7-7-2015

Machine Learning Approaches for Phenotype Refinement to Improve Genetic Association Analysis

Jiangwen Sun
javon@engr.uconn.edu

Machine Learning Approaches for Phenotype Refinement to Improve Genetic Association Analysis

Jiangwen Sun, Ph.D.

University of Connecticut, 2015

ABSTRACT

Unlike univariate phenotypes such as human height, multivariate phenotypes such as substance use disorders, are characterized by multiple low level phenotypic features. Due to the substantial variation in the multivariate features, these phenotypes are heterogeneous. This phenotypic heterogeneity substantially limits the success in uncovering genetic factors of the phenotype. The identification of homogeneous disease subtypes can be both necessary and beneficial. Despite great progress in molecular genetics that allows the genomewide identification of common and rare variants, there is considerably less progress in the refinement of phenotypes.

The most recent and sophisticated phenotype refinement approaches perform unsupervised cluster analysis to partition a sample population into subgroups based only on the phenotypic features. Since genotypic data are not used to guide the derivation of subtypes, the resultant subtypes may differ only in phenotypic features and thus have limited utility in genetic association analyses. In this thesis study, we propose to refine a multivariate phenotype by simultaneously modeling both phenotypic features and genotypic markers. Two integrative approaches are investigated.

In the first approach, we propose a multi-view cluster analysis to identify clusters of subjects that agree across the two views - phenotypic view and genotypic view. Two different algorithms have been developed along this line. Based on multi-objective

programming, the first algorithm integrates a cluster analysis on phenotypic data and classification on genotypic data by simultaneously optimizing two objectives: (1) the resultant clusters should differ significantly in phenotypic features; (2) these clusters can be well separated using genetic variants via classifiers. Based on sparse matrix decomposition methods, the second algorithm simultaneously decomposes the two data matrices of phenotypic features and genotypic markers into factorized components that share a common structure. This algorithm jointly groups rows (forming subject clusters) and columns (features that determine the subject clusters) of a matrix, and the resultant row groups are consistent across the two matrices.

In the second approach, we propose to use heritability to guide the subtype derivation. Heritability measures the genetic contribution to the variation of a trait, and is commonly estimated from related individuals in pedigrees. The availability of dense genomewide markers allows heritability to be directly estimated from unrelated individuals and their genomewide single nucleotide polymorphisms (SNPs). We have hence developed two algorithms that identify disease subtypes with high heritability. The first algorithm takes family pedigrees as genetic inputs whereas the second takes genomewide SNPs. Both algorithms derive subtypes as a linear combination of phenotypic features and this combination is obtained by maximizing the likelihood of observing a high pedigree-based or SNP-based heritability.

All proposed algorithms were first validated in simulation studies. The validated algorithms were then used in case studies to analyze real-life datasets that were aggregated from genetic studies of drug dependence including opioid and cocaine dependence. These empirical studies demonstrate the superior performance of the proposed approaches over the state of the art.

Machine Learning Approaches for Phenotype Refinement to Improve Genetic Association Analysis

Jiangwen Sun

Master of Engineering, Second Military Medical University, China, 2004

Bachelor of Medicine, Nanjing University, China, 2008

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2015

APPROVAL PAGE

Doctor of Philosophy Dissertation

Machine Learning Approaches for Phenotype Refinement to Improve Genetic Association Analysis

Presented by

Jiangwen Sun, B.M. Clinical Medicine , M.E. Computer Science

Major Advisor _____
Jinbo Bi

Associate Advisor _____
Henry R. Kranzler

Associate Advisor _____
Sanguthevar Rajasekaran

Associate Advisor _____
Dong-Guk Shin

Associate Advisor _____
Yufeng Wu

University of Connecticut

2015

ACKNOWLEDGMENTS

First, my heartfelt gratitude goes to my major advisor, Dr. Jinbo Bi, who I think is one of the best advisors that a graduate student can hope to have. It is really my fortune to be your Ph.D. student. I appreciate your encouragement to take on new opportunities and your motivation by giving me new responsibilities. Without your major guidance, there is no doubt that I would not achieve as much as I have in my doctoral program. You guidance have made me the researcher that I am today. I can only hope to become a researcher of your caliber in the future.

I would like to thank the members of my doctoral committee, Dr. Henry R. Kranzler, Dr. Sanguthevar Rajasekaran, Dr. Dong-Guk Shi and Dr. Yufeng Wu. Dr. Henry R. Kranzler, I appreciate that you taught me the domain knowledge of the real problem involved in my research and how to write good research papers. I hope one day I can write a good research paper as you do. Dr. Sanguthevar Rajasekaran, Dr. Dong-Guk Shi and Dr. Yufeng Wu, your guidance has been one of the most important sources that motivates me to do better research. I also would like to thank Dr. Don Sheehy for the insightful discussion on my research.

Next, my thanks go to my dear friends and colleagues in the health informatics Lab at UCONN: Tingyang Xu, Xin Wang, Jin Lu, Yu Wu and Arun Abraham. Working with you was one of the most enjoyable things in the last five years. I wish you all get nothing, but success in your goals and the best in your life.

I am truly grateful to my wife, Shengnan Zhang, who has been always there for me since my first day in UCONN. Without you, my journey in US would be much harder and totally different. Without your love and support, I would not have a

fruitful doctoral program as I have.

Last, but definitely not the least, I would like to give my deepest gratitude and love to my parents and sisters. Without your love and support, I would not even have the chance to come to US to pursue my dream.

To my family, I dedicate this dissertation.

The work was partially supported by research grants from NSF and NIH: IIS-1320586, DBI-1356655 and R01DA037349.

Contents

Chapter 1

Introduction

1.1 Motivation and Challenges

Complex phenotypes are often characterized by multiple low level phenotypic features, and are thus multivariate. For instance, the substance use disorder is a complex human disease with multivariate phenotypes. In order to diagnose whether a patient has a lifetime drug dependence, clinicians interview the patient to understand the patient's drug use behaviors, the negative consequences of drug use, treatment history and other co-occuring medical conditions. All of these clinical variables are used to arrive at a diagnosis of dependence on a certain drug [59]. In agriculture, breeding programs targeted at conceptual but economically important phenotypes, such as feed efficiency or heat tolerance of animals, are confronted with a wide variety of available multivariate measures [18, 12]. Residual body weight gain, residual feed intake or relative growth rate are feed efficiency measures for dairy cattle [7, 18]. All these measures are multivariate phenotypes that are defined by aggregating many low-level

variables, such as body weight, diet and feed energy intake, and days in milk.

Identifying genetic variation that underlies complex phenotypes has important implications in both biology and medicine. For example, it helps elucidate the biological processes that moderate or regulate a complex disease, such as cancer, heart disease, and substance use disorders, and thus facilitates the development of more effective treatments. To date, over 2,000 genome-wide association studies (GWAS) have been conducted by investigators world wide. These studies have identified a large number of common and rare genetic variants associated with common traits [44]. Despite these successes, a large portion of the trait heritability has not been explained by GWAS. Statistical genetics shows that the success of genetic correlation with a complex phenotype is dependent on two major factors: (1) the availability of comprehensive genetic variants, and (2) the quality of the phenotype. Phenotype refinement can hence play a significant role in enhancing genome-wide associations. Novel statistical and quantitative techniques are needed to refine the phenotypes of complex disorders, which is an important area of genetics research currently underdeveloped.

Case-control analyses are commonly used in the GWAS of complex diseases where the diagnosis itself induces a binary trait that partitions the population into cases (subjects with the disorder) and controls (subjects without). Although the disease indicators used to arrive at the diagnosis vary significantly among the individual cases, the binary trait cannot reflect this heterogeneity. This binary approach cannot distinguish the heterogeneous clinical manifestations that may be attributable to distinct causal effects from genetics and/or environment. Moreover, the diagnosis-induced traits often have low heritability and are suboptimal for genetic association analyses [41]. Hence, GWAS have had limited success in dissecting genetic etiology

of complex diseases. For substance use disorders, very few associations have been identified at the genome-wide significant level that can be replicated [25, 23, 71]. The substantial phenotypic heterogeneity has been one of the main factors limiting the success of GWAS [71].

In the effort of phenotype refinement to identify phenotypes that are suitable for genetic analysis, researchers can perform a simple phenome scan to assess the heritability of individual low-level phenotypic features that are used to characterize the overall complex phenotype. Using individual phenotypic features as traits in genetic association analysis cannot capture the phenotypic heterogeneity either because genetic associations with a single clinical feature may not explain the phenotypic variance in other disease-defining features. This univariate approach also cannot model the interplay between low-level phenotypic features. It cannot answer the question of whether a specific but unknown combination of several features would form a multivariate phenotype that has higher heritability and is more informative in genetic studies.

Multivariate data mining of low-level phenotypic features has seldom been utilized. Prior to the proposed research in this thesis study, the most sophisticated phenotypic refinement methods come from multivariate cluster analysis and latent class analysis that have been mainly used to *subtype* human disease phenotypes [15, 41, 53, 26, 27]. Traditional clustering algorithms such as k-means [61] and hierarchical clustering [60] have been extensively applied to phenotype complex diseases [79, 82, 14, 13]. Many of the studies lack a quantitative and objective measure to validate the clusters. Cluster analysis requires that the subtypes differ significantly on the disease-specific phenotypic parameters that are used.

More recently, heritability was used to assess the validity of the subtype clusters

[27, 15]. Because the power of most gene discovery studies is positively associated with the trait heritability [4], heritability can be a valid target for refinement. Heritability is a key population parameter that helps understand the genetic architecture of a phenotype. The narrow-sense heritability h^2 is defined by the percentage of phenotypic variance that is due to additive genetic effects [32]. The broad-sense heritability H^2 is defined as the overall genetic contribution to the phenotypic variation. The heritability of a quantitative trait is commonly estimated from related individuals in family pedigrees. Recent advances in acquiring dense genome-wide genetic markers have enhanced heritability estimation from apparently unrelated individuals using their genome-wide single nucleotide polymorphisms (SNPs). The SNP-based heritability is defined as the portion of the phenotypic variation that can be explained by the genotyped genetic markers [84]. If a phenotype has higher heritability, it is more genetically influenced, and thus there is greater chance to detect its causative variants. There are widely-used heritability estimation methods, such as sequential oligogenic linkage analysis routines (SOLAR) software [3] that can estimate the heritability of a trait from pedigrees, and the genome-wide complex trait analysis (GCTA) software [84] that can estimate the SNP-based heritability of a quantitative trait.

Figure 1.1.1 shows the flowchart of a common approach [26, 27, 57] for phenotypic subtyping. First, a standard clustering strategy, relying either on a single clustering method or k-means combined consecutively with hierarchical clustering, is applied to the phenotypic data to partition the sample. It assigns each subject to a specific cluster. Then, to form a quantitative trait (or a multivariate phenotype), a classification approach, typically logistic regression, is used to learn a classifier in the objective of separating subjects in different clusters. The learned classifier is a function of phenotypic features involved in previous cluster analysis and gives a membership score to

each subject with respect to a cluster. The classifier is expected to report a higher membership score for subjects in a specific cluster than those who are not. The score computed for each subject is regarded as a quantitative trait characterizing the specific cluster, and its heritability is estimated using software such as SOLAR [3]. This approach is limited by the fact that, although heritability is used to validate the clusters, it is not used in the creation of the clusters. Thus, the resultant quantitative traits are not guaranteed to achieve high heritability.

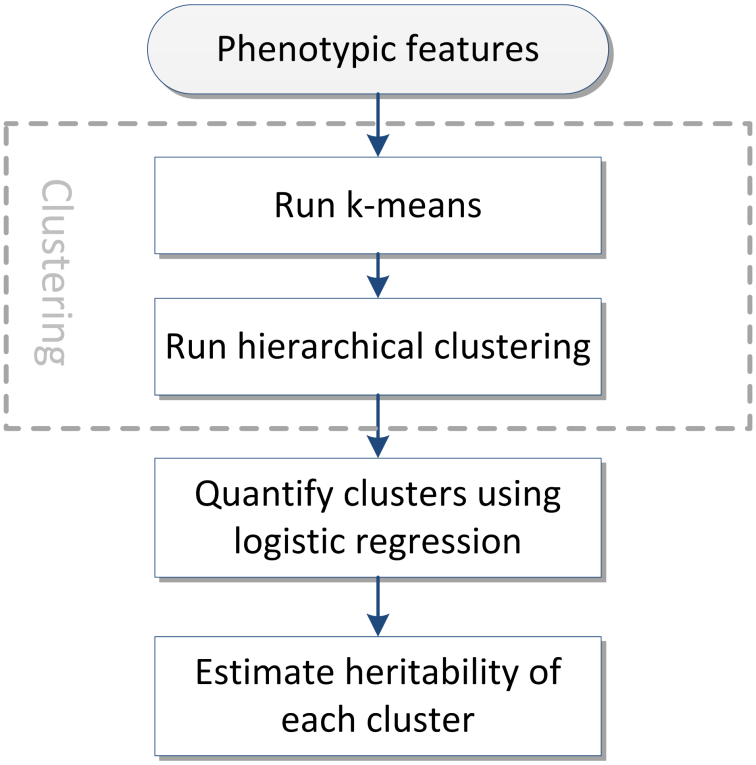

FIGURE 1.1.1: A common approach to phenotypic subtyping.

In our early work [66], an approach was proposed to identify stable and heritable subtypes of opioid use and related behaviors using a three-step sequence: variable selection, clustering, and classification. This approach advanced the subtyping

methodology by assuming that highly-heritable traits can be derived based on the clinical features that are also heritable. In the variable selection step, clinical features were selected based on their estimated heritability and used in cluster analysis. This method resulted in two highly-heritable opioid-use subtypes [66]. However, there are several limitations that may prevent successful applications of this approach to other data sets. First, some of the clinical features are binary traits, such as the response variable to a question of "have you used opiates more than 11 times in your lifetime?" is binary with two possible answers "Yes" or 'No". It is not straightforward to estimate the heritability of a binary trait. Second, it is unclear that simply combining highly heritable clinical features will necessarily lead to traits that are more heritable. Third, similar to the standard clustering approach reviewed above, heritability was not used directly in the clustering process.

Hence, in this dissertation work, we propose and investigate a series of new statistical models and algorithms that aim at identifying more genetically influenced traits and thus more suitable for use in genetic studies. Our approach consists of a set of integrative analytics that jointly models phenotypic features and genetic markers. The phenotypic features include clinical symptoms and disease indicators used in a diagnosis. The genetic inputs can either be family pedigrees of the study samples or their genetic variants (or genotyped markers).

1.2 Overview

Our approach aims to derive multivariate analytic algorithms based on quantitative genetics theory [4], statistical learning theory [31, 74] and theories of substance use

disorders that emphasize the multifaceted nature of substance use and related behaviors [5, 50, 80], and then use these new theory-driven algorithms to analyze empirical data. For instance, the proposed statistical methods to identify highly heritable traits will be derived from heritability estimation [21], and the joint analysis of clinical and genomic data will be based on multi-view data analytics [11, 42]. Figure 1.2.1 depicts an overview of the proposed techniques and algorithms.

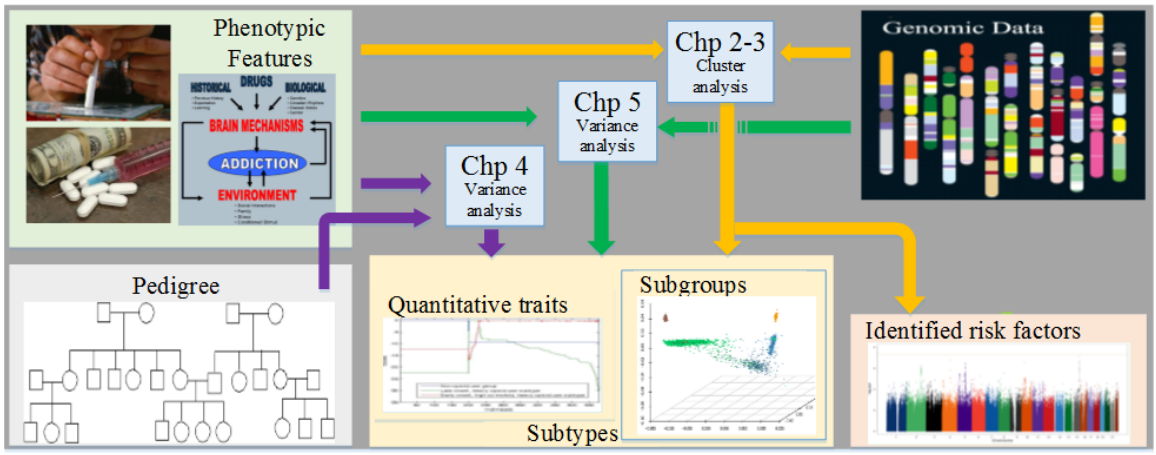

FIGURE 1.2.1: Overview of the proposed research.

First, we propose multi-view clustering methodologies to identify a grouping of subjects that agrees when subjects are clustered based on, respectively, their phenotypic data and genotypic data. Two algorithms have been developed for different purposes. The first one is used to create clusters and simultaneously quantify the predictive power of the genetic markers for the identified clusters. The second algorithm focuses on the identification of phenotypic clusters that can be associated with certain genetic markers, but will not be able to quantify the predictive power of the genetic markers for the clusters.

Based on multi-objective programming, the first algorithm is capable of clinically

categorizing a disease phenotype so as to discover genetically different subtypes. This method performs two tasks: a cluster analysis in the phenotypic view and classification in the genotypic view by jointly optimizing two objectives. One objective requires that the derived clusters differ significantly on clinical features. The other objective requires that these clusters can be well separated using genetic markers by classifiers. The classifiers are constructed as a function of genetic markers to separate subjects in the different clinical clusters. Extensive experiments have been conducted for two substance use disorders: opioid and cocaine dependence, using two populations: African Americans and European Americans. The experimental results demonstrate that the proposed method is superior to existing subtyping methods.

Based on multi-view matrix decomposition, the second algorithm integrates clinical features with genetic markers to detect subtypes with confirmatory evidence in both data sources. This approach groups subjects into clusters that are consistent between the phenotypic and genotypic dimensions of data, and also simultaneously finds the clinical features that define the cluster (subtype) and the genetic variants that are associated with the subtype. A simulation study validates that the proposed approach indeed identifies hypothesized subtypes and their associated features. The comparison with the latest biclustering and multi-view clustering methods on our real-life disease data shows that the proposed approach can identify genetically more separable *clinical subtypes* of a disease, thus demonstrating the superior performance of the proposed approach.

Second, we propose to derive subtypes or subphenotypes of a complex disease that are optimized with respect to heritability because heritability is a good indicator for the utility of the subtypes as traits in genetic association analysis. The resultant subphenotypes, characterized by quantitative traits, will achieve high heritability as

measured either by the traditional narrow-sense heritability or the recent SNP-based heritability. Two different such methods have been developed to accomodate the different types of genetic data - pedigrees or genetic markers.

In the first method, we consider family pedigrees of a study sample as the genetic input. This method searches for the linear combinations of phenotypic features that can maximize the heritability estimate of the resultant quantitative traits. A quadratic optimization problem is formulated by decomposing the traditional maximum likelihood method for estimating heritability of a quantitative trait. An efficient algorithm is developed to solve the proposed optimization problem following the framework of sequential quadratic programming (SQP). We have further extended our formulation to model covariates, such as age, sex and race, which can then identify subtypes of a complex disease that have high heritability even after correction for fixed effects of the covariates. We demonstrate the effectiveness of this method on synthetic data as well as in the analysis of real-world data. We applied our algorithm to identify highly heritable traits of substance use disorders including opioid and cocaine dependence. Our approach outperformed standard cluster analysis and several previous methods.

Due to the rapid advances in genotyping technologies, the availability of genome-wide genetic markers becomes much greater for genetic studies of multivariate phenotypes, such as substance use disorders. On the other hand, it is commonly difficult to recruit related family members in a genetic study of such a disorder. The majority of existing datasets contains unrelated individuals. The genetic relationship of unrelated individuals can be better estimated from the genome-wide sample of SNPs. Hence, our first method based on related individuals of pedigrees is not readily applicable to these datasets. In the second method, we take genome-wide genetic variants as the

genetic input when maximizing the heritability of a derived subtype. This method also searches for linear combinations of phenotypic features. The difference is that the resultant linearly-combined trait maximizes the SNP-based heritability (also referred to as the chip heritability). The objective function of this optimization problem is formulated by decomposing the restricted maximum likelihood method that is used to estimate chip heritability. We also take into account the covariate effects so that the derived traits will still have high chip heritability after correcting covariate effects. An efficient SQP based algorithm has also been developed to solve this optimization problem. Extensive empirical studies were conducted, demonstrating the effectiveness of this method as well.

Chapter 2

Multi-view Co-modeling to Improve Subtyping and Genetic Association of Complex Diseases

2.1 Introduction

In a subtyping study, an objective function may be used to evaluate how strongly the subtypes derived from the grouping are associated with a given set of genetic markers, or how well the subtypes can be separated by the genetic markers. Mathematically, given two sets of variables, clinical features Z and genetic markers X from the same sample, the goal is to partition the sample into subgroups based on pairwise similarities between subjects in Z so that the resultant subgroups y can be classified by X. This problem is different from traditional supervised or unsupervised machine learning problems where labels of subjects are either given precisely or not given at all. In our problem, the labels of subjects need to be derived from the clinical features

Z so they can be used to train a classifier with the genetic data X.

In the machine learning literature, the most related work might be the set of multi-view data analysis methods, co-training methods [11] and co-clustering methods [42] where multiple groups of input variables are collected for the same set of subjects. When only a small portion of the data is labeled, co-training improves the classification accuracy by enforcing consistency between the classification decisions of the unlabeled data determined by the models learned independently from each of the views. Nevertheless, co-training methods are not applicable to the subtyping problem because there are no labeled data to start with. Multi-view clustering methods seek groupings of subjects that are consistent across different views. These methods treat the data from the two views equally as the input variables. In the subtyping problem, however, the two views have to be treated differently in that one is used to define the subtypes y and the other is used to explain them. For instance, only a sparse set of genetic risk markers are identified to be associated with a subtype but the subtypes may be defined using many clinical features.

2.2 Proposed Methodology

We propose a multi-objective optimization framework to solve the subtyping problem. For a set of cluster labels y, each assigned to one subject, we construct a model as a function of a subject's genetic markers X to approximate the subject's label. The model M is built by minimizing a loss function $\ell(y, X|M_\theta)$ where M_θ is a specific inference model, such as the model of support vector machine (SVM), or logistic regression, and θ denotes the set of its parameters. Since the labels y of subjects are

not given beforehand, the labels themselves need to be derived. In other words, we optimize the objective as follows

$$\min_{y,\theta} \ell(y, X|M_\theta) + \lambda R(M_\theta) \tag{2.2.1}$$

for the best y and θ where $R(M_\theta)$ defines the regularization term that controls the complexity of the model M, and λ is a tuning factor to balance between ℓ and R. Notice that not every possible labeling y of subjects is a feasible solution of Problem (2.2.1). The search space of y is confined by the similarity measure defined on the features Z.

Suppose that the classification of subjects y is obtained by partitioning subjects based on a similarity measure that is pre-specified on Z. The parameters used in the similarity measure often need to be tuned, such as the parameter σ if a Gaussian similarity $\exp(-||Z_i - Z_j||^2/\sigma^2)$ is used where Z_i and Z_j are the two vectors of clinical features for Subjects i and j. Choosing different values of σ or other relevant parameters will produce different clusters of the subjects. In general, we expect that the resultant clusters will be well differentiated from each other and that subjects in the same cluster will be closer than those from other clusters in the Z space. Many metrics have been derived in the literature to measure the quality of clusters, such as the Dunn's Validity Index [20], Davies-Bouldin Validity Index [19], and Silhouette Validation [62]. If a metric $\epsilon(y|\sigma, Z)$ is employed to measure the quality of clusters when using a specific value of σ, the metric corresponds to another objective of the

subtyping problem. We hence optimize simultaneously two objectives as follows

$$\min_{y,\theta,\sigma} \begin{cases} Obj_1: & \epsilon(y|\sigma, Z) \\ Obj_2: & \ell(y, X|M_\theta) + \lambda R(M_\theta). \end{cases} \tag{2.2.2}$$

We assume that $\epsilon(y|\sigma, Z)$ is a metric to be minimized, or otherwise it can be inverted or negated. The two objectives of Problem (2.2.2) may not be optimized at the same solution. Thus, it formulates a multi-objective optimization problem.

Multi-objective programming (MOP) is a technique that was developed to solve optimization problems with multiple conflicting objectives. Solving a multi-objective program requires the search for Pareto-optimal solutions [8]. Traditional methods convert multiple objectives into a single objective using certain schemes and user-specified parameters. Two simple and widely used methods for such conversions are the weighted sum method and the constraint method [8]. The weighted sum method transforms two objectives into a single objective by multiplying each objective with a pre-defined weight and adding them together as follows

$$\min \quad c_1 Obj_1 + c_2 Obj_2 \tag{2.2.3}$$

where the weights c_1 and c_2 are non-negative and at least one of them is not zero. If the MOP is not convex, the non-convex frontier of the Pareto-optimal set cannot be obtained by the weighted sum method. The constraint method reformulates the MOP by keeping one of the objectives and restricting the rest of the objectives within user-specified limits, such as,

$$\min \quad Obj_2, \text{ subject to: } Obj_1 \leq \delta. \tag{2.2.4}$$

Our MOP-based subtyping framework follows the constraint method, and can be implemented using any proper cluster analysis algorithm to optimize Obj_1, and any suitable classification algorithm to optimize Obj_2. In the next section, we will instantiate this methodology by utilizing a spectral clustering method [48] and the one-norm SVM [86] in the MOP.

