Published in final edited form as:
IEEE/ACM Trans Comput Biol Bioinform. 2020 ; 17(5): 1671–1681. doi:10.1109/TCBB.2019.2899568.

Integration of imaging (epi)genomics data for the study of schizophrenia using group sparse joint nonnegative matrix factorization

Min Wang[1,2], **Ting-Zhu Huang**[1], **Jian Fang**[3], **Vince D. Calhoun**[4], **Yu-Ping Wang**[3,*]

[1]School of Mathematical Sciences/Research Center for Image and Vision Computing, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China

[2]School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, Jiangxi, 330013, China

[3]Department of Biomedical Engineering, Tulane University, New Orleans, LA 70118, USA

[4]The Mind Research Network, University of New Mexico, NM 87131, USA

Abstract

Schizophrenia (SZ) is a complex disease. Single nucleotide polymorphism (SNP), brain activity measured by functional magnetic resonance imaging (fMRI) and DNA methylation are all important biomarkers that can be used for the study of SZ. To our knowledge, there has been little effort to combine these three datasets together. In this study, we propose a group sparse joint nonnegative matrix factorization (GSJNMF) model to integrate SNP, fMRI and DNA methylation for the identification of multi-dimensional modules associated with SZ, which can be used to study regulatory mechanisms underlying SZ at multiple levels. The proposed GSJNMF model projects multiple types of data onto a common feature space, in which heterogeneous variables with large coefficients on the same projected bases are used to identify multi-dimensional modules. We also incorporate group structure information available from each dataset. The genomic factors in such modules have significant correlations or functional associations with several brain activities. At the end, we have applied the method to the analysis of real data collected from the Mind Clinical Imaging Consortium (MCIC) for the study of SZ and identified significant biomarkers. These biomarkers were further used to discover genes and corresponding brain regions, which were confirmed to be significantly associated with SZ.

Keywords

Nonnegative matrix factorization; Group sparsity; SNP; fMRI; Methylation; Feature selection

[*]Corresponding author. wyp@tulane.edu.

Conflict of interest: none declared.

I. Introduction

Schizophrenia (SZ) is a complex mental disorder which affects the way one thinks, feels and acts. It has been widely accepted that both genetic and environmental factors play an important role in the causes of SZ. The disorder tends to inherit in a family. In recent years, many studies have been conducted on exploring critical genes associated with SZ and many genetic variants have been identified, for example, the G72/G30 gene locus on chromosome 13q [1], copy number variations on gene GRIK3, EFNA5, AKAP5 and CACNG2 [2], [3] and gene DISC1 variation [4]. In addition, DNA methylation, one of the main epigenetic markers to regulate gene expression, has also been determined to be involved with the development of SZ. Davies *et al.* showed that the interindividual variations of DNA methylation are significantly correlated between the blood and the brain. Some studies have used blood DNA methylation to identify potential biomarkers for SZ [5], [6]. On the other hand, fMRI has been used to measure brain activity and to identify functional abnormalities within brain regions in SZ [7], [8]. These different datasets (e.g., SNP, fMRI and DNA methylation) represent the same biological sample from different views and provide partial while complementary information; their joint analysis has the potential to reveal the mechanism underlying SZ. Since these imaging and genomic data have different scales and formats, they cannot be simply aggregated for joint analysis. Most of existing works either use single dataset [9], [10], [11] or two datasets [12] and only a few studies [13], [14] exist that can take advantage of three or more datasets for a more comprehensive analysis of SZ.

Canonical correlation analysis (CCA) [15] and partial least squares (PLS) [16] are two popular methods for exploring the relationships between two data sets. The CCA (PLS) method maximizes the correlation (covariance) between the linear combinations of variables from two datasets to find the correlated components. To overcome the small sample size but large dimension of feature problem in imaging (epi)genomics, sparse CCA [17], sparse PLS [18] and sparse reduced rank regression [19] have been proposed by using sparse constraint in the model. To incorporate biological knowledge and group structures (e.g., SNPs within a gene, voxels within a region, and methylation sites within a gene), group sparse CCA [12] and network-regularized PLS [20] model were proposed. However, the above models were for the analysis of pairwise datasets and cannot handle three or more datasets. D.M. Witten *et al.* [21] proposed a sparse multiple CCA (SMCCA) model, which is an extension of two-way sparse CCA model when three or more datasets are considered for correlation analysis. In our recent work *et al.* [22] we proposed an adaptive SMCCA model to adaptively adjust the weight coefficients on the pairwise covariances in the SMCCA model. Despite of these efforts, there has been limited work to combine three or multiple imaging genomics data sets.

Nonnegative matrix factorization (NMF) [23] has been proposed to represent objects by incorporating the nonnegativity constraints, i.e., factorizing the representation matrix into two nonnegative ones. Specifically, it represents data as a linear combination of a set of nonnegative basis vectors. The nonnegativity constraints facilitate the interpretation of discovered latent factors, and many variants of NMF have been developed such as sparse NMF [24], [25], group sparse NMF (GSNMF) [26] and graph regularized NMF [27]. These methods incorporate both prior knowledge and specific data structure (e.g., sparsity, group and graph constrains). Zhang *et al.* [28] proposed a joint nonnegative matrix factorization (JNMF) framework, which simultaneously factorizes multiple data matrices to reveal hidden associations for pattern discovery in cancer genomic data. In JNMF framework, multiple data matrices were projected into a common subspace (i.e., the same set of basis vectors) to explore the correlation among multiple datasets. Since the correlated component pairs correspond to the same basis vectors, it is different from CCA-based or PLS-based model.

In [29], we have employed the JNMF model to extract correlative modules from SNP, fMRI and methylation for the SZ study. However, we haven't taken into account specific group structures in the data. In our preliminary work [30], we proposed a group sparse joint nonnegative matrix factorization (GSJNMF) model for multiple data integration. In current paper, we present the detailed procedure for the solution of the model and the convergence analysis. We also propose a parameter selection strategy based on variable stability selection for the stability of the results. Then we employ the GSJNMF model to identify correlative modules associated with SZ. Based on the modules, we can identify significant genes or biomarkers associated with SZ.

The rest of the paper is organized as follows. In Section II, we briefly review NMF and its variants. In Section III we describe the proposed GSJNMF model and the numerical algorithm for its solution. We then present the experimental results on both simulation and real SZ datasets in Section IV. We conclude the paper in Section V by summarizing the major contributions of the work.

II. Related work

In this section, we will briefly review NMF and some of its variants.

A. NMF

NMF [23] is a matrix factorization algorithm with many applications. Given m samples in $\mathbb{R}^n$, whose elements are all nonnegative and arranged in columns of a nonnegative matrix $X \in \mathbb{R}_+^{m \times n}$, NMF aims to find its approximation with two low-rank nonnegative matrices $W \in \mathbb{R}_+^{m \times r}$ and $H \in \mathbb{R}_+^{r \times n}$, with $r < \min(m, n)$. It can be formulated as

$$\min_{W, H} \|X - WH\|_F^2 \quad \text{s.t. } W, H \geq 0, \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm, $W \in \mathbb{R}_+^{m \times r}$ stores the basis column vectors and $H \in \mathbb{R}_+^{r \times n}$ stores the corresponding column coefficient vectors. r is the number of the basis vectors.

The objective function of NMF in Eq. (1) is nonconvex with respect to W and H, so it is impractical to find the global minimizer. Fortunately, it is convex in either W or H only, so an iterative optimization can be used to find a local minimizer. Lee *et al.* [31] developed the multiplicative update algorithm to solve the optimization problem in Eq. (1) as follows

$$W_{ik} \leftarrow W_{ik}\frac{\left(XH^T\right)_{ik}}{\left(WHH^T\right)_{ik}}, \quad H_{jk} \leftarrow H_{jk}\frac{\left(W^TX\right)_{jk}}{\left(W^TWH\right)_{jk}}. \tag{2}$$

For each column $X_{\cdot j}$, a linear, nonnegative combination of the basis vectors is given by

$$X_{\cdot j} = \sum_{i=1}^{r} W_{\cdot i}H_{ij} = WH_{\cdot j}, \tag{3}$$

where $W_{\cdot i}$ is the i-th column vector of W. Thus, the r basis vector $W_{\cdot i}$ can be considered as the skeleton of the data, while the r-dimensional coefficient vector $H_{\cdot j}$ gives the weights of each basis vectors on $X_{\cdot j}$. The basis vectors can often discover data structures that are latent in the X.

B. Variants of NMF

With the nonnegativity constraint, NMF can learn a parts-based representation of a dataset. However, NMF sometimes fails to discover intrinsic structures of the data, which is essential to the real-world applications. By incorporating prior information about the dataset or enforcing a sparsity constraint, we can further improve the model or make the results more interpretable.

For a nonnegative data matrix $X \geq 0$, Hoyer proposed nonnegative sparse coding (NSC) model [24] to ensure sparsity of the encoding matrix as follows

$$\min_{W,H}\frac{1}{2}\|X - WH\|_F^2 + \lambda\sum_{ij} H_{ij} \quad \text{s.t. } W, H \geq 0, \tag{4}$$

where $\lambda > 0$ is a parameter for balancing the two terms.

