Multi-view singular value decomposition for disease subtyping and genetic associations

Jiangwen Sun[1]
Email: javon@engr.uconn.edu

Jinbo Bi[1*]
*Corresponding author
Email: jinbo@engr.uconn.edu

Henry R Kranzler[2]
Email: kranzler@mail.med.upenn.edu

[1]Department of Computer Science and Engineering, University of Connecticut, 371 Fairfield Way, Storrs, CT 06269, USA

[2]Treatment Research Center, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 24105, USA

Abstract

Background

Accurate classification of patients with a complex disease into subtypes has important implications for medicine and healthcare. Using more homogeneous disease subtypes in genetic association analysis will facilitate the detection of new genetic variants that are not detectible using the non-differentiated disease phenotype. Subtype differentiation can also improve diagnostic classification, which can in turn inform clinical decision making and treatment matching. Currently, the most sophisticated methods for disease subtyping perform cluster analysis using patients' clinical features. Without guidance from genetic information, the resultant subtypes are likely to be suboptimal and efforts at genetic association may fail.

Results

We propose a multi-view matrix decomposition approach that integrates clinical features with genetic markers to detect confirmatory evidence for a disease subtype. This approach groups patients into clusters that are consistent between the clinical and genetic dimensions of data; it simultaneously identifies the clinical features that define the subtype and the genotypes associated with the subtype. A simulation study validated the proposed approach, showing that it identified hypothesized subtypes and associated features. In comparison to the latest biclustering and multi-view data analytics using real-life disease data, the proposed approach identified *clinical subtypes* of a disease that differed from each other more significantly in the genetic markers, thus demonstrating the superior performance of the proposed approach.

Conclusions

The proposed algorithm is an effective and superior alternative to the disease subtyping methods employed to date. Integration of phenotypic features with genetic markers in the subtyping analysis is a promising approach to identify concurrently disease subtypes and their genetic associations.

Keywords

Genotype-phenotype association, Multi-view data analysis, Subtyping, Biclustering, Matrix decomposition

Background

For complex diseases, such as substance dependence or psychiatric disorders, a variety of clinical features that collectively indicate or characterize the disease phenotype often vary substantially among individuals [1]. Studies of genetic association or those that aim to match patients with certain treatments for a complex disease can be impeded by this phenotypic heterogeneity [2]. Case-control association studies based on a binary trait, such as the diagnosis of a disease, which partitions the population into cases (subjects with the disease) and non-cases (subjects without the disease), cannot differentiate the heterogeneous manifestations of the disease. Although many candidate genes or genomic regions have been associated with complex diseases [3], the characteristics or subtypes of the disease for which the association exists remain to be specified. For instance, the specific addictive behaviors that underlie the associations with candidate genetic variants need to be elucidated to clarify the risk for addiction. [4].

Classification of a complex disease into homogeneous subcategories or subtypes may help to identify the genetic variants contributing to the effect of the subphenotypes [5,6]. However, prior studies have been limited to unsupervised cluster analysis or latent class analysis on clinical features to derive subtypes. Genotypic data have only been used to evaluate the validity of subtypes, such as in subsequent association tests with the derived subtypes, rather than to guide the creation of the subtypes. Consequently, the resultant subtypes may be of limited utility in genetic association analysis. Integration of data from both clinical and genomic dimensions also offers opportunities to find confirmatory evidence of a subtype based on both its genetic and clinical features. A few studies have examined the joint use of gene expression and genotypic data for cancer subtyping [7,8], but they did not identify a variable subspace (or a subset of features) in each data source so as to group subjects consistently across the two subspaces. Hence, they could not detect genetic variants associated with the identified clusters.

There has also been little research on this topic in the statistics literature. The most relevant area involves co-clustering [9] or multi-view data analysis [10], where samples are characterized or viewed in multiple ways, thus creating multiple sets of input variables. There are two types of co-clustering methods: (1) biclustering, also called two-mode clustering [11,12], which simultaneously clusters the rows and columns of a data matrix and (2) multi-view co-clustering [9,13], which seeks groupings that are consistent across different views. Biclustering is similar to another set of algorithms that search for subspaces and group subjects differently in each subspace [14].

Biclustering and subspace searching essentially identify different subgroups of subjects using different features (or markers), thus helping to identify genetic variants specific to a particular subgroup. However, this method can only be applied to one data matrix from a single view rather than data jointly from multiple views. Multi-view co-clustering, on the other hand, seeks a grouping of subjects that is consistent across different views (i.e., different sets of features), but the resultant clusters are defined using all of the available features, e.g., all of the studied genetic markers. Hence, it cannot be used to identify subtype-specific variants/features. Thus, to address our subtyping problem, we not only partitioned subjects in such a way that the subgroups differed in both clinical features and genetic markers, but also included a subspace search to identify the specific features or markers that defined the subgroups.

In this paper, we propose a multi-view matrix decomposition approach based on the sparse singular value decomposition (SSVD) technique [12] to classify a complex disease into subtypes using data both from the clinical and genetic views. The objective of this problem is to identify subject clusters that agree in the clinical and genetic views, and simultaneously identify features and markers that are associated with the clusters. Employing the *sparse* SVD in our approach is critical to its success, especially in terms of successfully detecting associated variants given that the number of truely associated variants are much fewer than the number of single nucleotide polymorphisms (SNPs) in the whole genome. The proposed approach was validated on synthetic datasets that were simulated to have subtype structures and several genetic markers associated with the subtypes and a real world clinical dataset that was aggregated from multiple genetic studies of substance dependence. We compared our approach to a biclustering approach [12] and the latest multi-view data analytics methods [9]. The results clearly show that the performance of our approach is superior to that of all other available methods.

Methods

We start with a presentation of the notations that are used throughout the paper. A vector is denoted by a bold lower case letter as in $\mathbf{v}$ and $||\mathbf{v}||_p$ represents its ℓ_p-norm, which is defined by $||\mathbf{v}||_p = (|\mathbf{v}_{(1)}|^p + \cdots + |\mathbf{v}_{(d)}|^p)^{1/p}$, where $\mathbf{v}_{(j)}$ is the j-th component of $\mathbf{v}$ and d is the length of $\mathbf{v}$, i.e., the total number of components in $\mathbf{v}$. We use $||\mathbf{v}||_0$ to represent the so-called *0-norm* of $\mathbf{v}$ that equals the number of non-zero components in $\mathbf{v}$. Denote $\mathbf{u} \odot \mathbf{v}$ the component-wise (Hadamard) products of $\mathbf{u}$ and $\mathbf{v}$. The set $\mathcal{B}_d$ contains all binary vectors of length d. A binary vector is a vector whose components equal either 0 or 1. A matrix is denoted by a bold upper case letter, e.g., $\mathbf{M}_{n\times d}$ is a n-by-d matrix, and $||\mathbf{M}||_F$ is its Frobenius norm defined by $(tr(\mathbf{M}^T\mathbf{M}))^{1/2}$ where $tr(\cdot)$ is the trace of a matrix. Rows and columns in $\mathbf{M}$ are denoted by $\mathbf{M}_{(i,\cdot)}$ and $\mathbf{M}_{(\cdot,j)}$, respectively.