2.3 Multi-objective Optimization Formulation

A spectral clustering method [48] is employed to search for the cluster assignments of subjects by varying the parameter σ in its Gaussian similarity measure. The Davies-Bouldin Validity Index [19], measuring how significantly the resultant clusters differ from each other, serves as Obj_1. The one-norm SVM [86] is used to build a classifier, as a function of the genetic variables X, that separates subjects in different clusters. The loss function used in the one-norm SVM serves as Obj_2. Notice that the framework (2.2.2) can be realized in conjunction with other choices of clustering and classification methods.

2.3.1 The Clustering Algorithm

Spectral clustering is a method based on undirected similarity graph $G = (V, E)$ in which each node in V represents a data point (a subject) and each edge in E is weighted by the similarity between the two connected data points. Partitions of data points represented in the similarity graph can be obtained by cutting the graph into unconnected components with the minimum cost. In a balanced cut, the sizes of these unconnected components should be comparable. Two methods have been proposed

to achieve this kind of balanced cut, RatioCut [30] and Ncut [64], that minimize the following objectives, respectively,

$$\begin{aligned} RatioCut(C_1,...,C_k) &:= \tfrac{1}{2}\sum_{i=1}^{k}\tfrac{A(C_i,\bar{C}_i)}{|C_i|} = Tr(H^T L H) \\ Ncut(C_1,...,C_k) &:= \tfrac{1}{2}\sum_{i=1}^{k}\tfrac{A(C_i,\bar{C}_i)}{vol(C_i)} \\ &= Tr(T^T D^{-1/2} L D^{-1/2} T) \end{aligned} \tag{2.3.1}$$

where C_i is one of the identified components (clusters), $|C_i|$ and $vol(C_i)$ denote the number of nodes and the sum of edge weights in C_i respectively, and $\bar{C}_i$ consists of the nodes that are not in C_i. The matrix $A = \{a_{ij}\}$ is the adjacency matrix and a_{ij} measures the similarity between the nodes i and j, D is a diagonal matrix whose i^{th} diagonal element $d_{ii} = \sum_{j:j\neq i} a_{ij}$, L is the graph Laplacian defined by $L = D - A$, $Tr(\cdot)$ means the trace norm, and both H and T are matrixes consisting of indicator vectors as columns defined as follows:

$$\begin{aligned} H &= [\tfrac{1}{\sqrt{|C_1|}}\mathbb{1}_1, ..., \tfrac{1}{\sqrt{|C_k|}}\mathbb{1}_k] \\ T &= D^{1/2}[\tfrac{1}{\sqrt{vol(C_1)}}\mathbb{1}_1, ..., \tfrac{1}{\sqrt{vol(C_k)}}\mathbb{1}_k] \end{aligned} \tag{2.3.2}$$

where $\mathbb{1}_i$ is an indicator vector whose entries equal 1 if the corresponding nodes are in C_i, or 0 otherwise. Finding the global optimal solution to either of these two objectives is NP hard [76]. Their relaxed versions have been defined by allowing the indicator vectors in H and T to take real values. It has been shown that the optimal solutions to the relaxed problems of RatioCut and Ncut are the matrices composed by the eigenvectors corresponding to the first k smallest eigenvalues of L and $D^{-1/2}LD^{-1/2}$, respectively [48].

In spectral clustering, the clusters are determined by the adjacency matrix A

which is further determined by a pre-chosen similarity measure. Spectral clustering is sensitive to changes in the similarity measure [48]. In our approach, we search for the most suitable similarity measure, more precisely, the best value of σ in the Gaussian similarity, to optimize Obj_1 and Obj_2.

2.3.2 The Objectives in Our Multi-objective Program

(1) First Objective. Spectral clustering requires an adjacency matrix A that encodes the pairwise similarities between subjects and the desired number of clusters k as its inputs, and outputs the clusters C_i of subjects, $i = 1, \cdots, k$. In our approach, we search for the best value of σ in the Gaussian similarity measure to optimize the Davies-Bouldin Validity Index (DBVI) [19] that measures the quality of the clusters. DBVI is a measure related to the ratio of within-cluster distance to between-cluster distance. The lower value of DBVI indicates better quality of the clusters. Hence, we minimize the DBVI as follows using Ncut [64] for the best σ

$$\min_{\sigma} \quad \text{DBVI} = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} \frac{Dist(C_i) + Dist(C_j)}{Dist(C_i, C_j)} \tag{2.3.3}$$

where $Dist(C_i)$ is the average distance from each data point in C_i to the cluster center, $Dist(C_i, C_j)$ is the distance between the center of C_i and the center of C_j. These distances are calculated in the Z dimension.

(2) Second Objective. For each cluster C_i, without loss of generality, we construct a classifier in the linear form of $f(X) = W^T X + b$ to separate the subjects in C_i from the remaining subjects. The model $W_i^T X + b_i$ specific for Cluster C_i is obtained by minimizing the regularized empirical error $\ell(y_i, X, W_i) + \lambda R(W_i)$ where we use

a binary vector y_i to indicate the cluster membership: $y_i^j = 1$ if subject X_j is in C_i, or otherwise $y_i^j = -1$, $j = 1, \cdots, n$, for all n subjects. We employ the hinge loss commonly used in SVMs, e.g., $\ell(y_i, X, W_i) = \sum_{j=1}^{n} \left[1 - y_i^j \left(W_i^T X_j + b_i\right)\right]_+$ where $[a]_+ = 0$ if $a < 0$, otherwise $[a]_+ = a$, and $R(W_i)$ takes a sparse-favoring form in order to select among features, in particular, ℓ_1-norm $||W_i||_1 = \sum_d |W_{id}|$. The ℓ_1-norm shrinks the coefficients W of irrelevant variables to zero [86]. Constructing all of the k classifiers together corresponds to minimizing the overall regularized error as follows:

$$\min_{W_i, b_i, i=1,\cdots,k} \sum_{i=1}^{k} \left[\ell(y_i, X, W_i) + \lambda R(W_i)\right] \tag{2.3.4}$$

(3) Constrained Conversion. Clearly, the first objective is not convex, which leads to a non-convex multi-objective program. The constraint conversion method is more suitable to find the Pareto-optimal solutions to this problem. As the subtyping problem seeks to obtain clusters that are interpretable in the X dimension (genetic markers), we model the first objective as a constraint. In other words, we search for solutions that minimize the second objective subject to an acceptable quality of clusters in the Z dimension (clinical features). The following problem (2.3.5) is the problem we will solve.

$$\begin{aligned} \min_{\substack{\sigma, W_i, b_i \\ i=1,\cdots,k}} \quad & \sum_{i=1}^{k} \left(\sum_{j=1}^{n} \left[1 - y_i^j \left(W_i^T X_j + b_i\right)\right]_+ + \lambda ||W_i||_1 \right) \\ \text{subject to} \quad & \tfrac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \frac{Dist(C_i) + Dist(C_j)}{Dist(C_i, C_j)} \leq \delta \\ & l_\sigma \leq \sigma \leq u_\sigma \end{aligned} \tag{2.3.5}$$

where δ, l_σ and u_σ are tuning parameters to bound σ.

2.3.3 The Proposed Algorithm

Traditional methods for finding the optimal solution to a constrained optimization problem include deterministic approaches such as gradient-based methods, Newton's methods, and non-deterministic approaches such as simulated annealing [39]. To avoid the difficulty of computing derivatives of the objective function, we design an efficient algorithm based on simulated annealing to solve the converted MOP (2.3.5) as depicted in Algorithm 1.

Algorithm 1 Simulated Annealing for MOP (2.3.5)

Input: Z, X, k, δ, M_I
Initialize: σ, T, $h = 0$;
for $t = 0$ **to** M_I **do**
  Calculate Temperature T;
  Find a neighbor of σ to obtain σ_{new} based on T;
  Construct adjacency matrix A using Z and the Gaussian similarity with σ_{new};
  Obtain clusters $C_i, i = 1, \cdots, k$, by running Ncut with A and k;
  Calculate Obj_1 in (2.3.3) and assign its value to q;
  if $q \leq \delta$ **then**
    Compute W_i, b_i for each C_i separately by the one-norm SVM;
    Calculate Obj_2 in (2.3.4) and assign its value to h_{new};
  else
    Continue;
  end if
  if $probability(h, h_{new}, T) > random(0, 1)$ **then**
    $h = h_{new}$, $\sigma = \sigma_{new}$;
  end if
end for
Output: clusters $C_{i:1,\ldots,k}$, the values of Obj_1 and Obj_2.

In Algorithm 1, the temperature T starts from a high value, and decreases gradually at each iteration. A probability density function defined according to T is used to search for σ_{new}. The first objective is evaluated after the clusters are obtained. If this objective is within the pre-specified limit δ, an SVM model is constructed for

each cluster, and the second objective is evaluated. The probability of accepting σ_{new} is calculated via the acceptance probability density function discussed in [37] and defined by the objective values h, h_{new} and the temperature T. If this probability is larger than a number randomly drawn from $[0, 1]$, σ_{new} is accepted; or otherwise the previous value of σ is retained. Readers can consult with [37] for more discussions on simulated annealing.

2.4 Computational Results

We applied the proposed algorithm to two real-world data sets that were aggregated from genetic studies of opioid dependence (OD) and cocaine dependence (CD) [15, 66, 41, 24]. We limited the analysis to European Americans to avoid confounding by population differences in allele frequencies and structure. We compared our approach to an existing subtyping method that performed a sequence of two separate steps: spectral clustering and one-norm SVM classification in the same fashion as in [15]. We refer to this as the sequential subtyping method. The two approaches were compared in terms of the separability of their resultant clusters based on genetic markers.

2.4.1 Data sets

Subjects were recruited from multiple sites, including 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.

TABLE 2.4.1: Summary of the OD and CD data sets

Dataset	#cases	#controls	#Vars	#MCA Dims	#SNPs
opioid	827	643	69	13	1185
cocaine	1279	187	68	25	1248

Opioid use and cocaine use behaviors were assessed by two separate components dedicated to the diagnosis of OD and CD respectively in a computer-assisted interview process, called the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA) [59]. The SSADDA variables selected by previous OD and CD subtyping studies [41, 15] were used in the present analysis. Multiple Correspondence Analysis (MCA) [51] was performed to reduce data. The top MCA dimensions that overall explained more than 80% of data variance were used in cluster analysis.

A total of 1350 single nucleotide polymorphisms (SNPs) selected from 130 candidate genes were genotyped for association tests [34]. For each dataset, we performed quality control as follows. 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 further analysis. The minor allele frequency (MAF) of each SNP was calculated within each population. SNPs with MAF less than 0.5% in a population were removed from the association tests for the respective population. The remaining missing entries in the SNP data were imputed.

For the OD data set, we treat opioid users as cases and healthy subjects as controls. For the CD data set, subjects who were diagnosed with cocaine dependence were treated as cases and healthy subjects who had been exposed to illicit drugs were regarded as controls. Table 2.4.1 summarizes the statistics of the two data sets in terms of the numbers of cases, controls, SSADDA variables (Vars), MCA dimensions (Dims), and SNPs used in the subtyping analysis.

2.4.2 Experimental settings

We utilized the CPLEX optimization package to solve the one-norm SVM, and implemented spectral clustering in MATLAB. Adaptive simulated annealing, an open source variant of simulated annealing, together with its MATLAB gateway (ASAMIN v1.39) was used to search for the value of σ that optimizes the multi-objective program (2.3.5). The parameters δ, λ were set to 0.7 and 0.08 respectively. The upper bound of σ, u_σ was set to a number that led to a pairwise similarity value of at least 0.99, and the lower bound of σ, l_σ was set to the value producing a similarity matrix of the median value less than 0.0001. These tuning steps were based on 3-fold cross validation.

A typical way to choose a value for σ is to use the median value of all entries in the pairwise distance matrix [48]. We fixed σ to the median value in the sequential method. For both the proposed and sequential methods, cluster analysis was only applied to cases. The resultant clusters were characterized based on important clinical features related to drug use and related behaviors. A generalized estimating equation (GEE) Wald Type 3 χ^2-test was employed to test the significance of the difference between the resultant clusters in these clinical variables with Bonferroni correction for multiple comparisons.

For each obtained cluster, an SVM model was built to separate cases in the cluster labeled as $+1$ from controls labeled as -1. SVM is sensitive to unbalanced data where the size of a sample with one label is significantly larger than that with another label. To address this problem, we duplicated subjects in the smaller group to make the sample size of the two groups comparable. Let a and b be the dominating and minor groups, respectively, n_a and n_b be their sample sizes, and $t = \lfloor n_a/n_b \rfloor$. We first

duplicated each subject labeled by b t times, and then randomly selected $n_a - t * n_b$ subjects from the sample pool composed by all subjects with label b. Ten-fold cross validation with stratified case-control split was conducted for every cluster, and receiver operating characteristic (ROC) curves were obtained using the test data combined from all folds to evaluate the classification performance. We provide the Area Under the ROC Curve (AUC) in our results to compare the two methods. The AUC reflects the cluster separability based on genetic markers.

Moreover, different analytic approaches, such as SVM, or logistic regression, may identify important SNPs of different associative effects. A larger coefficient for a SNP in the SVM models does not necessarily translate into a smaller p-value in logistic regression. We further tested each of the selected SNPs, i.e., those with none zero coefficients in the SVM models, by a separate logistic regression and evaluated their corresponding p-values to determine the significance of the association with the identified subtypes. Here, logistic regression models were obtained in the similar sampling scheme introduced early on to balance the data.

2.4.3 Opioid use subtypes

We set the desired number of clusters to 2, so that the resultant clusters were sufficiently large and gave adequate statistical power. The optimal value of σ found by our approach was 5.8.

Cluster Clinical Characteristics

We characterized the two clusters obtained with $\sigma = 5.8$ based on 11 important clinical variables depicting opioid use and its consequences. Table 2.4.2 shows that

TABLE 2.4.2: Clinical Opioid-related Characteristics of opioid user clusters [N(%)]

Behaviors	**Cluster 1** 657(79.44)	**Cluster 2** 170(20.56)	χ^2**(df)**	p**-value**
Age of first use [Mean (SD) in year]	21.15(6.59)	21.67(7.71)	0.58	0.45
Used opioids daily or almost daily	653(99.39)	107(62.94)	65.48	5.55×10^{-16}
Injected opioids intravenously	526(80.06)	50(29.41)	134.40	$< 1 \times 10^{-16}$
Stayed high from opioids for a whole day or more	599(91.17)	103(60.59)	78.05	$< 1 \times 10^{-16}$
Strong desire for opioids made it hard to think of anything else	617(93.91)	50(29.41)	245.63	$< 1 \times 10^{-16}$
Opioid use interfered with work, school, or home life	574(87.37)	39(22.94)	201.13	$< 1 \times 10^{-16}$
Family members, friends, doctor, clergy, boss, or people at work or school objected to opioid use	611(93.00)	52(30.59)	187.13	$< 1 \times 10^{-16}$
Been arrested or had trouble with the police because of opioid use	444(67.58)	23(13.53)	114.34	$< 1 \times 10^{-16}$
Give up or greatly reduced important activities due to opioid use	600(91.32)	48(28.24)	212.67	$< 1 \times 10^{-16}$
Ever treated for an opioid-related problem	610(92.85)	37(21.76)	260.89	$< 1 \times 10^{-16}$
Ever attended self-help group for opioid use	505(76.86)	23(13.53)	141.76	$< 1 \times 10^{-16}$

TABLE 2.4.3: Risk factors (SNPs) associated with opioid-use subtypes

	SNP	***p*-value**	**Odds Ratio**	**Gene**
Cluster 1	rs915906	5.32×10^{-5}	0.6595	*CYP2E1*
	rs10896065	3.32×10^{-4}	2.0537	*FOSL1*
	rs7940700	4.15×10^{-4}	2.2496	*FOSL1*
	rs755203	5.18×10^{-4}	0.7617	*CHRNA4*
	rs2581206	5.56×10^{-4}	0.7594	*SLC6A11*
	rs698	5.59×10^{-4}	0.7615	*ADH1C*
	rs4077851	7.69×10^{-4}	1.5542	*GABRB2*
	rs2515642	8.02×10^{-4}	0.7294	*CYP2E1*
Cluster 2	rs6957496	1.09×10^{-5}	2.25	*CHRM2*

the two clusters differ significantly on almost all of these clinical features, except the mean age of first opioid use. Subjects in Cluster 1 have used opioids more heavily than those in Cluster 2. For example, they had heavier daily use and more intravenous injections. The negative consequences of opioid use, such as "interfering with work" and "been arrested" among subjects in Cluster 1 were much more severe than those for subjects in Cluster 2. Thus, Cluster 1 was a heavy opioid user group whereas Cluster 2 was composed of moderate opioid users.

Associated Genetic Markers

Eight SNPs were associated with Cluster 1 at $p < 1 \times 10^{-3}$ as shown in Table 2.4.3. A SNP (rs915906) was very close to the empirical threshold ($p < 0.05/1154 = 4.34 \times 10^{-5}$) after Bonferroni correction was applied to address the inflation of type I error due to multiple tests. For Cluster 2, SNP rs6957496 on gene *CHRM2* was significant with a p-value close to 10^{-5}, and it remained significant after Bonferroni correction (empirical threshold: $p < 0.05/1154 = 4.34 \times 10^{-5}$). Odds ratios and the genes where the corresponding SNPs are located are also shown in Table 2.4.3.

TABLE 2.4.4: Comparison on genetic separability of opioid user clusters

	Optimal $\sigma = 5.8$		$\sigma = 6.07$	
	N(%)	AUC	N(%)	AUC
Cluster 1	657(79.4)	0.59	600(72.6)	0.50
Cluster 2	170(20.6)	0.85	227(27.4)	0.80

Comparison

For the sequential method, we followed the standard approach to selecting σ for spectral clustering [48] and computed the median value of the pairwise distances, which was 1.07. When $\sigma = 1.07$, a very unbalanced partition was resulted: 826 in one cluster and 1 in the other, which was not of practical value. In order to find a σ value that gives clusters of similar size, we increased the value of σ several times, and each time by 1 until a proper σ was found. The final value was 6.07. Two clusters were built to separate each subtype from controls. The AUC results were compared to evaluate the cluster separability in the genetic view as shown in Table 2.4.4. Genetic markers had better predictive power for those clusters obtained by the proposed approach than the sequential method with a larger supporting sample size. More significant associations were found for the clusters created by the proposed method. Thus it demonstrates the effectiveness of the proposed method.

2.4.4 Cocaine use subtypes

Since a large number of cases were available, we set the desired number of clusters to 3. The optimal value of σ found by our approach here was 1.76.

TABLE 2.4.5: Clinical cocaine-related Characteristics of cocaine user clusters [N(%)]

Behaviors	**Cluster 1** 340(33.11)	**Cluster 2** 328(31.94)	**Cluster 3** 359(34.96)	χ^2**(df)**	p**-value**
Age of first cocaine use [Mean (SD) in year]	17.61(4.13)	19.53(5.16)	21.28(6.22)	79.50(2)	$< 1 \times 10^{-16}$
Age of onset of heaviest cocaine use [Mean (SD) in year]	25.95(8.09)	25.82(8.12)	29.47(7.70)	44.48(2)	2.19×10^{-10}
Used cocaine daily or almost daily	329(96.76)	251(76.52)	340(94.71)	73.32(2)	1.11×10^{-16}
Injected cocaine intravenously	311(91.47)	132(40.24)	33(9.19)	298.77(2)	$< 1 \times 10^{-16}$
Stayed high from cocaine for a whole day or more	304(89.41)	210(64.02)	327(91.09)	83.49(2)	$< 1 \times 10^{-16}$
Strong desire for cocaine made it hard to think of anything else	308(90.59)	176(53.66)	332(92.48)	162.45(2)	$< 1 \times 10^{-16}$
Cocaine interfered with work, school, or home life	312(91.76)	139(42.38)	311(86.63)	198.06(2)	$< 1 \times 10^{-16}$
Family members, friends, doctor, clergy, boss, or people at work or school objected to cocaine use	310(91.18)	173(52.74)	324(90.25)	159.72(2)	$< 1 \times 10^{-16}$
Been arrested or had trouble with the police because of cocaine use	223(65.59)	69(21.04)	175(48.75)	127.35(2)	$< 1 \times 10^{-16}$
Give up or greatly reduced important activities due to cocaine use	321(94.41)	179(54.57)	340(94.71)	177.31(2)	$< 1 \times 10^{-16}$
Ever treated for an cocaine-related problem	264(77.65)	91(27.74)	249(69.36)	178.74(2)	$< 1 \times 10^{-16}$
Ever attended self-help group for cocaine use	250(73.53)	89(27.13)	227(63.23)	139.27(2)	$< 1 \times 10^{-16}$

Cluster Clinical Characteristics

The three clusters obtained with $\sigma = 1.76$ were characterized in Table 2.4.5 based on 12 important features related to cocaine use and its consequences. Table 2.4.5 shows that the three clusters differ significantly on all the 12 clinical features. Both Clusters 1 and 3 were heavy cocaine user groups compared to Cluster 2 as indicated by almost all of the features. For example, 96.76% and 94.71% of the subjects in Clusters 1 and 3, respectively, ever used cocaine daily or almost daily in comparison with only 76.52% of the subjects in Cluster 2. Even though Clusters 1 and 3 were both heavy user groups, they were distinct on several features, especially on the age of onset and on cocaine intravenous injection rates. Subjects in Cluster 1 started the initial and heavy use of cocaine at much younger age than those in Cluster 3. Cluster 1 had a high portion of subjects (91.47%) who had injected cocaine intravenously in contrast to a much lower rate of that (9.19%) in Cluster 3.

Associated Genetic Markers

The results from association tests for the three clusters are provided in Table 2.4.6, in which only those SNPs with tested p values less than 1×10^{-3} are shown. SNP rs3802280 on gene *OPRK1* was associated with Cluster 1 at $p < 1 \times 10^{-3}$. Four SNPs were identified to be nominally associated with Cluster 3 at $p < 1 \times 10^{-3}$. None of the SNPs was identified to be associated with Cluster 2 at $p < 1 \times 10^{-3}$.

Comparison

For the CD data, the median value of the pairwise distances was 1.45, which was used as the value of σ in the sequential method. We ran spectral clustering based

TABLE 2.4.6: Risk factors (SNPs) associated with cocaine-use subtypes

	SNP	***p*-value**	**Odds Ratio**	**Gene**
Cluster 1	rs3802280	7.98×10^{-4}	1.8265	*OPRK1*
Cluster 3	rs511895	3.03×10^{-4}	0.6456	*CAT*
	rs722651	4.95×10^{-4}	1.5062	*MPDZ*
	rs7940700	5.87×10^{-4}	0.6585	*CAT*
	rs494024	6.22×10^{-4}	0.6602	*CAT*

on the similarity matrix and also obtained three clusters. We compared these three clusters against those obtained by our approach in terms of the cluster separability based on genetic data. We built three classifiers, each used to separate subjects in one cluster from the controls. We averaged the AUC of the three classifiers with standard deviation over the 10-fold cross validation. A box plot was drawn for each method as shown in Figure 2.4.1. As shown in Figure 2.4.1, classifiers trained on the clusters obtained by the proposed method have a slightly better average AUC value (i.e., separability) and significantly smaller error bar than those obtained on the clusters from the sequential method, which implicates that the proposed approach is better in terms of finding genetically-separable clinical clusters than the existing approach.

2.5 Summary

Identifying genes that contribute to risk of complex diseases has been challenging due to two major issues: (1) The diseases have diverse clinical manifestations and complex etiology with both genetic and environmental risk factors. (2) Disease phenotypes are heterogeneous and homogeneous subtypes have not been optimized empirically. To address these issues, researchers have sought to leverage the technology of cluster analysis to identify clinically homogeneous subtypes that correlate to homogeneous genetic risk factors. Although encouraging results have been obtained, success re-

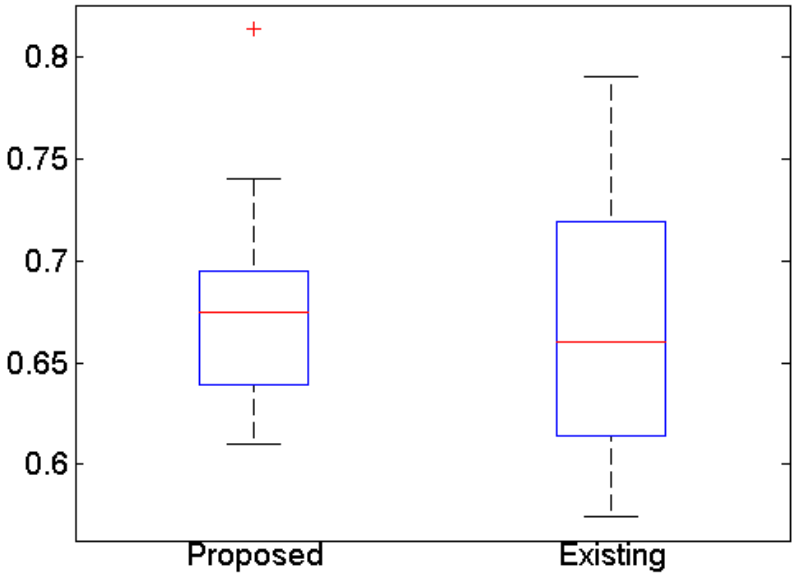

FIGURE 2.4.1: The comparison of genetic separability of the cocaine user clusters obtained by the proposed method and the sequential method in [15].

mains limited because existing methods mismatch the clinical cluster analysis to the goal of genetic association.

