Multi-view Biclustering for Genotype-Phenotype Association Studies of Complex Diseases

Jiangwen Sun, Jinbo Bi
Department of Computer Science and Engineering
University of Connecticut
Storrs, CT, USA
javon, jinbo@engr.uconn.edu

Henry R. Kranzler
Treatment Research Center
University of Pennsylvania
Philadelphia, PA, USA
kranzler@mail.med.upenn.edu

***Abstract*—Complex disorders exhibit great heterogeneity in both clinical manifestation and genetic etiology. This heterogeneity substantially limits the identification of genotype-phenotype associations. Differentiating homogeneous subtypes of a complex phenotype will enable the detection of genetic variants contributing to the effect of subtypes that cannot be detected by the non-differentiated phenotype. However, the most sophisticated subtyping methods available so far perform unsupervised cluster analysis or latent class analysis on only phenotypic features. Without guidance from the genetic dimension, the resultant subtypes can be suboptimal and genetic associations may fail. We propose a multi-view biclustering approach that integrates phenotypic features and genetic markers to detect confirming evidence in the two views for a disease subtype. This approach groups subjects in clusters that are consistent between the phenotypic and genetic views, and simultaneously identifies the phenotypic features that are used to define a subtype and the genotypes that are associated with the subtype. Our simulation study validates this approach, and our extensive comparison with several biclustering and multi-view data analytics on real-life disease data demonstrates the superior performance of the proposed approach.**

***Keywords*-multi-view data analysis; biclustering; subtyping; genotype-phenotype association; substance dependence**

I. Introduction

Identifying genetic variations that underlie complex phenotypes is important in genetics and systems biology. For complex phenotypes, such as substance dependence (SD), a variety of traits that collectively indicate or characterize the phenotype are often associated with substantial phenotypic variation [1]. Genotype-phenotype association analysis of such a complex phenotype is impeded by the phenotypic heterogeneity [2]. Case-control studies based on a binary trait, such as the diagnosis of a disease, that partitions the population into cases (subjects with the disease) and controls (subjects without the disease) does not differentiate the heterogeneous manifestation of the disease. On the other hand, many candidate genes or genomic regions have been shown to be associated to complex diseases [3]. The characteristics or subtypes of the disease for which the association exists often remain unclear. For instance, 130 genes involved in several biological systems have been shown to be related to addictions, but it remains to be determined which addictive behaviors are associated with specific genetic variants [4].

Differentiating homogeneous subtypes of a disease phenotype has shown promise in the identification of genetic variants contributing to the likelihood of subtype membership [5]. However, these studies perform unsupervised cluster analysis or latent class analysis [6] on phenotypic features only. Genotype data has only been used after-the-fact to evaluate subtypes, such as in subsequent association tests with the derived subtypes, rather than to guide the creation of subtypes. Consequently, the resultant subtypes are of little utility in genetic analysis, and genetic association analysis may fail. Integration of data from clinical and genomic dimensions offers benefits, such as new opportunities to find confirmatory evidence of a subtype based on its genetic basis and clinical manifestations. Clinical subtyping methods have not jointly used clinical and genomic data to define subtypes.

There has been little research on this topic in the statistics literature. The most closely related area involves multi-view data analysis [7], [8], where samples are characterized or viewed in multiple ways, thus creating multiple sets of input variables. Multi-view clustering [7] seeks groupings that are consistent across different views, but they use all of the phenotypic features and genetic markers to define clusters/subtypes and cannot be used to identify subtype-specific variants. Our subtyping problem, although similar to multi-view clustering, seeks to classify subjects in ways that are consistent in the clinical and genetic views, but modeling in both views requires subspace search so that the resulting subtypes rely on only subsets of variables, thus leading to genetically and clinically homogeneous subtypes. For a single view, biclustering methods classify samples and simultaneously identify features that determine the sample classification [9], [10], and subspace clustering methods search subspace and group samples differently in each subspace [11]. However, there is no algorithm to date that finds consensus sample grouping across multiple views based on subsets of variables from each view.