In Eq. (4), since $H \geq 0$, we can get

$$\sum_{ij} H_{ij} = \sum_{ij}\left|H_{ij}\right| = \left\|\text{vec}(H)\right\|_1, \tag{5}$$

which is the sparse constraint. vec(·) is a vectorization operator. In [25], Hoyer proposed a novel sparseness measure based on the relationship between the L_1 norm and L_2 norm. Given a vector $x \in \mathbb{R}^n$

$$\text{sparseness }(x) = \frac{\sqrt{n} - \|x\|_1/\|x\|_2}{\sqrt{n} - 1}. \tag{6}$$

This function equals to 1 if and only if x contains only a single non-zero element, and equals to 0 if and only if all elements are equal (up to signs), a trade-off between the two extremes. Hoyer [25] enforced this sparse constraint to the columns of W and rows of H with desired sparseness values as follows

$$\begin{aligned}&\mathrm{sparseness}(W_{.i}) = S_w, \forall i, i = 1, \ldots, r,\\&\mathrm{sparseness}(H_{i.}) = S_h, \forall i, i = 1, \ldots, r,\end{aligned} \tag{7}$$

where S_w and S_h are predefined.

Liu *et al.* [26] proposed a group sparse NMF (GSNMF) model to learn multiple linear manifolds for face recognition. GSNMF imposes the group sparsity constraint on the column vectors of the coefficient matrix to get a group sparse representation. For nonnegative coefficient matrix $H \in \mathbb{R}_+^{r \times n}$, let's assume there are K manifolds and the dimension of each manifold is p. For each row vector of H, it can be divided into K groups and each group has p coefficients. The i-th row, k-th group norm $\|H_{i.}\|_{\mathcal{G}_k}$ is defined as

$$\|H_{i.}\|_{\mathcal{G}_k} = \left(\sum_{\alpha \in \mathcal{G}_k} H_{i\alpha}^2\right)^{\frac{1}{2}}, \tag{8}$$

where $\mathcal{G}_k$ is the column index set of the k-th group and i is the row index. Then the group sparsity for H is defined by

$$\|H\|_{\mathcal{G}} = \sum_{i,k} \|H_{i.}\|_{\mathcal{G}_k} = \sum_{i,k}\left(\sum_{\alpha \in \mathcal{G}_k} H_{i\alpha}^2\right)^{\frac{1}{2}}. \tag{9}$$

If we consider each row as a group, $\|\cdot\|_{\mathcal{G}} = \|\cdot\|_{2,1}$.

All the variants of NMF model described above just factorize one single matrix and cannot handle multiple data matrices. Zhang *et al.* [28] proposed a joint NMF (JNMF) framework, which simultaneously projects multiple types of data matrices onto a common subspace. The common subspace is spanned by nonnegative basis vectors, which can be used to represent the heterogeneous variables with nonnegative weights. Based on the nonnegative weights on a particular basis vector, JNMF can identify the correlated variables and reveal the hidden associations. Given three nonnegative matrices $X_1 \in \mathbb{R}_+^{m \times n_1}$, $X_2 \in \mathbb{R}_+^{m \times n_2}$, $X_3 \in \mathbb{R}_+^{m \times n_3}$, the JNMF model is formulated as follows

$$\begin{aligned}&\min_{W, H_1, H_2, H_3} \sum_{q=1}^{3} \|X_q - WH_q\|_F^2\\&\text{s.t. } W, H_1, H_2, H_3 \geq 0,\end{aligned} \tag{10}$$

where $W \in \mathbb{R}_+^{m \times r}$ stores the common basis vectors shared by the three data matrices and $H_q \in \mathbb{R}_+^{r \times n_q}, (q = 1, 2, 3)$ are the corresponding coefficient vectors.

III. Proposed method

The nonnegative constraint in NMF model only allows additive combinations of the nonnegative basis vectors, which differs from other matrix factorization methods such as singular value decomposition (SVD). As a result, NMF can learn parts-based representation, which find good applications to many real-world problems such as document clustering [32] and DNA gene expression analysis [33]. JNMF simultaneously projects multiple data matrices into a common subspace to explore their correlations but overlook the prior knowledge or specific structure information in the data. In this section, we propose the group sparse JNMF (GSJNMF) model by enforcing group sparse constraint.

A. GSJNMF

We consider three types of datasets from the same samples. After preprocessing, we make the data matrices nonnegative and denote them as $X_1 \in \mathbb{R}_+^{m \times n_1}$, $X_2 \in \mathbb{R}_+^{m \times n_2}$, and $X_3 \in \mathbb{R}_+^{m \times n_3}$, where m is the sample size and n_i is the number of variables in data X_i. For data X_i ($i = 1, 2, 3$), assuming there are K_i disjoint groups in the n_i variables, we denote the group information of variables in X_i as

$$\begin{aligned} \mathscr{G}^i &= \left\{\mathscr{G}_1^i, \mathscr{G}_2^i, \cdots, \mathscr{G}_{K_i}^i\right\} \\ \text{s.t. } & \mathscr{G}_p^i \cap \mathscr{G}_q^i = \varnothing, p \neq q \\ & \bigcup_{j=1}^{K_i} \mathscr{G}_j^i = \{1, 2, \cdots, n_i\}, \end{aligned} \tag{11}$$

where $\mathscr{G}_j^i$ is the column index set of the j-th group in data X_i. The GSJNMF model is then formulated as

$$\begin{aligned} \min_{W, H_1, H_2, H_3} \mathscr{F} &= \sum_{i=1}^{3} \left(\frac{1}{2}\|X_i - WH_i\|_F^2 + \lambda_i \|H_i\|_{\mathscr{G}^i}\right), \\ \text{s.t. } & W, H_1, H_2, H_3 \geq 0, \\ & \|W_{\cdot j}\|_2^2 = 1 \ (j = 1, \cdots, r), \end{aligned} \tag{12}$$

where $W \in \mathbb{R}_+^{m \times r}$, $H_1 \in \mathbb{R}_+^{r \times n_1}$, $H_2 \in \mathbb{R}_+^{r \times n_2}$, $H_3 \in \mathbb{R}_+^{r \times n_3}$. r is a predefined rank and λ_1, λ_2, λ_3 are regularization parameters. $W_{.j}$ is the j-th column basis vector in matrix W. $\|H_i\|_{\mathscr{G}^i}$ is the group sparse penalty term of matrix H_i defined as follows

$$\|H_i\|_{\mathscr{G}^i} = \sum_{j,k} \left(\left|\mathscr{G}_k^i\right| \sum_{\alpha \in \mathscr{G}_k^i} (H_i)_{j\alpha}^2\right)^{\frac{1}{2}}, \tag{13}$$

where $\left|\mathscr{G}_k^i\right|$ is the number of elements in $\mathscr{G}_k^i$.

GSJNMF simultaneously factorizes multiple data matrices and obtains the shared basis vectors. Since the variables in each data matrix have group structure, the group sparse penalty on the coefficient vectors can yield simultaneously nonzero weights within the same group. The constraint on the basis vectors W is used to prevent the elements in W from growing arbitrarily large. In particular, the term $\left|\mathscr{G}_k^i\right|$ in (13) can reduce the effect of group size difference during the process of optimizing the coefficient values.

B. Multiplicative updating algorithm

The objective function of GSNMF in Eq. (12) is nonconvex in W, H_1, H_2, H_3. Therefore, it is difficult for an algorithm to find the global minimizer. In the following, we introduce an iterative algorithm to find a local minimizer. Let $\Psi \in \mathbb{R}^{m \times r}$, $\Phi_1 \in \mathbb{R}^{r \times n_1}$, $\Phi_2 \in \mathbb{R}^{r \times n_2}$, $\Phi_3 \in \mathbb{R}^{r \times n_3}$ be the Lagrange multipliers for constraint $W \geq 0$, $H_1 \geq 0$, $H_2 \geq 0$, $H_3 \geq 0$, respectively; the Lagrange function is then given by

$$\begin{aligned}\mathscr{L} = &\sum_{i=1}^{3}\left(\frac{1}{2}\|X_i - WH_i\|_F^2 + \lambda_i\|H_i\|_{\mathscr{G}^i}\right)\\ &+ \sum_{i=1}^{3}\operatorname{tr}\left(\Phi_i H_i^T\right) + \operatorname{tr}\left(\Psi W^T\right),\end{aligned} \tag{14}$$

where tr(·) is the trace of a matrix. The partial derivatives of $\mathscr{L}$ with respect to W and H_i (i = 1, 2, 3) are

$$\frac{\partial \mathscr{L}}{\partial W} = \sum_{i=1}^{3}\left((WH_i - X_i)H_i^T\right) + \Psi, \tag{15}$$

$$\frac{\partial \mathscr{L}}{\partial H_i} = W^T(WH_i - X_i) + \lambda_i\frac{\partial\|H_i\|_{\mathscr{G}^i}}{\partial H_i} + \Phi_i\,. \tag{16}$$

Based on the Karush-Kuhn-Tucker conditions $\Psi_{jk}W_{jk} = 0$, or specifically $(\Phi_i)_{jk}(H_i)_{jk} = 0$, we can get the following equations for W_{jk} and $(H_i)_{jk}$:

$$\begin{aligned}&\sum_{i=1}^{3}\left(WH_iH_i^T\right)_{jk}W_{jk} = \sum_{i=1}^{3}\left(X_iH_i^T\right)_{jk}W_{jk},\\ &\left(W^TWH_i + \lambda_i\frac{\partial\|H_i\|_{\mathscr{G}^i}}{\partial H_i}\right)_{jk}(H_i)_{jk} = \left(W^TX_i\right)_{jk}(H_i)_{jk}\,.\end{aligned}$$

Then we can get the following multiplicative updating rules:

$$W_{jk} \leftarrow W_{jk}\frac{\sum_{i=1}^{3}\left(X_iH_i^T\right)_{jk}}{\sum_{i=1}^{3}\left(WH_iH_i^T\right)_{jk}}, \tag{17}$$

$$(H_i)_{jk} \leftarrow (H_i)_{jk} \frac{\left(W^T X_i\right)_{jk}}{\left(W^T W H_i + \lambda_i \frac{\partial \|H_i\|_{\mathscr{G}^i}}{\partial H_i}\right)_{jk}}, \tag{18}$$

where

$$\left(\frac{\partial \|H_i\|_{\mathscr{G}^i}}{\partial H_i}\right)_{jk} = \frac{\sqrt{\left|\mathscr{G}_q^i\right|}(H_i)_{jk}}{\sqrt{\sum_{\alpha \in \mathscr{G}_q^i} (H_i)_{j\alpha}^2}}. \tag{19}$$

The iteration will terminate when the relative error of the value of the objective function in Eq. (12) between two iterations is smaller than a predefined tolerance $\tau > 0$. We summarize the algorithm to solve GSJNMF model in Algorithm 1. Since the objective function (12) is nonconvex on W, H_1, H_2, H_3 as a whole, the above algorithm may only find a local minimizer. We repeat the procedure for 100 times with different initialization settings. The solution with the lowest objective function value was used as the final result for further analysis.