We have developed a novel multi-objective programming approach that optimizes two objectives: (1) the cluster-derived subtypes should differ significantly in clinical features; (2) the subtypes can be classified using genetic markers. Our method forms a novel multi-view data analytic method that treats the different views differently instead of equally as input views. In our method, the view of clinical features was used to define and derive subtypes of the disease based on cluster analysis, and the view of genetic markers was used to interpret the subtypes based on sparse modeling. Two case studies of subtyping of opioid use and cocaine use, and related behaviors in aggregated samples of European Americans were performed. A comparison be-

tween our proposed approach and a typical subtyping method [15] demonstrated the superiority of our approach.

Chapter 3

Multi-view Singular Value Decomposition for Disease Subtyping and Genetic Associations

3.1 Introduction

Integration of data from the phenotypic and genomic dimensions offers benefits such as new opportunities to find confirming evidence of a subtype from its genetic basis and phenotypic manifestations. A few studies examined the joint use of gene expression and genotype data for cancer subtyping [17, 69], but they did not identify the variable subspaces in the two sources of data to group subjects into consistent clusters across the two subspaces. Hence, they cannot detect genetic variants that are associated with the identified clusters.

There has been little research on this topic in the statistics literature. The most

related area involves co-clustering [43] or multi-view data analysis [16], 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 [72, 46], which simultaneously clusters the rows and columns of a data matrix and (2) multi-view co-clustering [42, 43] which seeks groupings that are consistent across different views. Biclustering is similar to another set of algorithms [29] that search for subspaces and group subjects differently in each subspace.

Biclustering and subspace searching essentially find different subgroups of subjects using different features (or markers), thus helping to identify genetic variants specific to a particular subgroup. However, it can only be applied to one data matrix from a single view rather than data jointly from more 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 resultant clusters are defined using all given features. Hence, it cannot be used to identify subtype-specific variants/features. Our subtyping problem seeks a grouping of subjects that is consistent when using clinical features and using genetic markers, but also requires a subspace search to identify the specific features or markers that define the subgroups.

In this chapter, we propose a multi-view matrix decomposition approach based on the sparse singular value decomposition (SSVD) technique [46] 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 SSVD in our approach is critical to its success, especially in terms of successfully detecting 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 synthetic datasets that are simulated to have subtype structures and several genetic markers associated with the 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 [46] and the latest multi-view data analytics methods [43]. The results clearly shows the superior performance of our approach over all other compared methods.

3.2 Methods

We start with a presentation of the notations that are used throughout the chapter. 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 $\mathbf{v}_{(j)}$ 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 of components that 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 $\mathbf{M}_{(\cdot,j)}$, respectively.

3.2.1 Review of single-view biclustering

We briefly review the biclustering method with a single view of data based on the sparse singular value decomposition [46]. 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 resulted 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 $\mathbf{M}$ corresponding to non-zero components in $\mathbf{u}$ form a row subgroup. The columns in $\mathbf{M}$ corresponding to non-zero components in $\mathbf{v}$ form a column subgroup. The resulted row and column clusters help to define each other. The SSVD finds all singular vectors sequentially by repeatedly solving the following problem with a data matrix $\mathbf{M}$:

$$\begin{aligned} \min_{\sigma,\mathbf{u},\mathbf{v}} \quad & \|\mathbf{M} - \sigma\mathbf{u}\mathbf{v}^T\|_F^2 + \lambda_u\|\sigma\mathbf{u}\|_0 + \lambda_v\|\sigma\mathbf{v}\|_0 \\ \text{subject to} \quad & \|\mathbf{u}\|_2 = 1, \|\mathbf{v}\|_2 = 1. \end{aligned} \tag{3.2.1}$$

The regularization terms $\|\sigma\mathbf{u}\|_0$ and $\|\sigma\mathbf{v}\|_0$ are used to enforce the sparsity of $\mathbf{u}$ and $\mathbf{v}$. Notice 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 between 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 [46] 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.(3.2.1); (b), fix $\mathbf{v}$ and find $\mathbf{u}$ that optimizes the objective of Eq.(3.2.1).

Assume that the rows and columns of $\mathbf{M}$ represent subjects and their features, respectively. 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 do not overlap with subjects in any identified cluster, the SSVD solves Eq.(3.2.1) using a new matrix $\tilde{\mathbf{M}}$ which 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.(3.2.1) with the deflated $\tilde{\mathbf{M}} = \mathbf{M} - \sigma \mathbf{u}\mathbf{v}^T$ as used in the standard SVD.

3.2.2 The proposed formula for two-view joint biclustering

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

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, it requires $\mathbf{u}_1$ and $\mathbf{u}_2$ to have non-zero components at the same position. Notice that the two $\mathbf{u}$ vectors are not necessarily the same given 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:

$$\begin{aligned} \min_{\mathbf{z},\sigma_i,\mathbf{u}_i,\mathbf{v}_i,i=1,2} \quad & \|\mathbf{M}_1 - \sigma_1(\mathbf{z} \odot \mathbf{u}_1)\mathbf{v}_1^T\|_F^2 + \|\mathbf{M}_2 - \sigma_2(\mathbf{z} \odot \mathbf{u}_2)\mathbf{v}_2^T\|_F^2 \\ & + \lambda_z\|\mathbf{z}\|_0 + \lambda_{v_1}\|\sigma_1\mathbf{v}_1\|_0 + \lambda_{v_2}\|\sigma_2\mathbf{v}_2\|_0, \\ \text{subject to} \quad & \|\mathbf{u}_i\|_2 = 1, \|\mathbf{v}_i\|_2 = 1, \quad i = 1, 2, \\ & \mathbf{z} \in \mathcal{B}_n. \end{aligned} \tag{3.2.2}$$

where λ_z, λ_{v_1} and λ_{v_2} are tuning parameters that balance the approximation errors and regularization terms. Although the $\mathbf{u}$'s are constrained to be unit vectors, the $\mathbf{z} \odot \mathbf{u}$'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 only non-zero values 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 $j = 1 \cdots, n$. Then, we 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 $\bar{\mathbf{u}}$ together with the new σ will produce the same objective value as the original solution $\hat{\mathbf{u}}$, and hence is an optimal solution as well. We will design a fast algorithm to find such a $\bar{\mathbf{u}}$ for Eq.(3.2.2).

We discuss two alternatives to the proposed formula (3.2.2). A restricted version of Eq.(3.2.2) 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.(3.2.2), which leads to the following problem

$$\begin{aligned} \min_{\sigma_i, \mathbf{u}, \mathbf{v}_i, i=1,2} \quad & \|\mathbf{M}_1 - \sigma_1 \mathbf{u}\mathbf{v}_1^T\|_F^2 + \|\mathbf{M}_2 - \sigma_2 \mathbf{u}\mathbf{v}_2^T\|_F^2 \\ & + \lambda_u \|\mathbf{u}\|_0 + \lambda_{v_1} \|\sigma_1 \mathbf{v}_1\|_0 + \lambda_{v_2} \|\sigma_2 \mathbf{v}_2\|_0, \\ \text{subject to} \quad & \|\mathbf{u}\|_2 = 1, \|\mathbf{v}_i\|_2 = 1, \quad i = 1, 2. \end{aligned} \tag{3.2.3}$$

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 unnecessary 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 concerned. Our problem only seeks the indicators of whether or not a component of $\mathbf{u}$ is zero.

It is insightful to also discuss the relation between Eq.(3.2.3) 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

$$\begin{aligned} \min_{\sigma, \mathbf{u}, \mathbf{v}} \quad & \|[\mathbf{M}_1\ \mathbf{M}_2] - \sigma \mathbf{u}\mathbf{v}^T\|_F^2 + \lambda_u \|\sigma \mathbf{u}\|_0 + \lambda_v \|\sigma \mathbf{v}\|_0 \\ \text{subject to} \quad & \|\mathbf{u}\|_2 = 1, \|\mathbf{v}\|_2 = 1. \end{aligned} \tag{3.2.4}$$

where the $\mathbf{v}$ vector is of size $d_1 + d_2$. In comparison with Eq.(3.2.3), Eq.(3.2.4) 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 Eq.(3.2.3) can be feasible to Eq.(3.2.4) 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 unnecessarily an optimal solution to Eq.(3.2.4). An optimal $\mathbf{v}$ to Eq.(3.2.4) may have either $\mathbf{v}_1$ or $\mathbf{v}_2$ be zero, which is however not allowed in Eq.(3.2.3). When one of the $\mathbf{v}$'s 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 number of genetic markers outweighed that of clinical features, the resultant clusters differed significantly on the SNPs only, leading to disease subtypes that cannot be clinically recognized.

3.2.3 A fast algorithm for two-view joint biclustering

The proposed formulation (3.2.2), 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 $\mathbf{u}$'s; one set consists of $\mathbf{v}$'s; and the last set consists of the binary variables $\mathbf{z}$. 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 (3.2.2) 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 $\mathbf{z}$.

$$\begin{aligned} \min_{\sigma_1,\mathbf{u}_1,\mathbf{v}_1} \quad & \|\mathbf{M}_1 - \sigma_1(\mathbf{z}\odot\mathbf{u}_1)\mathbf{v}_1^T\|_F^2 + \lambda_{v_1}\|\sigma_1\mathbf{v}_1\|_0 \\ \text{subject to} \quad & \|\mathbf{u}_1\|_2 = 1, \|\mathbf{v}_1\|_2 = 1, \end{aligned} \tag{3.2.5}$$

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 $\mathbf{v}_1 = \tilde{\mathbf{v}}_1/\sigma_1$.

$$\min_{\tilde{\mathbf{v}}_1} \quad \|\mathbf{M}_1 - (\mathbf{z}\odot\mathbf{u}_1)\tilde{\mathbf{v}}_1^T\|_F^2 + \lambda_{v_1}\|\tilde{\mathbf{v}}_1\|_0. \tag{3.2.6}$$

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 [46]:

$$\tilde{\mathbf{v}}_{1(j)} = \begin{cases} \alpha_{(j)} - \beta, & \text{if } \alpha_{(j)} > \beta, \\ 0, & \text{if } |\alpha_{(j)}| \le \beta, \\ \alpha_{(j)} + \beta, & \text{if } \alpha_{(j)} < -\beta, \end{cases} \quad j = 1, \cdots, d. \tag{3.2.7}$$

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

After $\mathbf{v}_1$ is obtained and fixed, we optimize Problem (3.2.5) 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 (3.2.5).

$$\min_{\tilde{\mathbf{u}}_1} \quad \|\mathbf{M}_1 - (\mathbf{z} \odot \tilde{\mathbf{u}}_1)\mathbf{v}_1^T\|_F^2. \tag{3.2.8}$$

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

$$\tilde{\mathbf{u}}_{1(i)} = \begin{cases} \dfrac{\mathbf{M}_{1(i,\cdot)}\mathbf{v}_1}{\mathbf{z}_{(i)}}, & \text{if } \mathbf{z}_{(i)} \neq 0 \\ 0, & \text{if } \mathbf{z}_{(i)} = 0. \end{cases} \qquad i = 1, \cdots, n. \tag{3.2.9}$$

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

When all $\mathbf{u}$'s and $\mathbf{v}$'s are fixed in Problem (3.2.2), the optimization problem becomes:

$$\min_{\mathbf{z} \in \mathcal{B}_n, \sigma_1, \sigma_2} \quad \|\mathbf{M}_1 - \sigma_1(\mathbf{z} \odot \mathbf{u}_1)\mathbf{v}_1^T\|_F^2 + \|\mathbf{M}_2 - \sigma_2(\mathbf{z} \odot \mathbf{u}_2)\mathbf{v}_2^T\|_F^2 + \lambda_z \|\mathbf{z}\|_0. \tag{3.2.10}$$

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 $\tilde{\mathbf{z}}$:

$$\min_{\tilde{\mathbf{z}}} \quad \|\mathbf{M}_1 - (\tilde{\mathbf{z}} \odot \mathbf{u}_1)\mathbf{v}_1^T\|_F^2 + \|\mathbf{M}_2 - (\hat{\sigma}_2/\hat{\sigma}_1)(\tilde{\mathbf{z}} \odot \mathbf{u}_2)\mathbf{v}_2^T\|_F^2 + \lambda_z \|\tilde{\mathbf{z}}\|_1 \tag{3.2.11}$$

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. Notice that Problem (3.2.11) can be rewritten into the following problem

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

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 $\text{diag}(\tilde{\mathbf{z}})$ converts $\tilde{\mathbf{z}}$ into a diagonal matrix. Then, each component of an optimal $\tilde{\mathbf{z}}$ can be analytically computed as follows:

$$\tilde{\mathbf{z}}_{(i)} = \begin{cases} \gamma_{(i)} - \theta, & \gamma_{(i)} > \theta \\ 0, & |\gamma_{(i)}| \leq \theta \\ \gamma_{(i)} + \theta, & \gamma_{(i)} < -\theta \end{cases} \quad i = 1, \cdots, n. \tag{3.2.12}$$

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.(3.2.12) is derived based on the same calculation in [46] as how Eq.(3.2.7) is derived.

After obtaining $\tilde{\mathbf{z}}$, the solution $\mathbf{z}$ to Problem (3.2.10) can be calculated as follows:

$$\mathbf{z}_{(i)} = \begin{cases} 1, & \text{if } \tilde{\mathbf{z}}_{(i)} \neq 0 \\ 0, & \text{if } \tilde{\mathbf{z}}_{(i)} = 0. \end{cases} \quad i = 1, \cdots, n. \tag{3.2.13}$$

To preserve the same objective value of Problem (3.2.2) after updating $\mathbf{z}$, we update $\mathbf{u}_1$ and $\mathbf{u}_2$ accordingly as follows:

$$\mathbf{u}_{(i)} = \begin{cases} \mathbf{u}_{(i)}/\tilde{\mathbf{z}}_{(i)}, & \text{if } \tilde{\mathbf{z}}_{(i)} \neq 0, \\ 0, & \text{if } \tilde{\mathbf{z}}_{(i)} = 0, \end{cases} \quad i = 1, \cdots, n. \tag{3.2.14}$$

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 (3.2.6), (3.2.8) and (3.2.10) 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. In our experiment both on 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.

3.2.4 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 an analysis. The optimization problem (3.2.2) 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 similarly solved by decomposing it into several sub-problems and alternatively solving each sub-problem. 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:

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

Notice that when $\mathbf{z}$ is fixed, the optimization of $\mathbf{u}_i$ and $\mathbf{v}_i$ is independent from each other among different views. Thus, these singular vectors can be computed in parallel, which can bring down 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 2 summarizes all 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 choose to initialize $\mathbf{z}$ with a vector of all ones which assumes that all subjects have potentials to be in the cluster if no prior is given.

3.3 Computational Results and Discussion

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

Algorithm 2 Multi-view Singular Value Decomposition

Input: $\mathbf{M}_i$, λ_z, λ_{v_i}, $i = 1, \cdots, m$
Output: $\mathbf{z}$, σ_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.(3.2.7).
   Compute $\mathbf{v}_i$ from $\tilde{\mathbf{v}}_i$ and update σ_i.
   Compute $\tilde{\mathbf{u}}_i$ by Eq.(3.2.9).
   Compute $\mathbf{u}_i$ from $\tilde{\mathbf{u}}_i$ and update σ_i.
4. Compute $\tilde{\mathbf{z}}$ by Eq.(3.2.12).
5. Compute $\mathbf{z}$ from $\tilde{\mathbf{z}}$ by Eq.(3.2.13).
6. Update σ_i, $\mathbf{u}_i$, $i = 1, \cdots, m$ by Eq.(3.2.14) accordingly.

Repeat Steps 3 to 6 until $\mathbf{z}$ reaches a fixed point.

(CD).

Normalized mutual information (NMI) was used to measure the consistency between 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 $\mathcal{C}_i$ indexes the subjects in the i-th cluster. NMI computes the mutual information between the two clusterings normalized by the cluster entropies. In other words,

$$\mathrm{NMI}(\mathcal{C}^{(1)}, \mathcal{C}^{(2)}) = \frac{I(\mathcal{C}^{(1)}, \mathcal{C}^{(2)})}{(H(\mathcal{C}^{(1)}) + H(\mathcal{C}^{(2)}))/2} \tag{3.3.1}$$

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 index set $\mathcal{C}_i$. Since the true clusters are known in synthetic data, we computed NMI to measure the consistency between the true cluster assignments and the cluster assignments resulted from cluster analysis. Hence, 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) [22] 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 [85] 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 values on both synthetic and real world data. The existing methods that were used in our comparison study are given in the following list:

- **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 in [43] 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:** Radial basis function (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 and used in spectral clustering to obtain row clusters.

3.3.1 A simulation study

Two disease subtypes, *subtype 1* and *subtype 2*, were simulated. Each of them was not only defined by a set of phenotypic/clinical features but also associated with a set of genetic markers. However, the clinical features and genetic markers are different for the two subtypes. Each subtype hence corresponded to a cluster of subjects that presented the specific clinical features and had minor alleles at the associated SNP markers (here we assumed that minor alleles were risk variants). The goal of the simulation was to create an agreed partition of subjects in both views of their genetic markers and clinical features.

Genetic data was obtained from the 1000 Genome Project [2] in which 1092 subjects were genotyped with several million genetic markers. We randomly selected 1000 markers from chromosome 5 that had a minor allele frequency (MAF) 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 chosen SNPs for that subtype. This amounts to an additive genetic model for each subtype (i.e., adding up risk variants). Subjects who did not belong to any 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 the clarity in the partition. In total, 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 which were consistent to certain degree with the genotypic clusters. Notice that many diseases, although highly heritable, are multifactorial both 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, which aimed 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 that subject i had, 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 different but very similar clusters of subjects to the genotypic clusters (and the noise parameter e determined the level of similarity).

We named these clusters the phenotypic clusters because these clusters were used to synthesize clinical features so that the cluster structure was reflected in the clinical data. Similarly, we removed from the analysis the subjects that overlapped in the two phenotypic clusters. Fewer than 15 subjects were excluded in any simulated dataset in the experiments. In addition to the 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 realistic. The two additional clusters each included 200 subjects that were randomly selected from 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 probability. *Subtype 1* and *subtype 2* each was associated with three features. Subjects in each simulated phenotypic cluster 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 clusters was associated with two features, and subjects in each of the two subtypes obtained the value of 1 in 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 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 compared methods were used to obtain three subject clusters. Table 3.3.1 provides the NMI calculated by comparing subject clusters obtained from each approach to true phenotypic clusters simulated. 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 subject clusters 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 subtypes from controls. The subject cluster from each method containing the largest number of controls was considered as the control group. The average AUC values and their interquantiles obtained by all compared approaches on each dataset are plotted in Figure 3.3.1. The proposed method achieved the second best performance on this measurement. Although the feature concatenation method

TABLE 3.3.1: Comparison between different approaches on normalized mutual information (NMI) values. The NMI values for different approaches with different genotypic effect e to the phenotypic variation in simulated data.

	$e = 1$	$e = 0.8$	$e = 0.6$	$e = 0.4$
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**

found the clusters that were most separable genetically with the best AUC, these clusters were not recognizable in clinical features. As shown in Table 3.3.1, they were most disparate from the simulated true phenotypic clusters.

A significant advantage of the proposed method is that it can simultaneously identifies 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 3.3.2, 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.

3.3.2 A case study: cocaine use and related behaviors

A total of 1474 subjects were phenotyped and genotyped for genetic studies of cocaine dependence [1]. 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 sub-

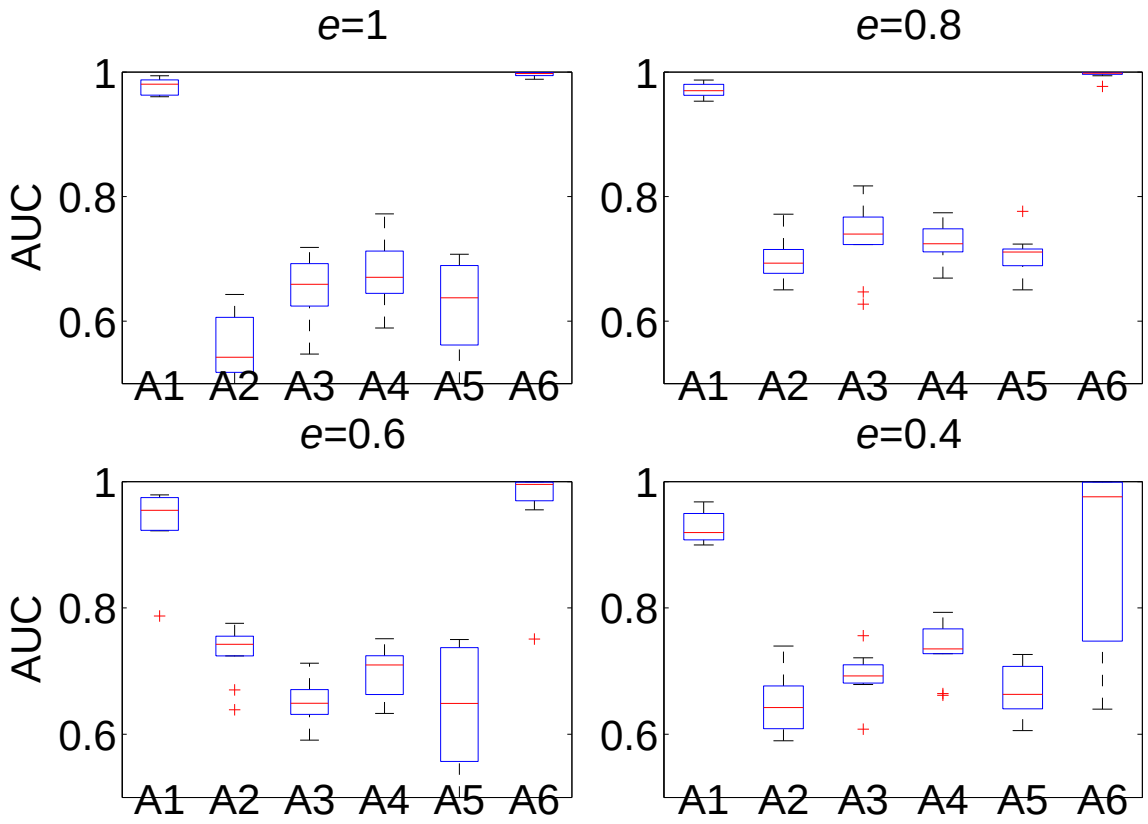

FIGURE 3.3.1: Comparison between different approaches in terms of AUC values on simulated data. The box plot of AUC values obtained from all comparison approaches on simulated data. A1 - the proposed method, A2 - single-view SSVD, A3 - co-regularized spectral, A4 - kernel addition, A5 - kernel product, A6 - feature concatenation. e is the genetic effect to the phenotypic variation in simulated data.

jects 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) [59], a polydiagnostic instrument 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 genetic association studies [41, 9]. These variables were used as the phenotypic fea-

TABLE 3.3.2: Statistics of features identified by the proposed method for both subgroups in the two views. e is the genetic effect to the phenotypic variation in simulated data. TF is the number of True Features that specify a population subgroup. TPF (True Positive Features) is the number of features correctly identified. FPF (False Positive Features) is the number of features incorrectly identified.

		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

tures. Of the 1474 subjects, 1287 were diagnosed with cocaine dependence. Subjects were genotyped with 1350 SNPs that were selected from 130 candidate genes [33]; 1248 SNPs with minor allele frequency (MAF) of at least 1% were used as genetic markers in our analysis.

The feature concatenation method overlooked the information in the clinical or phenotypic view as observed in both the simulation study and this case study. We hence excluded the kernel concatenation method from the further comparison. Three subject clusters were obtained from each of the methods in the comparison. Classification models were built and tested in a manner similar to that used for synthetic data. Figure 3.3.2 shows the box plot of the AUC values. As shown in Figure 3.3.2, our approach significantly outperformed all other methods with respect to the genetic separability of the resultant clusters. A paired t-test was performed to compare the AUC values from our method with those from a compared method, yielding p-values < 0.05 for all comparisons.

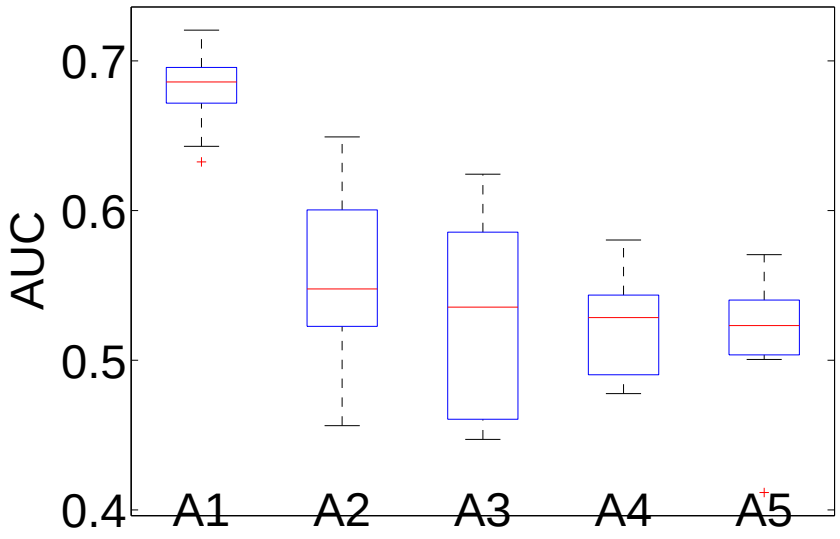

FIGURE 3.3.2: Comparison between different approaches in terms of AUC values on cocaine data. The box plot of AUC values obtained from all comparison approaches on dataset of cocaine use and related behaviors. A1 - the proposed method, A2 - single-view SSVD, A3 - co-regularized spectral, A4 - kernel 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, respectively. *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. We hence used this group as a control group in our association analysis. The number of clinical features identified to be associated with *Group 1* and *Group 2* were 18 and 17, respectively. Figures 3.3.3 and 3.3.4 compare the three subject clusters in terms of 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 with the features in the figures with a correlation coefficient > 0.7.