Review of single-view biclustering

We briefly review the biclustering method with a single view of data based on the sparse singular value decomposition [12]. For a single data matrix $\mathbf{M}$ of size n-by-d, a subgroup of its rows and a subgroup of its columns can be simultaneously obtained by the SSVD. The SSVD requires both the left and right singular vectors to be sparse. Let $\mathbf{u}$ of size n and $\mathbf{v}$ of size d be a pair of singular vectors resulting from the SSVD. Their outer product forms a sparse low-rank approximation of the original matrix, i.e., $\mathbf{M} = \sigma\mathbf{u}\mathbf{v}^T$ where σ is the corresponding singular value. Then, the rows in $\mathbf{M}$ that correspond to non-zero components in $\mathbf{u}$ form a row subgroup. The columns in $\mathbf{M}$ that correspond to non-zero components in $\mathbf{v}$ form a column subgroup. The resultant row and column clusters help to define one another. The SSVD finds all singular vectors sequentially by repeatedly solving the following problem with a data matrix $\mathbf{M}$:

$$\begin{aligned} \min_{\sigma,\mathbf{u},\mathbf{v}} \quad & ||\mathbf{M} - \sigma\mathbf{u}\mathbf{v}^T||_F^2 + \lambda_u||\sigma\mathbf{u}||_0 + \lambda_v||\sigma\mathbf{v}||_0 \\ \text{subject to} \quad & ||\mathbf{u}||_2 = 1, ||\mathbf{v}||_2 = 1. \end{aligned} \tag{1}$$

The regularization terms $||\sigma\mathbf{u}||_0$ and $||\sigma\mathbf{v}||_0$ are used to enforce the sparsity of $\mathbf{u}$ and $\mathbf{v}$. Note that the scalar σ will not affect the value of the regularization terms. The parameters λ_u and λ_v are two hyper-parameters to balance the approximation performance and the regularization terms. If both λ_u and λ_v equal 0, the optimal solution to this problem is the left and right singular vectors of $\mathbf{M}$ that correspond to its largest singular value. An alternating algorithm has been proposed in [12] to solve this problem effectively when λ_u and λ_v are not 0. This algorithm first initiates $\mathbf{u}$ and $\mathbf{v}$ by the first left and right singular vectors of $\mathbf{M}$, then alternates between solving two sub-problems until it converges. The two sub-problems are: (a), fix $\mathbf{u}$ and find $\mathbf{v}$ that optimizes the objective of Eq.(1); (b), fix $\mathbf{v}$ and find $\mathbf{u}$ that optimizes the objective of Eq.(1).

Assume that each row of $\mathbf{M}$ represents a subject and each column corresponds to a feature. Once a pair of vectors $\mathbf{u}$ and $\mathbf{v}$ is obtained, a subject (row) cluster as indicated by the non-zero components of $\mathbf{u}$ is obtained. At the same time, the features on which the subjects in the cluster show high similarity are also identified in a column cluster as indicated by the non-zero components of $\mathbf{v}$. More clusters can be obtained by repeating the optimization process with modified data matrices. To obtain subsequent clusters that are disjoint from any identified cluster in terms of subjects, the SSVD solves Eq.(1) using a new matrix $\mathbf{M}$ that excludes subjects (rows) already included in a row cluster. To obtain subsequent clusters that allow overlapping of subjects with identified clusters, the SSVD can solve Eq.(1) with the deflated $\mathbf{M} = \mathbf{M} - \sigma \mathbf{u}\mathbf{v}^T$ that removes the identified SVD components as used in the standard SVD.

The proposed formula for two-view joint biclustering

In this section, we extend the single-view SSVD to find a consistent grouping of subjects across two data matrices. In a later section, the resulting method will be extended to incorporate more than two data matrices.

Assume that two data matrices denoted by $\mathbf{M}_1$ of size n-by-d_1 and $\mathbf{M}_2$ of size n-by-d_2 characterize the same set of n subjects from two different views. We can obtain $\mathbf{u}_1$, $\mathbf{v}_1$, and $\mathbf{u}_2$, $\mathbf{v}_2$ by a separate SSVD of $\mathbf{M}_1$ and $\mathbf{M}_2$, respectively. However, it will not guarantee that the row clusters specified by $\mathbf{u}_1$ and $\mathbf{u}_2$ agree. To make them consistent, $\mathbf{u}_1$ and $\mathbf{u}_2$ must have non-zero components at the same position. Note that the two $\mathbf{u}$ vectors are not necessarily the same, because they may be derived from very different features in the views, such as real-valued clinical features but discrete values in genetic markers.

We propose to use a binary vector $\mathbf{z}$ of size n that serves as a common factor to link the two views. Each component of $\mathbf{u}$ is then multiplied by the corresponding component of $\mathbf{z}$, i.e., $u_i = u_i z_i$. In other words, we represent each $\mathbf{u}$ vector by $\mathbf{z} \odot \mathbf{u}$ in the objective function of SSVD to construct the sparse, rank one approximation matrices of $\mathbf{M}_1$ and $\mathbf{M}_2$, simultaneously. When $\mathbf{z}$ is sparse, both $\mathbf{z} \odot \mathbf{u}_1$ and $\mathbf{z} \odot \mathbf{u}_2$ will be sparse. Thus, we enforce the sparsity of $\mathbf{z}$ rather than individual $\mathbf{u}$ and solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{z},\sigma_i,\mathbf{u}_i,\mathbf{v}_i,i=1,2} \quad & \|\mathbf{M}_1 - \sigma_1(\mathbf{z} \odot \mathbf{u}_1)\mathbf{v}_1^T\|_F^2 + \|\mathbf{M}_2 - \sigma_2(\mathbf{z} \odot \mathbf{u}_2)\mathbf{v}_2^T\|_F^2 \\ & + \lambda_z \|\mathbf{z}\|_0 + \lambda_{v_1}\|\sigma_1 \mathbf{v}_1\|_0 + \lambda_{v_2}\|\sigma_2 \mathbf{v}_2\|_0, \\ \text{subject to} \quad & \|\mathbf{u}_i\|_2 = 1, \|\mathbf{v}_i\|_2 = 1, \quad i = 1,2, \\ & \mathbf{z} \in \mathcal{B}_n. \end{aligned} \tag{2}$$

where λ_z, λ_{v_1} and λ_{v_2} are tuning parameters that balance the approximation errors and regularization terms. Although the values of $\mathbf{u}$'s are constrained to be unit vectors, the values of $\mathbf{z} \odot \mathbf{u}$'s are not necessarily unit vectors. However, a careful examination reveals that for any optimal solution $\hat{\mathbf{u}}$, we can find another optimal solution $\bar{\mathbf{u}}$ that has non-zero values only at the entries indicated by the binary vector $\mathbf{z}$, which ensures that $\mathbf{z} \odot \bar{\mathbf{u}}$ is also a unit vector. We first set $\bar{\mathbf{u}}_{(j)} = \hat{\mathbf{u}}_{(j)}$, if $\mathbf{z}_{(j)} \neq 0$, or $\bar{\mathbf{u}}_{(j)} = 0$ otherwise, for $j = 1 \cdots, n$. We then update the corresponding singular value $\sigma = \sigma \|\bar{\mathbf{u}}\|_2$ and rescale $\bar{\mathbf{u}} = \bar{\mathbf{u}} / \|\bar{\mathbf{u}}\|_2$. This new vector $\bar{\mathbf{u}}$ satisfies the constraints of Eq.(2), and together with the new σ will produce the same objective value as the original solution $\hat{\mathbf{u}}$, thus corresponding to an optimal solution as well. We design a fast algorithm in a later section to find such a sparse $\bar{\mathbf{u}}$ for Eq.(2).

We discuss two alternatives to the proposed formula (2). A restricted version of Eq.(2) may require $\mathbf{u}_1 = \mathbf{u}_2 = \mathbf{u}$ and then replace $\mathbf{z} \odot \mathbf{u}_1$ and $\mathbf{z} \odot \mathbf{u}_2$ by the same $\mathbf{u}$ in the objective function of Eq.(2), which leads to the following problem

$$\begin{aligned} \min_{\sigma_i,\mathbf{u},\mathbf{v}_i,i=1,2} \quad & \|\mathbf{M}_1 - \sigma_1\mathbf{u}\mathbf{v}_1^T\|_F^2 + \|\mathbf{M}_2 - \sigma_2\mathbf{u}\mathbf{v}_2^T\|_F^2 \\ & + \lambda_u\|\mathbf{u}\|_0 + \lambda_{v1}\|\sigma_1\mathbf{v}_1\|_0 + \lambda_{v2}\|\sigma_2\mathbf{v}_2\|_0, \\ \text{subject to} \quad & \|\mathbf{u}\|_2 = 1, \|\mathbf{v}_i\|_2 = 1, \quad i = 1, 2. \end{aligned} \tag{3}$$

By requiring $\mathbf{u}$ to be sparse, it can also identify consistent row clusters between two views. The resultant optimization problem is easier to solve without integer variables in $\mathbf{z}$. However, it is an unnecessarily stringent constraint to limit the search space to $\mathbf{u}_1 = \mathbf{u}_2$, which rules out a number of potential solutions that may include the optimal row clusters. Another alternative is to minimize the difference between $\mathbf{u}_1$ and $\mathbf{u}_2$, which suffers from the same over-constrained problem because the exact values of the difference are not involved. Our problem only seeks to identify the indicators of whether or not a component of $\mathbf{u}$ is zero.