C. Convergence analysis

From Eq. (12), we can know that the objective function $\mathscr{F}$ is bounded from below by zero. If we can prove that the $\mathscr{F}$ is nonincreasing under the multiplicative update rule given in Eq. (17) and Eq. (18), the objective function $\mathscr{F}$ will be invariant if and only if W and H_i (i = 1, 2, 3) are at a stationary point. The final solution will be a local minimizer. To simplify the proof, we just prove that $\mathscr{F}$ is nonincreasing under the update rule for H_i (i = 1,2,3). The proof for $\mathscr{F}$ being nonincreasing under the update rule for W can follow a similar way. We introduce an auxiliary function to prove the convergence as Lee *et al.* used in [31] and the definition of the auxiliary function is given in the following.

Definition 1: $G(h, h^*)$ is an auxiliary function for $F(h)$ if the condition $G(h, h^*) \geq F(h)$, $G(h, h) = F(h)$ is satisfied.

Algorithm 1

The algorithm to solve GSJNMF model

```
Input: $X_i \in \mathbb{R}_+^{m \times n_i}$, $\mathscr{G}^i$, $\lambda_i$, ($i$ = 1, 2, 3), $r$, $\tau$.
Initialize $W$ and $H_i$, ($i$ = 1, 2, 3) to random positive matrices.
$k$ = 1.
while 1 do
  Calculate the current function value $f_k$ on Eq. (12).
  update $W$ by using Eq. (17).
  for $j$ = 1 : $r$
    $W_{.j} = W_{.j}/\|W_{.j}\|_2$.
  end for
```

```
    Update $H_i$ ($i$ = 1, 2, 3) by using Eq. (18).
    Calculate the current function value $f_{k+1}$ on Eq. (12).
    **if** $|(f_k - f_{k+1})/f_{k+1}| < \tau$
        **break.**
    **end if**
    $k = k + 1$.
**end while**
Output: $W$, $H_i$, ($i$ = 1, 2, 3).
```

For H_i, considering any element $(H_i)_{jk}$, we use Fi_{jk} to denote the part of $\mathcal{F}$ which is only relevant to $(H_i)_{jk}$. The first order and second order derivative of Fi_{jk} are given as

$$\begin{aligned} Fi'_{jk} &= \left(W^T(WH_i - X_i)\right)_{jk} + \lambda_i\left(\frac{\partial\|H_i\|_{\mathcal{G}^i}}{\partial H_i}\right)_{jk}, \\ Fi''_{jk} &= \left(W^TW\right)_{jj} + \lambda_i\left(\frac{\partial^2\|H_i\|_{\mathcal{G}^i}}{\partial^2 H_i}\right)_{jk}. \end{aligned}$$

Lemma 1: Function

$$\begin{aligned} &G(h, (H_i^{(t)})_{jk}) \\ &= Fi_{jk}((H_i^{(t)})_{jk}) + Fi'_{jk}((H_i^{(t)})_{jk})(h - (H_i^{(t)})_{jk}) \\ &+ \frac{(W^TWH_i^{(t)})_{jk} + \lambda_i\left(\frac{\partial\left\|H_i^{(t)}\right\|_{\mathcal{G}^i}}{\partial H_i^{(t)}}\right)_{jk}}{2(H_i^{(t)})_{jk}}(h - (H_i^{(t)})_{jk})^2 \end{aligned} \tag{20}$$

is an auxiliary function for Fi_{jk}.

Proof: Obviously, $G(h, h) = Fi_{jk}(h)$, we only need to prove $G(h, (H_i^{(t)})_{jk}) \geq Fi_{jk}(h)$. The Taylor series expansion of $Fi_{jk}(h)$ in $(H_i^{(t)})_{jk}$ is given as

$$\begin{aligned} &Fi_{jk}(h) \\ &= Fi_{jk}((H_i^{(t)})_{jk}) + Fi'_{jk}((H_i^{(t)})_{jk})(h - (H_i^{(t)})_{jk}) \\ &+ \frac{Fi''_{jk}((H_i^{(t)})_{jk})}{2}(h - (H_i^{(t)})_{jk})^2. \end{aligned} \tag{21}$$

Comparing Eq. (20) with Eq. (21), we can find that $G(h, (H_i^{(t)})_{jk}) \geq Fi_{jk}(h)$ is equivalent to

$$\begin{aligned} &(W^TWH_i^{(t)})_{jk} + \lambda_i\left(\frac{\partial\left\|H_i^{(t)}\right\|_{\mathcal{G}^i}}{\partial H_i^{(t)}}\right)_{jk} \\ &\geq \left(\left(W^TW\right)_{jj} + \lambda_i\left(\frac{\partial^2\left\|H_i^{(t)}\right\|_{\mathcal{G}^i}}{\partial^2 H_i^{(t)}}\right)_{jk}\right)(H_i^{(t)})_{jk}. \end{aligned} \tag{22}$$

Obviously, we have

$$\begin{aligned}&\left(W^TWH_i^{(t)}\right)_{jk}\\&=\sum_v\left(W^TW\right)_{jv}\left(H_i^{(t)}\right)_{vk}\geq\left(W^TW\right)_{jj}\left(H_i^{(t)}\right)_{jk}.\end{aligned}\tag{23}$$

From Eq. (19), we have

$$\left(\frac{\partial\left\|H_i^{(t)}\right\|_{\mathcal{G}^i}}{\partial H_i^{(t)}}\right)_{jk}=\frac{\sqrt{\left|\mathcal{G}_q^i\right|}\left(H_i^{(t)}\right)_{jk}}{\sqrt{\sum_{\alpha\in\mathcal{G}_q^i}\left(H_i^{(t)}\right)_{j\alpha}^2}}.\tag{24}$$

We can also get

$$\begin{aligned}&\lambda_i\left(\frac{\partial^2\left\|H_i^{(t)}\right\|_{\mathcal{G}^i}}{\partial^2H_i^{(t)}}\right)_{jk}\left(H_i^{(t)}\right)_{jk}\\&=\lambda_i\left(\frac{\partial\left\|H_i^{(t)}\right\|_{\mathcal{G}^i}}{\partial H_i^{(t)}}\right)_{jk}\left(1-\frac{\left(H_i^{(t)}\right)_{jk}^2}{\sum_{\alpha\in\mathcal{G}_q^i}\left(H_i^{(t)}\right)_{j\alpha}^2}\right)\\&\leq\lambda_i\left(\frac{\partial\left\|H_i^{(t)}\right\|_{\mathcal{G}^i}}{\partial H_i^{(t)}}\right)_{jk}.\end{aligned}\tag{25}$$

Adding Eq. (23) and Eq. (25), we can get Eq. (22). Thus $G(h,(H_i^{(t)})_{jk})\geq Fi_{jk}(h)$ holds and function $G(h,(H_i^{(t)})_{jk})$ is an auxiliary function for Fi_{jk}. ∎

Lemma 2: If G is an auxiliary function of F, then F is nonincreasing under the update $h^{(t+1)}$ = arg min $G(h, h^{(t)})$.

Proof:

$$\begin{aligned}&G\left(h^{(t+1)},h^{(t+1)}\right)=F\left(h^{(t+1)}\right)\\&\leq G\left(h^{(t+1)},h^{(t)}\right)\leq G\left(h^{(t)},h^{(t)}\right)=F\left(h^{(t)}\right),\end{aligned}$$

then $F(h^{(t+1)})\leq F(h^{(t)})$. So F is nonincreasing under the update $h^{(t+1)}$ = arg min $G(h, h^{(t)})$. ∎

Based on Lemma 2, we only show that the multiplicative updating rules given in Eq. (18) is the optimum of $G(h,(H_i^{(t)})_{jk})$.

$$\begin{aligned}&\frac{\partial G(h,(H_i^{(t)})_{jk})}{\partial h}\\&=\frac{\left(W^TWH_i^{(t)}\right)_{jk}+\lambda_i\left(\frac{\partial\left\|H_i^{(t)}\right\|_{\mathcal{G}^i}}{\partial H_i^{(t)}}\right)_{jk}}{\left(H_i^{(t)}\right)_{jk}}(h-(H_i^{(t)})_{jk})\\&+\left(W^T(WH_i-X_i)\right)_{jk}+\lambda_i\left(\frac{\partial\left\|H_i\right\|_{\mathcal{G}^i}}{\partial H_i}\right)_{jk}=0.\end{aligned}\tag{26}$$

Then we can get

$$h = \left(H_i^{(t)}\right)_{jk} \frac{\left(W^T X_i\right)_{jk}}{\left(W^T W H_i^{(t)} + \lambda_i \frac{\partial \left\| H_i^{(t)} \right\|_{\mathcal{G}^i}}{\partial H_i^{(t)}}\right)_{jk}}, \tag{27}$$

which is the updating rule of H_i given in Eq. (18). Due to the property of the auxiliary function, Fi_{jk} is nonincreasing under this updating rule.