From these two figures, we can see that *Group 1* is distinctively associated with

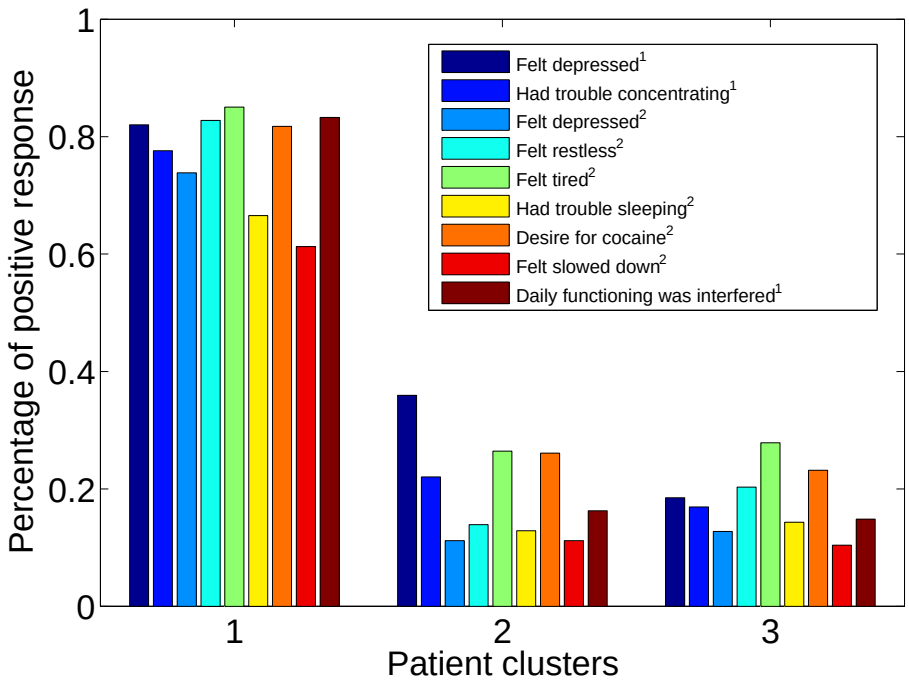

FIGURE 3.3.3: Comparison between the three cocaine user subgroups on features linked to *Group 1*. [1] indicates symptoms due to cocaine use. [2] denotes symptoms due to stopping, cutting down or going without cocaine.

several withdraw symptoms, such as felt 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 3.3.4. Besides the subjects in *Group 2*, the subjects in *Group 1* also showed high values on these features. Note that these subjects were already 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.

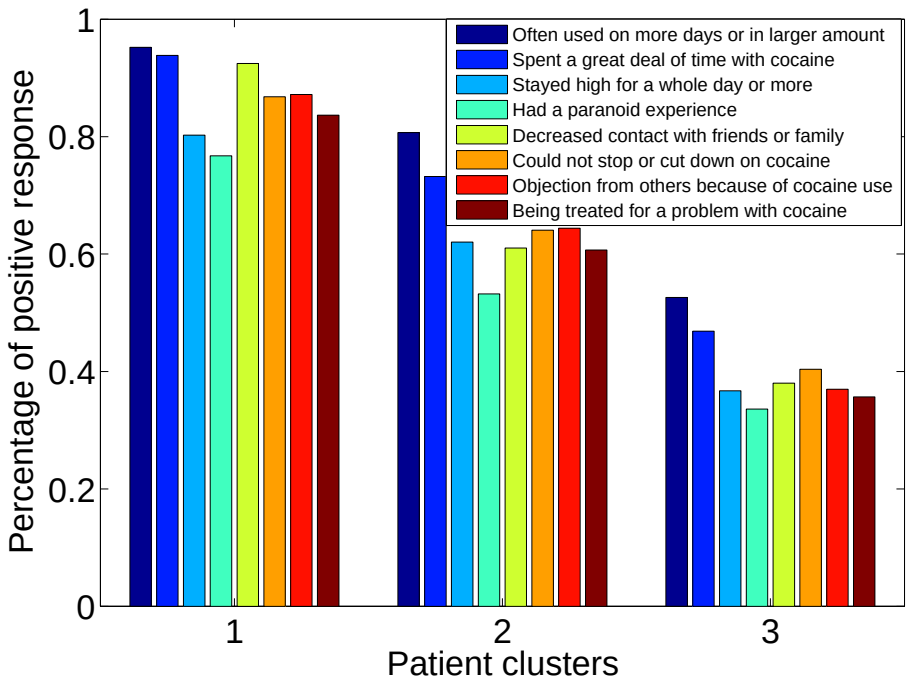

FIGURE 3.3.4: Comparison between the three cocaine user subgroups on features linked to *Group 2*.

There were 114 and 237 genetic markers identified for *Group 1* and *Group 2* respectively. Two classification models were built to further identify the top markers that had the highest predictive power in separating subjects in *Group 1* or in *Group 2*, from those in the control group. Table 3.3.3 gives the top 5 SNPs that showed the highest influence in terms of distinguishing the case groups from the control group. It is also interesting to point out that the *HTR2C* gene significantly associated with *Group 1* in our study (with a p-value $< 10^{-5}$) was previously identified for a heavy use, early-onset and high comorbidity subtype of cocaine dependence [9].

TABLE 3.3.3: Top five SNPs with the largest magnitude of weights in the two classification models. These two models are built to separates controls in *Group 3* from cases in *Group 1* and *Group 2*, respectively.

	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*

3.4 Summary

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 [24, 41, 15, 66, 67]. 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 chapter, we have proposed a multi-view biclustering approach to perform the phenotype refinement 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 when a subtype is determined,

which is critical to its success. Hence, associative genetic variants are likely to be revealed. 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. Hence, it may identify markers that are relevant to a disease subtype only due to population structure rather than the specific disease, causing spurious association findings. Racial populations are often stratified before an association analysis is performed. Our method will need to be extended to deal with the three-way relationship among population subgroups, genotypes and phentoypes in order to further improve the phenotype-genotype associations. Further, the proposed method has been used in our empirical study to identify the first two major subgroups of data, for which we have not observed any invalid clusters caused by random noise. When more subsequent clusters are to identified, the two methods we have designed either by excluding subjects in the identified subgroups or by deflating singular value components from the data matrix both suffer a higher risk to detect invalid clusters because singular values will decrease in subsequent decomposition. More empirical studies may be needed to thoroughly examine the signal-to-noise pattern for the proposed method.

Chapter 4

Identifying Heritable Composite Traits with Pedigree for Complex Phenotypes

4.1 Introduction

Complex phenotypes consist of a variety of low level traits that are often highly variable. Association analysis of such a complex phenotype is impeded by this phenotypic heterogeneity [28]. For example, the diagnosis of drug dependence, such as dependence on cocaine or opioids, is determined by various patterns of drug use, their effects, and related behaviors, including the age of first drug use, drug use frequency, negative consequences of drug use, withdrawal symptoms, and treatment history [59]. A binary composite trait defined based on the low level traits, such as the diagnosis of cocaine dependence, which partitions the population into cases (subjects with the disease) and controls (subjects without the disease), cannot differentiate the het-

erogeneous manifestations of the disease. Because of this, the success of identifying underlying genetic variants in the association study using this binary trait as respondent variable [77, 25]. Identifying highly heritable composite traits or subphenotypes of the disease could permit the detection of genetic variants that are not detectable using the standard diagnosis-based traits [36, 9]. Examination of multiple low-level traits that characterize a complex phenotype has shown the potential to discover composite traits of high heritability [41, 15, 66]. Efforts have been made to enhance the binary trait by capturing more phenotypic variation, such as defining a composite trait as symptom count [25]. However, this kind of composite trait can have low heritability and may thus be sub-optimal for association analysis.

Besides enhancing association analysis as discussed above, identified high heritable composite traits can be of direct use in animal and plant breeding. In agricultural science, breeding programs targeted at conceptual but economically important phenotypes, such as feed efficiency or heat tolerance, are confronted with a wide variety of available measures for the phenotype [18, 12]. Residual body weight gain, residual feed intake and relative growth rate are all feed efficiency measures for dairy cattle with heritabilities that range from 0.28 to 0.45 [7, 18]. Moreover, each of these measures forms a composite trait that is defined by a linear function of low-level traits, such as body weight, diet and feed energy intake, and days in milk. Identifying more heritable measures for conceptual phenotypes could enhance animal and plant breeding.

Recently, there are methods being proposed in the literature to directly target at high heritable composite traits by identifying optimal linear combinations of low-level traits [55, 78, 40, 56]. All current methods decompose the identification of heritable composite traits into solving two separate subproblems in sequence. They first esti-

mates two covariance matrices of the low-level traits: $\mathbf{\Sigma}_a$, the variance due to additive genetic effects; and $\mathbf{\Sigma}$, the covariance matrix due to effects other than additive genetic effects. If there are d low level traits in $\mathbf{x}$, this means that two d-by-d matrices need to be estimated from the sample. Once the two covariance matrices are computed, a generalized eigenproblem is solved to identify the combination coefficients $\mathbf{w}$ so that the ratio of $\mathbf{w}^\top \mathbf{\Sigma}_a \mathbf{w} / \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}$ is maximized, leading to high heritability for the composite trait $\mathbf{w}^\top \mathbf{x}$.

A few methods have been developed in the literature to estimate the two covariance matrices. In [78, 40], the two matrices are estimated based on the genetic effect of a single quantitative-trait locus to all the low level traits. This method has limited utility when the variance-covariance of the low level traits is due to multiple genetic loci (which is often the case for complex phenotypes). In [55, 56, 38], the two covariance matrices are estimated from family pedigrees of the sample. The approach used in [55] takes only siblings in a family, so it is inadequate to handle general (complex) pedigrees. The two approaches in [56] and [38] can handle general pedigrees. The first one derives an analytic formula for the covariance matrices based on Analysis of Variance (ANOVA). Although reducing the computational cost, the analytic formula makes it unable to take into account the fixed effects from covariates such as sex, age or race, which is also a problem for the method in [55]. Currently, the most comprehensive approach is a maximum likelihood method [38] that can estimate the fixed effects and covariance matrices together, but this method is computationally prohibitive as discussed in [56]. Even when $d = 20$ low level traits are used, this method can run hours to report performance and further as observed in our experiments, the method may not converge even with a large limit on the number of iterations. Moreover, it requires very large sample set in order to obtain reliable estimates of two

covariance matrices and d combination coefficients, totally $2d^2 + d$ parameters, from a sample set.

By solving the inverse problem of heritability estimation, we show that, to obtain highly heritable multivariate traits, the estimation of two covariance matrices is not necessary. We propose an optimization approach that directly identifies a linear combination of low level traits whose estimated heritability is high. This optimization problem is formulated by decomposing the maximum likelihood method for estimating the heritability of a quantitative trait. We develop an efficient *sequential quadratic programming* algorithm to optimize the proposed formulation. We then extend the basic formulation to take into account the effects of covariates so that the identified trait has high heritability even after correction for the fixed effects. Because we do not estimate any covariance matrix, our approach is computationally much more efficient than those in [55, 38].

The proposed approach is validated in both simulations and a case study on cocaine dependence. The effectiveness of the approach is demonstrated not only by the higher cross-validated heritability of the derived traits than the existing methods but also by a follow-up association study that compares the utility of the derived traits with the commonly used phenotype. Specifically, a highly heritable multivariate trait was derived for cocaine dependence. More statistically significant associations were found for this trait than for a symptom-count phenotype, with successful replications in an independent sample.

4.2 Methods

Our formulation aims to solve the inverse problem of heritability estimation. Hence, we first introduce the standard methods for heritability estimation. We then derive our formulation that defines a linearly-combined trait by maximizing its heritability. An efficient algorithm is then developed to optimize the formulation. At last, we extend the approach to take into consideration of fixed-effect covariates.

4.2.1 Background: heritability estimation

We briefly review the well established maximum likelihood method that is based on linear mixture models to estimate the heritability of a quantitative trait y [45, 4]. The method assumes that the phenotype $\mathbf{y}^i$ of a family i follows a multivariate normal distribution with covariance $\mathbf{\Omega}_i$ and separate means for male and female family members, μ_m and μ_f, respectively. Separate means are used for males and females based on the general observation that males and females present differences in quantitative traits, such as height and weight. The (j,k)-th entry of $\mathbf{\Omega}_i$ is the phenotypic covariance of two family members j and k, given by

$$cov(y_j^i, y_k^i) = 2\sigma_a^2 \mathbf{\Phi}_{jk}^i + \sigma_d^2 \mathbf{\Delta}_{jk}^i + \sigma_e^2 \mathbf{\Gamma}_{jk}^i \tag{4.2.1}$$

where σ_a^2 and σ_d^2 are the variance components due to additive and dominant genetic effects, respectively, and σ_e^2 denotes the variance component due to environmental factors. Eq. (4.2.1) can be extended to include other effects, such as an epistatic genetic effect σ_I^2. The quantity $\mathbf{\Phi}_{jk}^i$ is the kinship coefficient between members j and k. It is the probability that two alleles randomly drawn from j and k at a genetic locus

TABLE 4.2.1: Elements of the matrices $\mathbf{\Phi}$ and $\mathbf{\Delta}$ for selected relationships in a family when random mating is assumed.

Relationship	$\mathbf{\Phi}$	$\mathbf{\Delta}$
Same person	1/2	1
Parent-Child	1/4	0
Full-siblings	1/4	1/4
Half-siblings	1/8	0
Monozygous twins	1/2	1
Grandparent-grandchild	1/8	0
Uncle/aunt-nephew/niece	1/8	0
First cousins	1/16	0
Double first cousins	1/8	1/16
Spouses	0	0

are identical by descent (IBD), i.e., that these two alleles are identical copies of the same ancestral allele. An allele is one of the alternative forms at a genetic locus. As the human genome is diploid, each individual has two copies of an allele that may differ at a genetic locus. The quantity $\mathbf{\Delta}^i_{jk}$ is the probability that members j and k share both alleles at a genetic locus. Both matrices $\mathbf{\Phi}_i$ and $\mathbf{\Delta}_i$ can be calculated from the family pedigrees [4]. Exemplar entries of $\mathbf{\Phi}$ and $\mathbf{\Delta}$ between selected family members are illustrated in Table 4.2.1 where random mating is assumed. The parameter $\mathbf{\Gamma}^i_{jk}$ is an environmental indicator that encodes whether j and k live together ($\mathbf{\Gamma}^i_{jk} = 1$) or apart ($\mathbf{\Gamma}^i_{jk} = 0$).

The narrow sense heritability $h^2 = \sigma_a^2/\sigma_p^2$ where σ_p^2 is the total variance in y, i.e., $\sigma_p^2 = \sigma_a^2 + \sigma_d^2 + \sigma_e^2$. The broad sense heritability $H^2 = (\sigma_a^2 + \sigma_d^2)/\sigma_p^2$. In this thesis study, we target quantitative traits with higher narrow sense heritability, which we henceforth refer to simply as heritability. However, our formulation can be easily modified to derive a quantitative trait of high H^2.

The five parameters, μ_m, μ_f, σ_a^2, σ_d^2 and σ_e^2, are estimated by maximizing the log

likelihood of pedigrees over all sample families [45]. The log likelihood is computed by

$$LL = \sum_i -\frac{1}{2}\ln|\mathbf{\Omega}_i| - \frac{1}{2}(\mathbf{y}^i - \mu^i)^\top \mathbf{\Omega}_i^{-1}(\mathbf{y}^i - \mu^i), \tag{4.2.2}$$

where μ^i denotes a vector of the means μ_m and μ_f for male or female members, respectively, in the family i. The gradient and Hessian of Eq.(4.2.2) with respect to μ_m, μ_f, σ_a^2, σ_d^2 and σ_e^2 can be calculated, and a Newton-Raphson algorithm or a scoring method [45] can be applied to maximize the log likelihood Eq.(4.2.2).

The heritability of a quantitative trait y is often estimated with correction for the effects of covariates $\mathbf{z}$, such as age, sex, or race. These covariate effects are often modeled as fixed effects on y. Thus, a linear regression model $y = \mathbf{z}^\top\mathbf{v} + \epsilon$ can be built where $\mathbf{v}$ indicates the combination weights for the covariates. The heritability of the residual ϵ is then estimated using the described maximum likelihood method and treated as the heritability of y after adjusting for covariate effects.

4.2.2 Proposed quadratic optimization

In heritability estimation, a trait is given, and we search for the values of σ_a^2, σ_d^2 and σ_e^2 that maximize the likelihood of observing the trait values and compute the heritability as $\sigma_a^2/(\sigma_a^2+\sigma_d^2+\sigma_e^2)$. In our study, we solve the inverse problem that a trait must be derived so that its heritability is maximized when estimated by the method described in the above section. For a given set of d phenotypic features $\mathbf{x}$, we find a linearly combined trait $y : y = \mathbf{x}^\top\mathbf{w}$ that has a high estimate of heritability. Here, we limit our discussion to linear combinations. However, the proposed method can be used to create non-linear combinations by applying the so-called kernel mappings to $\mathbf{x}$ [63].

According to the discussion in the background section, if a trait y has the highest possible heritability, the covariance of y among any family members in family i, $cov(y^i_j, y^i_k)$, should be due to the additive effect σ^2_a only, and $\sigma^2_d = \sigma^2_e = 0$. In other words, for such a trait, the covariance matrix of the phenotype $\mathbf{y}^i$ of a family i relies only on the additive effect parameter σ^2_a and the kinship matrix $\mathbf{\Phi}^i$, i.e., $\mathbf{\Omega}_i = 2\sigma^2_a\mathbf{\Phi}^i$. Thus σ^2_a is equal to the total variance σ^2_p of y. Now, this trait y is composed by $\mathbf{x}^\top\mathbf{w}$ from a set of features $\mathbf{x}$, and $\mathbf{w}$ is the parameter to be determined. We thus need to search for the values of $\mathbf{w}$ that maximize the likelihood of observing $\sigma^2_d = \sigma^2_e = 0$, or in other words, that maximize the dependence of the covariance $\mathbf{\Omega}_i$ of a pedigree i on the additive effect.

Let $\mathbf{X}_i$ be the data matrix on the d features for the subjects in family i. Then the family's trait values $\mathbf{y}^i = \mathbf{X}_i^\top\mathbf{w}$. Note that we can scale $\mathbf{w}$ so that the sample variance of the derived trait is 1, which implies that $\sigma^2_p = \sigma^2_a = 1$. Substituting the values of $\mathbf{\Omega}_i$, $\mathbf{y}^i$ and σ^2_a in the log likelihood in Eq.(4.2.2) yields the following maximization problem:

$$\max_{\mathbf{w},\mu_m,\mu_f} \sum_i -\frac{1}{2}\ln|2\mathbf{\Phi}_i| - \frac{1}{4}(\mathbf{X}_i^\top\mathbf{w} - \mu^i)^\top\mathbf{\Phi}_i^{-1}(\mathbf{X}_i^\top\mathbf{w} - \mu^i),$$

which is equivalent to the following minimization problem after eliminating constants

$$\min_{\mathbf{w},\mu_m,\mu_f} \sum_i (\mathbf{X}_i^\top\mathbf{w} - \mu^i)^\top\mathbf{\Phi}_i^{-1}(\mathbf{X}_i^\top\mathbf{w} - \mu^i). \tag{4.2.3}$$

Let $\boldsymbol{\beta} = [\mathbf{w}^\top, \mu_m, \mu_f]^\top$, and $\mathbf{H}_i$ be defined by

$$\mathbf{H}_i = [\mathbf{X}_i^\top, [-1/0]_{im}, [-1/0]_{if}]^\top$$

where $[1]_i$, $[-1/0]_{im}$ and $[-1/0]_{if}$ are column vectors with length equal to the number

of members in family i; $[1]_i$ consists of all 1's. For males in the family, -1 is assigned at their corresponding entries in $[-1/0]_{im}$ and 0 at other positions of the vector. The vector of $[-1/0]_{if}$ is similarly defined for female family members. Then, we have $\mathbf{X}_i^\top \mathbf{w} - \mu^i = \mathbf{H}_i^\top \boldsymbol{\beta}$, and Problem (4.2.3) can be rewritten as follows:

$$\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^\top (\sum_i \mathbf{H}_i \boldsymbol{\Phi}_i^{-1} \mathbf{H}_i^\top) \boldsymbol{\beta} \tag{4.2.4}$$

Let $\mathbf{H}$ be the matrix by stacking all $\mathbf{H}_i$ in columns, then the sample variance of the resulting trait is calculated as $(1/n)\boldsymbol{\beta}^\top \mathbf{H}\mathbf{H}^\top \boldsymbol{\beta}$. The scaling of $\mathbf{w}$ in that $\sigma_p^2 = 1$ discussed above corresponds to a constraint $\boldsymbol{\beta}^\top \mathbf{H}\mathbf{H}^\top \boldsymbol{\beta} - n = 0$ in the formulation . In fact, μ_m and μ_f are not free parameters, as they are determined once $\mathbf{w}$ is determined. They are equal to the sample means of the trait, i.e., Mean$(\mathbf{x}^\top \mathbf{w})$, for male and female, respectively, when the optimal $\mathbf{w}$ is found. Let $\boldsymbol{\mu}_m$, $\boldsymbol{\mu}_f$ be the two mean vectors of respective male and female samples on the features $\mathbf{x}$. Both $\boldsymbol{\mu}_m$ and $\boldsymbol{\mu}_f$ have a length of d. Let

$$\mathbf{a}_m = [\boldsymbol{\mu}_m^\top, -1, 0]^\top, \quad \mathbf{a}_f = [\boldsymbol{\mu}_f^\top, 0, -1]^\top,$$

then the equality of $\mu_m = \text{Mean}(\mathbf{x}^\top \mathbf{w})$ on all male subjects is translated into $\mathbf{a}_m^\top \boldsymbol{\beta} = 0$. Similarly, we also have $\mathbf{a}_f^\top \boldsymbol{\beta} = 0$.

Imposing all of these constraints on Problem (4.2.4) yields an optimization problem where a quadratic objective needs to be minimized subject to both quadratic and

linear equality constraints as follows:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \boldsymbol{\beta}^\top (\sum_i \mathbf{H}_i \boldsymbol{\Phi}_i^{-1} \mathbf{H}_i^\top) \boldsymbol{\beta}, \\ \text{subject to} \quad & \boldsymbol{\beta}^\top \mathbf{H}\mathbf{H}^\top \boldsymbol{\beta} - n = 0, \\ & \mathbf{a}_m^\top \boldsymbol{\beta} = 0, \quad \mathbf{a}_f^\top \boldsymbol{\beta} = 0. \end{aligned} \tag{4.2.5}$$

According to statistical learning theory [74], optimizing only the empirical heritability on the training sample as in Eq.(4.2.5) will lead to the so-called overfitting problem, which means that the resultant model $y = \mathbf{x}^\top \mathbf{w}$ has low validation heritability despite a high training heritability. To enhance the generalizability of the derived model to new samples, a regularization condition on $\mathbf{w}$, $R(\mathbf{w})$, is required to control the complexity of the model. The objective function in Problem (4.2.5) thus becomes

$$\boldsymbol{\beta}^\top (\sum_i \mathbf{H}_i \boldsymbol{\Phi}_i^{-1} \mathbf{H}_i^\top) \boldsymbol{\beta} + \lambda R(\mathbf{w}), \tag{4.2.6}$$

where λ is a pre-specified tuning parameter for balancing the two terms in the objective function, and $R(\mathbf{w})$ can be realized in different forms and be application-specific. For example, $R(\mathbf{w})$ can be implemented with the ℓ_1 vector norm: $||\mathbf{w}||_1 = \sum_{j=1}^{d} |w_j|$, which is known to create shrinkage effects on $\mathbf{w}$ as shown in the Least Absolute Shrinkage and Selection Operator (LASSO) method [70]. When features in $\mathbf{x}$ are clustered in multiple groups and sparsity in each feature group level is desirable, $R(\mathbf{w})$ can be implemented by the $\ell_{2,1}$ vector norm as used in the group LASSO [49] and defined by $||\mathbf{w}||_{2,1} = \sum_{\ell=1}^{L} \sqrt{\sum_{j \in \mathcal{G}_\ell} \mathbf{w}_j^2}$ where $\mathcal{G}_\ell$ contains the indices of the features in the group ℓ.

Specifically, we develop an algorithm in the next section to solve the following optimization problem with the ℓ_1 norm regularization condition

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \boldsymbol{\beta}^\top (\sum_i \mathbf{H}_i \boldsymbol{\Phi}_i^{-1} \mathbf{H}_i^\top) \boldsymbol{\beta} + \lambda ||\mathbf{w}||_1, \\ \text{subject to} \quad & \boldsymbol{\beta}^\top \mathbf{H}\mathbf{H}^\top \boldsymbol{\beta} - n = 0, \\ & \mathbf{a}_m^\top \boldsymbol{\beta} = 0, \quad \mathbf{a}_f^\top \boldsymbol{\beta} = 0. \end{aligned} \tag{4.2.7}$$

Note that Problem (4.2.5) is a special case of Problem (4.2.7) when $\lambda = 0$. Hence, the solver for Problem (4.2.7) serves as a solver for Problem (4.2.5) as well.