In this paper, we propose a multi-view biclustering approach based on sparse singular value decomposition

(SSVD) technique [10]. The objective of this problem is to identify subject clusters that are consistent in both the clinical and genetic views, and simultaneously identify features and markers that determine the clusters. Employing *sparse* SVD in our approach is critical to its success, especially to successfully detect associative variants given the number of true associative variants are much fewer than the single nucleotide polymorphisms (SNPs) in the whole genome. The proposed approach has been validated on both synthetic datasets that are simulated so that few genetic markers are associated with specific subtypes and a real world clinical dataset that is aggregated from multiple genetic studies of cocaine dependence. We compared our approach to the biclustering approach in [10] and multiple existing multi-view data analytic methods. The results clearly show that the performance of our approach is superior to the other methods examined. This paper is organized as follows. We introduce the proposed multi-view biclustering method in Section II, followed by the computational results in Section III. We provide conclusions in Section IV.

II. METHOD

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 that is defined by $\|\mathbf{v}\|_p = (|\mathbf{v}_{(1)}|^p + \cdots + |\mathbf{v}_{(d)}|^p)^{1/p}$, where $\mathbf{v}_{(i)}$ is a 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 $\mathbf{u} \odot \mathbf{v}$ as a vector whose components are the multiplications of respective components in $\mathbf{u}$ and $\mathbf{v}$. The set $\mathcal{B}_d$ contains all binary vectors of length d. A binary vector means a vector with 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 noted by $\mathbf{M}_{(i,\cdot)}$ and $\mathbf{M}_{(\cdot,i)}$ respectively.

Given a matrix $\mathbf{M}$, a subgroup of its rows and a subgroup of its columns can be simultaneously achieved by sparse singular decomposing $\mathbf{M}$ [10], that is approximating $\mathbf{M}$ with a sparse rank one matrix $\tilde{\mathbf{M}}$. The resulted row and column subgroups help to define each other. Let $\mathbf{u}$ and $\mathbf{v}$ be the two vectors resulted from the SSVD , i.e., $\tilde{\mathbf{M}} = \mathbf{u}\mathbf{v}^T$, rows in $\mathbf{M}$ corresponding to non-zero components in $\mathbf{u}$ form the row subgroup and columns in $\mathbf{M}$ corresponding to non-zero components in $\mathbf{v}$ form the column subgroup. For two data matrices $\mathbf{M}_1$ of size n-by-d_1 and $\mathbf{M}_2$ of size n-by-d_2 that characterize the same set of subjects from two different views, we can obtain $\mathbf{u}_1$, $\mathbf{v}_1$ for $\mathbf{M}_1$, and $\mathbf{u}_2$, $\mathbf{v}_2$ for $\mathbf{M}_2$ by sparse singular value decomposition of $\mathbf{M}_1$ and $\mathbf{M}_2$ separately. However, it will not guarantee the two row clusters specified, respectively, by $\mathbf{u}_1$ and $\mathbf{u}_2$ be consistent. To make them consistent, it requires $\mathbf{u}_1$ and $\mathbf{u}_2$ to have non-zero components at the same positions. Notice that $\mathbf{u}_1$ and $\mathbf{u}_2$ are not necessarily the same given they may be derived from very different features from two views, such as real-valued features in the clinical view and SNP genotypes in the genetic view. We propose to use a binary vector $\mathbf{z}$ that serves as a common factor to link the two views and represent each $\mathbf{u}$ by $\mathbf{z} \odot \mathbf{u}$ in the objective function. When $\mathbf{z}_{(i)} = 0$, the i-th components of both $\mathbf{u}_1$ and $\mathbf{u}_2$ are 0, and consequently, the i-th subject will be excluded from the subgroup in both views. We hence require the sparsity of $\mathbf{z}$ instead of $\mathbf{u}_1$ and $\mathbf{u}_2$ in the optimization problem as follows:

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

where λ_z, λ_{v_1} and λ_{v_2} are hyper-parameters to balance the errors and sparsity penalty.