D. Identification of modules

After GSJNMF, the three datasets were projected onto a common subspace whose basis vectors were stored in the basis matrix W. The basis vectors can be considered as the skeleton of the three datasets. For j-th basis vector, we can select the variables in X_q with large coefficients in j-th row of H_q to form a membership variable set $S_j^q, (j = 1, 2, \cdots, r; q = 1, 2, 3)$. Since S_j^1, S_j^2 and S_j^3 all have large coefficients on the j-th basis vector in the common subspace, the j-th basis vector is the bridge that link the three member variable sets. In other words, each basis vector can define a module and the member variables across the three datasets in the module are correlated. For example, $S_j = \{S_j^1, S_j^2, S_j^3\}$ is a module corresponding to the j-th basis vector. In [33], researchers have used the maximum of each column of coefficient matrix to determine the variable's membership. In this way, each variable can belong to only one module. However, some variables may be either inactive in any module or active in multiple modules.

Considering above facts, we calculated the z-score [28] for each element in each row of coefficient matrices H_q (q=1, 2, 3) by

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \tag{28}$$

where μ_i is the mean value of i-th row vector in H_q and σ_i is the standard deviation. For H_q, $z_{ij} > T$ means that the j-th variable in dataset X_q is a member of $S_i^q, q = 1, 2, 3$ and $T > 0$ is a given threshold. Since GSJNMF extract the correlated variables from multiple datasets based on the shared basis vectors from the same sample, this model can only be used for multiple datasets from the same subject.

E. Significance estimation

For module $S = \{S^1, S^2, S^3\}$, we expect that the variables in S^1, S^2 and S^3 are correlated. To check if such relationship is statistically significant, we employ a permutation test to estimate the P-value of the identified modules. Assuming the number of elements in S^1, S^2 and S^3 are l_1, l_2 and l_3, respectively, we then denote them by $A = \left[a_1, a_2, \cdots, a_{l_1}\right],\ B = \left[b_1, b_2, \cdots, b_{l_2}\right], C = \left[c_1, c_2, \cdots, c_{l_3}\right]$, where a_i, b_j, c_k are column vectors from X_1, X_2, X_3, respectively, and the length of the vector is the number of samples. We use $\rho(x, y)$ to represent the Pearson correlation between x and y. Based on the above assumption, the mean correlation among the three datasets in a module can be given by

$$\rho^* = \frac{1}{3}\left(\frac{1}{l_1 l_2}\sum_{i=1}^{l_1}\sum_{j=1}^{l_2}\left|\rho\left(a_i, b_j\right)\right| + \frac{1}{l_1 l_3}\sum_{i=1}^{l_1}\sum_{k=1}^{l_3}\left|\rho(a_i, c_k)\right| + \frac{1}{l_2 l_3}\sum_{j=1}^{l_2}\sum_{k=1}^{l_3}\left|\rho\left(b_j, c_k\right)\right|\right). \tag{29}$$

We permutate the row order of matrices A and B while keep the matrix C unchanged for Θ times. For each permutation, the mean correlation $\rho_\theta^*(\theta = 1, 2, \cdots, \Theta)$ can be calculated by Eq. (29), which is used to build the null distribution of the mean correlation. By large number of permutations, the significance of the mean correlation can be evaluated by

$$\text{P-value} = \left|\left\{\theta \mid \rho_\theta^* \geq \rho^*, \theta = 1, 2, \cdots, \Theta\right\}\right| / \Theta, \tag{30}$$

where $|\cdot|$ is the number of elements in the set. Variables with P-values smaller than $0.05/r$ were considered to be significant.

F. Parameter selection

For NMF-based model, how to determine the number of basis vectors r is still a challenging problem. If r is too large, the matrices will be over-factorized and we cannot achieve the goal of dimension reduction. If r is too small, e.g., $r = 1$, the limited basis vectors will represent the original data with a large residual error and we cannot discover the hidden skeleton in the datasets. A common method is to choose r based on the stability of the corresponding solutions [33]. Most of the time, we prefer a smaller value and let $r \ll \min(\mathrm{m}, n_1, n_2, n_3)$. The convergence tolerance was set to $\tau = 1 \times 10^{-6}$. As for the regularization parameters λ_1, λ_2, λ_3, we apply grid search method based on variable stability selection to find the optimal value, which were proposed by Sun *et al.* in [34]. In Sun *et al.*'s work, there is only one sparse coefficient vector used to select variables. However, in our model, based on multiple coefficient vectors, we select multiple set of variables for the basis vectors simultaneously. Not all of the modules will be adopted in the following analysis and we just keep the modules with significant mean correlations defined in Eq. (29). In a word, we cannot use the pipeline in our model directly. Given a decreasing sequence for λ_1, λ_2, λ_3, we assume $\Lambda = \{\Lambda^1, \Lambda^2, \cdots, \Lambda^v\}$, where $\Lambda^j = \left[\lambda_1^j, \lambda_2^j, \lambda_3^j\right]$. is the j-th parameter combination ($j = 1, 2, 3, \cdots, v$). The procedure of selecting the optimal combination of parameters is as follows:

1. The sample set was denoted as Ω and $|\Omega| = m$, where m is the number of matrices rows. We randomly partition the Ω into two disjoint sets Ω^1 and Ω^2 and $|\Omega^1| = |\Omega^2| = m/2$. If m is an odd number, we delete one sample randomly. For Λ^j, we perform the GSJNMF on Ω^1 and Ω^2 and obtain two sets of significant modules represented as M_j^1, M_j^2, respectively. The modules in these two sets were sorted by ascending order of P-values.

2. Assume $\delta^j = \min\left(\left|M_j^1\right|, \left|M_j^2\right|\right)$, $\delta = \max\left(1, \min\left\{\delta^1, \delta^2, \cdots, \delta^v\right\}\right)$. Once we get δ, it will not change in the following operations.

3. For Λ^j ($j = 1, 2, \cdots, \nu$), we rerun Step 1) and extract the top δ modules in M_j^1 and M_j^2. For each dataset and each sample set, we compute the union set of variables in the δ modules. For example, $\mathscr{A}_j^i$ is the union set of variables in the first dataset of Ω^i; $\mathscr{B}_j^i$ is the union set of variables in the second dataset of Ω^i; and $\mathscr{C}_j^i$ is the union set of variables in the third dataset of Ω^i ($i = 1, 2$).

4. The variable selection stability of the first dataset can be measured by Cohen's kappa coefficient [35] as follows

$$\kappa(\mathscr{A}_j^1, \mathscr{A}_j^2) = \frac{P_a\left(\mathscr{A}_j^1, \mathscr{A}_j^2\right) - P_c\left(\mathscr{A}_j^1, \mathscr{A}_j^2\right)}{1 - P_c\left(\mathscr{A}_j^1, \mathscr{A}_j^2\right)}, \tag{31}$$

where P_a is the relative probability of observed agreement and P_c is the hypothetical probability of chance agreement, which can be calculated by using the method in [34]. Analogically, the variable selection stability of the second and third datasets can also be measured. We use the mean value of these three Cohen's kappa coefficient κ_j to represent the variable selection stability on the parameters combination Λ_j.

5. Repeat Step 3), 4) for D times and the d-th mean Cohen's kappa coefficient for Λ^j is κ_j^d. The average variable selection stability of these D times repeats is given by

$$\bar{\kappa}_j = \frac{1}{D}\sum_{d=1}^{D} \kappa_j^d, j = 1, 2, \cdots, \nu\,. \tag{32}$$

We select the parameter combination, corresponding to the largest average variable selection stability.

IV. Materials and results

A. Simulation study

To assess the performance of the proposed GSJNMF model, we simulate three datasets with correlated components and then we compare JNMF and GSJNMF based on their abilities to identify the hidden associations within these simulated datasets. Since we want to simulate the group effect in the datasets, each dataset consists of some disjoint groups. The variables in a group are generated based on one single seed vector. For example, a group component with n variables can be denoted as

$$\alpha[n] = \{\beta_i \mid \beta_i = \alpha + \sigma\eta_i, i = 1, 2, \ldots, n\}, \tag{33}$$

where $\alpha \in \mathbb{R}^m$ is a seed vector with entries randomly chosen from the standard normal distribution and $\eta_i \in \mathbb{R}^m$ is a Gaussian noise vector. $\sigma \in \mathbb{R}^+$ indicates the noise level. We arrange the column vectors β_i, $(i = 1, 2, \ldots, n)$ to form a sub-matrix of size $m \times n$. The group components generated from the same seed vector are considered as correlated. Then these sub-matrices are concatenated to form the final data matrix. Based on the number of variables in each group and the number of groups in each data matrix, we generate four types of cases for the three data matrices in Table I, where "√" means equal while "×" means unequal. Two correlated group components are set across the three data matrices in all cases.