4.2.3 Solving the proposed optimization problem

The objective function in Problem (4.2.7) is not differentiable because of the one-norm regularization term. To convert it into an equivalent differentiable form so gradient based solvers can be used, we introduce two sets of variables $\mathbf{u} \geq 0$ and $\mathbf{v} \geq 0$ both of length equal to that of $\mathbf{w}$. We set $\mathbf{w} = \mathbf{u} - \mathbf{v}$. Correspondingly, we replace the variables in $\boldsymbol{\beta}$ by:

$$\boldsymbol{\gamma} = [\mathbf{u}^\top, \mathbf{v}^\top, \mu_m, \mu_f]^\top.$$

Because $\mathbf{X}_i^\top \mathbf{w} = \mathbf{X}_i^\top \mathbf{u} - \mathbf{X}_i^\top \mathbf{v}$, we replace $\mathbf{H}_i$ by

$$\mathbf{K}_i = [\mathbf{X}_i^\top, -\mathbf{X}_i^\top, [-1/0]_{im}, [-1/0]_{if}]^\top,$$

and let

$$\mathbf{b}_m^\top = [\boldsymbol{\mu}_m, -\boldsymbol{\mu}_m, -1, 0]^\top, \quad \mathbf{b}_f^\top = [\boldsymbol{\mu}_f, -\boldsymbol{\mu}_f, 0, -1]^\top, \mathbf{J} = [\mathbf{I}_{2d \times 2d}, [0]_{2d}, [0]_{2d}],$$

where $\mathbf{I}_{2d\times 2d}$ is the identity matrix of dimension $2d\times 2d$, and $[0]_{2d}$ is a column vector of all zero entries with length of $2d$. Then, we rewrite Problem (4.2.7) into the following optimization problem

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \quad & f : \boldsymbol{\gamma}^\top (\sum_i \mathbf{K}_i \boldsymbol{\Phi}_i^{-1} \mathbf{K}_i^\top)\boldsymbol{\gamma} + \lambda \sum_{j=1}^{2d} \gamma_j \\ \text{subject to} \quad & g_1 : \boldsymbol{\gamma}^\top \mathbf{K}\mathbf{K}^\top \boldsymbol{\gamma} - n = 0, \\ & g_2 : \mathbf{b}_m^\top \boldsymbol{\gamma} = 0, \\ & g_3 : \mathbf{b}_f^\top \boldsymbol{\gamma} = 0, \\ & g_{4:e} : \mathbf{J}\boldsymbol{\gamma} \succeq 0, \end{aligned} \tag{4.2.8}$$

where $e = 2d + 3$ is the total number of constraints in the problem, and $\mathbf{K}$ is the matrix by stacking all $\mathbf{K}_i$ in columns.

It can be mathematically proved that the optimal solution of Problem (4.2.8) is identical to the optimal solution of Problem (4.2.7) in the sense that the optimal $\mathbf{w} = \mathbf{u} - \mathbf{v}$. Note that the regularizer in Eq.(4.2.8), $\sum_{j=1}^{2d} \gamma_j$, is just equal to $\sum_{j=1}^{d}(u_j + v_j)$. At optimality of Problem (4.2.8), either $u_j = 0$ or $v_j = 0$ for the jth feature because otherwise they would not be optimal. If both $u_j > 0$ and $v_j > 0$ and assume $u_j \geq v_j$, then we have another solution, $(\tilde{u}_j = u_j - v_j,\ \tilde{v}_j = 0)$, that achieves lower objective value than (u_j, v_j) because the first term of f remains the same whereas the second term of f becomes smaller. Thus, at optimality, the regularizer of Eq.(4.2.8) $\sum_{j=1}^{2d} \gamma_j = \sum_{j=1}^{d}(u_j + v_j) = \sum_{j=1}^{d} |w_j|$.

Although Problem (4.2.8) is not a convex problem due to the quadratic equality constraint g_1, we can solve it efficiently by the framework of sequential quadratic programming (SQP) [54]. The gradients of the objective function f and the constraints

$g_{i:i=1:e}$ with respect to γ can be calculated as follows:

$$\begin{aligned}
\nabla f &= 2(\sum_i \mathbf{K}_i \mathbf{\Phi}_i^{-1} \mathbf{K}_i^\top)\boldsymbol{\gamma} + \lambda \mathbf{c}, \\
\nabla g_1 &= 2(\mathbf{K}\mathbf{K}^\top)\boldsymbol{\gamma}, \quad \nabla g_2 = \mathbf{b}_m, \\
\nabla g_3 &= \mathbf{b}_f, \quad \nabla g_{4:e} = c
\end{aligned}$$

where $\mathbf{c} = [[1]_{2d}^\top, 0, 0]^\top$ and $[1]_{2d}$ is a column vector of all ones with length of $2d$. The Lagrangian function of this problem is:

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\alpha}) = f(\boldsymbol{\gamma}) - \sum_i \alpha_i g_i(\gamma) \tag{4.2.9}$$

where $\boldsymbol{\alpha}$ contains all Lagrange multipliers. The Hessian of $\mathcal{L}$ with respect to $\boldsymbol{\gamma}$ is calculated as:

$$\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \mathcal{L} = 2 \sum_i \mathbf{K}_i \mathbf{\Phi}_i^{-1} \mathbf{K}_i^\top - 2\alpha_1 \mathbf{K}\mathbf{K}^\top. \tag{4.2.10}$$

A SQP algorithm solves iteratively a sequence of quadratic programming subproblems, each formulated based on the current solution $\boldsymbol{\gamma}_t$ and Lagrange multipliers $\boldsymbol{\alpha}_t$. At the iteration $t+1$, it finds the searching directions for both $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ by solving the following quadratic program

$$\begin{aligned}
\min_{\mathbf{p}} \quad & f(\boldsymbol{\gamma}_t) + \nabla f(\boldsymbol{\gamma}_t)^\top \mathbf{p} + \frac{1}{2}\mathbf{p}^\top \nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \mathcal{L}(\boldsymbol{\gamma}_t, \boldsymbol{\alpha}_t)\mathbf{p} \\
\text{subject to} \quad & \nabla g_i(\boldsymbol{\gamma}_t)^\top \mathbf{p} + g_i(\boldsymbol{\gamma}_t) = 0, i \in [1:3] \\
& \nabla g_i(\boldsymbol{\gamma}_t)^\top \mathbf{p} + g_i(\boldsymbol{\gamma}_t) \succeq 0, i \in [4:e]
\end{aligned} \tag{4.2.11}$$

where $\mathbf{p}$ is the searching direction of $\boldsymbol{\gamma}$, along which the objective function can be decreased. Let $\hat{\mathbf{p}}_t$ be the solution to this subproblem and $\hat{\mathbf{q}}_t$ be the corresponding

Algorithm 3 A sequential quadratic programming approach to solving Problem (4.2.8)

Input: $\mathbf{K}_i$, $\mathbf{\Phi}_i$, $\mathbf{a}'_m$, $\mathbf{a}'_f$, λ
Output: $\boldsymbol{\gamma}$
1. Initialize $\boldsymbol{\gamma}$ with $\mathbf{u} = \mathbf{1}$, $\mathbf{v} = \mathbf{0}$, and μ_m, μ_f equal to the sample male and female means of the obtained trait when $\mathbf{w} = \mathbf{1}$.
2. Initialize $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha} = 1$.
3. Evaluate f, ∇f, ∇g_i and $\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathcal{L}$ with the current $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$.
4. Solve Problem (4.2.11) to obtain $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$.
5. Perform line search to find the learning step size s.
6. Update $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ as in Eq.(4.2.12).
Repeat 3-6 until $\boldsymbol{\gamma}$ reaches a fixed point.

optimal Lagrange multipliers of $\hat{\mathbf{p}}_t$, the searching direction of $\boldsymbol{\alpha}$ is calculated as $\hat{\mathbf{q}}_t - \boldsymbol{\alpha}_t$. Then, a line search method, such as those described in [54], can be used to determine the step size of moving along the directions. It then updates $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ as follows:

$$\boldsymbol{\gamma}_{t+1} = \boldsymbol{\gamma}_t + s\hat{\mathbf{p}}_t, \boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + s(\hat{\mathbf{q}}_t - \boldsymbol{\alpha}_t). \tag{4.2.12}$$

Algorithm 3 summarizes the SQP algorithm that we developed to solve Problem (4.2.8), and hence Problem (4.2.7).

4.2.4 Correction for covariates

As discussed in the background section, the heritability of a quantitative trait y with effects from covariates $\mathbf{z}$ is equal to the heritability of the residual ϵ of the linear model $y = \mathbf{z}^\top\mathbf{v} + \epsilon$. Therefore, our objective here is to find $\hat{\mathbf{w}}$ and $\hat{\mathbf{v}}$ that optimize the heritability estimate of $\epsilon : \epsilon = \mathbf{x}^\top\mathbf{w} - \mathbf{z}^\top\mathbf{v}$, as $y = \mathbf{x}^\top\mathbf{w}$. Let $\mathbf{Z}_{p\times n}$ be the data matrix on $\mathbf{z}$ of length p for the n subjects, the residual is calculated for all the subjects as $\boldsymbol{\epsilon} = \mathbf{X}^\top\mathbf{w} - \mathbf{Z}^\top\mathbf{v}$.

Given the data $\mathbf{Z}$ and $\mathbf{y}$, a linear regression model $y = \mathbf{z}^\top \mathbf{v}$ is typically obtained through a least squares method which has an analytical solution, $\hat{\mathbf{v}} = (\mathbf{Z}\mathbf{Z}^\top)^{-1}\mathbf{Z}\mathbf{y}$. As $\mathbf{y} = \mathbf{X}^\top \mathbf{w}$, we have $\hat{\mathbf{v}} = (\mathbf{Z}\mathbf{Z}^\top)^{-1}\mathbf{Z}\mathbf{X}^\top \mathbf{w}$ and

$$\boldsymbol{\epsilon} = (\mathbf{X}^\top - (\mathbf{Z}\mathbf{Z}^\top)^{-1}\mathbf{Z}\mathbf{X}^\top)\mathbf{w}.$$

Let $\mathbf{M} = (\mathbf{X}^\top - (\mathbf{Z}\mathbf{Z}^\top)^{-1}\mathbf{Z}\mathbf{X}^\top)^\top$, which can be pre-calculated from data, the calculation of $\boldsymbol{\epsilon}$ can be rewritten as $\boldsymbol{\epsilon} = \mathbf{M}^\top \mathbf{w}$. Then, the objective of optimizing the heritability of ϵ can be translated to finding the optimal $\mathbf{w}$ that gives an $\boldsymbol{\epsilon}$ of highest estimate of heritability. Comparing to the problem of finding $\mathbf{w}$ that gives a trait $y = \mathbf{x}^\top \mathbf{w}$ with highest possible heritability, the only difference we have here is that the design matrix has been changed from $\mathbf{X}$ to $\mathbf{M}$ for the parameters $\mathbf{w}$. Therefore, we can use the same algorithm that we have developed in the formulation section to find the $\mathbf{w}$ that optimizes the heritability of ϵ. An interesting observation in our derivation is that correcting a quantitative trait to account for covariant effects is equivalent to correcting the data matrix that used to derive the trait.

4.3 Computational Results

The proposed approach was first validated in simulations where we compared it with the current two-step approaches, i.e., estimating the two covariance matrices from pedigrees first and then solving an eigenproblem. We compared with all the three different methods that can be used to estimate the covariance matrices, which were referred to, respectively, as Ott [55], Anova [56] and ML [38]. After the proposed approach was validated in simulations, it was then used in a case study to analyze

a real-world dataset that was aggregated from genetic studies of cocaine dependence (CD) [25, 23]. Our algorithm was able to derive a quantitative trait with higher heritability than that of commonly used CD phenotypes. To show how our approach helped the association analysis, we compared the utility of the derived trait against that of the symptom-count phenotype as traits in association analysis and replicated the findings on a separate sample. The narrow sense heritability of all of the tested traits in this study was estimated by the widely-used *polygenic* function in the Sequential Oligogenic Linkage Analysis Routines (SOLAR) program [3].

4.3.1 Synthetic data

In order to make our synthetic data more realistic but with known pattern, we created the synthetic data based on family structures in the SSADDA dataset. In this dataset, there were totally 6810 subjects, of which 1915 were from small nuclear families and the rest subjects were unrelated individuals. Based on the family structures in this data, we synthesized a quantitative trait y_1 following the assumptions used in the maximum likelihood method for heritability estimation. Specifically, the values of y_1 for each family were randomly drawn from a multivariate Gaussian distribution: $N(\boldsymbol{\mu}, \boldsymbol{\Omega})$. The dimension of the distribution was determined by the number of subjects in the family, such that each dimension corresponded to an individual in the family. Note that the $\boldsymbol{\mu}$ used in the simulation of each family may vary between families according to the gender of the members. Precisely, if a family member is male, μ was set to μ_m; otherwise it was set to μ_f. The covariance matrix Ω was given as follows:

$$\boldsymbol{\Omega} = 2\sigma_a^2\boldsymbol{\Phi} + \sigma_d^2\boldsymbol{\Delta} + \sigma_e^2\mathbf{I}, \tag{4.3.1}$$

where $\mathbf{\Phi}$ and $\mathbf{\Delta}$ were composed according to Table 4.2.1. Without loss of generality, in this study we used identity matrix $\mathbf{I}$ as the matrix $\mathbf{\Gamma}$ in Eq.(4.2.1). The quantitative trait y_1 was simulated with the following choices of the parameters:

$$[\sigma_a^2, \sigma_d^2, \sigma_e^2, \mu_m, \mu_f] = [0.8,\ 0.1,\ 0.1,\ 0.9,\ 0.3]. \tag{4.3.2}$$

Hence, 80% of the phenotypic variance was due to additive genetic effects, and the ideal heritability is 0.8 according to Eq.(4.3.2). By the random nature of the simulation, the actual heritability of the simulated trait may vary a little.

In order to evaluate if our approach can correct for fixed effects of covariates, we created another quantitative trait y_2 based on y_1 by adding effects from age and race. Let α_1 and α_2 measure the effects of age and race respectively on y_2, we calculated y_2 as follows: $y_2 = y_1 + \alpha_1 \times age + \alpha_2 \times race$. The values of the two α's were arbitrarily set to $\alpha_1 = 1.1$ and $\alpha_2 = 0.7$, (which can certainly be set to any other values). Using SOLAR, we estimated the heritability of y_1 with sex as covariate ($h^2 = 0.796$) and the heritability of y_2 with sex, age, race as covariates ($h^2 = 0.797$).

We next simulated phenotypic features (i.e., the low level traits) for the two quantitative traits y_1 and y_2. We synthesized five datasets consisting of d =10, 20, 30, 40 and 50 phenotypic features respectively for each simulated trait. For each dataset, we used the following procedure to create the features $\mathbf{x}$ and assign their weights $\mathbf{w}$: the values of the first $d-1$ features were randomly drawn from the standard Gaussian distribution; then we assigned weights for each of the d features randomly; and then the values of the last feature were computed by $(y - \sum_{j=1}^{d-1} w_j x_j)/w_d$ (assuming $w_d \neq 0$). This procedure guaranteed that $y = \sum_{j=1}^{d} w_j x_j$. In practice, a multivariate trait may not depend on all of the considered phenotypic features. In order to test if

our approach can recover the relevant features, the assigned weights in $\mathbf{w}$ had 1, $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$ and $\frac{1}{5}$ of the entries nonzero, respectively, for the five datasets. For instance, for the dataset that had 40 features, only 10 of them were randomly assigned nonzero weights and other weights were set to 0.

All the three methods which we compared with employed a regularizer in their eigenproblem, so they also had a tuning parameter λ. On each dataset, the parameter λ of all methods were tuned in the same three-fold cross validation process. More specifically, for each dataset, we randomly split the sample into three groups, and each group had the same amount of unrelated individuals and families with multiple members whenever it was possible. Samples in each group were used in one of the three folds, respectively, as the validation data to test the heritability of the trait derived by a method from the rest of the samples. We repeated this three-fold cross validation with 10 random splits for each choice of the λ values on each dataset. The choices of λ were pre-specified to the range of $[0, 50]$ with a step size 1. For each method, the choice of λ that gave the best cross validated heritability was used in the subsequent analysis.

We first examined some algorithmic behavior of the proposed approach. Figure 4.3.1 shows the three-fold cross validated heritability of the traits derived by our approach from the five datasets in the experiments with y_1. On all the five datasets, the proposed method was able to identify the linearly-combined traits that were close to y_1 with a relatively wide range of λ choices. From Figure 4.3.1, we see that when $\lambda = 1$, 1, 13, 18, and 18 respectively for the five datasets, the best validation heritability was obtained. This observation shows that when the underlying model gets sparse, lager λ is favorable to prevent overfitting. We had similar observations in the experiments with y_2 as shown in Figure 4.3.2. Figure 4.3.2 plots the cross validated

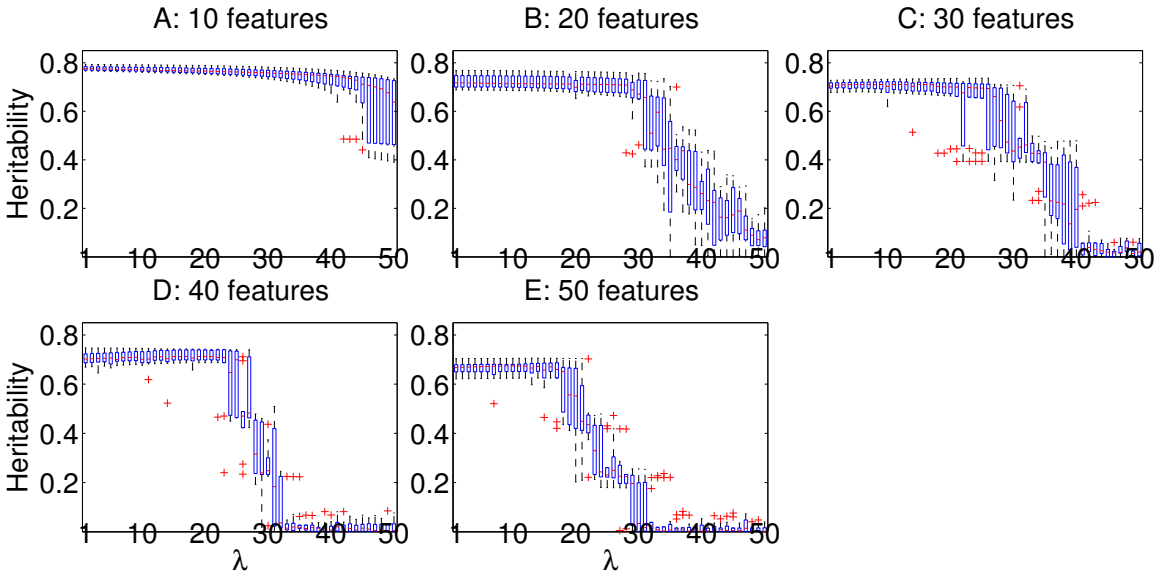

FIGURE 4.3.1: Cross validated heritability of derived quantitative traits in the experiment with simulated trait y_1 when λ varies from 0 to 50 with step size 1.

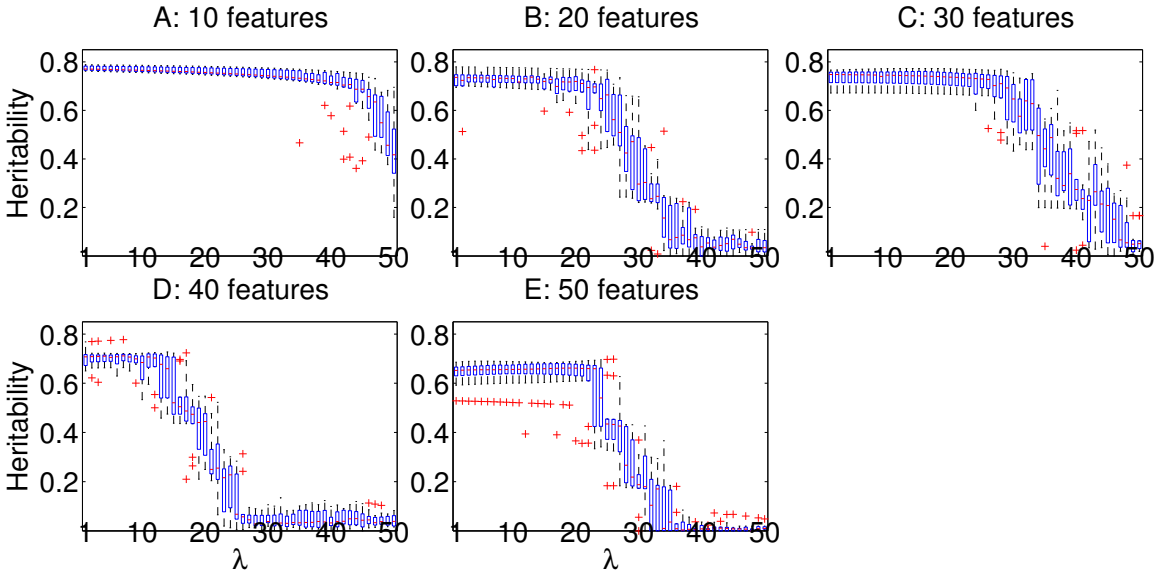

FIGURE 4.3.2: Cross validated heritability of derived quantitative traits in the experiment with simulated trait y_2 when λ varies from 0 to 50 with step size 1.

heritability obtained in the experiments with covariates and y_2. The validation heritability of the derived traits are high (with a little decrease when more irrelevant features were experimented), which demonstrates that the proposed approach can effectively correct for covariates in finding the heritable components.

TABLE 4.3.1: Cross validated heritability of the traits derived by the different methods in the experiments without and with covariates.

Method	10 features	20 features	30 features	40 features	50 features
Proposed	**0.777**(0.009)	**0.724**(0.027)	**0.707**(0.018)	**0.717**(0.021)	**0.670**(0.024)
Anova	0.638(0.063)	0.581(0.043)	0.430(0.042)	0.551(0.050)	0.447(0.060)
Ott	0.378(0.049)	0.465(0.080)	0.292(0.048)	0.398(0.036)	0.352(0.065)
ML	0.755(0.020)	0.046(0.032)	–	–	–
Proposed	0.775(0.010)	0.735(0.023)	0.738(0.030)	0.708(0.031)	0.644(0.051)
ML	0.708(0.097)	0.044(0.037)	–	–	–

The "–" sign indicates that those experiments were not conducted due to prohibitive computation cost.

We next examined the comparison of our approach against the state of the art. To be more thorough, we compared all four methods using four different metrics including validated heritablity, sum of squared residuals to the target trait y_1 or y_2 (SE(trait)), squared difference between the learned weights $\hat{\mathbf{w}}$ and the true weights $\mathbf{w}$, i.e., $||\mathbf{w} - \hat{\mathbf{w}}||^2$ (SE($\mathbf{w}$)), as well as the computation cost. Table 4.3.1 gives the cross validated heritability of the traits derived by each of the methods in the two sets of experiments with y_1 and y_2. The performance was reported when the best λ choice was used by each method. It is clear that the traits derived by our approach always achieved the highest heritability.

Table 4.3.2 compares the values of SE(trait), SE($\mathbf{w}$), and the computation time in seconds. In particular, the computational cost was measured by running each of the methods on the full datasets without splits when the best λ value was used. Across all

TABLE 4.3.2: Comparison of the methods on the sum of squared residuals (SE(trait)), squared difference of the true weights and the learned weights (SE($\mathbf{w}$)), and the computation time (in seconds) in the experiments without and with covariates.

Dataset	SE(trait)				SE($\mathbf{w}$)				Computation Time (sec.)			
	Proposed	Anova	Ott	ML	Proposed	Anova	Ott	ML	Proposed	Anova	Ott	ML
10 features	**10.89**	59.03	67.44	57.97	**0.09**	1.35	1.38	1.34	0.61	0.17	0.11	8.24e+02
20 features	**16.62**	60.83	63.08	128.01	**0.17**	1.37	1.39	2.54	0.85	0.19	0.15	1.16e+04
30 features	**19.69**	63.03	72.46	–	**0.21**	1.38	1.48	–	0.90	0.19	0.14	–
40 features	**23.31**	62.71	68.39	–	**0.27**	1.39	1.44	–	0.98	0.29	0.23	–
50 features	**25.23**	64.22	67.23	–	**0.29**	1.40	1.43	–	2.13	0.30	0.26	–
10 features	**13.61**	*	*	85.98	**0.11**	*	*	1.35	0.86	*	*	8.85e+02
20 features	**16.14**	*	*	173.40	**0.18**	*	*	2.58	1.07	*	*	1.20e+04
30 features	**26.60**	*	*	–	**0.31**	*	*	–	1.30	*	*	–
40 features	**26.81**	*	*	–	**0.29**	*	*	–	1.61	*	*	–
50 features	**25.87**	*	*	–	**0.31**	*	*	–	2.52	*	*	–

The "–" sign indicates that those experiments were not conducted due to prohibitive computation cost. The "$*$" sign indicates that the corresponding methods were not tested due to the limitation of the methods that could not handle covariates. The computation time reported for the ML method was measured when the maximum number of iterations was set to 200.

of the datasets, our approach obtained the smallest errors among all the methods as measured by SE(trait) and SE($\mathbf{w}$). Because Anova used analytic formula to compute covariance matrices, and Ott used a single locus in the covariance estimation, both methods required slightly less computation cost than our approach. However, they were limited only to the situations that had no confounding factors in heritability calculation. Between the two more comprehensive methods, our approach was significantly more efficient than the ML method in computation, making the heritable component analysis with a large number of phenotypic features feasible.

Our approach could identify multivariate traits of much higher heritability than the commonly used traits. We compared the heritability of the traits derived by our approach with that of individual phenotypic features and the simple average of all features. We used the traits derived by our approach from the cross validation process when the best λ values were used. As shown in Figure 4.3.3 (without covariates) and 4.3.4 (with covariates), the validation heritability of the derived traits were significantly higher than the heritability of any individual feature and the trait by averaging all features.

