SMCCA, conventional SMCCA and weighted SMCCA in terms of ROC: (a) The ROC curve for data X_1 ; (b) The ROC curve for data X_2 ; (c) The ROC curve for data X_3 . Each line represents the average performance of 200 experiments.

Fig. 4.

The comparison of TPR and FDR under different noise levels. (a–c) The TPR for X_1 , X_2 and X_3 respectively; (d–f) The FDR for X_1 , X_2 and X_3 respectively.

Fig. 5.

The ROC comparison between adaptive SMCCA, conventional SMCCA and weighted SMCCA. (a–c) ROC for X_1 , X_2 , and X_3 respectively; Data X_1 , X_2 , and X_3 contain signals following $\Gamma(2,2)$ and white noise. (d–f) ROC for X_1 , X_2 , X_3 respectively; The correlations between X_1 , X_2 , X_3 are set to be unequal: $\text{corr}(X_1, X_2) = 0.4$, $\text{corr}(X_1, X_3) = 0.4$, $\text{corr}(X_2, X_3) = 0.1$.

Fig. 6.

Maps showing the comparison of the detected fMRI brain regions between by SMCCA and adaptive SMCCA. (a) shows the brain region of interests (ROIs) detected by conventional SMCCA; (b) shows the ROIs detected by adaptive SMCCA. The color bar indicates different ROIs.

Fig. 7.

Heat-maps showing the absolute Pearson correlation between the selected fMRI ROIs, SNPs (gene), and CpG sites (gene). (a–c) heat-maps for conventional SMCCA; (d–f) heat-maps for adaptive SMCCA. The lists of detected SNPs (and corresponding genes) and CpG sites (and corresponding genes) are listed in S. Table I and S. Table II.

Table I**Iterative Algorithm for adaptive SMCCA**

Input: Standardized data X_i (size $n \times p_i$), sparsity parameter k_i , $i = 1, 2, \dots, m$. Scale X_i to have unit spectral norm.	
<hr/>	
1.	Compute the initial loading vectors u_i^0 ($i=1, 2, \dots, m$), $\ u_i^0\ =1$.
2.	do
3.	$\Lambda_{ij} = \ X_i u_i\ _2^{-1} \ X_j u_j\ _2^{-1}, i, j=1, 2, \dots, m.$ <p>for i from 1 to m</p> $u_i = \sum_{j \neq i} \Lambda_{ij} X_i^t X_j u_j;$ $u_i = \mathcal{S}(u_i, u_i(k_i + 1)); u_i = u_i / \text{norm}(u_i);$ <p>end</p>
4.	until convergence
5.	Output: First canonical vectors $\mu_1, \mu_2, \dots, \mu_m$
6.	To obtain more canonical vectors, update the covariance matrix
	$X_i^t X_j = X_i^t X_j - \text{tr}(X_i^t X_j u_i u_i^t) u_i u_i^t.$

Table II

Detected brain ROIs (fMRI) and corresponding areas.

SMCCA		Adaptive SMCCA	
<i>ROI name</i>	<i>L/R volume (cm³)</i>	<i>ROI name</i>	<i>L/R volume (cm³)</i>
Frontal_Mid	1.57/*	Frontal_Inf_Tri	1.03/*
Hippocampus	1.59/*	Hippocampus	1.65/*
Fusiform	1.40/1.11	Occipital_Inf	*/1.22
Temporal_Inf	1.94/*	Fusiform	1.86/1.00
Cerebellum_Crus2	1.43/*	Temporal_Mid	*/1.30
Vermis_8	1.22	Temporal_Inf	2.11/*

Table III

Gene enrichment analysis of DNA methylation results

SMCCA			
Pathway name/Gene ontology term	p-value	q-value	Pathway source
Hexose transport	0.00111	0.0127	Reactome
Cell-cell junction organization	0.00146	0.0127	Reactome
SLC-mediated transmembrane transport	0.00214	0.0127	Reactome
Methionine and cysteine metabolism	0.0025	0.0127	EHMN
Cell junction organization	0.00302	0.0127	Reactome
Plasma membrane part	0.00327	0.0621	Gene Ontology
Adaptive SMCCA			
NRF2 pathway	6.40E-05	0.00211	Wikipathways
Arachidonic acid metabolism	0.000481	0.00594	EHMN
Leukotriene metabolism	0.00054	0.00594	EHMN
Tube morphogenesis	0.000851	0.0422	Gene Ontology
SLC-mediated transmembrane transport	0.000918	0.00635	Reactome
Neural tube development	0.000974	0.0422	Gene Ontology