In our simulation test, the length of variable vector $m = 40$. We generate the seed vectors randomly from the standard normal distribution and form the three data matrices as follows,

1. Case1: $X_1 \leftarrow \{\alpha_1[20], \alpha_2[20], \alpha_3[20], \alpha_4[20], \alpha_5[20]\}$, $X_2 \leftarrow \{\alpha_3[20], \alpha_6[20], \alpha_1[20], \alpha_7[20], \alpha_8[20]\}$, $X_3 \leftarrow \{\alpha_9[20], \alpha_{10}[20], \alpha_{11}[20], \alpha_3[20], \alpha_1[20]\}$.
2. Case2: $X_1 \leftarrow \{\alpha_1[15], \alpha_2[5], \alpha_3[9], \alpha_4[12], \alpha_5[10]\}$, $X_2 \leftarrow \{\alpha_3[11], \alpha_6[19], \alpha_1[11], \alpha_7[7], \alpha_8[2]\}$, $X_3 \leftarrow \{\alpha_9[6], \alpha_3[8], \alpha_{10}[17], \alpha_{11}[18], \alpha_1[6]\}$.
3. Case3: $X_1 \leftarrow \{\alpha_1[20], \alpha_2[20], \alpha_3[20], \alpha_4[20], \alpha_5[20]\}$, $X_2 \leftarrow \{\alpha_2[20], \alpha_6[20], \alpha_1[20]\}$, $X_3 \leftarrow \{\alpha_7[20], \alpha_8[20], \alpha_9[20], \alpha_{10}[20], \alpha_1[20], \alpha_{11}[20], \alpha_2[20]\}$.
4. Case4: $X_1 \leftarrow \{\alpha_1[10], \alpha_2[13], \alpha_3[8], \alpha_4[9], \alpha_5[11]\}$, $X_2 \leftarrow \{\alpha_1[8], \alpha_6[11], \alpha_3[13]\}$, $X_3 \leftarrow \{\alpha_7[8], \alpha_8[9], \alpha_3[12], \alpha_9[10], \alpha_{10}[7], \alpha_1[13], \alpha_{11}[11]\}$.

In each case, there are two correlated modules across the three data matrices. The corresponding variable indices of the correlated group components are displayed in Table II.

We normalize the three data matrices and make them to fit the constraints of nonnegativity with the following transformations. First, we standardize each column vector in the matrices to make the mean value be 0 and variance be 1. Second, we use the function $F(x) = x - \min(x)+\epsilon$ to make each column be nonnegative, where x is the objective column vector and $\epsilon \sim \mathrm{unif}(0, 10^{-3})$. Third, we normalize each column vector to make the L_2 norm equal to 1. Fourth, we scale all the matrices so that the three data matrices have the same Frobenius norm. It is worth noting that the operations from the second to fourth step are all linear transformations, which will not change the Pearson correlation between the column vectors.

In the synthetic data experiments, because there are only two correlated modules in our design, we set $r = 5 > 2$. Since we identify the modules based on the z-score of each variables in the coefficient vectors, in each case we assess the performance of the GSJNMF and JNMF models by comparing the distribution of the z-scores of the variables in their identified modules. Based on the variable indices of the correlated module in Table II, we draw Figure A.1(a), A.2(a), A.3(a), A.4(a) (see Appendix A) to show the active variable indices in the four cases. The color of a line indicates a corresponding dataset. For example, red, blue and green means that the variables are from X_1, X_2 and X_3, respectively. In each case, we calculate the z-score of the 5 modules from JNMF and GSJNMF models and select the modules that are most likely to be the true correlated modules we generated. Three noise levels ($\sigma = 0.5, 1, 1.5$) are considered in each case and the z-scores of the variables of the two selected modules from JNMF and GSJNMF model are displayed in Figure A.1–A.4 (see Appendix A). A variable with a large z-score means that it is active in the module.

From the results of the four cases, we can know that the z-score of the variables in the modules from GSJNMF model can reflect the true correlated modules. When the noise level $\sigma = 0.5$, the z-scores of the variables in the same group are at the same level, and the z-score line looks very flat. If we increase the noise level to $\sigma = 1$, the GSJNMF model can still find the true correlated modules, but the z-score line becomes fluctuant. When the $\sigma = 1.5$, the fluctuation gets more severe. It's worth noting that for all noise levels, the high z-score variables are all in the true correlated modules and the variables with Z-scores outside of the correlated modules are all very small. In this way, the GSJNMF model can identify the true variables in the modules.

As to the JNMF model, in the experiments of the four cases, even though the z-score of the variables in the true module are large, some variables outside of the true modules are still very large because of the fluctuations in the lines. These variables will be easily identified as wrong active variables. For example, in Figure A.1(a) (see Appendix A), the z-score of the variables (index from 21 to 40) from X_2 in JNMF-module2 are large, but are not the members of module2. When the noise level $\sigma = 0.5$, the distance between the lines in the large z-score region is relatively large. When the noise level $\sigma = 1$, the z-score lines become closer due to the increase of fluctuation and it is difficult to separate the true modules from the whole variables. When the noise level $\sigma = 1.5$, the fluctuations become more severe and we can hardly get any module information from the lines.

From this comparison, we can know that JNMF model is more sensitive to noise. In GSJNMF model, even though there are some fluctuations, the z-score lines still reflect the true variables in each module. When there exists group structure in the datasets, GSJNMF can employ this group information and improve performance.

B. Data preparation and preprocessing

Participants in this study were from the Mind Clinical Imaging Consortium (MCIC). 80 SZ patients (age: 34±11, 20 females) and 104 healthy controls (age: 32 ± 11, 38 females) were analyzed here. We used three types of datasets (e.g., SNP, fMRI, DNA methylation data) of the 184 samples. Each SNP was categorized into three clusters based on their genotype and was represented with discrete numbers: 0 for 'BB' (no minor allele), 1 for 'AB' (one minor allele) and 2 for 'AA' (two minor alleles). The fMRI data were extracted with 53×63×46 voxels and all the voxels with missing measurements were excluded. 116 ROIs were extracted based on the AAL brain atlas. DNA from blood samples was assessed by the Illumina Infinium Methylation27 Assay. A methylation value represents the ratio of the methylated probe intensity to the total probe intensity. We followed the same preprocessing procedures in [12] for SNP and fMRI, as well as the one [11] for DNA methylation, resulting in 722,177 SNPs, 41,236 fMRI voxels and 27,508 methylation sites, respectively. Since we want to find the biomarkers only associated with SZ, we applied the t-test to these three datasets between SZ and healthy samples, and only selected those variables with P-value < 0.05. For SNP data, we only keep the SNPs included in the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway [36]. After variable selection, we obtained 10,351 SNPs, 2,428 fMRI voxels and 2,724 methylation sites in 2,006 genes, 76 brain regions, 2,134 genes from 184 samples, which were represented in three matrices $X_1 \in \mathbb{R}^{184 \times 10351}$, $X_2 \in \mathbb{R}^{184 \times 2428}$ and $X_3 \in \mathbb{R}^{184 \times 2724}$, respectively. We used the same procedure to process the three data matrices and normalize them to the same value level. We then applied the GSJNMF model to the three matrices.

C. Module discovery and validation

The variables in SNP, fMRI and DNA methylation datasets were grouped based on the genes and brain regions of interest (ROIs) (e.g., SNPs within the same gene, voxels within the same region, and methylation sites within the same gene). As a result, there were 2,006 groups in X_1, 76 groups in X_2, 2134 groups in X_3 and these groups have different sizes. We then performed the GSJNMF model on the preprocessed datasets to identify multi-dimensional modules. In our test, we set the number of basis vectors $r = 20$ and the threshold for z-score $T = 3$ corresponding to the P-value = 0.0013 < 0.01. We searched the regularization parameters in $\{0.1 \times \frac{1}{2^n} \mid n = 1, 2, \cdots, 10\}$ and set $\lambda_1 = 1.9531 \times 10^{-4}$, and $\lambda_2 = \lambda_3 = 0.025$ by using the proposed selection procedure. The three data matrices were broken down into 20 basic building blocks, which capture the major information embedded in the original data. In other words, the variables in the three datasets can be linearly represented with the 20 basis vectors. For each basis vector, we identified a module, which consists of multiple variables from the three datasets. Within the 20 modules, four of them were significantly correlated, and these four modules were used for the subsequent analysis. Table III and IV provide the gene lists within the four modules identified from the SNP variables and DNA methylation sites, respectively. The brain ROIs identified from the fMRI voxels are displayed in Table V, where ‘*’ means null and each voxel’s volume is 3×3×3 mm^3.

In the 1-st module, there are 18 genes identified from SNPs, 6 genes from DNA methylations and 1 brain ROIs from fMRI, which can be further validated. Among them, DNMT3B may increase the risk for SZ because of the gene-gene interaction with DRD1 [37]. DCC is a promising novel candidate gene that may contribute to the genetic basis behind individual differences in susceptibility to SZ [38]. PRKG1 [39] has shown its association with SZ with the 21-st most significant SNP in the CATIE GWAS [40]. PRKG1 also interacts with RGS2 and GABRR1, which has modest association with SZ symptoms [41] and schizoaffective disorder [42]. The PLA2G4A gene has been found to be associated with negative symptoms of SZ [43]. The abnormalities of PLA2G4A may be involved in a subgroup of the illness. C10orf26 as one of the target gene of miR-137 was also reported to have genome-wide significant associations with SZ [44]. CDH13 has been implicated in the susceptibility to a variety of psychiatric diseases, which may contribute to the genetic risk of SZ [45], [46].