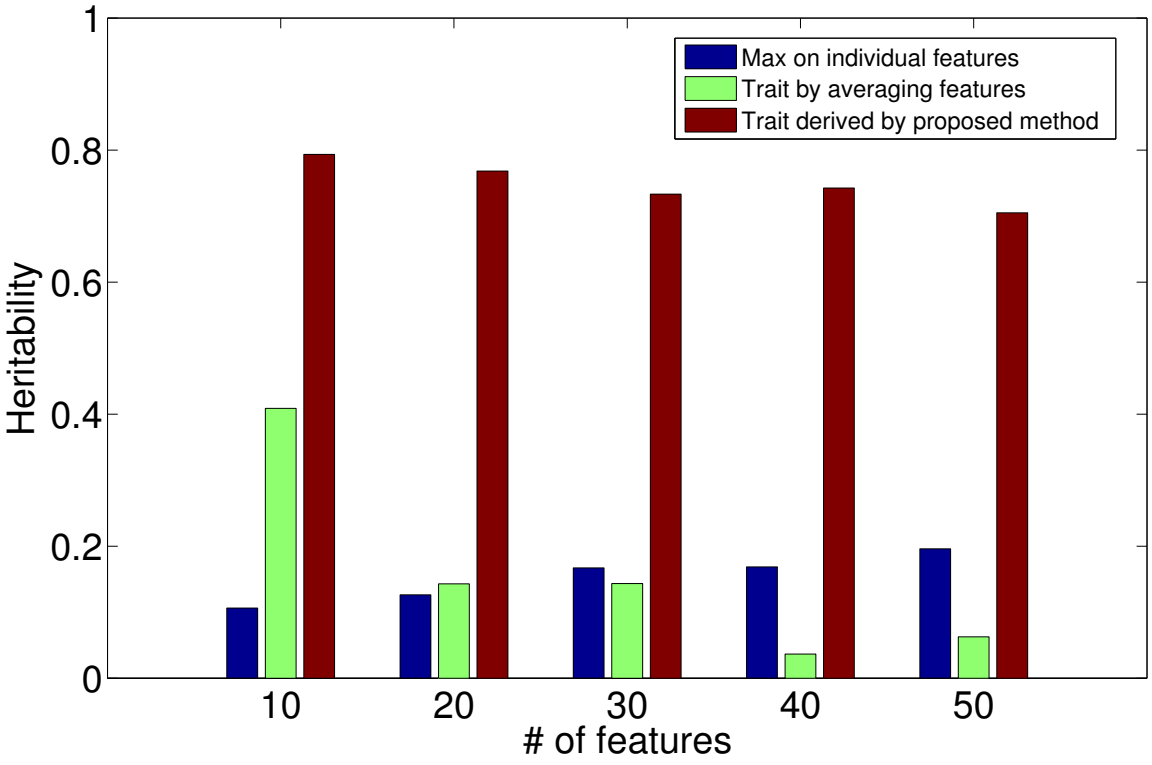

FIGURE 4.3.3: Comparison between the trait derived by the proposed approach, individual features and the simple average of features in the experiment with simulated trait y_1. Heritability was estimated without taking into account covariates. The maximum among the heritabilities of individual features is taken and shown in the figure.

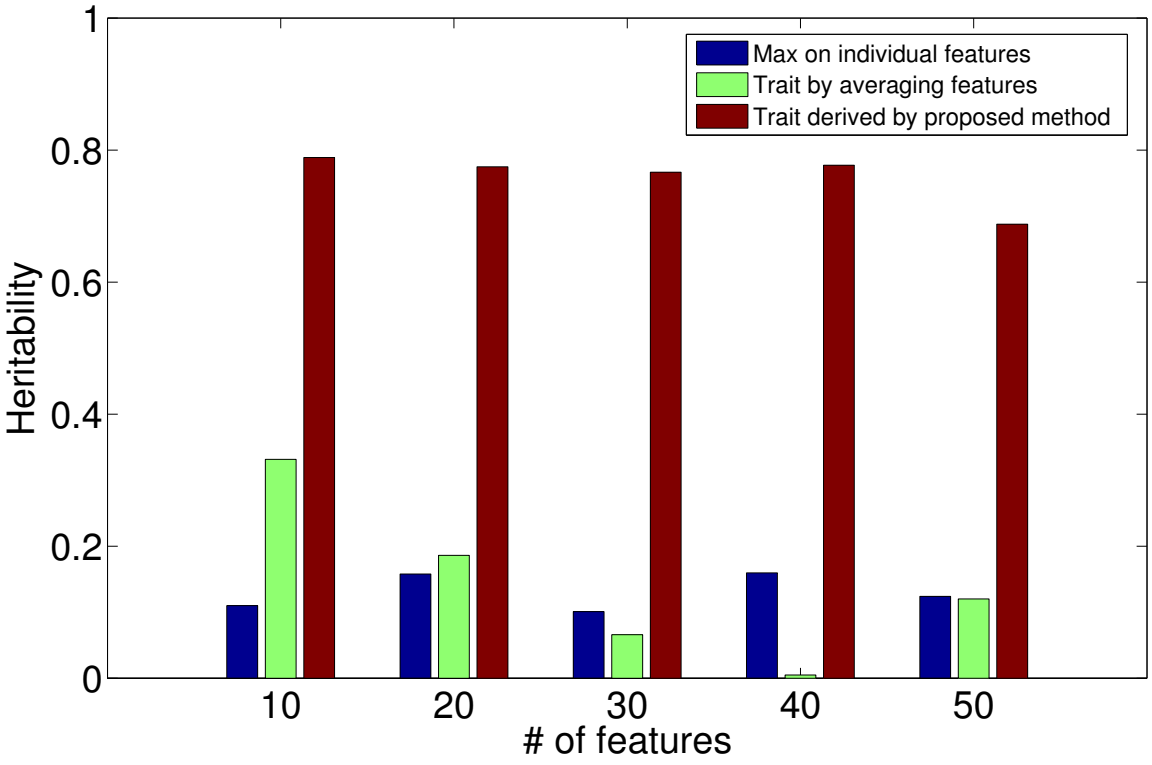

FIGURE 4.3.4: Comparison between the trait derived by the proposed approach, individual features and the simple average of features in the experiment with simulation trait y_2. Heritability was estimated taking into account the covariate effects. The maximum among the heritabilities of individual features is taken and shown in the figure.

4.3.2 A case study: cocaine use and related Behaviors

We applied the proposed approach to a real-world problem represented by genetic studies of cocaine use and related behaviors. Two independent sets of samples were used in our analysis: the *Semi-Structured Assessment for Drug Dependence and Alcoholism* (SSADDA) dataset [25], which was used for discovery and was the one our simulation study in previous section based on; and the *Study of Addiction: Genetics and Environment* (SAGE) dataset [52], which was used for replication. The SSADDA subjects were recruited from multiple sites, including the University of Connecticut Health Center, Yale University School of Medicine, the University of Pennsylvania School of Medicine, McLean Hospital and the 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. A total of 6810 subjects were used in our analysis, including 1922 individuals from small nuclear families, with the remainder being unrelated individuals. We included unrelated individuals in our analysis to estimate phenotypic variance even though they had no effect on the covariance estimate. The SAGE data were aggregated from multiple NIH-funded projects [10], and downloaded from a public domain [52]. It consisted of 58 individuals from nuclear families and 1603 unrelated individuals. All subjects in the two datasets were from one of two populations: African American (AA) or European American (EA).

All involved subjects were reported to have used cocaine in their lives, which were assessed on the following 13 features of cocaine use and related behaviors:

- *F1* - tolerance to cocaine;
- *F2* - withdrawal from cocaine;
- *F3* - using cocaine in larger amounts or over longer period than intended;

- *F4* - persistent desire or unsuccessful efforts to cut down or control cocaine use;
- *F5* - great amount of time spent in activities necessary to obtain, use or recover from the effects of cocaine;
- *F6* - gave up or reduced important social, occupational, or recreational activities because of cocaine use;
- *F7* - cocaine use despite knowledge of persistent or recurrent physical or psychological problems likely to have been caused or exacerbated by cocaine;
- *F8* - number of cocaine symptom endorsed;
- *F9* - age when first used cocaine;
- *F10* - age when last used cocaine;
- *F11* - age when first being diagnosed with DSM4 cocaine dependence;
- *F12* - age when last being diagnosed with DSM4 cocaine dependence;
- *F13* - transition time in years between the first time cocaine use and the first cocaine dependence diagnosis.

Features *F1-7* were binary variables that take a value of "yes=1" or "no=0". *F8-13* were continuous variables, which we normalized to the range of $[0, 1]$ in the analysis.

The majority of the 6810 subjects interviewed with SSADDA, were genotyped on a illumina microarray for 988,306 autosomal single-nucleotide polymorphisms (SNPs). Genotypes for additional 37,427,733 SNPs were imputed using IMPUTE2 [35] with genotyped SNPs and 1000 Genomes reference panel released in June 2011 (http://www.1000genomes.org). Both subjects and SNPs were under through stringent quality control (refer to [25] for details). There were in total 4,845 subjects (2674 AAs, 2171 EAs) and 30,078,279 SNPs (695,308 genotyped) remained for analysis. Top three ancestral principal components were computed using 145,472 SNPs

that were common to discovery samples and Hapmap panel. All of the 1661 subjects (640 AAs, 1021 EAs) in replication dataset were genotyped for 1,072,657 SNPs.

We derived a composite trait based on the 13 features of cocaine use and related behaviors. This trait was derived using discovery data and Algorithm 1 with correction of the effects of covariates age and race. As in simulation study three-fold cross validation was performed to find optimal λ, which was subsequently used to find a linear combination of the 13 features and derive a trait using the entire discovery data. The heritability of obtained trait was estimated and compared to that of individual quantitative features in the data, including cocaine symptom count(F8), which was recognized as a better trait than the binary cocaine dependence diagnosis in the study hunting for genetic risk factors [25]. We applied the learned linear model to replication data. The heritability of derived trait on replication samples was not estimated because 97% of the sample consisted of unrelated individuals. Associate tests were firstly performed on discovery samples for both derived trait and cocaine symptom count. The associations with p-value less than 5×10^{-6} were further tested using replication samples. Association tests were performed separately for AA and EA. All tests included age, sex and the first three ancestral principle components as covariates. The association test results on discovery and replication data were combined by performing meta analysis using Metal [81].

The heritability of generated traits for λs from 1 to 50 with step size 1 is plotted in Figure 4.3.5. When $\lambda = 2$, the resulted traits have the highest heritability on average in the cross validation comparing to other choices of λ. Hence, we chose $\lambda = 2$ and used it in Algorithm 1, which was then applied to the entire discovery sample to obtain a composite trait and its linear model. The heritability of resulted trait together with that of all individual quantitative features are reported in Table 4.3.3.

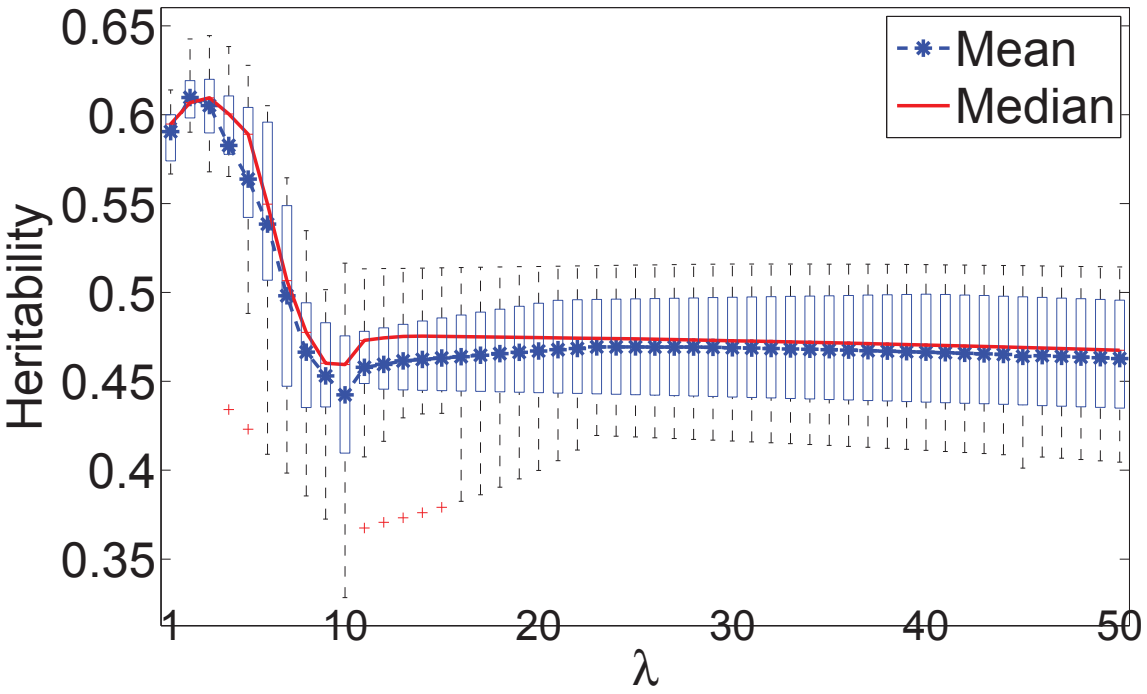

FIGURE 4.3.5: Validation heritability of the composite traits derived by our approach for cocaine use and related behaviors using different values of λ.

TABLE 4.3.3: Heritability estimates for the composite trait derived by the proposed method and all individual quantitative features in the data.

Traits	heritability	p-value	standard deviation
Trait derived by proposed method	0.70	4.36×10^{-22}	0.06
Cocaine symptom count	0.41	1.52×10^{-08}	0.07
Age when first used cocaine	0.39	2.41×10^{-09}	0.07
Age when last used cocaine	0.45	6.70×10^{-06}	0.10
Age when first CD diagnosis	0.43	1.15×10^{-10}	0.07
Age when last CD diagnosis	0.38	5.99×10^{-09}	0.07
Transition time between first cocaine use and CD diagnosis	0.42	8.09×10^{-10}	0.07

The composite trait derived by our approach has the highest heritability estimate among all the traits.

Due to the use of the sparsity-favoring ℓ_1-norm regularization, our approach selects features to use in the linear model. Figure 4.3.6 shows the weights of the features obtained in our model. Five of 13 features had weight 0, thus were removed from the model. The feature of age when first used cocaine received the largest positive weights and therefore had the largest impact on the obtained trait. The other four features that had significant impact on the obtained trait were *F11* - age onset of DSM4 CD diagnosis; *F4* - persistent desire or unsuccessful efforts to cut down or control cocaine use; *F5* - great amount of time spent in activities necessary to obtain, use or recover from the effects of cocaine; and *F3* - using cocaine in larger amounts or over longer period than intended. Features *F6*, *F1* and *F2* had limited effect on the resulted trait. The values of the derived trait computed on the discovery sample are shown in Figure 4.3.7.

We identified three SNPs for the AA population and four SNPs for the EA population that passed our p-value threshold in the genomewide association tests with

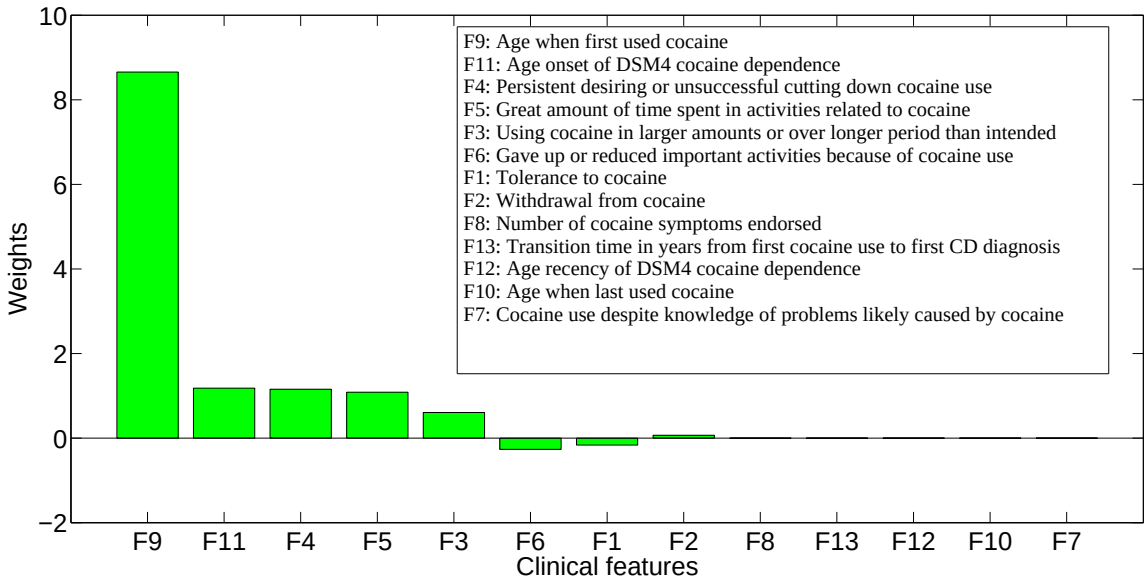

FIGURE 4.3.6: Weights of the eight clinical features in the linear model of the composite trait derived by our approach to the evaluation of cocaine use and related behaviors.

the discovery sample. The top findings are listed in Table 4.3.4. These SNPs were found significantly or nominally significantly associated with the trait derived by the proposed approach at the genomewide level, but not significantly associated with the cocaine symptom count. In other words, using the standard phenotype in the association tests would not discover these SNPs that are associated with a specific subtype (a quantitative subphenotype) of cocaine dependence. Among the seven SNPs in Table 4.3.4, four (one: rs833936 for AA, three: rs11079045, rs7224135 and rs10490394 for EA) were significantly associated with the derived trait in the replication study with meta-analysis p-values $< 10^{-7}$. The marker rs833936 is located at the *TXNIP* gene whose expression is suppressed by synaptic activity in brain [6]. Both markers rs11079045 and rs7224135 are located at the *PTRF* gene which was identified to be associated with cocaine abuse in an early transcriptional change study [47]. The

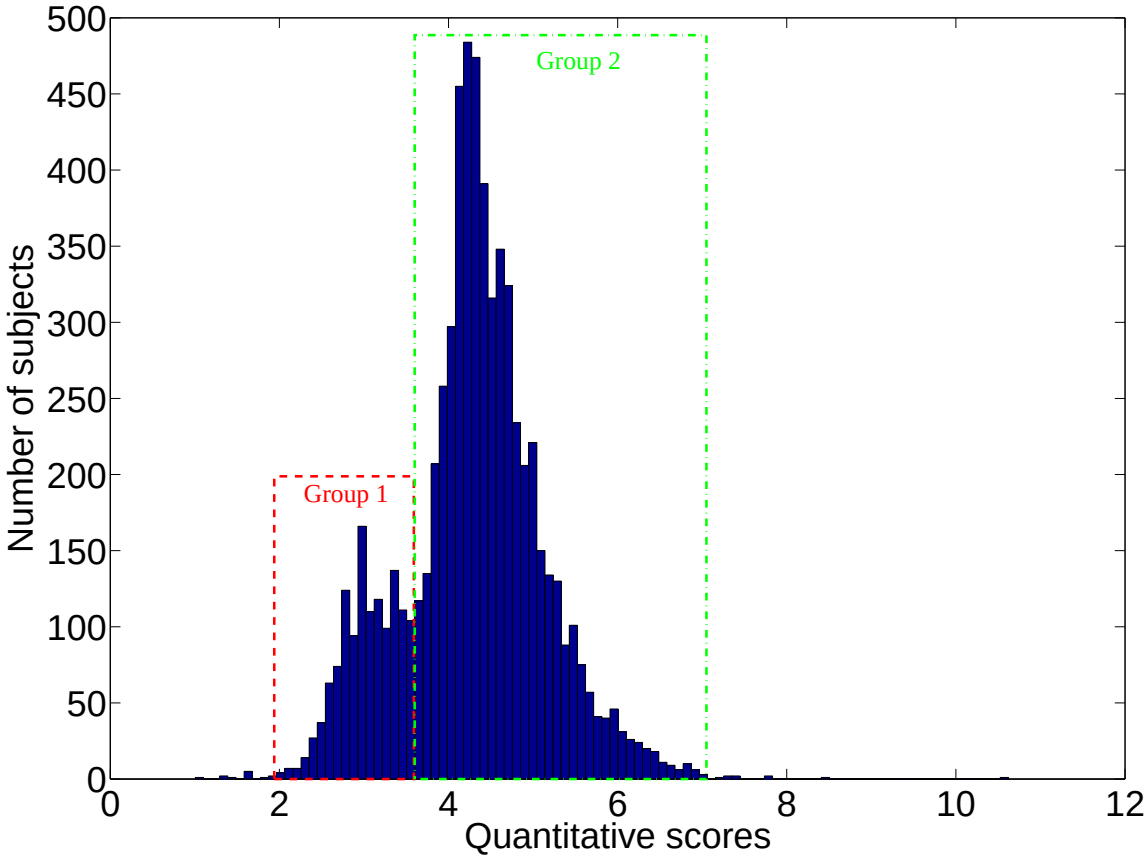

FIGURE 4.3.7: Distribution of the trait values computed on the discovery sample.

EFEMP1 gene has not been reported in the genetic analysis of cocaine dependence. Since all the identified SNP markers have not been thoroughly studied in genetics of cocaine dependence, our findings may promote subsequent investigations for these genes as well as subtypes of cocaine dependence. The proposed heritable component analysis for multivariate phenotypes may provide a new strategy to improve genomewide association studies of complex disorders.

TABLE 4.3.4: Top findings in genome-wide association study.

	SNP	Chr	Gene	Discovery			Replication			Meta	
				MAF	$p_{derived}$	p_{symp}	MAF	$p_{derived}$	p_{symp}	$p_{derived}$	p_{symp}
AA	rs769065	6	*DNAH8*	0.26	6.14×10^{-6}	9.62×10^{-2}	0.03	8.74×10^{-3}	3.58×10^{-2}	1.85×10^{-7}	1.57×10^{-2}
	rs833936	1	*TXNIP*	0.36	7.90×10^{-8}	2.51×10^{-2}	0.12	2.22×10^{-2}	1.76×10^{-2}	$\mathbf{5.59 \times 10^{-9}}$	2.43×10^{-3}
	rs75621732	11	*MLSTD2*	0.06	1.89×10^{-6}	1.85×10^{-1}	0.35	4.95×10^{-2}	5.60×10^{-1}	2.70×10^{-7}	1.48×10^{-1}
EA	**rs11079045**	17	*PTRF*	0.40	2.48×10^{-6}	2.24×10^{-1}	0.42	1.48×10^{-3}	2.24×10^{-1}	$\mathbf{1.33 \times 10^{-8}}$	1.82×10^{-1}
	rs7224135	17	*PTRF*	0.40	7.61×10^{-7}	1.50×10^{-1}	0.41	2.29×10^{-3}	1.50×10^{-1}	$\mathbf{6.51 \times 10^{-9}}$	1.08×10^{-1}
	rs10490394	2	*EFEMP1*	0.20	8.78×10^{-7}	1.53×10^{-1}	0.19	9.15×10^{-3}	1.53×10^{-1}	$\mathbf{3.22 \times 10^{-8}}$	2.33×10^{-1}
	rs7330895	13	*DACH1*	0.39	7.50×10^{-6}	6.00×10^{-2}	0.34	2.81×10^{-2}	6.00×10^{-2}	8.00×10^{-7}	2.80×10^{-3}

Chr, chromosome; MAF, minor allele frequency; $p_{derived}$, p-value on the trait derived by the proposed method; p-value on cocaine symptom count. SNPs: rs769065, rs833936 and rs75621732 were tested using their close linkage equilibrium (LD) proxies: rs12215108 (602 bp downstream), rs4636400 (1004 bp upstream) and rs1730833 (80 bp upstream), respectively in replication data, because these SNPs were not available in this data. For these SNPs, the corresponding MAFs in the table belong to their proxies. SNPs and p-values that reach genome-wide significant level ($< 5 \times 10^{-8}$) are in bold font.

4.4 Summary

In this chapter, we have proposed a quadratic optimization formulation that is capable of identifying highly heritable composite traits of complex phenotypes. The composite trait is derived as a linear function $y = \mathbf{x}^\top \mathbf{w}$ of lower level traits $\mathbf{x}$ by explicitly maximizing its heritability. Specifically, we search for the optimal $\mathbf{w}$ that maximizes the likelihood of observing a high value of heritability. This is equivalent to finding the best $\mathbf{w}$, so that the projected trait $\mathbf{x}^\top \mathbf{w}$ will be best aligned with the kinship matrix $\mathbf{\Phi}$ of the pedigree. An efficient algorithm based on sequential quadratic programming has been developed to optimize the proposed formulation. The algorithm is extended to allow the correction for covariate effects when deriving a composite trait.

Our simulation study proves the effectiveness of the proposed approach as a means to find highly heritable composite traits. Then a case study on the complex phenotypes of cocaine use and dependence was conducted. A composite trait was identified based on eight cocaine use symptoms and behaviors. The trait had a heritability estimate of 0.52 (with $p = 6.47 \times 10^{-12}$, std = 0.07), which was much higher than a standard cocaine-use phenotype, i.e., the symptom-count trait, with heritability of 0.4. The subsequent phenotype-genotype association study demonstrated greater utility of the derived trait than the standard phenotype for use in association analysis. Our results show that three of the four associated SNPs were more significantly associated with the derived trait, and two of them were replicated in an independent data set.

Our formulation has a hyper-parameter λ. Using a hyper-parameter is common in machine learning algorithms such as support vector machines [73]. As a hyper-parameter, λ is not determined by solving the formulation itself and instead needs

to be pre-specified. Both our simulation study and our case study showed that our formulation is fairly robust to the value of λ when it is chosen from a reasonably wide range. In real-world applications, hyper-parameters are often determined by a cross-validation process, which was used in our experiments.