It is also useful to discuss the relation between Eq.(3) and the feature concatenation method, which simply merges the features from the two views in a cluster analysis. The feature concatenation method finds a single set of $\mathbf{u}$ and $\mathbf{v}$ for the data matrix $[\mathbf{M}_1\ \mathbf{M}_2]$ by solving the following problem

$$\begin{aligned} \min_{\sigma,\mathbf{u},\mathbf{v}} \quad & \|[\mathbf{M}_1\ \mathbf{M}_2] - \sigma\mathbf{u}\mathbf{v}^T\|_F^2 + \lambda_u\|\sigma\mathbf{u}\|_0 + \lambda_v\|\sigma\mathbf{v}\|_0 \\ \text{subject to} \quad & \|\mathbf{u}\|_2 = 1, \|\mathbf{v}\|_2 = 1. \end{aligned} \tag{4}$$

where the $\mathbf{v}$ vector is of size $d_1 + d_2$. In comparison with Eq.(3), Eq.(4) uses a single σ for the two views, and the concatenated $\mathbf{v}$ is constrained to be a unit vector rather than individual $\mathbf{v}_1$ and $\mathbf{v}_2$. It is easy to show that any optimal solution to Problem (3) can become a feasible solution to Problem (4) by properly rescaling $\mathbf{v}_1$ and $\mathbf{v}_2$ and absorbing the scaling factors by σ_1 and σ_2 to make $\sigma_1 = \sigma_2$, but is not necessarily an optimal solution to Problem (4). An optimal $\mathbf{v}$ to Problem (4) may have either $\mathbf{v}_1$ or $\mathbf{v}_2$ be zero, which is however not allowed in Eq.(3). When one of the $\mathbf{v}$ vectors is zero, the resultant clusters differ only on one view of the features. As an example, we concatenated 64 clinical features to 1248 SNPs in a disease subtyping analysis. Because the genetic markers outweighed the clinical features, the resultant clusters differed significantly only on the SNPs, leading to disease subtypes that could not be clinically recognized.

A fast algorithm for two-view joint biclustering

The proposed formulation (2), although is a mixed-integer program, can be effectively solved after proper relaxations. We design an alternating optimization algorithm to solve it by splitting the variables into three working sets: one set consists of the $\mathbf{u}$ vectors; one set consists of the $\mathbf{v}$ vectors; and the last set consists of the binary variables in $\mathbf{z}$. We optimize the variables in one working set at a time in alternative steps.

(1) Find the optimal $\mathbf{u}_1$, $\mathbf{v}_1$, $\mathbf{u}_2$, and $\mathbf{v}_2$ with fixed $\mathbf{z}$

When $\mathbf{z}$ is fixed, Problem (2) can be decomposed into two sub-problems that optimize with respect to each individual view. Without loss of generality, we show how to optimize $\mathbf{u}_1$ and $\mathbf{v}_1$ by solving the following sub-problem with a fixed $\mathbf{z}$.

$$\begin{aligned} \min_{\sigma_1,\mathbf{u}_1,\mathbf{v}_1} \quad & \|\mathbf{M}_1 - \sigma_1(\mathbf{z} \odot \mathbf{u}_1)\mathbf{v}_1^T\|_F^2 + \lambda_{v1}\|\sigma_1\mathbf{v}_1\|_0 \\ \text{subject to} \quad & \|\mathbf{u}_1\|_2 = 1, \|\mathbf{v}_1\|_2 = 1, \end{aligned} \tag{5}$$

which can be solved by alternating between optimizing for $\mathbf{u}$ and for $\mathbf{v}$.

(a) Solve for $\mathbf{v}_1$ when $\mathbf{u}_1$ is fixed

We solve the following equivalent problem for the optimal $\tilde{\mathbf{v}}_1$ by relaxing the unit length constraint on $\mathbf{v}_1$, and then setting $\sigma_1 = \|\tilde{\mathbf{v}}_1\|_2$ and $\mathbf{v}_1 = \tilde{\mathbf{v}}_1/\sigma_1$.

$$\min_{\tilde{\mathbf{v}}_1} \quad \|\mathbf{M}_1 - (\mathbf{z} \odot \mathbf{u}_1)\tilde{\mathbf{v}}_1^T\|_F^2 + \lambda_{v_1}\|\tilde{\mathbf{v}}_1\|_0. \tag{6}$$

Similar to the single-view SSVD, we relax the *0-norm* to have the ℓ_1 vector norm, and solve for $\mathbf{v}$ by minimizing $\|\mathbf{M}_1 - (\mathbf{z} \odot \mathbf{u}_1)\tilde{\mathbf{v}}_1^T\|_F^2 + \lambda_{v_1}\|\tilde{\mathbf{v}}_1\|_1$. Each component $\tilde{\mathbf{v}}_{1(j)}$ in $\tilde{\mathbf{v}}_1$ can be computed independently from the others by solving

$$\min_{\tilde{\mathbf{v}}_{1(j)}} \quad \tilde{\mathbf{v}}_{1(j)}^2 - 2\alpha_{(j)}\tilde{\mathbf{v}}_{1(j)} + 2\beta|\tilde{\mathbf{v}}_{1,(j)}|,$$

where $\alpha_{(j)} = \mathbf{u}_1^T\mathbf{M}_{1(\cdot,j)}$, and $\beta = \lambda_{v_1}/2$. This problem can be solved analytically by soft-thresholding [12]:

$$\tilde{\mathbf{v}}_{1(j)} = \begin{cases} \alpha_{(j)} - \beta, & \text{if } \alpha_{(j)} > \beta, \\ 0, & \text{if } |\alpha_{(j)}| \le \beta, \\ \alpha_{(j)} + \beta, & \text{if } \alpha_{(j)} < -\beta, \end{cases} \quad j = 1, \cdots, d. \tag{7}$$

(b) Solve for $\mathbf{u}_1$ when $\mathbf{v}_1$ is fixed

After $\mathbf{v}_1$ is obtained and fixed, we optimize Problem (5) with respect to σ_1 and $\mathbf{u}_1$. We let $\tilde{\mathbf{u}}_1 = \sigma_1\mathbf{u}_1$, and solve the following problem to obtain $\tilde{\mathbf{u}}_1$. By setting $\sigma_1 = \|\tilde{\mathbf{u}}_1\|_2$ and $\mathbf{u}_1 = \tilde{\mathbf{u}}_1/\sigma_1$, we obtain a solution to Problem (5).

$$\min_{\tilde{\mathbf{u}}_1} \quad \|\mathbf{M}_1 - (\mathbf{z} \odot \tilde{\mathbf{u}}_1)\mathbf{v}_1^T\|_F^2. \tag{8}$$

Each component $\mathbf{u}_{1(i)}$ in an optimal $\mathbf{u}_1$ can be independently and analytically computed as follows:

$$\tilde{\mathbf{u}}_{1(i)} = \begin{cases} \dfrac{\mathbf{M}_{1(i,\cdot)}\mathbf{v}_1}{\mathbf{z}_{(i)}}, & \text{if } \mathbf{z}_{(i)} \neq 0 \\ 0, & \text{if } \mathbf{z}_{(i)} = 0. \end{cases} \quad i = 1, \cdots, n. \tag{9}$$

(2) Find the optimal $\mathbf{z}$ with fixed $\mathbf{u}_1$, $\mathbf{v}_1$, $\mathbf{u}_2$, and $\mathbf{v}_2$

When all values of $\mathbf{u}$'s and $\mathbf{v}$'s are fixed in Problem (2), the optimization problem becomes:

$$\min_{\mathbf{z} \in \mathcal{B}_n, \sigma_1, \sigma_2} \quad \|\mathbf{M}_1 - \sigma_1(\mathbf{z} \odot \mathbf{u}_1)\mathbf{v}_1^T\|_F^2 + \|\mathbf{M}_2 - \sigma_2(\mathbf{z} \odot \mathbf{u}_2)\mathbf{v}_2^T\|_F^2 + \lambda_z\|\mathbf{z}\|_0. \tag{10}$$