As an alternative, a restricted version of Eq(1) is to require $\mathbf{u}_1 = \mathbf{u}_2 = \mathbf{u}$ and solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{u},\sigma_i,\mathbf{v}_i,i=1,2} \quad & \|\mathbf{M}_1 - \sigma_1\mathbf{u}\mathbf{v}_1^T\|_F^2 + \|\mathbf{M}_2 - \sigma_2\mathbf{u}\mathbf{v}_2^T\|_F^2 \\ & + \lambda_u\|\sigma_1\mathbf{u}\|_0 + \lambda_{v_1}\|\sigma_1\mathbf{v}_1\|_0 + \lambda_{v_2}\|\sigma_2\mathbf{v}_2\|_0 \\ \text{subject to} \quad & \|\mathbf{u}\|_2 = 1, \|\mathbf{v}_1\|_2 = 1, \|\mathbf{v}_2\|_2 = 1. \end{aligned}$$

This problem is easier to solve without integer variables as in $\mathbf{z}$. It can be similarly solved using the approach proposed in [10]. However, it is an unnecessary constraint to require $\mathbf{u}_1 = \mathbf{u}_2$, 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 problem as the exact values of the difference are not concerned. Our problem concerns with the indicator of whether a component is zero.

We hence focus on developing effective alternating optimization algorithms to solve Problem (1).

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

When $\mathbf{z}$ is fixed, optimal $\mathbf{u}_1$, $\mathbf{v}_1$ and optimal $\mathbf{u}_2$, $\mathbf{v}_2$ that minimize (1) can be computed separately and in the same fashion. Thus, we only discuss how to compute $\mathbf{u}_1$ and $\mathbf{v}_1$ for a given $\mathbf{z}$. This sub-problem can be written as follows:

$$\begin{aligned} \min_{\sigma_1,\mathbf{u}_1,\mathbf{v}_1} \quad & \|\mathbf{M}_1 - \sigma_1(\mathbf{z}\odot\mathbf{u}_1)\mathbf{v}_1^T\|_F^2 + \lambda_{v_1}\|\sigma_1\mathbf{v}_1\|_0 \\ \text{subject to} \quad & \|\mathbf{u}_1\|_2 = 1, \|\mathbf{v}_1\|_2 = 1 \end{aligned} \quad (2)$$

which can be solved by alternating in solving the following two sub-problems:

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

We solve the following equivalent problem for the optimal $\tilde{\mathbf{v}}_1$ and then set $\sigma_1 = \|\tilde{\mathbf{v}}_1\|_2$ and $\mathbf{v}_1 = \tilde{\mathbf{v}}_1/\sigma_1$.

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

Unfortunately, it has shown that this problem is NP-Hard [12]. There are several approximation methods [13]. We use a simple approximation, i.e., $\|\cdot\|_1$ that is also in favor of sparsity [14]. With this approximation, we can obtain $\tilde{\mathbf{v}}_1$ 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 $\tilde{\mathbf{v}}_{1,(i)}$ in $\tilde{\mathbf{v}}_1$ can be analytically computed by soft-thresholding as shown in Eq.(3) [10] where $\alpha = (\mathbf{z} \odot \mathbf{u}_1)^T\mathbf{M}_{1,(\cdot,i)}$ and $\beta = \lambda_{v_1}/2$.

$$\hat{\tilde{\mathbf{v}}}_{(i)} = \begin{cases} \alpha - \beta, & \alpha > \beta \\ 0, & |\alpha| \leq \beta \\ \alpha + \beta, & \alpha < -\beta \end{cases}. \tag{3}$$

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

We now optimize (2) with respect to σ_1 and $\mathbf{u}_1$. By setting $\sigma_1 = \|\tilde{\mathbf{u}}_1\|_2$ and $\mathbf{u}_1 = \tilde{\mathbf{u}}_1/\sigma_1$, solving Problem (2) is equivalent to computing $\tilde{\mathbf{u}}_1$ by

$$\min_{\tilde{\mathbf{u}}_1} \quad \|\mathbf{M}_1 - (\mathbf{z} \odot \tilde{\mathbf{u}}_1)\mathbf{v}_1^T\|_F^2, \tag{4}$$

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

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

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

When we solve Problem (1) with respect to $\mathbf{z}$ only, it is equivalent to solving the following problem:

$$\begin{aligned} \min_{\tilde{\mathbf{z}}} \quad & \|\mathbf{M}_1 - (\tilde{\mathbf{z}} \odot \mathbf{u}_1)\mathbf{v}_1^T\|_F^2 + \\ & \|\mathbf{M}_2 - (\hat{\sigma}_2/\hat{\sigma}_1)(\tilde{\mathbf{z}} \odot \mathbf{u}_2)\mathbf{v}_2^T\|_F^2 + \lambda_z\|\tilde{\mathbf{z}}\|_0 \end{aligned} \tag{5}$$

where $\hat{\sigma}_i$ is the value of σ_i from previous iteration. After obtaining $\tilde{\mathbf{z}}$, $\mathbf{z}$ can be calculated as:

$$\mathbf{z}_{(i)} = \begin{cases} 1, & \text{if } \tilde{\mathbf{z}}_i \neq 0 \\ 0, & \text{if } \tilde{\mathbf{z}}_i = 0. \end{cases}.$$