The proposed approach may be useful to enhance animal or plant breeding programs that aim to improve genetic selection of a conceptual, economically important phenotypes such as feed efficiency or heat tolerance. Our approach can be used to derive new composite traits by combining a variety of lower-level traits that are used to measure the conceptual phenotype. These new composite traits will have much higher heritability than that of currently used. The heritability of the breeding trait is considered to be one of the most important factors for the performance of a breeding program.

There are limitations of our proposed technique. The non-convex quadratic optimization formulation requires a complex solver, such as sequential quadratic programming. For a sample that contains millions of subjects, it may become computationally prohibitive. More efficient solvers or approximations may be needed to scale up the proposed approach. In some applications, complex grouping structures may exist in the data between different lower-level traits. A formulation that takes into account the special data structure may be more useful in producing biologically and clinically meaningful traits. As discussed in the chapter, alternative regularization conditions exist, including some that may deal well with complex data structures, such as the one based on $\ell_{2,1}$ vector norm. Algorithms that can solve the formulations with alternative regularization terms need to be developed. Additional empirical studies across different disciplines are needed to evaluate the power of the proposed approach.

Chapter 5

Identifying Heritable Composite Traits with Genome-Wide SNPs for Complex Phenotypes

5.1 Introduction

In order to use the approaches we have described in Chapter 4 to identify heritable subtypes from phenotypic features, family members are required in the study. However, on one hand it is difficult to collect data for multi-member families in large scale. For example, in the largest GWAS for CD to date [25], only around 1,200 subjects with African American ancestry are from nuclear families, while there are in total 4,121 African Americans in the study. On the other hand, with the availability of dense genotype data, now it is possible to estimate genetic relationship among subjects using genome-wide SNPs, and also the narrow-sense heritability, h^2 of a trait under interest. The h^2 estimated from unrelated individuals using genome-wide

single nucleotide polymorphisms (SNPs) is so called "chip heritability". In fact, it has been argued that estimating h^2 from unrelated individuals has advantage over traditional pedigree-based approaches because the estimated h^2 corresponds only to the causal-variant heritability that is tagged by genotyped SNPs [83, 65]. Thus, it is of great interest to develop a method that can identify heritable composite trait from phenotypic variables using unrelated individuals and their genotype data for genome-wide SNPs.

In this chapter, we propose such an approach. To estimate chip heritability of a given trait, recently-published methods use the restricted maximum likelihood (REML) method assuming the trait follows a mixed effect model with random genetic effects and fixed effects due to covariates [83, 65]. To identify a trait of high chip heritability, we solve the inverse problem of (chip) heritability estimation. In other words, we search for a linearly-combined trait when estimating the trait's chip heritability using the REML method, the estimate is high. Directly solving the inverse problem leads to a quadratic optimization problem that can be optimized efficiently via a sequential quadratic programming algorithm. We validate the proposed approach on both synthetic and real world data. Our experimental results show the effectiveness and generalizability of the proposed approach.

5.2 Method

5.2.1 Background: chip heritability estimation

Given a set of n subjects, we use a vector $\mathbf{y}$ of length n to denote their trait values for a quantitative trait y, a matrix $\mathbf{Z}_{n\times m}$ to represent their standardized genotypic

data at m genetic variants, and $\mathbf{C}_{n\times p}$ to represent their data on p covariates. The matrix $\mathbf{Z}$ is calculated from the genotypic data as follows. Let f_j be the frequency of reference allele at the j-th genetic variant, r_{ij} be the number of copies of reference allele that i-th subject has at j-th variant, and then the standardized genotype z_{ij} is calculated as $(r_{ij} - 2f_j)/\sqrt{2f_j(1-f_j)}$. The well known chip heritability estimation method [83] considers the following linear mixture model that characterizes how a phenotype is related to genotypes and covariates:

$$\mathbf{y} = \mathbf{C}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{5.2.1}$$

where $\boldsymbol{\varepsilon}$ is a vector of length n, which specifies residual effects. In Eq.(5.2.1), all covariates have fixed effects (fixed $\boldsymbol{\beta}$) on the phenotype whereas genetic effects are random (random $\mathbf{u}$). Assume $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ follow Gaussian distributions: $\mathbf{u} \sim N(\mathbf{0}, \mathbf{I}\sigma_u^2)$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$. Then, the variance of $\mathbf{y}$, denoted by $\boldsymbol{\Omega}$, can be calculated as:

$$\boldsymbol{\Omega} = \mathbf{Z}\mathbf{Z}^T\sigma_u^2 + \mathbf{I}\sigma_e^2. \tag{5.2.2}$$

Let σ_g^2 be the phenotypic variance attributable to all m genetic causal variants in $\mathbf{Z}$. Then, we have $\sigma_g^2 = m\sigma_u^2$. Let $\mathbf{G} = \mathbf{Z}\mathbf{Z}^T/m$ be the genetic relationship matrix (GRM) among subjects determined by the causal variants, Eq. (5.2.2) can be re-written as:

$$\boldsymbol{\Omega} = \mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_e^2. \tag{5.2.3}$$

σ_g^2 and σ_e^2 can be estimated by the REML method [58, 84]; and the heritability contributed by the m causal variants is computed as $h^2 = \sigma_g^2/\sigma_p^2$, where $\sigma_p^2 = \sigma_g^2 + \sigma_e^2$ is the overall phenotypic variance. Because the causal variants are usually unknown

for a trait, recent research has proposed to use genome-wide SNPs in estimating a GRM [83, 65].

The main idea of REML is to project $\mathbf{y}$ with a set of n basis represented by columns in a matrix $\mathbf{L}_{n\times n}$. This matrix can be decomposed into two matrices: $\mathbf{L}_{1n\times p}$ and $\mathbf{L}_{2n\times(n-p)}$ with $\mathbf{L} = [\mathbf{L}_1\ \mathbf{L}_2]$, $\mathbf{L}_1^T\mathbf{C} = \mathbf{I}_{p\times p}$ and $\mathbf{L}_2^T\mathbf{C} = 0$. Then $\mathbf{L}^T\mathbf{y}$ follows the multivariate Gaussian distribution as following:

$$\begin{bmatrix}\mathbf{L}_1^T\mathbf{y}\\ \mathbf{L}_2^T\mathbf{y}\end{bmatrix} \sim N\left(\begin{bmatrix}\boldsymbol{\beta}\\ 0\end{bmatrix}, \begin{bmatrix}\mathbf{L}_1^T\boldsymbol{\Omega}\mathbf{L}_1 & \mathbf{L}_1^T\boldsymbol{\Omega}\mathbf{L}_2\\ \mathbf{L}_2^T\boldsymbol{\Omega}\mathbf{L}_1 & \mathbf{L}_2^T\boldsymbol{\Omega}\mathbf{L}_2\end{bmatrix}\right).$$

Let $\tilde{\mathbf{y}} = \mathbf{L}^T\mathbf{y}$, $\tilde{\mathbf{y}}_1 = \mathbf{L}_1^T\mathbf{y}$ and $\tilde{\mathbf{y}}_2 = \mathbf{L}_2^T\mathbf{y}$, we have $\tilde{\mathbf{y}}_2 \sim N(0, \mathbf{L}_2^T\boldsymbol{\Omega}\mathbf{L}_2)$ and the conditional distribution:

$$\tilde{\mathbf{y}}_1|\tilde{\mathbf{y}}_2 \sim N\big(\boldsymbol{\beta} + \mathbf{L}_1^T\boldsymbol{\Omega}\mathbf{L}_2(\mathbf{L}_2^T\boldsymbol{\Omega}\mathbf{L}_2)^{-1}\tilde{\mathbf{y}}_2, (\mathbf{C}^T\boldsymbol{\Omega}\mathbf{C})^{-1}\big)$$

Then, the log likelihood of $\tilde{\mathbf{y}}$ can be decomposed into:

$$\ell(\sigma_g^2, \sigma_e^2; \tilde{\mathbf{y}}) = \ell_2(\sigma_g^2, \sigma_e^2; \tilde{\mathbf{y}}_2) + \ell_1(\sigma_g^2, \sigma_e^2; \tilde{\mathbf{y}}_1|\tilde{\mathbf{y}}_2).$$

Notice that here, ℓ_2 is not a function of $\boldsymbol{\beta}$. The two variance components, i.e., σ_g^2 and σ_e^2 are estimated by maximizing ℓ_2, because $\boldsymbol{\beta}$ and $\tilde{\mathbf{y}}_1$ are of equal length, so there is no further information in ℓ_2 for estimating the variance components.
ℓ_2 is calculated as (excluding constants, i.e., terms without σ_g^2 and σ_e^2):

$$\ell_2(\sigma_g^2, \sigma_e^2; \tilde{\mathbf{y}}_2) = -\frac{1}{2}(\ln|\mathbf{L}_2^T\boldsymbol{\Omega}\mathbf{L}_2| + \tilde{\mathbf{y}}_2^T(\mathbf{L}_2^T\boldsymbol{\Omega}\mathbf{L}_2)^{-1}\tilde{\mathbf{y}}_2).$$

It has been shown in [75] that when $\mathbf{L}_1^T\mathbf{C} = \mathbf{I}_{p\times p}$ and $\mathbf{L}_2^T\mathbf{C} = 0$, we have:

$$\mathbf{\Omega} - \mathbf{\Omega}\mathbf{L}_2(\mathbf{L}_2^T\mathbf{\Omega}\mathbf{L}_2)^{-1}\mathbf{L}_2^T\mathbf{\Omega} = \mathbf{C}(\mathbf{C}^T\mathbf{\Omega}^{-1}\mathbf{C})^{-1}\mathbf{C}^T.$$

With this equality, the calculation of ℓ_2 can be written as:

$$\ell_2(\sigma_g^2, \sigma_e^2; \tilde{\mathbf{y}}_2) = -\frac{1}{2}(\ln|\mathbf{\Omega}| + \ln|\mathbf{C}^T\mathbf{\Omega}^{-1}\mathbf{C}| + \mathbf{y}^T\mathbf{P}\mathbf{y}), \tag{5.2.4}$$

where $\mathbf{P} = \mathbf{\Omega}^{-1} - \mathbf{\Omega}^{-1}\mathbf{C}(\mathbf{C}^T\mathbf{\Omega}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{\Omega}^{-1}$. Maximizing Eq.(5.2.4) leads to the estimates of σ_g^2 and σ_e^2, which are denoted by $\hat{\sigma_g}^2$ and $\hat{\sigma_e}^2$, respectively [84]. Once σ_g^2 and σ_e^2 are estimated, a generalized least square estimate of $\boldsymbol{\beta}$ can be obtained as:

$$\hat{\boldsymbol{\beta}} = \tilde{\mathbf{y}}_1 - \mathbf{L}_1^T\mathbf{\Omega}\mathbf{L}_2(\mathbf{L}_2^T\mathbf{\Omega}\mathbf{L}_2)^{-1}\tilde{\mathbf{y}}_2,$$

which can be reduced to:

$$\hat{\boldsymbol{\beta}} = (\mathbf{C}^T\mathbf{\Omega}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{\Omega}^{-1}\mathbf{y}. \tag{5.2.5}$$

Also the chip heritability of the trait y can then be computed using the two variance estimates, i.e., $\hat{\sigma_g}^2$ and $\hat{\sigma_e}^2$.

5.2.2 Proposed problem formulation

When estimating chip heritability, we find optimal $\hat{\sigma_g}^2$ and $\hat{\sigma_e}^2$ that maximizes Eq.(5.2.4) with given y, $\mathbf{C}$ and $\mathbf{Z}$. However, in the inverse problem, a definitive quantitative trait y is not known beforehand but needs to be derived from a set of phenotypic

variables. Let $\mathbf{X}_{n\times d}$ be the data matrix of d phenotypic variables for the same n subjects as in $\mathbf{Z}$, a trait y is defined by a linear function $\mathbf{y} = \mathbf{X}\mathbf{w}$. Unlike the heritability estimation process that finds the best values of σ_g^2 and σ_e^2 to maximize the likelihood of observing the values of y, the inverse problem searches for the best $\mathbf{w}$ so to form a trait y that maximizes the likelihood, (or equivalently the log likelihood $\ell(\sigma_g^2, \sigma_e^2; \mathbf{y}, \mathbf{C}, \mathbf{Z})$), of observing a large heritability, i.e., a large σ_g^2 but small σ_e^2. For simplicity and easy interpretation of the resultant model, here we only consider linear models, but the proposed method can be easily extended to construct non-linear models through kernel mapping [74].

Notice that the highest possible heritability of a trait y is 1 when $\sigma_g^2 = 1$ and $\sigma_e^2 = 0$. We propose to formulate an optimization problem, in which we search for an optimal $\mathbf{w}$ that maximizes $\ell(\sigma_g^2, \sigma_e^2; \mathbf{y}, \mathbf{C}, \mathbf{Z})$ where $\mathbf{y} = \mathbf{X}\mathbf{w}$, and $\sigma_g^2 = 1$ and $\sigma_e^2 = 0$. Substituting the values of these parameters into the log likelihood and removing constants yield the following objective function:

$$\min_{\mathbf{w}} \quad \mathbf{w}^T(\mathbf{X}^T\mathbf{P}\mathbf{X})\mathbf{w} \tag{5.2.6}$$

where $\mathbf{P}$ is calculated as:

$$\mathbf{P} = \mathbf{G}^{-1} - \mathbf{G}^{-1}\mathbf{C}(\mathbf{C}^T\mathbf{G}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{G}^{-1}. \tag{5.2.7}$$

Since $\sigma_g^2 = 1$ and $\sigma_e^2 = 0$, the phenotypic covariance matrix $\mathbf{\Omega} = \mathbf{G}$ (based on Eq.(5.2.3)).

Because $\sigma_p^2 = \sigma_g^2 + \sigma_e^2$, when $\sigma_g^2 = 1$ and $\sigma_e^2 = 0$, we have $\sigma_p^2 = 1$. This requires to impose a constraint to the optimization problem so that the total phenotypic variance

that is due to either genetic or environmental effect should be 1. An estimate of σ_p^2 can be obtained by calculating the sample variance after correcting for the covariates effect as follows:

$$\hat{\sigma}_p^2 = \frac{1}{n}(\mathbf{X}\mathbf{w} - \mathbf{C}\boldsymbol{\beta})^T(\mathbf{X}\mathbf{w} - \mathbf{C}\boldsymbol{\beta})$$

As an estimate of $\boldsymbol{\beta}$ can be calculated as in Eq.(5.2.5), substituting its value and also letting

$$\mathbf{J} = \mathbf{I} - \mathbf{C}(\mathbf{C}^T\boldsymbol{\Omega}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\boldsymbol{\Omega}^{-1},$$

the calculation of $\hat{\sigma}_p^2$ can be written as $\hat{\sigma}_p^2 = (1/n)\mathbf{w}^T\mathbf{X}^T(\mathbf{J}^T\mathbf{J})\mathbf{X}\mathbf{w}$. To further simplify the notation, letting

$$\mathbf{Q} = (1/n)\mathbf{J}^T\mathbf{J}, \tag{5.2.8}$$

we have $\hat{\sigma}_p^2 = \mathbf{w}^T(\mathbf{X}^T\mathbf{Q}\mathbf{X})\mathbf{w}$. Combining the objective function and the constraint together, the proposed optimization problem is formulated as:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^T(\mathbf{X}^T\mathbf{P}\mathbf{X})\mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^T(\mathbf{X}^T\mathbf{Q}\mathbf{X})\mathbf{w} = 1 \end{aligned} \tag{5.2.9}$$

We now regularize the linear model by introducing a regularizer on $\mathbf{w}$ which aims to avoid the overfitting problem. If overfitting occurs, the optimal $\mathbf{w}$ of Problem (5.2.9) may correspond to a trait that has high heritability on the data that is used to train the linear model, but when the model is applied to a new sample, the resultant trait has low heritability. In order to prevent overfitting and identify a trait with high heritability that can generalize, we incorporate a regularizer $R(\mathbf{w})$ in the formulation.

The optimization problem becomes:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^T(\mathbf{X}^T\mathbf{P}\mathbf{X})\mathbf{w} + \lambda R(\mathbf{w}) \\ \text{subject to} \quad & \mathbf{w}^T(\mathbf{X}^T\mathbf{Q}\mathbf{X})\mathbf{w} = 1, \end{aligned} \tag{5.2.10}$$

where λ is a hyper-parameter and needs to be pre-determined. It can either be chosen by users according to domain knowledge or determined using cross-validation as in the experiments conducted in this paper. According to learning theory [74], minimizing $\mathbf{w}^T(\mathbf{X}^T\mathbf{P}\mathbf{X})\mathbf{w}$ corresponds to empirical risk minimization, whereas minimizing the objective in Eq.(5.2.10) corresponds to structural risk minimization that improves the generalizability of the resultant model. There are many different ways to realize $R(\mathbf{w})$ [68]. The $\|\mathbf{w}\|_2^2$ norm defined by $\|\mathbf{w}\|_2^2 = \sum_i w_i^2$ is a common choice. The $\|\mathbf{w}\|_1$ norm defined by $\|\mathbf{w}\|_1 = \sum_i |w_i|$ is a better choice when more model sparsity is required to select less variables for use in the model. In more complicated applications where variables may be grouped and feature selection among groups is expected, a structured regularizer, such as the group lasso $||\mathbf{w}||_{2,1} = \sum_{\ell=1}^{L} \sqrt{\sum_{i \in \mathcal{G}_\ell} w_i^2}$, can be used where $\mathcal{G}_\ell$ contains the indices of variables belonging to a group ℓ.

5.2.3 Solving proposed problem

In this thesis study, we realize $R(\mathbf{w})$ by the $\|\mathbf{w}\|_1$ norm and develop an efficient algorithm to solve the resultant optimization problem as follows:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^T(\mathbf{X}^T\mathbf{P}\mathbf{X})\mathbf{w} + \lambda\|\mathbf{w}\|_1 \\ \text{subject to} \quad & \mathbf{w}^T(\mathbf{X}^T\mathbf{Q}\mathbf{X})\mathbf{w} = 1. \end{aligned} \tag{5.2.11}$$

The algorithm we will describe next, although is designed for Problem (5.2.11), can be modified to solve Problem (5.2.10) that takes another form of the above-discussed regularization.

Due to the $\|\mathbf{w}\|_1$ norm, the objective function in Problem (5.2.11) is not continuously differentiable and a gradient decent type of approach cannot be applied directly. A well known strategy to overcome this obstacle is to decompose $\mathbf{w}$ into two parts: $\mathbf{w} = \mathbf{u} - \mathbf{v}$, both $\mathbf{u}$ and $\mathbf{v}$ are vectors of the same size as that of $\mathbf{w}$; and all the components in $\mathbf{u}$ and $\mathbf{v}$ are required to be non-negative. Let $\boldsymbol{\gamma} = [\mathbf{u}^T, \mathbf{v}^T]^T$ and $\mathbf{H} = [\mathbf{X}, -\mathbf{X}]$. By the change of variables, Problem (5.2.11) can be equivalently re-written as:

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \quad & f : \boldsymbol{\gamma}^T(\mathbf{H}^T\mathbf{P}\mathbf{H})\boldsymbol{\gamma} + \lambda \sum_{i=1}^{2d} \gamma_i \\ \text{subject to} \quad & g_1 : \boldsymbol{\gamma}^T(\mathbf{H}^T\mathbf{Q}\mathbf{H})\boldsymbol{\gamma} - 1 = 0 \\ & g_{2:e} : \boldsymbol{\gamma} \succeq 0, \end{aligned} \tag{5.2.12}$$

where f denotes an objective function, g's denote constraints, and $e = 2d + 1$, indicating the number of constraints in that group. It is easy to show the equivalence between Problems (5.2.12) and (5.2.11). When Problem (5.2.12) reaches optimality, at least one of the two components u_i and v_i at any i-th position of the two vectors will exactly be 0. Otherwise, by setting $\tilde{u}_i = u_i - v_i$ and $\tilde{v}_i = 0$ if $u_i \geq v_i$, or $\tilde{u}_i = 0$ and $\tilde{v}_i = v_i - u_i$ if $u_i < v_i$, we obtain a better solution with $\tilde{u}_i$ and $\tilde{v}_i$. Therefore, the optimal $\hat{\mathbf{w}}$ to Problem (5.2.11) can be derived from the optimal $\hat{\boldsymbol{\gamma}}$ to Problem (5.2.12) by setting $\hat{\mathbf{w}} = \hat{\mathbf{u}} - \hat{\mathbf{v}}$.

Problem (5.2.12) is not a convex problem because of the quadratic equality constraint. However, it can be efficiently solved using sequential quadratic programming

(SQP) algorithm [54] because both of the objective and constraints are either in a quadratic or a linear form. The gradient of the objective and constraint functions with respect to γ can be calculated as:

$$\begin{aligned}\nabla f &= 2(\mathbf{H}^T\mathbf{P}\mathbf{H})\boldsymbol{\gamma} + \lambda\mathbb{1},\\ \nabla g_1 &= 2(\mathbf{H}^T\mathbf{Q}\mathbf{H})\boldsymbol{\gamma},\\ \nabla g_{2:e} &= \mathbf{I}.\end{aligned}$$

Let $\boldsymbol{\alpha}$ be the Lagrange multipliers, the Lagrangian function of this problem can be written as:

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\alpha}) = f(\boldsymbol{\gamma}) + \sum_i \alpha_i g_i(\gamma);$$

and the Hessian with respect to γ is computed as:

$$\nabla^2_{\mathcal{L}} = \mathbf{H}^T 2(\mathbf{P} + \alpha_1\mathbf{Q})\mathbf{H}.$$

We iteratively search for the optimal solution to Problem (5.2.12). In each iteration, we first solve the following quadratic program to find the moving direction for $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$,

$$\begin{aligned}\min_{\mathbf{p}} \quad & f(\boldsymbol{\gamma}_t) + \nabla f(\boldsymbol{\gamma}_t)^\top \mathbf{p} + \frac{1}{2}\mathbf{p}^\top \nabla^2 \mathcal{L}(\boldsymbol{\gamma}_t, \boldsymbol{\alpha}_t)\mathbf{p}\\ \text{subject to} \quad & \nabla g_1(\boldsymbol{\gamma}_t)^\top \mathbf{p} + g_1(\boldsymbol{\gamma}_t) = 0,\\ & \nabla g_i(\boldsymbol{\gamma}_t)^\top \mathbf{p} + g_i(\boldsymbol{\gamma}_t) \succeq 0, i \in [2:e].\end{aligned} \tag{5.2.13}$$

The optimal solution $\hat{\mathbf{p}}$ to this problem will give the next moving direction for $\boldsymbol{\gamma}$, along which the objective of Problem (5.2.12) can be decreased. Let $\hat{\mathbf{q}}$ be the optimal

Algorithm 4 A sequential quadratic programming approach to solving Problem (5.2.11)

Input: $\mathbf{Z}$, $\mathbf{C}$, $\mathbf{X}$, λ
Output: $\boldsymbol{\gamma}$
1. Calculate $\mathbf{P}$ according to Eq.(5.2.7), and $\mathbf{Q}$ according to Eq.(5.2.8).
2. Initialize $\boldsymbol{\gamma}$ with $\mathbf{u} = \mathbf{1}$, $\mathbf{v} = \mathbf{0}$.
3. Initialize $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha} = 1$.
4. Evaluate f, ∇f, ∇g_i and $\nabla^2 \mathcal{L}$ with the current $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$.
5. Solve Problem (5.2.13) to obtain $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$.
6. Perform a line search to find the learning step size s.
7. Update $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ as in Eq.(5.2.14).

Repeat 4-7 until $\boldsymbol{\gamma}$ reaches a fixed point.

Lagrange multipliers corresponding to $\hat{\mathbf{p}}$. The next moving direction of $\boldsymbol{\alpha}$ is calculated as $\hat{\mathbf{q}} - \boldsymbol{\alpha}_t$. With the moving direction being calculated, we then employ a line search method [54] to find the optimal learning step size s and update $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ as follows:

$$\boldsymbol{\gamma}_{t+1} = \boldsymbol{\gamma}_t + s\hat{\mathbf{p}}_t, \boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + s(\hat{\mathbf{q}}_t - \boldsymbol{\alpha}_t). \tag{5.2.14}$$

We summarize the proposed algorithm in Algorithm 4.

5.3 Computational Results

We validated the proposed approach in both simulations and the analysis of a real-world data set that was aggregated from multiple genetic studies of cocaine dependence (CD). Following the design principle for the chip heritability simulations in [83], we used real genotypic data in the CD study in our simulations but synthesized phenotypes. We first give a thorough description about our CD study data.

In the CD study, subjects were recruited from multiple sites, including the Uni-

versity of Connecticut, Yale University School of Medicine, the University of Pennsylvania School of Medicine, McLean Hospital and the Medical University of South Carolina. All subjects participated using procedures approved by the institutional review board at each participating site. There were 6,621 subjects genotyped at totally 1,140,420 SNPs genome-wide. Among them, 2,674 were stratified into the African American population, and only these subjects were used in our experiments to avoid spurious findings due to population structure. We removed 537 subjects who have other family members in the data so GRM was computed for unrelated individuals.

A series of data cleaning steps were performed to insure the quality of genotypic markers. Markers that meet any of the following conditions were excluded: low call rate ($< 98\%$), G/C and A/T markers (to avoid strand issues), deviating from Hardy-Weinberg equilibrium ($p < 1e-8$), significant cohort calling discrepancy and monomorphic. We also removed non-autosomal markers, so only markers from the 22 autosomal chromosomes were used in the analysis. After these data cleaning steps, 690,864 were SNPs remained. Genetic relationship was estimated for each pair of subjects using the software called GCTA [84] and all the 690,864 SNPs. We then further excluded 385 subjects whose relatedness with some subjects was greater than 0.025 (corresponding to the relatedness of second cousins). Eventually, 1,752 subjects were used in our analysis.