In the 2-nd module, there are 7 genes identified from SNP, 5 genes identified from DNA methylation and 3 brain ROIs identified from fMRI. CD28 gene polymorphisms may not only act in immune deregulation observed in SZ, but may also influence the course of the illness by modifying the susceptibility to the co-occurrence of psychotic and affective symptoms [47]. In [48], fMRI results showed reduced clusters of activation in left lingual gyrus in SZ subjects as compared to controls during empathy task, which means that the left lingual gyrus is associated with empathy in SZ patients. The bilateral reduction in fusiform gyrus [49] and progressive reduction in left superior temporal gyrus [50] gray matter volume are associated with SZ patients in first-episode.

In the 3-rd module, 15 genes from SNP, 7 genes from DNA methylation and 6 brain ROIs were identified. A real effect of variation on CACNG5 may modify the susceptibility to SZ, which means CACNG5 might contribute to the risk of SZ [51]. The maternal GRIK2 transmission disequilibrium previously reported for autism supports that GRIK2 is a susceptibility gene for SZ [52]. NRG1 and ERBB4, critical neurodevelopmental genes, are implicated in SZ [53]. FUT8 may associate with SZ because of its lower expression [54]. Within insula, abnormalities in gray matter volume, cortical thickness, cellular structure and the expression of proteins can be observed in SZ, which means insula may play an important role in the development of SZ [55]. The SZ patients' gray matter in left postcentral gyrus significantly decreases relative to the control group, indicating that the left postcentral gyrus maybe associated to the SZ [56].

In the 4-th module, we identified 7 genes from SNP, 5 genes from DNA methylation and 1 brain ROI from fMRI dataset. ATM was considered as one of the biomarker genes to discriminate SZ from controls. The combination of ATM and ADSS may confer susceptibility to the development of SZ [57]. Since the TGFBR2 mRNA levels in the peripheral leukocytes may be a potential state marker for SZ, TGFBR2 gene may be involved in the pathogenesis of SZ [58]. CTNNA2 is differentially regulated by smoking in SZ patients and it represents a promising candidate gene for SZ based on previous genetic linkage and expression study [59]. PLXNA2 is involved in axonal guidance during development and may modulate neuronal plasticity and regeneration and the PLXNA2 ligand semaphorin 3A is upregulated in the cerebellum of SZ patients, which means PLXNA2 is likely a candidate susceptibility gene for SZ [60]. We also plot the selected fMRI voxels corresponding to the four modules in Figure 1.

From the above analysis, each module contains significant biomarkers, which correspond to genes and brain ROIs related to SZ supported with existing literatures. Moreover, the genes and brain ROIs within the same module are significant correlated numerically. It indicates that these biomarkers may also have some functional associations with SZ. For real data integration, since we don't have the ground truth, the genes and ROIs not reported in the literatures may contribute to new candidate biomarkers associated with SZ. These biomarkers in the same module and their correlation with clinical outcomes need to be further verified by the biologists. We also found that there are some genes and ROIs overlapping between the modules. For example, NCOR2 and PLA2G4A correspond to both 1-st and 2-nd modules' SNP components. PRKG1 corresponds to the 1-st and 3-rd modules' SNP component. DCC is derived from the 1-st module's SNP component and 3-rd module's DNA methylation component. The left postcentral gyrus is shown in 3-rd and 4-th modules' fMRI component. The left superior temporal gyrus appears in 2-nd and 3-rd modules' fMRI component. These overlaps among the multi-dimensional modules may infer that these genes and ROIs are active and involved with multiple biological and brain functions. All these findings therefore demonstrate the implications of the selected modules related to SZ.

V. Conclusions

SNP, fMRI and DNA methylation provide important and complementary information about SZ, but most existing approaches either focus on one or two datasets analysis. If we represent the three datasets as a joint matrix with rows for the same subjects, JNMF simultaneously projects the matrices into a lower dimension subspace shared by the three datasets and the nonnegative coefficient values for each matrix can be used to select significantly correlated features among the three datasets. Since the SNP, fMRI and DNA methylation datasets have group structures (e.g., multiple SNPs spanning a gene, a group of voxels within a ROI, and multiple methylation sites within a gene), we can take advantage of the group information to improve the JNMF model. In other words, our GSJNMF model can incorporate prior knowledge by enforcing group sparse constrains into the corresponding coefficient matrices in the model. As a result, the hidden dependence structures can be identified and the data heterogeneity in the datasets can also be reflected. In addition, the new GSJNMF model will render the results to be more easily interpretable. The model is finally validated by applying to the real imaging genomic data from MCIC to identify significant genes or biomarkers associated with SZ. In the future, we will incorporate gene networks and/or brain region network information into our analysis model. Although in this work we focus on the study of SZ, the model can be applicable to the study of many other diseases, where multi-omics data are ubiquitous.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

This work is supported by 973 Program (2013CB329404), NSFC (61370147), the NIH (R01GM109068, R01MH104680, R01MH107354, P20GM103472) and NSF (1539067).

Biography

Min Wang received the B.S. degree from the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, China, in 2012, where he is currently pursuing the PhD degree with the School of Mathematical Sciences. His current research interests include image processing, biomedical imaging and bioinformatics.

Ting-Zhu Huang is currently a Professor with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, China. He received the B.S. degree, the M.S. degree, and the PhD degree from Xi’an Jiaotong University, Xi’an, China, in 1986, 1992, and 2001, respectively. His current research interests include numerical linear algebra and scientific computation with applications in electromagnetics, modeling, and algorithms for image processing. He has authored over 100 papers in international journals, including the SIAM Journal on Scientific Computing, the SIAM Journal on Matrix Analysis and Applications, the IMA Journal of Numerical Analysis, the Journal of Computational Physics, Computer Physics Communications, Numerical Linear Algebra with Applications, Automatica, the IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, Information Sciences, the Journal of the Optical Society of America A, Computing, Linear Algebra and its Applications, Applied Mathematics Letters, Computers and Mathematics with Applications, Applied Mathematical Modelling, the Journal of The Franklin Institute, the Journal of Computational and Applied Mathematics, and Communications in Nonlinear Science and Numerical Simulation. Dr. Huang received the Science and Technology Progress Award of Sichuan Province from the Chinese Information Ministry for several times. He has served on the Editorial Board of several international journals.

Jian Fang received the B.S. degree in mathematics from Nanjing Normal University, Nanjing, China, in 2008, and the PhD degree in Applied Mathematics from Xi'an Jiaotong University, China, in 2016. He is currently a Postdoctoral Fellow of Biomedical Engineering at Tulane University School of Science and Engineering. His research interests include signal processing, sparse modeling, machine learning, and their applications to the integration of multi-omics and biomedical imaging data.

Vince D. Calhoun received the bachelor degree in electrical engineering from the University of Kansas, Lawrence, Kansas, in 1991, master degree in biomedical engineering and information systems from Johns Hopkins University, Baltimore, in 1993 and 1996, respectively, and the PhD degree in electrical engineering from the University of Maryland Baltimore County, Baltimore, in 2002. He worked as a senior research engineer at the psychiatric neuroimaging laboratory at Johns Hopkins from 1993 until 2002. He then served as the director of medical image analysis at the Olin Neuropsychiatry Research Center and as an associate professor at Yale University. He is currently the chief technology officer and the director of Image Analysis and MR Research at the Mind Research Network and is a professor in the Departments of Electrical and Computer Engineering (primary), Neurosciences, Psychiatry and Computer Science at the University of New Mexico. He is the author of more than 650 full journal articles and more than 300 technical reports, abstracts, and conference proceedings. Much of his career has been spent on the development of data driven approaches for the analysis of brain imaging data. He has won more than $18 million in NSF and NIH grants on the incorporation of prior information into independent component analysis (ICA) for functional magnetic resonance imaging, data fusion of multimodal imaging and genetics data, and the identification of biomarkers for disease. He is a chartered grant reviewer for NIH. He has organized workshops and special sessions at multiple conferences. He is currently serving on the IEEE Machine Learning for Signal Processing (MLSP) technical committee and previously served as the general chair of the 2005 meeting. He is a reviewer for many journals and is on the editorial board of the Human Brain Mapping and Neuroimage journals. He is a senior member of the IEEE, the Organization for Human Brain Mapping, the International Society for Magnetic Resonance in Medicine, and the American College of Neuropsychopharmacology.

Yu-Ping Wang received the B.S. degree in applied mathematics from Tianjin University, China, in 1990, and the M.S. 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 with the University of Missouri-Kansas City. He is currently a professor of biomedical engineering and biostatistics & bioinformatics with the Tulane University School of Science and Engineering & School of Public Health and Tropical Medicine. He is also a member of the Tulane Center of Bioinformatics and Genomics, Tulane Cancer Center, and Tulane Neuroscience Program. His research interests include computer vision, signal processing, and machine learning with applications to biomedical imaging and bioinformatics, where he has over 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. He is a senior member of the IEEE.