In order to keep the objective unchanged, we update $\mathbf{u}_1$ and $\mathbf{u}_2$ accordingly as follows:

$$\mathbf{u}_{j,(i)} = \begin{cases} \mathbf{u}_{j,(i)}/\tilde{\mathbf{z}}_i, & \text{if } \tilde{\mathbf{z}}_i \neq 0 \\ 0, & \text{if } \tilde{\mathbf{z}}_i = 0 \end{cases} j = 1, 2,$$

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 $\mathbf{u}_2$ as in $\mathbf{u}_1 = \mathbf{u}_1/\|\mathbf{u}_1\|_2$, $\mathbf{u}_2 = \mathbf{u}_2/\|\mathbf{u}_2\|_2$. Again, here we use ℓ_1-norm of $\tilde{\mathbf{z}}$ to approximate its 0-norm and we obtain $\tilde{\mathbf{z}}$ by solving the following problem:

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

Overall, we alternate between solving above sub-problems until a local minimizer is reached. The overall objective is monotonically non-increasing when minimizing each sub-problem, the convergence of this iterative process is guaranteed. In our experiment both on synthetic and real world data, this process reached a convergent point in about 10 iterations. Algorithm 1 summarizes all related steps. To derive another row subgroup, we repeat algorithm 1 using new matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ that exclude the rows corresponding to the subjects in the identified subgroup. By repeating this procedure, the desired number of population subgroups can be achieved.

Algorithm 1 Joint Multi-view Biclustering

Input: M_1, M_2, λ_z, λ_{v_1}, λ_{v_2}
Output: $\mathbf{z}$, σ_1, σ_2, $\mathbf{u}_1$, $\mathbf{v}_1$, $\mathbf{u}_2$, $\mathbf{v}_2$
1. Initialize $\mathbf{z}$ with all ones.
2. Compute σ_1, $\mathbf{u}_1$ and $\mathbf{v}_1$ by solving (2).
3. Set 0 to components of $\mathbf{z}$ at the positions where corresponding components in $\mathbf{u}_1$ are 0.
4. Compute σ_2, $\mathbf{u}_2$ and $\mathbf{v}_2$ in the same way as how σ_1, $\mathbf{u}_1$ and $\mathbf{v}_1$ are calculated in (2).
5. Compute $\mathbf{z}$ by solving (5) and update σ_i, $\mathbf{u}_i$, $i = 1, 2$ accordingly.
Repeat 2, 4, 5 until $\mathbf{u}_1$ reaches a fixed point.

III. Computational Results

We first validated the proposed method using synthetic data that was simulated with known association structures. Then we evaluated our approach on a real world disease dataset aggregated from multiple genetic studies of cocaine dependence (CD) disorder. Normalized mutual information (NMI) was used to measure the consistency between two cluster solutions. Since the true clusters are known in synthetic data, we computed NMI to measure the consistency between the true clusters and the clusters resulted from clustering methods. A higher NMI value indicates better performance. In addition to NMI, classifiers were built based on genetic markers to separate subjects in different clusters. We used the Area Under the receiver operating characteristic Curve (AUC) in a 10-fold cross validation setting to measure the genetic separability or homogeneity of the resultant clusters. We used a regularized logistic regression as the classification model throughout these experiments.

Extensive comparison of the proposed approach against biclustering and multi-view analytics was conducted. We calculated NMI for different methods on synthetic data and AUC on both synthetic and real world data. The existing methods that were used in our comparison study are given in the following list:

- **Biclustering via 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 in [7] for finding consistent row clusters among multiple views by applying spectral clustering alternatively on each view together with a co-regularization factor applied to the cluster indicator vector.
- **Kernel addition:** RBF kernels were calculated for each view and combined by adding them together. Then spectral clustering was applied to the combined kernel to obtain row clusters.
- **Kernel product:** This is the same procedure 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 simply put together by feature concatenation and a kernel matrix was computed based on the combined dataset with spectral clustering to obtain row clusters.