Denote the values of σ_i's from the previous iteration by $\hat{\sigma}_1$ and $\hat{\sigma}_2$. We temporarily relax the binary $\mathbf{z}$ variables to be real-valued and then let $\tilde{\mathbf{z}} = \hat{\sigma}_1\mathbf{z}$. Again, we use the ℓ_1-norm of $\tilde{\mathbf{z}}$ to approximate its 0-norm and solve the following problem for $\tilde{\mathbf{z}}$:

$$\min_{\tilde{\mathbf{z}}} \quad \|\mathbf{M}_1 - (\tilde{\mathbf{z}} \odot \mathbf{u}_1)\mathbf{v}_1^T\|_F^2 + \|\mathbf{M}_2 - (\hat{\sigma}_2/\hat{\sigma}_1)(\tilde{\mathbf{z}} \odot \mathbf{u}_2)\mathbf{v}_2^T\|_F^2 + \lambda_z\|\tilde{\mathbf{z}}\|_1 \tag{11}$$

The normalization step for $\tilde{\mathbf{z}}$ by σ_1 is used to contrast the different singular values for the different views so re-scaling $\mathbf{z}$ will not cause an issue. Note that Problem (11) can be rewritten as follows:

$$\min_{\tilde{\mathbf{z}}} \quad \|\mathbf{M} - \text{diag}(\tilde{\mathbf{z}})\mathbf{E}\|_F^2 + \lambda_z\|\tilde{\mathbf{z}}\|_1$$

where $\mathbf{M} = [\mathbf{M}_1\ \mathbf{M}_2]$ is obtained by concatenating the data matrices in columns, $\mathbf{E} = [\mathbf{u}_1\mathbf{v}_1^T\ \ (\hat{\sigma}_2/\hat{\sigma}_1)\mathbf{u}_2\mathbf{v}_2^T]$, and diag($\tilde{\mathbf{z}}$) converts $\tilde{\mathbf{z}}$ into a diagonal matrix. Then, each component of an optimal $\tilde{\mathbf{z}}$ can be analytically computed as follows:

$$\tilde{\mathbf{z}}_{(i)} = \begin{cases} \gamma_{(i)} - \theta, & \gamma_{(i)} > \theta \\ 0, & |\gamma_{(i)}| \leq \theta \\ \gamma_{(i)} + \theta, & \gamma_{(i)} < -\theta \end{cases} \quad i = 1, \cdots, n. \tag{12}$$

where $\gamma_{(i)} = \frac{\mathbf{E}_{(i,\cdot)}\mathbf{M}_{(i,\cdot)}^T}{\|\mathbf{E}_{(i,\cdot)}\|_2^2}$ and $\theta = \frac{\lambda_z}{2\|\mathbf{E}_{(i,\cdot)}\|_2^2}$. Eq.(12) is derived based on the same calculation in [12] which was used to derive Eq.(7).

After obtaining $\tilde{\mathbf{z}}$, the solution $\mathbf{z}$ to Problem (10) can be calculated as follows:

$$\mathbf{z}_{(i)} = \begin{cases} 1, & \text{if } \tilde{\mathbf{z}}_{(i)} \neq 0 \\ 0, & \text{if } \tilde{\mathbf{z}}_{(i)} = 0. \end{cases} \quad i = 1, \cdots, n. \tag{13}$$

To preserve the same objective value of Problem (2) after updating $\mathbf{z}$, we update $\mathbf{u}_1$ and $\mathbf{u}_2$ as follows:

$$\mathbf{u}_{(i)} = \begin{cases} \mathbf{u}_{(i)}/\tilde{\mathbf{z}}_{(i)}, & \text{if } \tilde{\mathbf{z}}_{(i)} \neq 0, \\ 0, & \text{if } \tilde{\mathbf{z}}_{(i)} = 0, \end{cases} \quad i = 1, \cdots, n. \tag{14}$$

and σ_1, σ_2 are recalculated as: $\sigma_1 = \|\mathbf{u}_1\|_2$, $\sigma_2 = (\hat{\sigma}_2/\hat{\sigma}_1)\|\mathbf{u}_2\|_2$; then we normalize $\mathbf{u}_1$ and $\mathbf{u}_2$ by $\mathbf{u}_1 = \mathbf{u}_1/\|\mathbf{u}_1\|_2$, and $\mathbf{u}_2 = \mathbf{u}_2/\|\mathbf{u}_2\|_2$.

The proposed algorithm alternates between solving the three sub-problems (6), (8) and (10) until a local minimizer is reached. The overall objective is monotonically non-increasing when minimizing each sub-problem, so the convergence of this iterative process is guaranteed. When applied to both synthetic and real world datasets, this process reached a convergent point in about 10 iterations. To derive another row subgroup, we repeat the algorithm using new matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ that either exclude the rows corresponding to the subjects in the identified subgroup or are deflated by subtracting the identified singular value components $\sigma\mathbf{u}\mathbf{v}^T$. By repeating this procedure, the desired number of subject groups can be achieved.

Extension to more than two views

In some applications, more than two views of data can be available. For example, besides data on clinical features and genetic markers, gene expression data may also be used in the analysis. The optimization problem (2) can be readily extended to incorporate m separate data matrices, e.g., $\mathbf{M}_i$, $i = 1, \cdot, m$, as follows:

$$\begin{aligned} \min_{\mathbf{z},\sigma_i,\mathbf{u}_i,\mathbf{v}_i,i=1,\ldots,m} \quad & \sum_{i=1}^{m} \|\mathbf{M}_i - \sigma_i(\mathbf{z} \odot \mathbf{u}_i)\mathbf{v}_i^T\|_F^2 + \lambda_z\|\mathbf{z}\|_0 + \sum_{i=1}^{m} \lambda_{v_i}\|\sigma_i\mathbf{v}_i\|_0, \\ \text{subject to} \quad & \|\mathbf{u}_i\|_2 = 1, \|\mathbf{v}_i\|_2 = 1, \quad i = 1, ..., m, \\ & \mathbf{z} \in \mathcal{B}_n. \end{aligned}$$

This problem can be solved similarly by decomposing it into several sub-problems and solving each sub-problem in turn. We obtain the singular vectors of the data matrix in the view i, i.e., $\mathbf{u}_i$ and $\mathbf{v}_i$ while fixing $\mathbf{z}$ and other $\mathbf{u}$'s and $\mathbf{v}$'s by optimizing:

$$\begin{aligned} \min_{\sigma_i,\mathbf{u}_i,\mathbf{v}_i} \quad & \|\mathbf{M}_i - \sigma_i(\mathbf{z} \odot \mathbf{u}_i)\mathbf{v}_i^T\|_F^2 + \lambda_{v_i}\|\sigma_i\mathbf{v}_i\|_0, \\ \text{subject to} \quad & \|\mathbf{u}_i\|_2 = 1, \|\mathbf{v}_i\|_2 = 1. \end{aligned}$$

Note that when $\mathbf{z}$ is fixed, the optimization of $\mathbf{u}_i$ and $\mathbf{v}_i$ is independent from one another among different views. Thus, these singular vectors can be computed in parallel, which can reduce the computation time significantly when more computational resources are available. When $\mathbf{u}_i$ and $\mathbf{v}_i$ are fixed for all views, we solve the following problem to obtain $\tilde{\mathbf{z}}$ and rescale $\tilde{\mathbf{z}}$ to obtain $\mathbf{z}$:

$$\min_{\tilde{\mathbf{z}}} \quad \sum_{i=1}^{m} \|\mathbf{M}_i - (\hat{\sigma}_i/\hat{\sigma}_1)(\tilde{\mathbf{z}} \odot \mathbf{u}_i)\mathbf{v}_i^T\|_F^2 + \lambda_z \|\tilde{\mathbf{z}}\|_1.$$

Algorithm 1 summarizes all of the related steps to solve a multi-view SVD. Again, this algorithm can be repeated to obtain subsequent clusters in iterations. Although a good initialization can be problem-specific, we chose to initialize $\mathbf{z}$ with a vector of all ones, which assumes that all subjects have the potential to be in the cluster if no prior is given.