References

[1]. Badner JA and Gershon ES, "Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia," Molecular psychiatry, vol. 7, no. 4, pp. 405–411, 2002. [PubMed: 11986984]

[2]. Wilson GM, Flibotte S, et al., "Dna copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling," Human molecular genetics, vol. 15, no. 5, pp. 743–749, 2006. [PubMed: 16434481]

[3]. Sutrala SR, Goossens D, et al., "Gene copy number variation in schizophrenia," Schizophrenia research, vol. 96, no. 1, pp. 93–99, 2007. [PubMed: 17826036]

[4]. Callicott JH, Straub RE, et al., "Variation in disc1 affects hippocampal structure and function and increases risk for schizophrenia," Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 24, pp. 8627–8632, 2005. [PubMed: 15939883]

[5]. Chen Y, Zhang J, et al., “Effects of maoa promoter methylation on susceptibility to paranoid schizophrenia,” Human genetics, vol. 131, no. 7, pp. 1081–1087, 2012. [PubMed: 22198720]

[6]. Liu J, Chen J, et al., “Methylation patterns in whole blood correlate with symptoms in schizophrenia patients,” Schizophrenia bulletin, vol. 40, no. 4, pp. 769–776, 2013. [PubMed: 23734059]

[7]. Szycik GR, Münte TF, et al., “Audiovisual integration of speech is disturbed in schizophrenia: an fmri study,” Schizophrenia research, vol. 110, no. 1, pp. 111–118, 2009. [PubMed: 19303257]

[8]. Damaraju E, Allen EA, et al., “Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia,” NeuroImage: Clinical, vol. 5, pp. 298–308, 2014. [PubMed: 25161896]

[9]. Gur RE, McGrath C, et al., “An fmri study of facial emotion processing in patients with schizophrenia,” American Journal of Psychiatry, vol. 159, no. 12, pp. 1992–1999, 2002.

[10]. Li T, Ball D, et al., “Family-based linkage disequilibrium mapping using snp marker haplotypes: application to a potential locus for schizophrenia at chromosome 22q11,” Molecular psychiatry, vol. 5, no. 1, pp. 77–84, 2000. [PubMed: 10673772]

[11]. Liu J, Chen J, et al., “Methylation patterns in whole blood correlate with symptoms in schizophrenia patients,” Schizophrenia bulletin, vol. 40, no. 4, pp. 769–776, 2014. [PubMed: 23734059]

[12]. Lin D, Calhoun VD, and Wang Y-P, “Correspondence between fmri and snp data by group sparse canonical correlation analysis,” Medical image analysis, vol. 18, no. 6, pp. 891–902, 2014. [PubMed: 24247004]

[13]. Sui J, He H, et al., “Three-way (n-way) fusion of brain imaging data based on mcca+ jica and its application to discriminating schizophrenia,” Neuroimage, vol. 66, pp. 119–132, 2013. [PubMed: 23108278]

[14]. Vergara VM, Ulloa A, et al., “A three-way parallel ica approach to analyze links among genetics, brain structure and brain function,” Neuroimage, vol. 98, pp. 386–394, 2014. [PubMed: 24795156]

[15]. Hotelling H, “Relations between two sets of variates,” Biometrika, vol. 28, no. 3/4, pp. 321–377, 1936.

[16]. Wold H, “Partial least squares,” Encyclopedia of statistical sciences, 1985.

[17]. Parkhomenko E, Tritchler D, and Beyene J, “Sparse canonical correlation analysis with application to genomic data integration,” Statistical Applications in Genetics and Molecular Biology, vol. 8, no. 1, pp. 1–34, 2009.

[18]. Chun H and Keles S¸, “Sparse partial least squares regression for simultaneous dimension reduction and variable selection,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 72, no. 1, pp. 3–25, 2010. [PubMed: 20107611]

[19]. Vounou M, Nichols TE, et al., “Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach,” Neuroimage, vol. 53, no. 3, pp. 1147–1159, 2010. [PubMed: 20624472]

[20]. Chen J and Zhang S, “Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data,” Bioinformatics, vol. 32, no. 11, pp. 1724–1732, 2016. [PubMed: 26833341]

[21]. Witten DM and Tibshirani RJ, “Extensions of sparse canonical correlation analysis with applications to genomic data,” Statistical applications in genetics and molecular biology, vol. 8, no. 1, pp. 1–27, 2007.

[22]. Hu W, Lin D, et al., “Adaptive sparse multiple canonical correlation analysis with application to imaging (epi)genomics study of schizophrenia,” IEEE Transactions on Biomedical Engineering, vol. 65, no. 2, pp. 390–399, 2018. [PubMed: 29364120]

[23]. Lee DD and Seung HS, “Learning the parts of objects by non-negative matrix factorization,” Nature, vol. 401, no. 6755, pp. 788–791, 1999. [PubMed: 10548103]

[24]. Hoyer PO, “Non-negative sparse coding,” in Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on IEEE, 2002, pp. 557–565.

[25]. Hoyer PO, “Non-negative matrix factorization with sparseness constraints,” Journal of machine learning research, vol. 5, no. Nov, pp. 1457–1469, 2004.

[26]. Liu X, Lu H, and Gu H, “Group sparse non-negative matrix factorization for multi-manifold learning.,” in BMVC, 2011, pp. 1–11.

[27]. Cai D, He X, et al., “Graph regularized nonnegative matrix factorization for data representation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 8, pp. 1548–1560, 2011. [PubMed: 21173440]

[28]. Zhang S, Liu C-C, et al., “Discovery of multi-dimensional modules by integrative analysis of cancer genomic data,” Nucleic acids research, vol. 40, no. 19, pp. 9379–9391, 2012. [PubMed: 22879375]

[29]. Wang M, Huang T-Z, et al., “Integration of multiple genomic imaging data for the study of schizophrenia using joint nonnegative matrix factorization,” in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on IEEE, 2017, pp. 1083–1087.

[30]. Wang M, Huang T-Z, and Deng L-J, “A group sparse joint nonnegative matrix factorization model for multiple data integration,” in Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2017 International Computer Conference on IEEE, 2017, pp. 20–23.

[31]. Lee DD and Seung HS, “Algorithms for non-negative matrix factorization,” in Advances in neural information processing systems, 2001, pp. 556–562.

[32]. Xu W, Liu X, and Gong Y, “Document clustering based on non-negative matrix factorization,” in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval ACM, 2003, pp. 267–273.

[33]. Brunet JP, Tamayo P, et al., “Metagenes and molecular pattern discovery using matrix factorization,” Proceedings of the national academy of sciences, vol. 101, no. 12, pp. 4164–4169, 2004.

[34]. Sun W, Wang J, and Fang Y, “Consistent selection of tuning parameters via variable selection stability,” Journal of Machine Learning Research, vol. 14, no. 1, pp. 3419–3440, 2013.

[35]. Cohen J, “A coefficient of agreement for nominal scales,” Educational and psychological measurement, vol. 20, no. 1, pp. 37–46, 1960.

[36]. Kanehisa M and Goto S, “Kegg: kyoto encyclopedia of genes and genomes,” Nucleic acids research, vol. 28, no. 1, pp. 27–30, 2000. [PubMed: 10592173]

[37]. Zhang C, Xie B, et al., “Gene-gene interaction between dnmt3b and drd1 in schizophrenia,” Zhonghua yi xue za zhi, vol. 90, no. 43, pp. 3059–3062, 2010. [PubMed: 21211326]

[38]. Grant A, Fathalli F, et al., “Association between schizophrenia and genetic variation in dcc: a case–control study,” Schizophrenia research, vol. 137, no. 1, pp. 26–31, 2012. [PubMed: 22418395]

[39]. Zhao Z, Webb BT, et al., “Association study of 167 candidate genes for schizophrenia selected by a multi-domain evidence-based prioritization algorithm and neurodevelopmental hypothesis,” PloS one, vol. 8, no. 7, pp. e67776, 2013. [PubMed: 23922650]

[40]. Sullivan PF, Lin D, et al., “Genomewide association for schizophrenia in the catie study: results of stage 1,” Molecular psychiatry, vol. 13, no. 6, pp. 570–584, 2008. [PubMed: 18347602]

[41]. Campbell DB, Lange LA, et al., “Association of rgs2 and rgs5 variants with schizophrenia symptom severity,” Schizophrenia research, vol. 101, no. 1, pp. 67–75, 2008. [PubMed: 18262772]

[42]. Green EK, Grozeva D, et al., “Variation at the gabaa receptor gene, rho 1 (gabrr1) associated with susceptibility to bipolar schizoaffective disorder,” American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, vol. 153, no. 7, pp. 1347–1349, 2010.

[43]. Tao R, Wei J, et al., “The pla2g4a gene and negative symptoms in a chinese population,” Schizophrenia research, vol. 86, no. 1, pp. 326–328, 2006. [PubMed: 16843642]

[44]. Kwon E, Wang W, and Tsai L-H, “Validation of schizophrenia-associated genes csmd1, c10orf26, cacna1c and tcf4 as mir-137 targets,” Molecular psychiatry, vol. 18, no. 1, pp. 11–13, 2013. [PubMed: 22182936]

[45]. Otsuka I, Watanabe Y, et al., “association analysis of the cadherin13 gene with schizophrenia in the japanese population,” Neuropsychiatric disease and treatment, vol. 11, pp. 1381–1393, 2015. [PubMed: 26082635]

[46]. Wright C, Calhoun VD, et al., “Meta gene set enrichment analyses link mir-137-regulated pathways with schizophrenia risk,” Frontiers in genetics, vol. 6, 2015.

[47]. Frydecka D, Beszłej JA, et al., "Ctla4 and cd28 gene polymorphisms with respect to affective symptom domain in schizophrenia," Neuropsychobiology, vol. 71, no. 3, pp. 158–167, 2015. [PubMed: 25998553]

[48]. Singh S, Modi S, et al., "Functional and structural abnormalities associated with empathy in patients with schizophrenia: An fmri and vbm study," Journal of biosciences, vol. 40, no. 2, pp. 355–364, 2015. [PubMed: 25963262]

[49]. Lee CU, Shenton ME, et al., "Fusiform gyrus volume reduction in first-episode schizophrenia: a magnetic resonance imaging study," Archives of General Psychiatry, vol. 59, no. 9, pp. 775–781, 2002. [PubMed: 12215076]

[50]. Kasai K, Shenton ME, et al., "Progressive decrease of left superior temporal gyrus gray matter volume in patients with first-episode schizophrenia," American Journal of Psychiatry, vol. 160, no. 1, pp. 156–164, 2003.