A. Synthetic data

Two disease subtypes, i.e., *subtype 1* and *subtype 2* were simulated. They had different sets of associative genetic factors that corresponded to different sets of phenotypic/clinical features. We started from simulating genotypic subtypes (population subgroups based on genetic markers), which were subsequently used to generate phenotypic subtypes along with random noise introduced to reflect environmental effects.

Genetic data was obtained from the 1000 Genome Project [15] and 1092 subjects were genotyped with several million genetic markers in this project. We randomly selected 1000 markers from chromosome 5 that had a minor allele frequency (MAF) of at least 5% as genetic inputs in the experiment. For each subtype, 10 markers were randomly chosen to be associated with each subtype. To assign subjects to subtypes, we assumed that the minor allele at each locus was the risk variant. We assigned subjects to a subtype if they had over 8 risk variants for that particular subtype. Subjects who did not belong to any of the subtypes were treated as controls. We removed from the analysis subjects who belonged to both of the two subtypes to ensure the clarity of comparison results. In total, 1013 subjects were retained for subsequent analysis. Of that number, 247 and 167 subjects were assigned to *subtype 1* and *subtype 2*, respectively and 599 were controls. We named the above population partition the genotypic subgroups.

To create population subgroups in the phenotypic view, we introduced random noise to reflect the environmental effects on phenotypic features. We used a parameter e to indicate the relative effect that genetic variation contributed to the effect of the phenotype. Denote r_i^j the number of risk variants of *subtype j* that subject i had, so $0 \leq r_i^j \leq 10$. If $r_i^j * e + N(0,1) > 7.5 * e$, we assigned subject i to *subtype j*. In contrast to genotypic subgroups, we named this population partition the phenotypic subgroups. Similarly, we removed from the analysis the subjects that overlapped in the two phenotypic subgroups. Fewer than 15 subjects were excluded in any simulated dataset in the experiment. In addition to the two phenotypic subgroups that had their counterparts in the genotypic view, two additional phenotypic subgroups were created to make the simulated data mimic real situations. The two additional subtypes each included 200 subjects that were randomly selected and assigned to them.

Table I
NMI COMPARISONS BETWEEN DIFFERENT APPROACHES WITH DIFFERENT EFFECTS e

	$e = 1.0$	$e = 0.8$	$e = 0.6$	$e = 0.4$
Biclustering via 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**

After phenotypic subgroups were created, we simulated 10 binary phenotypic features. A subject was assigned a value of 0 or 1 for each of the features according to a probability. *Subtype 1* and *subtype 2* each was associated with three features. Subjects in each simulated phenotypic subgroup obtained the value of 1 with probabilities of 0.6, 0.5, 0.4, respectively for the three designated features. Each of the two additional phenotypic subgroups was associated with two features, and subjects in each of the two subtypes obtained the value of 1 on the two features, with probabilities of 0.6 and 0.5, respectively. A subject obtained the value of 1 with a probability of 0.1 on any other features.

To evaluate how the proposed method performs when the genetic effect on phenotypic variation varies, four phenotypic datasets were generated with $e = 1, 0.8, 0.6, 0.4$ and analysed. Note that the genetic effect on phenotypic variation decreases with decreasing e. Decreased effects lead to a higher level of disagreement between genotypic and phenotypic subgroups.

All of the compared methods were used to obtain three population subgroups. Table I provides the NMI calculated by comparing population subgroups obtained from each approach to true phenotypic subgroups from each dataset. The proposed method has the greatest NMI on all of the four datasets. Along with the decreasing e, NMI obtained by the proposed method decreases gradually as expected, but the population subgroups consistent between the two views can still be uncovered.