Algorithm 1 Multi-view Singular Value Decomposition

```
Input: $\mathbf{M}_i$, $\lambda_z$, $\lambda_{v_i}$, $i = 1, \cdots, m$
Output: $\mathbf{z}$, $\sigma_i$, $\mathbf{u}_i$, $\mathbf{v}_i$, $i = 1, \cdots, m$
1. Initialize $\mathbf{z}$ with a vector of all ones.
2. Initialize $\mathbf{u}_i$'s by the corresponding left singular vectors of $\mathbf{M}_i$, $i = 1, \cdots, m$.
3. For $i = 1, \cdots, m$,
     Compute $\tilde{\mathbf{v}}_i$ by Eq.(7).
     Compute $\mathbf{v}_i$ from $\tilde{\mathbf{v}}_i$ and update $\sigma_i$.
     Compute $\tilde{\mathbf{u}}_i$ by Eq.(9).
     Compute $\mathbf{u}_i$ from $\tilde{\mathbf{u}}_i$ and update $\sigma_i$.
4. Compute $\tilde{\mathbf{z}}$ by Eq.(12).
5. Compute $\mathbf{z}$ from $\tilde{\mathbf{z}}$ by Eq.(13).
6. Update $\sigma_i$, $\mathbf{u}_i$, $i = 1, \cdots, m$ by Eq.(14) accordingly.
Repeat Steps 3 to 6 until $\mathbf{z}$ reaches a fixed point.
```

Results and discussion

We first validated the proposed method using synthetic data that were simulated with known cluster and association structures. We then evaluated our approach on a real world disease dataset aggregated from multiple genetic studies of cocaine dependence (CD).

Normalized mutual information (NMI) was used to measure the agreement between any two cluster solutions. Denote two clusterings by $\mathcal{C}^{(1)}$ and $\mathcal{C}^{(2)}$ where each clustering contains a number of clusters as a partition of a given sample, and $\mathcal{C}_i$ is a set containing indexes of the subjects in the i-th cluster. NMI computes the *mutual information* between the two clusterings normalized by the cluster entropies. In other words,

$$\text{NMI}(\mathcal{C}^{(1)}, \mathcal{C}^{(2)}) = \frac{I(\mathcal{C}^{(1)}, \mathcal{C}^{(2)})}{(H(\mathcal{C}^{(1)}) + H(\mathcal{C}^{(2)}))/2} \tag{15}$$

where $I(\mathcal{C}^{(1)}, \mathcal{C}^{(2)}) = \sum_{i,j} \frac{|\mathcal{C}_i^{(1)} \cap \mathcal{C}_j^{(2)}|}{n} \log \frac{n|\mathcal{C}_i^{(1)} \cap \mathcal{C}_j^{(2)}|}{|\mathcal{C}_i^{(1)}||\mathcal{C}_j^{(2)}|}$, $H(\mathcal{C}) = -\sum_i \frac{|\mathcal{C}_i|}{n} \log \frac{|\mathcal{C}_i|}{n}$, and $|\mathcal{C}_i|$ denotes the number of subjects in the cluster $\mathcal{C}_i$. Because the true clusters are known in synthetic data, we computed NMI to measure the agreement between the true cluster assignments and the cluster assignments resulting from cluster analysis. A higher NMI value indicates better performance.

In addition to NMI, for each clustering, classifiers were constructed based on genetic markers to separate subjects in different clusters. We used the Area Under the receiver operating characteristic Curve (AUC) [15] in a 10-fold cross-validation setting to measure the genetic separability or homogeneity of the clusters in a clustering and compared it between different clusterings. We used a regularized logistic regression [16] as the classification model in these experiments.

We compared the proposed approach extensively against biclustering and multi-view analytics. We calculated NMI for different methods on synthetic data and AUC values on both synthetic and real world data. Our comparison study included the following existing methods:

- **Single-view SSVD:** Clusters were included in the comparison by running the method of SSVD-based biclustering in the clinical view, as the biclustering method does not handle multiple views. Applying this method to genetic data created completely different clusters from those obtained in the clinical view.
- **Co-regularized spectral:** This method was proposed previously [9] to find consistent row clusters across multiple views by applying spectral clustering to each view in turn together with a co-regularization factor applied to the cluster indicator vector.
- **Kernel addition:** Radial basis function (RBF) kernels were calculated for each view and combined by summing them. Then spectral clustering was applied to the combined kernel to obtain row clusters.
- **Kernel product:** This is the same procedure as in the kernel addition described above except that kernel matrices were combined by multiplying their components in the same position.
- **Feature concatenation:** Data from the two views were combined by feature concatenation and a kernel matrix was computed based on the combined features. It was then used in spectral clustering to obtain row clusters.

A simulation study

Two disease subtypes, *subtype 1* and *subtype 2*, were simulated. Each of the subtypes was both defined by a set of phenotypic/clinical features and associated with a set of genetic markers. However, the clinical features and genetic markers differed for the two subtypes. Thus, each subtype corresponded to a cluster of subjects with the specific clinical features and the associated SNP markers (here we assumed that minor alleles at each locus were risk variants). The goal of the simulation was to create a reference partitioning of subjects in both views (i.e., genetic markers and clinical features).

Genetic data were obtained from the 1000 Genome Project [17], in which 1092 subjects were genotyped for several million genetic markers. We randomly selected 1000 markers from chromosome 5 that had a minor allele frequency of at least 5% as genetic inputs in our experiments. Ten markers (different for each subtype) were randomly chosen to be associated with each subtype. Thus, a cluster of subjects was formed for each subtype, and we assigned subjects to a cluster if they had ≥ 8 risk variants out of the 10 SNPs chosen for that subtype. This amounts to an additive genetic model for each subtype (i.e., derived by adding the risk variants). Subjects who did not belong to either of the subtypes were treated as controls, forming the third subject cluster. We removed from the analysis subjects who belonged to both subtypes to ensure clarity in the partition. A total of 1013 subjects were retained. Of these, 247 and 167 were assigned to *subtype 1* and *subtype 2*, respectively, and 599 were controls. We named these clusters the genotypic clusters.

We then created clusters of the same subjects in the clinical view to be consistent to a certain degree with the genotypic clusters. Note that many diseases, although highly heritable, are multifactorial genetically and environmentally. To reflect the environmental effects on the clinical features, we introduced random noise to the synthesized clinical data so that the clinical clusters were not exactly the same as the genotypic clusters, so as to test the robustness of the proposed approach. We used a parameter e to indicate the relative effect that genetic variation contributed to the phenotypic variation. Denote r_i^j the number of risk variants of *subtype j* shared by subject i, so $0 \leq r_i^j \leq 10$ according to our definition of genotypic clusters. If $r_i^j * e + N(0,1) > 7.5 * e$, we assigned subject i to *subtype j*. This process created clusters of subjects that were different but similar to the genotypic clusters (with the parameter e reflecting the level of similarity).

We named these clusters the phenotypic clusters because they were used to synthesize clinical features such that the clinical data represented these clusters. Similarly, we removed from the analysis subjects that overlapped in the two phenotypic clusters. Fewer than 15 subjects were excluded in any simulated dataset in the experiments. In addition to these two phenotypic clusters, two additional phenotypic clusters, independent of any genetic variant and based on clinical features only, were created to make the simulated data more difficult but more realistic. Each of the two additional clusters included 200 subjects that were randomly selected among the controls. This design aimed to reflect the observation that multiple clinical clusters may exist in a sample, but only some clusters (two in our simulations) are associated with genetic factors.

We simulated 10 binary phenotypic/clinical features that exhibited the phenotypic clusters. A subject was assigned a value of 0 or 1 for each of the features according to a pre-defined probability. *Subtype 1* and *subtype 2* each were associated with three features. Subjects in each simulated phenotypic cluster were assigned a value of 1 with probabilities of 0.6, 0.5, and 0.4, respectively, for the three designated features. Each of the two additional phenotypic clusters was associated with two features, and subjects in each of the two subtypes were assigned a value of 1 in the two features, with probabilities of 0.6 and 0.5, respectively. A subject was assigned a value of 1 with a probability of 0.1 on any other features.

To evaluate how the proposed method performed when the genetic effect varied, four phenotypic datasets with $e = 1,\ 0.8,\ 0.6$, and 0.4 were generated and analysed. The genetic effect on phenotypic variation decreases with decreasing e, which leads to a lower level of agreement between genotypic and phenotypic clusters.