[51]. Curtis D, Vine AE, et al., "Case-case genome wide association analysis reveals markers differentially associated with schizophrenia and bipolar disorder and implicates calcium channel genes," Psychiatric genetics, vol. 21, no. 1, pp. 1–4, 2011. [PubMed: 21057379]

[52]. Bah J, Quach H, et al., "Maternal transmission disequilibrium of the glutamate receptor grik2 in schizophrenia," Neuroreport, vol. 15, no. 12, pp. 1987–1991, 2004. [PubMed: 15305151]

[53]. Law AJ, Wang Y, et al., "Neuregulin 1-erbb4-pi3k signaling in schizophrenia and phosphoinositide 3-kinase-p110δ inhibition as a potential therapeutic strategy," Proceedings of the National Academy of Sciences, vol. 109, no. 30, pp. 12165–12170, 2012.

[54]. Mueller TM, Yates SD, et al., "Altered fucosyltransferase expression in the superior temporal gyrus of elderly patients with schizophrenia," Schizophrenia research, vol. 182, pp. 66–73, 2017. [PubMed: 27773385]

[55]. Wylie KP and Tregellas JR, "The role of the insula in schizophrenia," Schizophrenia research, vol. 123, no. 2, pp. 93–104, 2010. [PubMed: 20832997]

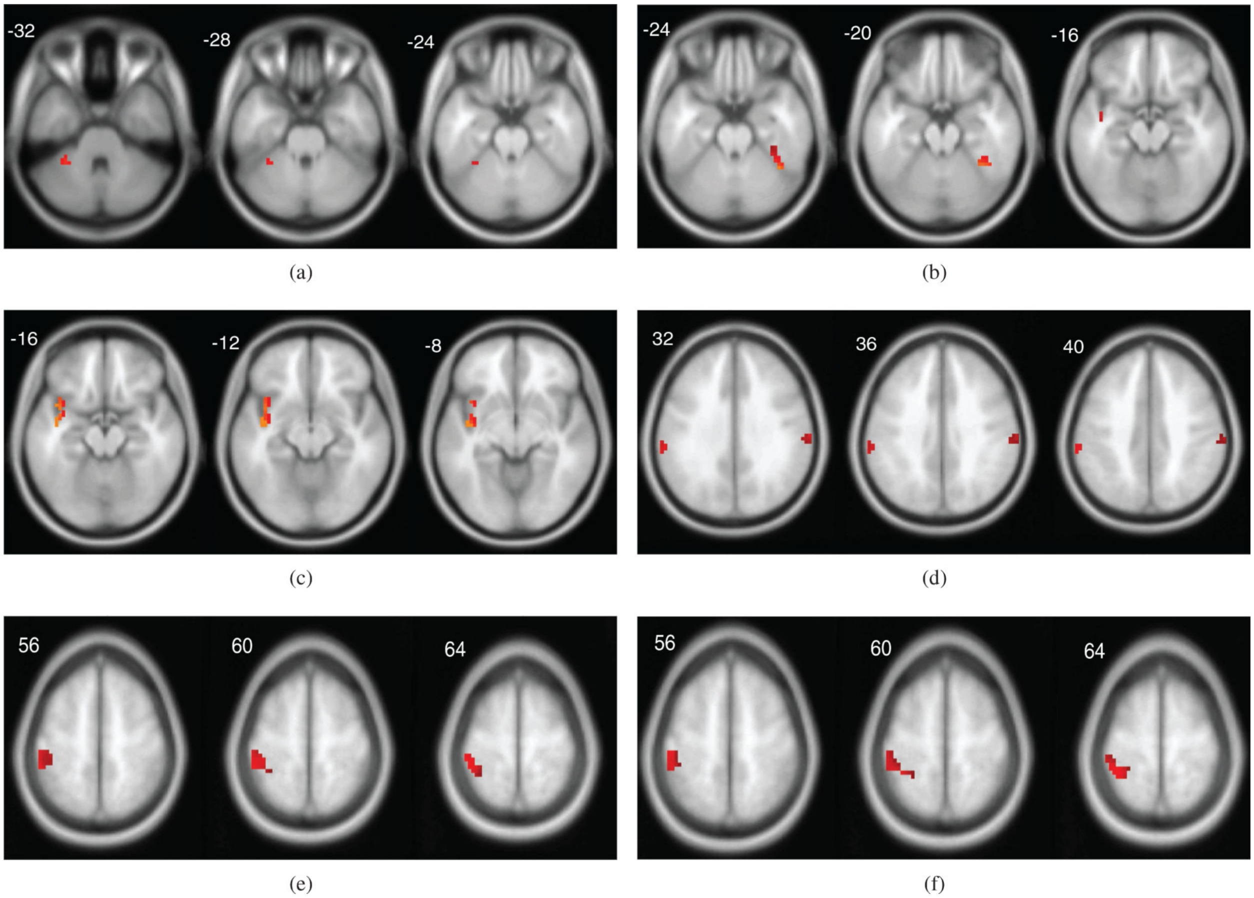
[56]. Job DE, Whalley HC, et al., "Structural gray matter differences between first-episode schizophrenics and normal controls using voxel-based morphometry," Neuroimage, vol. 17, no. 2, pp. 880–889, 2002. [PubMed: 12377162]

[57]. Zhang F, Xu Y, et al., "Association analyses of the interaction between the adss and atm genes with schizophrenia in a chinese population," BMC medical genetics, vol. 9, no. 1, pp. 119, 2008. [PubMed: 19115993]

[58]. Numata S, Ueno S, et al., "Tgfbr2 gene expression and genetic association with schizophrenia," Journal of psychiatric research, vol. 42, no. 6, pp. 425–432, 2008. [PubMed: 17560608]

[59]. Mexal S, Berger R, et al., "Regulation of a novel αn-catenin splice variant in schizophrenic smokers," American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, vol. 147, no. 6, pp. 759–768, 2008.

[60]. Mah S, Nelson MR, et al., "Identification of the semaphorin receptor plxna2 as a candidate for susceptibility to schizophrenia," Molecular psychiatry, vol. 11, no. 5, pp. 471–478, 2006. [PubMed: 16402134]



(a) (b) (c) (d) (e) (f)

Fig. 1.

(a) Brain ROIs in module 1 (b) Brain ROIs in module 2. (c,d,e) Brain ROIs in module 3. (f) Brain ROIs in module 4. The color indicates the z-score value of the selected voxels.

TABLE I.

The experiment setting of different cases

Case index	**1**	**2**	**3**	**4**
# groups in three data matrices	√	√	×	×
# variables in each group	√	×	√	×

TABLE II.

The variable indices of modules in each case

Case index	Module index	Variable index		
		X_1	X_2	X_3
1	1	1 ~ 20	41 ~ 60	81 ~ 100
	2	41 ~ 60	1 ~ 20	61 ~ 80
2	1	1 ~ 15	31 ~ 41	50 ~ 55
	2	21 ~ 29	1 ~ 11	7 ~ 14
3	1	1 ~ 20	41 ~ 60	81 ~ 100
	2	21 ~ 40	1 ~ 20	121 ~ 140
4	1	1 ~ 10	1 ~ 8	47 ~ 59
	2	24 ~ 31	20 ~ 32	18 ~ 29

TABLE III.

The list of genes selected from the significantly identified modules

Module	Gene ID (SNP)
1	VTI1B, DNMT3B, NFX1, NUMB, SEC61G, ASS1, ACTG2, ADCY2, FOXO3, ABCA12, PAK3, COL4A2, DIAPH2, GOT2, NCOR2, DCC, PRKG1, PLA2G4A
2	NLN, AVPR1A, NCOR2, IL7, MAN1A1, PLA2G4A, CD28
3	CACNG5, SGPP2, MKNK1, MASP1, PRKCQ, FUCA2, NDST4, GRIK2, ERBB4, FUT8, CBLB, RYR2, PRKG1, NEGR1, NRG1
4	ATM, ACP6, TGFBR2, GNA14, PPP1R12A, CTNNA2, PLXNA2

TABLE IV.

The list of genes selected from significantly identified modules

Module	Gene ID (DNA methylation)
1	FLJ22746, FLJ11155, CDKN2A, C10orf26, C10orf10, CDH13,
2	BRDT, EEF1E1, IKIP, MYO9A, LOC400696
3	C1orf26, FLJ11017, MICA, GMPR, COMMD5, WDR68, DCC,
4	FCGR3B, FLJ20245, NTNG2, TUB, DGAT2L6

TABLE V.

Brain regions detected from significantly identified modules

Brain region \ Module	L/R volumn(cm^3)			
	1	2	3	4
Insula	*	*	1.296/*	*
Lingual Gyrus	*	0.432/*	*	*
Fusiform Gyrus	*	*/0.678	*	*
Postcentral Gyrus	*	*	1.944/*	2.403/*
Supramarginal Gyrus	*	*	0.648/1.107	*
Superior Temporal Gyrus	*	0.486/*	0.486/*	*
Superior Temporal Pole	*	*	0.405/*	*
Lobule VI of Cerebellar Hemisphere	0.486/*	*	*	*