For each cluster solution, two classification models were built to separate subjects, in each of the two subgroups from controls. The population subgroup from each method containing the largest number of controls was considered

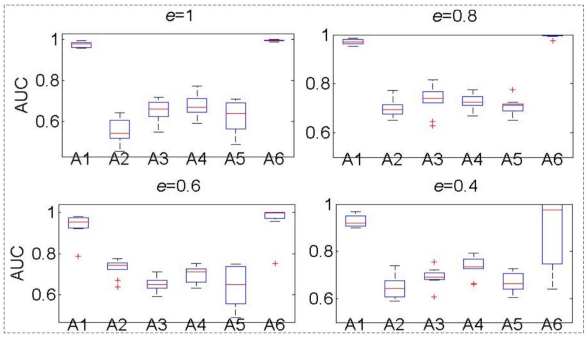

Figure 1. The box plot of AUC values obtained by the proposed method and other comparison approaches. A1 - proposed method, A2 - biclustering via SSVD, A3 - co-regularized spectral, A4 - kernel addition, A5 - kernel product, A6 - feature concatenation

the control group. The average AUC values and their interquantiles obtained from all of the compared approaches on each dataset are plotted in Figure 1. The proposed method achieved the second best performance on this measurement whereas the feature concatenation method performed best. Our further experimental examinations showed that the genetic view had much more features/markers than the phenotypic view, and when concatenating the data from two views to perform cluster analysis, the genotypic view outweighed the phenotypic view. Thus, the resultant clusters were genetically separable but not phenotypically separable. However, our approach created subtypes (population subgroups) that were both genetically and phenotypically separable (or homogeneous) as shown in the NMI comparisons in Table I.

A significant advantage of the proposed method is that features that specify the population subgroups can be simultaneously identified during population partition. 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 II, which shows that our approach correctly identified all true associated features in both views with a very low false discovery rate when taking into account the total number of features used in the analysis.

B. Cocaine use and related behaviors

A total of 1474 subjects were phentyped and genotyped for genetic studies of CD. 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) [16], a polydiagnostic instrument

Table II
THE NUMBER OF FEATURES IDENTIFIED BY THE PROPOSED METHOD TO LINK TO THE TWO POPULATION SUBGROUPS

		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

TF is the number of True Features that specify a population subgroup. TPF (True Positive Features) and FPF (False Positive Features) are the numbers of features that correctly and incorrectly identified, respectively.

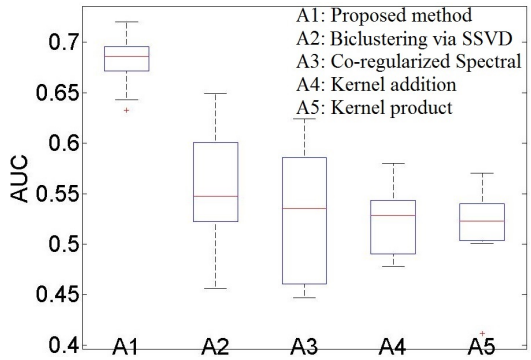

Figure 2. The box plot of AUC values obtained from all comparison methods on the dataset of cocaine use and related behaviors.

that was used to generate diagnoses of cocaine and other substance dependence. Sixty-four yes-or-no variables were generated by this survey, which were also used in previous studies [1], [17]. These variables were used as the phenotypic features. Of the 1474 subjects, 1287 were diagnosed with CD. Subjects were genotyped with 1350 SNPs from 130 candidate genes [4]; 1248 SNPs with minor allele frequency (MAF) of at least 1% were used as genetic markers.

The feature concatenation method overlooked the information in the phenotypic view because the number of features in the genetic view dominated, so clusters were created mainly using genetic information. We hence excluded the concatenation method from the further comparison. Three population subgroups were obtained from each of the other comparison methods. Classification models were built and tested in a manner similar to that used for synthetic data. Figure 2 shows the box plot of the AUC values. Our approach outperformed all other methods.

IV. CONCLUSION

It is challenging to uncover the genetic causes of complex disorders such as substance dependence, due to the

heterogeneity in their clinical manifestation, genetic causes, and environmental/genetic interactions. Phenotype refinement that leads to homogeneous subtypes has been shown to be a promising approach to solve this problem [5], [1], [18], [19], [17]. However, most of the methods used for phenotype refinement take into consideration only the phenotypic information even though genotypic information is usually available in genetic studies of a complex disorder. Thus, these approaches have limited success in finding a phenotypic subtype that is also genetically homogeneous. In this paper, we have proposed a multi-view biclustering approach to perform phenotype refinement by jointly taking into account genetic and phenotypic information. The proposed method is distinct from existing multi-view approaches in that the relevant features can be identified in the determination of a subtype, which is critical to its success. The proposed method is distinct from existing biclustering methods in that it harmonizes the subject groupings in two or more views. The results from extensive experimental comparisons on both synthetic data and real world datasets support the effectiveness and superior performance of the proposed approach.