All of the available methods were used to obtain three subject clusters. Table 1 provides the NMI calculated by comparing subject clusters obtained from each approach to the simulated phenotypic clusters. The proposed method has the highest NMI on all four of the datasets. With decreasing e, the NMI obtained by the proposed method decreases gradually, as expected, but the subject clusters consistent between the two views can still be discerned.

Table 1 Comparison of different methods on their cluster validity in the simulation

	$\boldsymbol{e = 1}$	$\boldsymbol{e = 0.8}$	$\boldsymbol{e = 0.6}$	$\boldsymbol{e = 0.4}$
Single-view SSVD	0.0821	0.1798	0.2432	0.2286
Co-regularized Spectral	0.2306	0.2477	0.2338	0.2549
Kernel addition	0.2587	0.2295	0.2350	0.2566
Kernel product	0.1917	0.2432	0.2302	0.2310
Feature concatenation	0.1569	0.1576	0.1532	0.1211
Proposed method	*0.7949*	*0.7693*	*0.6815*	*0.6329*

The normalized mutual information (NMI) values are shown, measuring the agreement between the clusters resulting from an approach and the simulated phenotypic clusters. The genetic contribution to the phenotypic variation varied according to different e values. A greater e value indicates a higher agreement between the simulated phenotypic clusters and genotypic clusters, making it easier for a clustering approach to recover the simulated phenotypic clusters. Italic fonts indicate the best performance in the experiments with each of the e values.

For each cluster solution, two classification models were built to separate subjects in each of the two subtypes from controls. The subject cluster from each method containing the largest number of controls was considered the control group. The average AUC values and their interquantiles obtained by all compared approaches on each dataset are plotted in Figure 1. The proposed method achieved the second best performance on this measurement. Although the feature concatenation method obtained the clusters that were most separable genetically (i.e., with the best AUC), the clusters were not clinically recognizable. As shown in Table 1, they were the most disparate from the simulated true phenotypic clusters.

Figure 1 Comparison of different methods on AUC values in the simulation. The box plot of AUC values obtained from all approaches in the comparison is shown for the simulated data. The methods were: A1 - the proposed method, A2 - single-view SSVD, A3 - co-regularized spectral clustering, A4 - kernel addition, A5 - kernel product, and A6 - feature concatenation. The parameter e reflects the level of genotypic effect to the phenotypic variation in the simulated data. The AUC values characterize the genetic separability of the clusters resulting from each method.

A significant advantage of the proposed method is that it can simultaneously identify the features that specify the subject clusters. We calculated the number of features that were correctly and incorrectly identified by the proposed method to measure its performance in this regard. The results are summarized in Table 2, which shows that our approach correctly identified all true associated features in both views with a very low false discovery rate ($\sim 15/1000$) when taking into account the total number of features used in the analysis.

Table 2 The features identified by the proposed method in both views in the simulation

		Phenotypic view			**Genotypic view**		
		TF	**TPF**	**FPF**	**TF**	**TPF**	**FPF**
Subtype 1	$e = 1$	3	3	1	10	10	4
	$e = 0.8$		3	1		10	5
	$e = 0.6$		3	2		10	15
	$e = 0.4$		3	0		10	10
Subtype 2	$e = 1$	3	3	0	10	10	4
	$e = 0.8$		3	0		10	4
	$e = 0.6$		3	0		10	2
	$e = 0.4$		3	0		10	5

The parameter e reflects the level of genotypic effect to the phenotypic variation in the simulated data. TF is the number of True Features used in the simulation to specify a subject cluster. TPF (True Positive Features) is the number of features correctly identified. FPF (False Positive Features) is the number of features incorrectly identified.

A disease study: cocaine use and related behaviors

A total of 1,474 African Americans were phenotyped and genotyped for genetic studies of cocaine dependence (CD) [18]. Subjects were recruited from the Yale University School of Medicine, University of Connecticut Health Center, University of Pennsylvania School of Medicine, McLean Hospital and Medical University of South Carolina. All subjects gave written, informed consent to participate, using procedures approved by the institutional review board at each participating site. Subjects were phenotyped using a computer-assisted interview, called the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA) [19], a polydiagnostic instrument that was used to generate diagnoses of dependence on cocaine and other substances. Sixty-four yes-or-no variables were generated by this survey, which were also used in previous genetic association studies [1,20,21]. These variables were used as the phenotypic features. Of the 1,474 subjects, 1,287 were diagnosed with cocaine dependence. Subjects were genotyped for 1,350 SNPs selected from 130 candidate genes [4] and 186 ancestry informative markers (AIMs) using the Illumina GoldenGate Assay platform (Illumina, Inc., San Diego, CA).

The original dataset aggregated from two studies was preprocessed with a sequence of steps for data cleaning and to address population stratification. Race was classified using STRUCTURE v2.3 [22] and AIMs, which stratified the study subjects into two population groups: African Americans (AAs) and European Americans (EAs). The AA group was used in the present analysis. Of the 1,474 AAs, 93.78% had AA as their self-reported race. We excluded other population groups from the analysis. Principal components analysis (PCA) was performed on the 186 AIMs for the stratified AA population. The first PCA dimension was used in the subsequent association tests as a covariate to correct for the residual population structure. SNPs for which data were available for less than 95% of the subjects, or for which the P value for Hardy-Weinberg equilibrium was less than 10^{-7}, were excluded from our analysis. The minor allele frequency (MAF) of each SNP was calculated within this AA population group. SNPs with a MAF $< 1\%$ were removed. The remaining 1,248 SNPs were used as the genetic markers in the multi-view biclustering experiment. The SNPs selected by the proposed Algorithm 1 were then used in the association test that was based on the logistic regression model.

The feature concatenation method overlooked the information in the clinical or phenotypic view as observed in both the simulation study and the case study. Thus, we excluded the kernel concatenation method from further comparisons. Three subject clusters were obtained from each of the methods in the comparison. Logistic regression models were built with sex, age and the first PCA dimension as covariates and tested in a manner similar to that used for synthetic data. Figure 2 shows the box plot of the AUC values. As shown there, our approach significantly outperformed all other methods with respect to the genetic separability of the resultant clusters. A paired t-test to compare the AUC values from our method with each of the other methods yielded a p-values $<$ 0.05 for all comparisons.

Figure 2 Comparison of different methods on AUC values in the CD study. The box plot of AUC values were obtained from all methods on the data of cocaine use and related behaviors. A1 - the proposed method, A2 - single-view SSVD, A3 - co-regularized spectral clustering, A4 - kernel addition, A5 - kernel product.

For the proposed method, the three identified subject clusters contained 795 (*Group 1*), 295 (*Group 2*) and 384 (*Group 3*) subjects. *Group 1* and *Group 2* were identified consecutively, and *Group 3* contained the remaining subjects. *Group 3* contained more than 80% of the control subjects; thus, we used this group as a control group in our association analysis. The number of clinical features identified as associated with *Group 1* and *Group 2* were 18 and 17, respectively. Figures 3 and 4 compare the three subject clusters on the percentage of positive responses to the identified clinical features. A few identified features are not shown in the figures, because they are highly correlated ($r > 0.7$) with the features shown.

Figure 3 Comparison among the three cocaine user groups on the features identified for *Group 1*. Cocaine use symptoms are identified by the superscript ^{1}, and the symptoms due to stopping, cutting down or going without cocaine are identified by the superscript ^{2}. The percentage of individuals endorsing any of the features are reported for each user group.

Figure 4 Comparison among the three cocaine user groups on the features identified for *Group 2*. The percentage of individuals endorsing any of the features are reported for each user group.

From these two figures, we can see that *Group 1* is distinctively associated with several withdrawal symptoms, such as feeling depressed, restless, or tired when the subject stopped, cut down or went without cocaine. When *Group 2*, the second row cluster, was identified, the corresponding column cluster contained 17 clinical features. We plotted the percentage of positive responses to eight of these features for all three cocaine user groups in Figure 4. Subjects in both *Group 2* and *Group 1* showed high values on these features. Note that subjects in *Group 1* were excluded when the second cluster was derived. From these observations, we can conclude that *Group 1* is a heavy user group with many negative consequences of cocaine use, *Group 2* is a moderate cocaine user group, and *Group 3* is a low cocaine user group.

There were 114 and 237 genetic markers identified for *Group 1* and *Group 2*, respectively, by Algorithm 1. Based on these markers, two logistic regression models were built to identify the markers that had the highest predictive power in distinguishing subjects in *Group 1* or in *Group 2*, from those in the control group. Table 3 gives the 5 SNPs that received the largest magnitude of weights in the models. It is interesting to note that the *HTR2C* gene was significantly associated with *Group 1* in our study (p-value $< 10^{-5}$), having previously been identified with a heavy use, early-onset and high comorbidity subtype of cocaine dependence [20].

Table 3 Top five SNPs associated with each of the two CD subtypes

	SNP	Chr	MAF	HWE	Gene
	rs6318	chrX	0.3643	1.00	*HTR2C*
Group 1	rs2427400	chr20	0.1280	0.22	*NTSR1*
vs.	rs460401	chr21	0.3500	0.18	*GRIK1*
Group 3	rs10485058	chr06	0.0585	0.38	*OPRM1*
	rs2279423	chr15	0.0237	0.81	*CHRM5*
	rs897692	chr11	0.3972	0.86	*HTR3A*
Group 2	rs9996854	chr04	0.5436	0.61	*GABRB1*
vs.	rs481036	chr01	0.5582	0.21	*CHRM3*
Group 3	rs6092933	chr20	0.2070	0.17	*SLC32A1*
	rs9371781	chr06	0.3687	0.49	*OPRM1*

The five SNP markers that received the largest magnitude of weights in the two classification models that separate the subtype cases, in *Group 1* and *Group 2*, respectively, from the controls in *Group 3*. The SNP name, the SNP location (chromosome i.e., Chr), the name of the gene (Gene), the minor allele frequency (MAF) and the P-value for Hardy Weinberg equilibrium (HWE) are provided for each SNP.

Conclusion

It is challenging to identify the genetic causes of complex disorders such as substance dependence, due to their heterogeneous clinical manifestations and complex genetic etiologies, which include gene x environment interactions. Phenotype refinement that leads to homogeneous subtypes is a promising approach to solve this problem [1,5,23-25]. However, most of the methods used to refine phenotypes take into consideration only the phenotypic information, despite the availability of genotypic information in genetic studies of a complex disorder. Thus, existing approaches have had limited success in finding a phenotypic subtype that is genetically homogeneous. In this paper, we propose a multi-view biclustering approach to refine the phenotype by jointly taking into account genetic and phenotypic information.

The proposed method is distinct from existing multi-view data analytics in that the relevant features can be identified at the same time that a subtype is determined, which is critical to its success. This increases the likelihood of finding genetic associations. The proposed method is distinct from existing biclustering methods in that it harmonizes the subject groupings in two or more views. The developed algorithm is highly scalable with large datasets because at each iteration it calculates closed-form solutions for different groups of working variables. The results from extensive experimental comparisons on both synthetic data and real world datasets demonstrate the effectiveness and superior performance of the proposed approach.

This study has a number of limitations. The proposed multi-view biclustering method, in its current form, does not simultaneously handle population stratification and phenotype-genotype association. It may spuriously identify markers that are relevant to a disease subtype due to population structure rather than being truly associated with the specific disease. Thus, population groups need to be stratified in additional steps such as those performed in our experiments. It is desirable to extend our method to address the three-way relationship among population subgroups, genotypes and phentoypes to ensure the validity of the identified phenotype-genotype associations. Further, the proposed method was used in our empirical study to identify the first two major subgroups of subjects, for which no invalid clusters caused by random noise were identified. When larger numbers of clusters are to be identified, the two methods we designed to find subsequent clusters (by either excluding subjects in the identified subgroups or deflating singular value components from the data matrix) become susceptible to the detection of invalid clusters because singular values will decrease in subsequent decomposition. Empirical studies may be needed to examine more thoroughly the signal-to-noise pattern of the proposed method.

Competing interests

JS and JB declare that they have no competing interests. Although unrelated to this study, HRK has been a consultant or Advisory Board Member for the following pharmaceutical companies: Alkermes, Lilly, Lundbeck, Pfizer, and Roche. He is also a member of the American Society of Clinical Psychopharmacology's Alcohol Clinical Trials Initiative, supported by AbbVie, Ethypharm, Lilly, Lundbeck, and Pfizer.

Authors' contributions

JB and JS designed the algorithm and all authors designed the study together. JS implemented the algorithm in Matlab and performed the experiments. HRK provided the substance dependence datasets and helped to interpret the results. JB and JS wrote the first manuscript, and HRK revised and edited it. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by NSF grant IIS-1320586 and NIH grant DA030976. We thank Joel Gelernter, M.D. from Yale University who was instrumental in recruiting, characterizing, and genotyping the subjects in the datasets used here. Kathleen Brady, M.D., Ph.D. of the Medical University of South Carolina, Roger Weiss, M.D., of McLean Hospital and Harvard Medical School, and David Oslin, M.D., of the University of Pennsylvania Perelman School of Medicine oversaw study recruitment at their respective sites. That work was supported by NIH grants AA011330, AA017535, DA12690, DA18432, and DA12849.

References

1. Kranzler HR, Wilcox M, Weiss RD, Brady K, Hesselbrock V, Rounsaville B, Farrer L, Gelernter J: **The validity of cocaine dependence subtypes.** *Addict Behav* 2008, **33**(1):41–53.

2. Babor TF, Caetano R: **Subtypes of substance dependence and abuse: implications for diagnostic classification and empirical research.** *Addiction (Abingdon, England)* 2006, **101:**104–110.

3. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN: **Genome-wide association studies for complex traits: consensus, uncertainty and challenges.** *Nat Rev Genet* 2008, **9**(5):356–369.

4. Hodgkinson CA, Yuan Q, Xu K, Shen PH, Heinz E, Lobos EA, Binder EB, Cubells J, Ehlers CL, Gelernter J, Mann J, Riley B, Roy A, Tabakoff B, Todd RD, Zhou Z, Goldman D: **Addictions biology: haplotype-based analysis for 130 candidate genes on a single array.** *Alcohol Alcohol* 2008, **43**(5):505–515.

5. Gelernter J, Panhuysen C, Wilcox M, Hesselbrock V, Rounsaville B, Poling J, Weiss R, Sonne S, Zhao H, Farrer L, Kranzler HR: **Genomewide linkage scan for opioid dependence and related traits.** *Am J Hum Genet* 2006, **78**(5):759–769.

6. Schwartz B, Wetzler S, Swanson A, Sung SC: **Subtyping of substance use disorders in a high-risk welfare-to-work sample: a latent class analysis.** *J Subst Abuse Treat* 2010, **38**(4):366–374.

7. Chen P, Hung YS, Fan Y, Wong STC: **An integrative bioinformatics approach for identifying subtypes and subtype-specific drivers in cancer.** In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. New York: IEEE; 2012:169–176.

8. Tay ST, Leong SH, Yu K, Aggarwal A, Tan SY, Lee CH, Wong K, Visvanathan J, Lim D, Wong WK, Soo KC, Kon OL, Tan P: **A combined comparative genomic hybridization and expression microarray analysis of gastric cancer reveals novel molecular subtypes.** *Cancer Res* 2003, **63**(12):3309–3316.

9. Kumar A, Rai P, Daume H III: **Co-regularized multi-view spectral clustering.** In *Advances in Neural Information Processing Systems 24.* Edited by Shawe-Taylor J, Zemel RS, Bartlett P, Pereira FCN, Weinberger KQ. Cambridge, MA: MIT Press; 2011:1413–1421.

10. Chaudhuri K, Kakade SM, Livescu K, Sridharan K: **Multi-view clustering via canonical correlation analysis.** In *Proceedings of the 26th International Conference on Machine Learning.* New York: ACM; 2009:129–136.

11. Van Mechelen I, Bock H-H, De Boeck P: **Two-mode clustering methods: a structured overview.** *Stat Methods Med Res* 2004, **13**(5):363–394.

12. Lee M, Shen H, Huang JZ, Marron JS: **Biclustering via sparse singular value decomposition.** *Biometrics* 2010, **66**(4):1087–1095.