ACKNOWLEDGMENT

Leah Zindel, MALS, provided assistance in editing the manuscript. This work was supported by NIH grants DA12849, DA12690, AA013736 and NSF grant IIS-1320586.

REFERENCES

[1] H. R. Kranzler, M. Wilcox, R. D. Weiss, K. Brady, V. Hesselbrock, B. Rounsaville, L. Farrer, and J. Gelernter, "The validity of cocaine dependence subtypes," *Addict Behav*, vol. 33, no. 1, pp. 41–53, 2008.

[2] T. F. Babor and R. Caetano, "Subtypes of substance dependence and abuse: implications for diagnostic classification and empirical research," *Addiction (Abingdon, England)*, vol. 101, pp. 104–10, 2006.

[3] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn, "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nature Reviews Genetics*, vol. 9, no. 5, pp. 356–369, 2008.

[4] C. A. Hodgkinson, Q. Yuan, K. Xu, and et al., "Addictions biology: haplotype-based analysis for 130 candidate genes on a single array," *Alcohol Alcohol*, vol. 43, no. 5, pp. 505–15, 2008.

[5] J. Gelernter, C. Panhuysen, M. Wilcox, and et al., "Genomewide linkage scan for opioid dependence and related traits," *Am J Hum Genet*, vol. 78, no. 5, pp. 759–69, 2006.

[6] B. Schwartz, S. Wetzler, A. Swanson, and S. C. Sung, "Subtyping of substance use disorders in a high-risk welfare-to-work sample: a latent class analysis," *Journal of Substance Abuse Treatment*, vol. 38, no. 4, pp. 366–374, 2010.

[7] A. Kumar, P. Rai, and H. D. III, "Co-regularized multi-view spectral clustering," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., 2011, pp. 1413–1421.

[8] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 129–136.

[9] I. Van Mechelen, H.-H. Bock, and P. De Boeck, "Two-mode clustering methods: a structured overview," *Statistical Methods in Medical Research*, vol. 13, no. 5, pp. 363–394, 2004.

[10] M. Lee, H. Shen, J. Z. Huang, and J. S. Marron, "Biclustering via sparse singular value decomposition," *Biometrics*, vol. 66, no. 4, pp. 1087–95, 2010.

[11] Y. Guan, J. Dy, and M. I. Jordan, "A unified probabilistic model for global and local unsupervised feature selection," in *Proceedings of the International Conference on Machine Learning*, 2011, pp. 1073–1080.

[12] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, no. 12, pp. 237 – 260, 1998.

[13] J. Weston, A. Elisseeff, B. Schlkopf, and P. Kaelbling, "Use of the zero-norm with linear models and kernel methods," *Journal of Machine Learning Research*, vol. 3, pp. 1439–1461, 2003.

[14] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[15] G. R. Abecasis, A. Auton, L. D. Brooks, and et al., "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

[16] A. Pierucci-Lagha, J. Gelernter, G. Chan, A. Arias, J. F. Cubells, L. Farrer, and H. R. Kranzler, "Reliability of dsm-iv diagnostic criteria using the semi-structured assessment for drug dependence and alcoholism (ssadda)," *Drug Alcohol Depend*, vol. 91, no. 1, pp. 85–90, 2007.

[17] J. Sun, J. Bi, and H. Kranzler, "A multi-objective program for quantitative subtyping of clinically relevant phenotypes," *BIBM*, 2012.

[18] G. Chan, J. Gelernter, D. Oslin, L. Farrer, and H. R. Kranzler, "Empirically derived subtypes of opioid use and related behaviors," *Addiction*, vol. 106, no. 6, pp. 1146–1154, 2011.

[19] J. Sun, J. Bi, G. Chan, D. Oslin, L. Farrer, J. Gelernter, and H. Kranzler, "Improved methods to identify stable, highly heritable subtypes of opioid use and related behaviors," *Addictive Behaviors*, 2012.