13. Kumar A, Daume H III: **A co-training approach for multi-view spectral clustering.** In *Proceedings of the 28th International Conference on Machine Learning.* Edited by Getoor L, Scheffer T. New York: ACM; 2011:393–400.

14. Guan Y, Dy J, Jordan MI: **A unified probabilistic model for global and local unsupervised feature selection.** In *Proceedings of the 28th International Conference on Machine Learning*. New York: ACM; 2011:1073–1080.

15. Fawcett T: **An introduction to ROC analysis.** *Pattern Recogn Lett* 2006, **27**(8):861–874.

16. Yuan G-X, Ho C-H, Lin C-J: **An improved glmnet for l1-regularized logistic regression.** *J Mach Learn Res* 2012, **13:**1999–2030.

17. The 1000 Genomes Project Consortium: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491**(7422):56–65.

18. American Psychiatric Association: *Diagnostic and Statistical Manual of Mental Disorders: Fourth Edition (DSM-IV)*. Washington, DC: American Psychiatric Press Inc; 1994.

19. Pierucci-Lagha A, Gelernter J, Chan G, Arias A, Cubells JF, Farrer L, Kranzler HR: **Reliability of dsm-iv diagnostic criteria using the semi-structured assessment for drug dependence and alcoholism (ssadda).** *Drug Alcohol Depend* 2007, **91**(1):85–90.

20. Bi J, Gelernter J, Sun J, Kranzler HR: **Comparing the utility of homogeneous subtypes of cocaine use and related behaviors with DSM-IV cocaine dependence as traits for genetic association analysis.** *Am J Med Genet B* 2013, **165B**(2):148–156.

21. Sun J, Bi J, Kranzler HR: **Multi-view co-modeling to improve subtyping and genetic association of complex diseases.** *IEEE J Biomed Health Inf* 2013, **18**(2):548–554.

22. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155**(2):945–959.

23. Chan G, Gelernter J, Oslin D, Farrer L, Kranzler HR: **Empirically derived subtypes of opioid use and related behaviors.** *Addiction* 2011, **106**(6):1146–1154.

24. Sun J, Bi J, Chan G, Anton RF, Oslin D, Farrer L, Gelernter J, Kranzler HR: **Improved methods to identify stable, highly heritable subtypes of opioid use and related behaviors.** *Addict Behav* 2012, **37**(10):1138–1144.

25. Sun J, Bi J, Kranzler HR: **A multi-objective program for quantitative subtyping of clinically-relevant phenotypes.** In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM2012)*. New York: ACM; 2012:256–261.

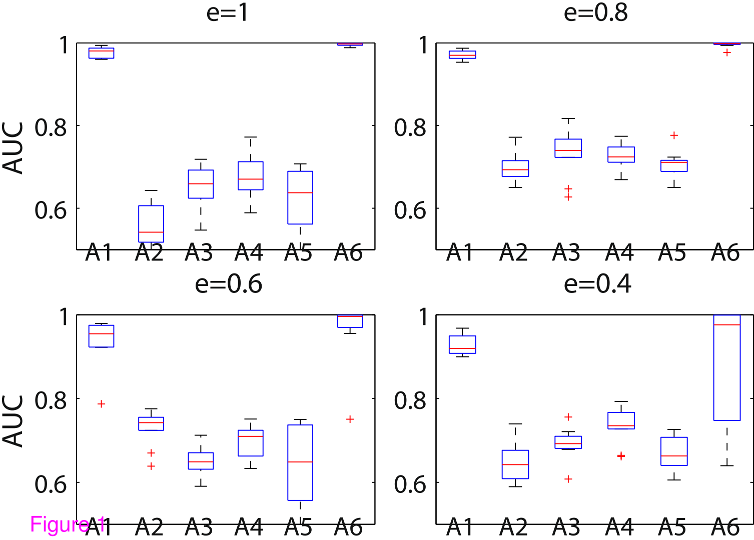

Figure1

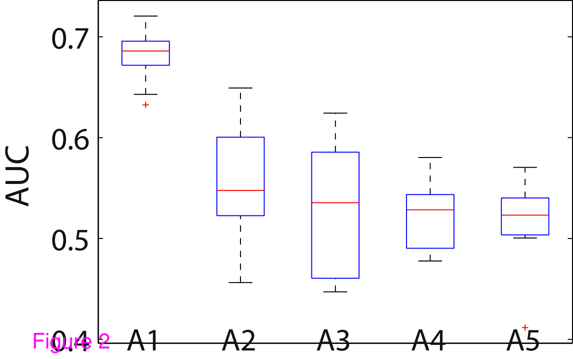

Figure 2

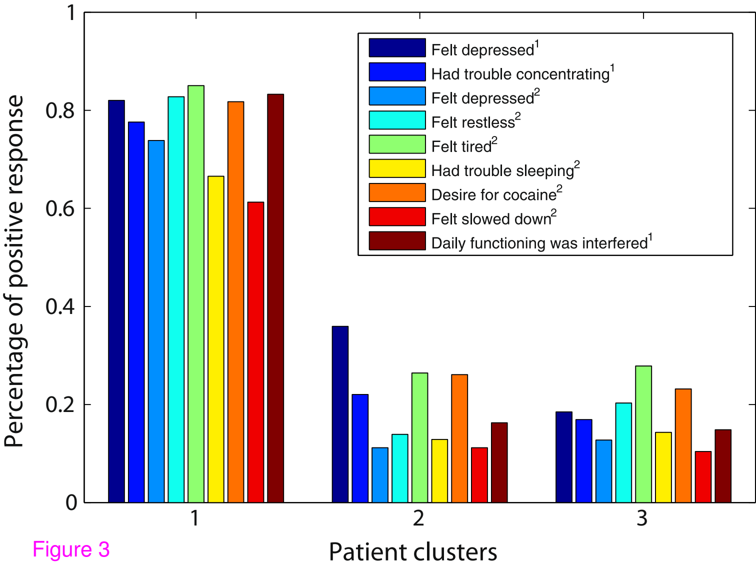
1
0.8
0.6
0.4
0.2
0
Percentage of positive response
1
2
3
Patient clusters
Felt depressed[1]
Had trouble concentrating[1]
Felt depressed[2]
Felt restless[2]
Felt tired[2]
Had trouble sleeping[2]
Desire for cocaine[2]
Felt slowed down[2]
Daily functioning was interfered[1]

Figure 3

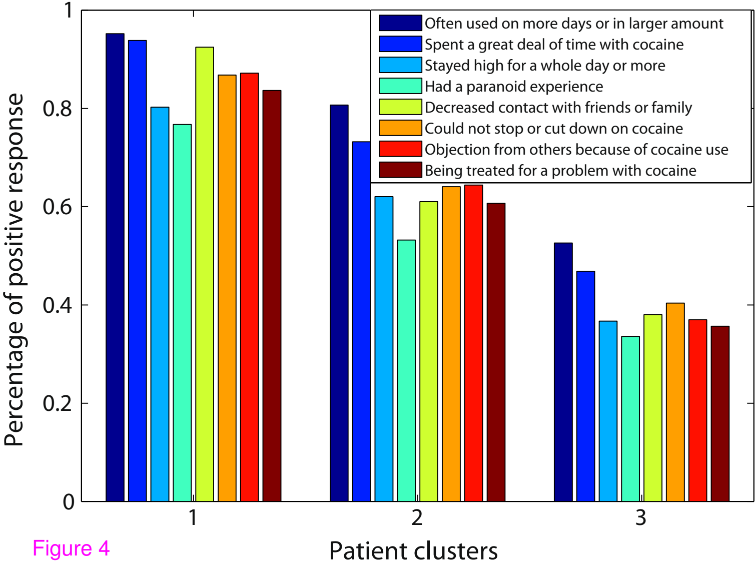
1
0.8
0.6
0.4
0.2
0
Percentage of positive response
1
2
3
Patient clusters
Often used on more days or in larger amount
Spent a great deal of time with cocaine
Stayed high for a whole day or more
Had a paranoid experience
Decreased contact with friends or family
Could not stop or cut down on cocaine
Objection from others because of cocaine use
Being treated for a problem with cocaine

Figure 4