We first validated the proposed approach using synthetic data that was created based on the cleaned CD study data. Then we applied our approach to the real clinical features that characterize the cocaine use behavior of the subjects. For both simulation and the CD case study, we ran 10 times three-fold cross validation (CV) to determine a proper value of λ. At each fold of the CV, a linear model was derived by running the proposed method on training data, and then applied to test data.

Testing h^2 of the composite trait model was estimated only using subjects in the test data. We ran the CV for multiple choices of λ (see figures for the choices we used) and chose the one that gave a trait with the highest testing h^2. Once λ was determined, we ran the proposed method on the entire sample set to identify the final trait. All the reported h^2 was estimated using GCTA with a GRM computed with the 690,864 SNPs. We compared the heritability of the trait derived by our approach with that of the cocaine symptom counting phenotype which has been recognized as a better composite trait than CD diagnosis in genetic studies [25]. Moreover, in order to understand better the composite trait resulted from our approach, we studied the quantitative scores that all the subjects obtained for the trait; and we show the characteristics of subject subgroups partitioned based on their quantitative scores on important cocaine use related variables.

5.3.1 Synthetic data

We first simulated a quantitative trait y based on the linear mixed model as shown in Eq.(5.2.1). From the 690,864 SNPs, we randomly picked 2,000 to use as the causal variants of the simulated trait. Each component u_i in $\mathbf{u}$ was generated independently by sampling from the standard normal distribution, i.e., $N(0, 1)$. Each component ε_i in ε was created from the normal distribution with mean of 0 and variance of $\text{var}(\mathbf{z}_i\mathbf{u})(1/h^2 - 1)$, where $\mathbf{z}_i$ is the i-th row in $\mathbf{Z}$, var(.) is the sample variance of a random variable and h^2 is the implanted heritability. In our simulation, we set $h^2 = 0.8$. Once $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ were determined, we added the effect from the covariates to generate the final trait. Three covariates were used in our study, including: age, sex and the first principle component (PC) of the GRM in the CD study data, their

effect was arbitrarily specified as -0.2, 0.1, and 0.1, respectively. The resultant y has a h^2 estimate of $0.86(s.e. = 0.27)$.

Then we simulated phenotypic data on seven variables (i.e., the $\mathbf{X}$) based on y created above. In order to simulate the real world situation where $\mathbf{w}$ is sparse, three out of the seven variables were created with impact on y, while the other four with no impact. The $\mathbf{w}$ in $\mathbf{y} = \mathbf{Xw}$ was created by assigning random values in the range of $(0, 2)$ to the three variables with impact and 0 to the rest four variables. We randomly drew values for all the variables except one with impact on y (which was randomly picked) from the standard normal distribution. Date for the left out variable was generated with value of $(y_i - \sum_{j=1}^{2} w_j x_{ij})/w_3$, where w_1 and w_2 were weights of the two variables with impact, and for which data were randomly created, w_3 was the weight of the left out variable. We estimated the h^2 for all the seven simulated variables. The results are reported in table 5.3.1. The maximum heritability was 0.17. By simulating the data in this way, we know that there is at least one linear combination of variables in the data that results in a composite trait with h^2 of 0.86. Hence, if our approach works, it should at least find this linear combination if there is no any other combination that gives a trait with even higher h^2, so our approach should find a trait of $h^2 \geq 0.86$.

We performed 10 times three-fold cross validation for 20 different λ's ranging from 1 to 20 with step size 1. Besides the test h^2 was estimated for each trait model in the CV, we also estimated their training h^2 to have a full view of how h^2 changes when λ varies. The training and testing h^2's are plotted in Figures 5.3.1 and 5.3.2, respectively. Both of them have the highest value when $\lambda = 1$, with a median of 0.85 for training and 0.6 for testing h^2. When λ increases, both training and testing h^2 drop, but the testing h^2 changes only slightly when λ varies between 1-6. We

TABLE 5.3.1: Heritability estimates of the seven simulated variables

Variable	h^2	s.e.
V1	0.12	0.27
V2	0	0.26
V3	0	0.26
V4	0	0.26
V5	0.17	0.27
V6	0.14	0.26
V7	0	0.26

also observed from Figure 5.3.3 that less number of variables remained in the model when λ increased. This showed that more redundant variables were excluded from the model.

When $\lambda = 6$, the traits derived in the CV had training h^2 very close to those developed when $\lambda < 6$ on average, but used less number of variables in the model, thus reducing the risk of overfitting. We hence chose $\lambda = 6$ when we developed the final trait with the entire sample set. The h^2 estimate of the resultant trait is 0.94 (s.e. 0.27), and it is larger than that of the trait we implanted. Moreover, it is larger than the h^2 estimate of any individual trait. These results demonstrate the effectiveness of the proposed approach in identifying heritable composite trait from complex multivariate phenotype.

5.3.2 A case study: cocaine use and related behaviors

All the 1,752 subjects were interviewed with a computer-assisted form, the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA) [59], which consists of questions designed for cocaine use and related behaviors. The responses to these questions lead to the definition of seven important cocaine use related variables,

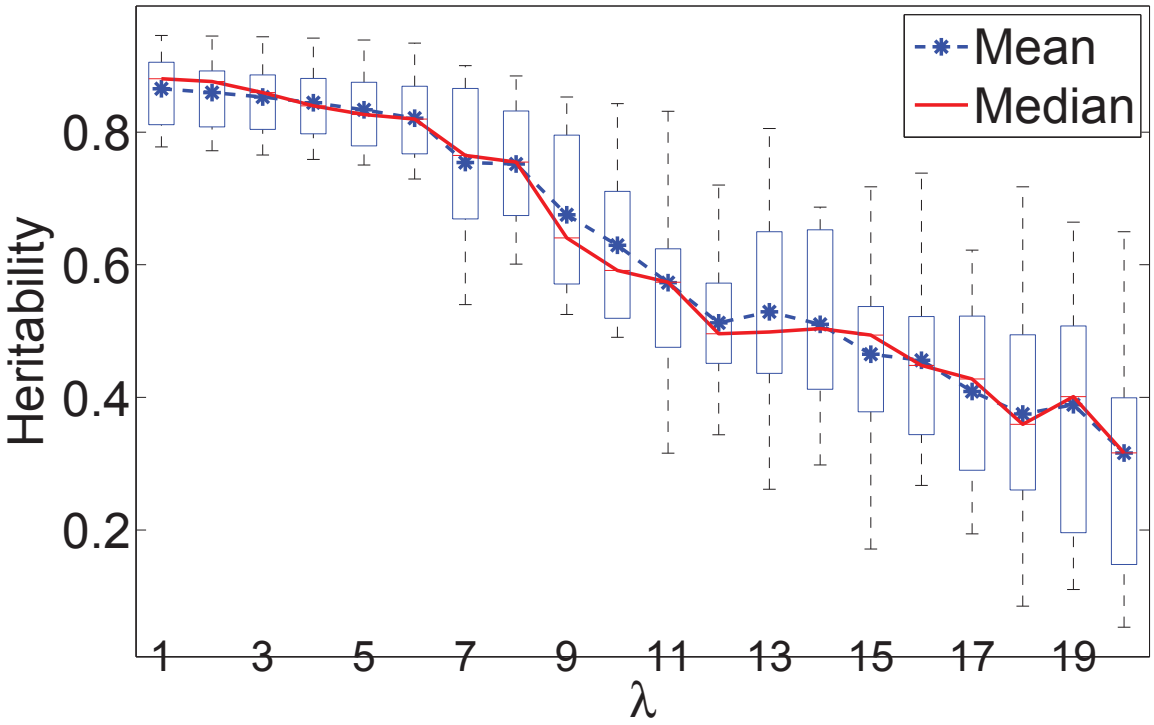

FIGURE 5.3.1: Simulation study: the **training** h^2 of the composite traits developed in three-fold cross validation with varying λ.

based on which a diagnosis of CD is determined. These seven variables are listed below:

- *F1* - tolerance to cocaine;
- *F2* - withdrawal from cocaine;
- *F3* - using cocaine in larger amounts or over longer period than intended;
- *F4* - persistent desire or unsuccessful efforts to cut down or control cocaine use;
- *F5* - great amount of time spent in activities necessary to obtain, use or recover from the effects of cocaine;
- *F6* - gave up or reduced important social, occupational, or recreational activities because of cocaine use;

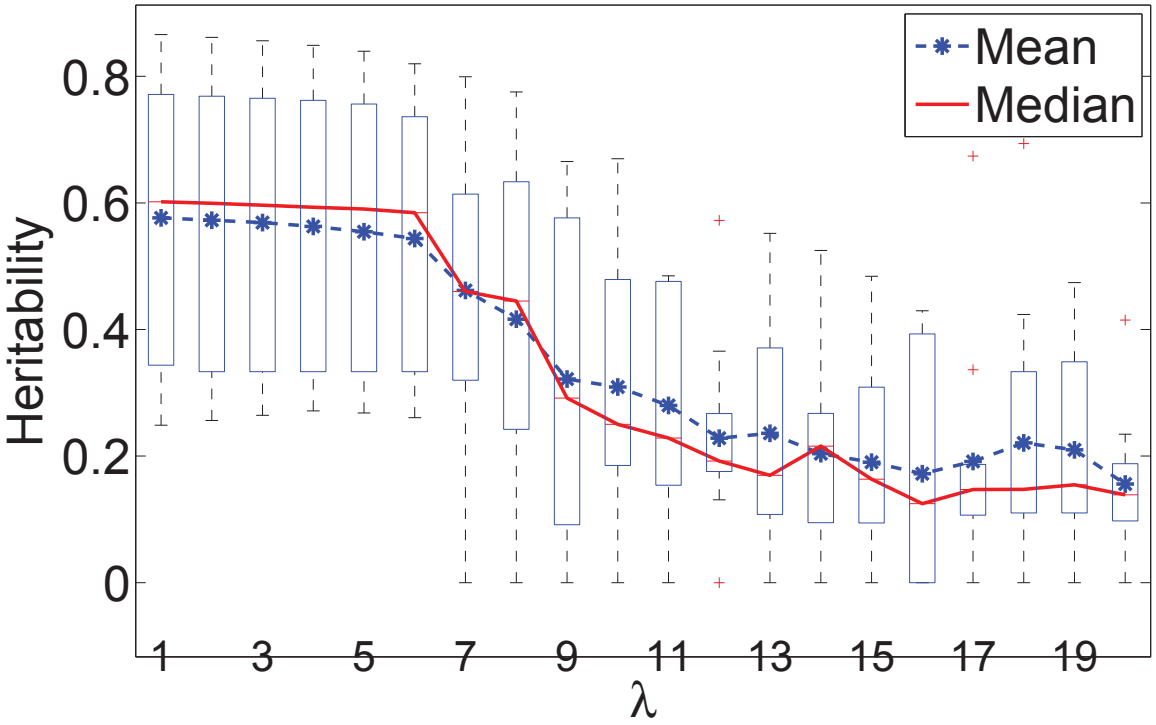

FIGURE 5.3.2: Simulation study: the **testing** h^2 of the composite traits developed in three-fold cross validation with varying λ.

- *F7* - cocaine use despite knowledge of persistent or recurrent physical or psychological problems likely to have been caused or exacerbated by cocaine.

In our experiments, positive responses to the seven variables are coded with 1 otherwise 0. In previous study [25], the counting of symptoms related to cocaine use has been shown as a better trait than the binary trait induced by the diagnosis of CD for use in genetic association studies. Cocaine symptom counting is defined as the number of positive responses to the seven variables. In other words, it is a composite trait resulted from the linear combination of the seven variables with equal weights. Our objective was to identify a linear combination of the same seven variables that led to a trait with a higher h^2 estimate than that of the cocaine symptom counting. Since all the seven variables are binary, their h^2s were not estimated here.

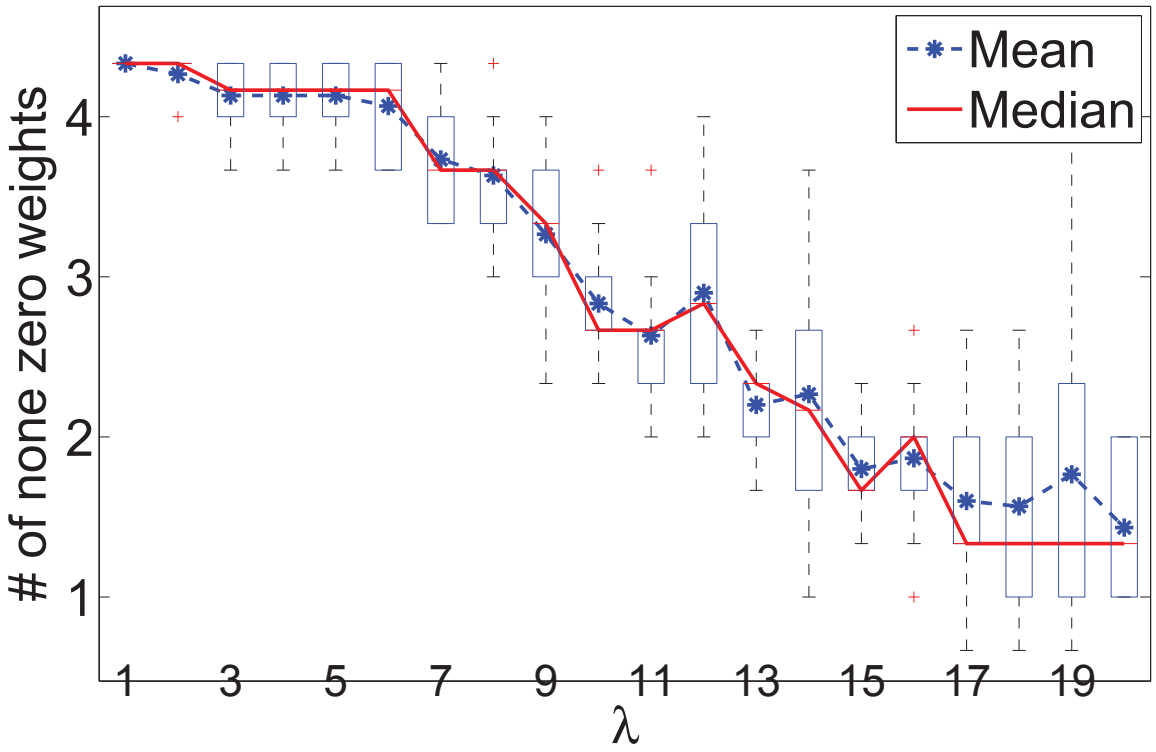

FIGURE 5.3.3: Simulation study: the number of variables used in linear combinations constructed by the proposed method in cross validation with different choices of λ.

We also performed 10 times three-fold CV for λ's ranging from 1 to 20 with step size 1. In all these expriments, we used age, sex and the first three PCs of the GRM as covariates. The testing h^2s of all traits derived in the CV are plotted in Figure 5.3.4. When $\lambda = 4$, resultant traits have the highest heritability estimates on average with a mean of 0.29. We set $\lambda = 4$ and developed a trait by running the proposed method on the entire sample set. The h^2 estimate of the resulted trait is 0.30 (s.e. 0.27). We also estimated the heritability for cocaine symptom counting using exactly the same sample set, GRM and covariates. It has a h^2 estimate close to zero. These results again demonstrate the effectiveness of our approach in identifying heritable composite trait from complex multivariate phenotype.

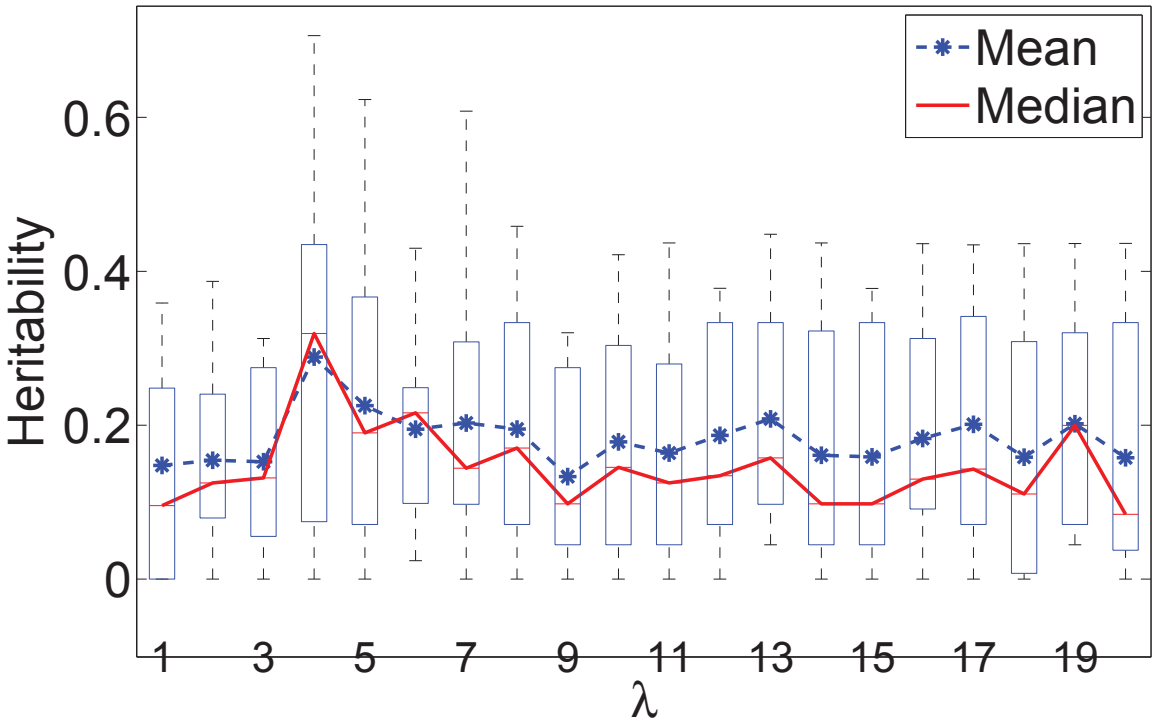

FIGURE 5.3.4: Case study on cocaine dataset: the testing h^2 of the composite traits derived in three-fold cross validation with varying λ.

The weights that each variable obtained in the linear combination obtained by our approach with $\lambda = 4$ and the entire sample set are show in Figure 5.3.3. Out of the seven variables, three (F3, F5 and F7) received zero weight, and were completely ruled out from the model. Another three variables (F2, F4, F6) received weights that were significantly deviated from 0, thus had the most impact on the resultant trait. Among these three variables, F2 and F4 had positive weight, which implies that positive responses to these two variables would result in high scores for the resultant trait; whereas F6 had a negative weight, which means a positive response to this variable would result in a low score. Variable F1 had a weight of 0.0026, and had very limited impact on the trait comparing to F2, F4 and F6.

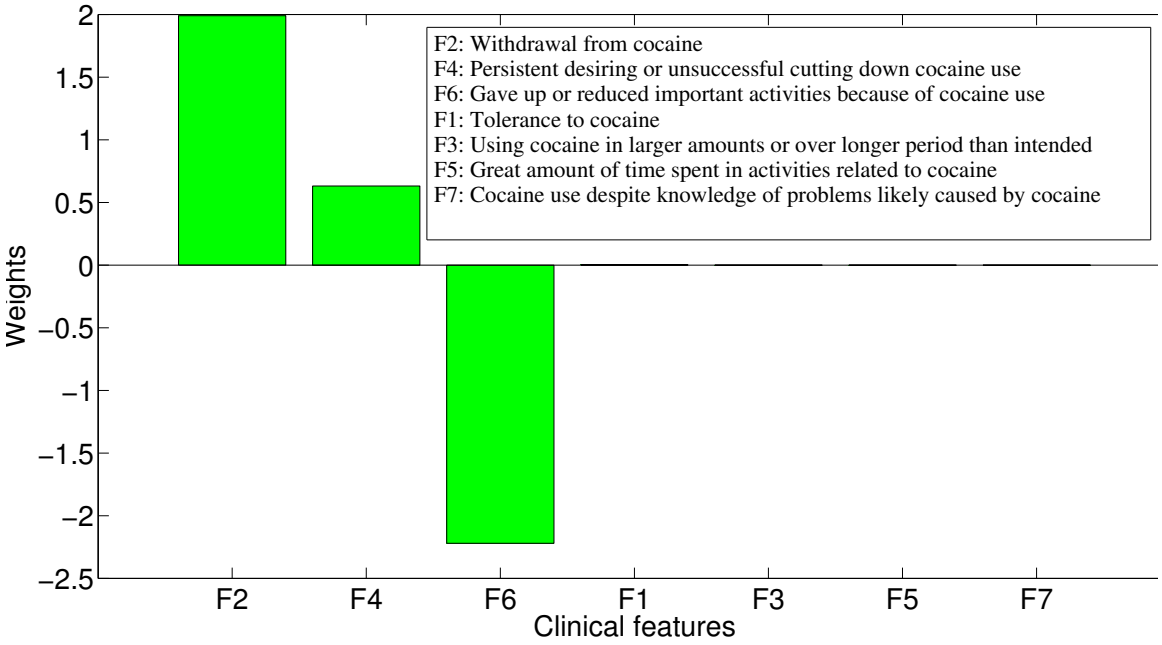

FIGURE 5.3.5: Case study on cocaine dataset: the weights of variables in the linear combination learned by the proposed method with $\lambda = 4$ and the entire sample set.

Figure 5.3.6 shows the distribution of scores (trait values) that subjects obtained for the composite trait derived by our approach with $\lambda = 4$ and the entire sample set. Based on the scores, the sample could be partitioned into three groups as shown in Figure 5.3.6. There were 222 subjects (12.67% of total) in group 1, which had a mean score of -1.61. Group 2 was the largest one consisting of 1358 subjects, and took up 77.51% of the entire sample set. The mean score of this group was 0.39. Group 3 had the least number of subjects, that was 172 (9.82%), with a mean score of 2.59.

In order to understand better the derived trait, we characterized the three groups using important clinical variables related to cocaine use. Besides the seven variables that we have used in deriving the trait, we included another four important variables: (1) total number of cocaine symptoms endorsed (i.e., the cocaine symptom counting); (2) age when first time used cocaine; (3) age when first being diagnosed with DSM-IV

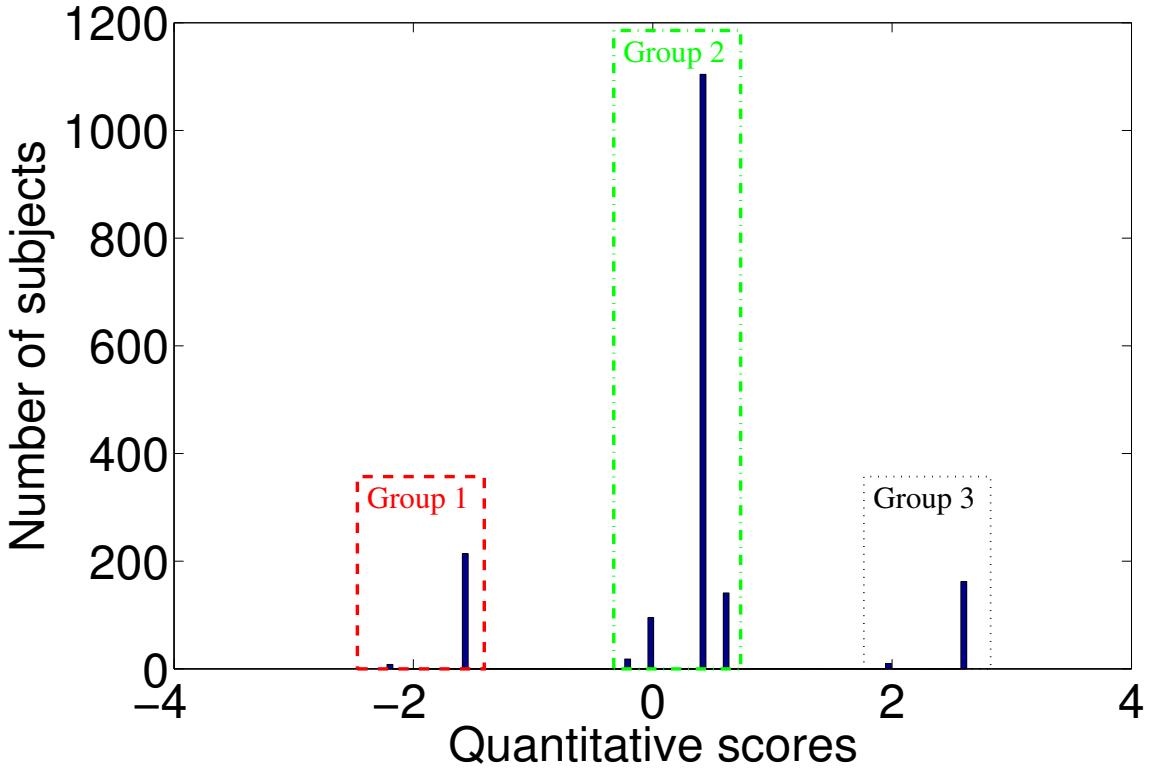

FIGURE 5.3.6: Case study on cocaine dataset: the distribution of scores for the composite trait derived by the proposed method with $\lambda = 4$ and the entire sample set.

CD; (4) the transition time in years from first time cocaine use to first DSM-IV CD diagnosis. The results are summarized in Table 5.3.2. Noteworthy findings include: no subject in Group 1 had withdraw symptom from cocaine while all subjects in Group 3 had it; no subject in Group 3 ever gave up or reduced important activities because of cocaine use while all subjects in Group 1 had such an experience; and subjects in Group 2 had the largest number of cocaine symptoms endorsed, and longest transition time from the first cocaine use to first diagnosis of CD among the three groups.

TABLE 5.3.2: Characteristic of the three subject groups on important clinical variables related to cocaine use.

Variable	Group1 222(12.67)	Group2 1358(77.51)	Group3 172(9.82)
Tolerance to cocaine	111(50.00)	939(69.15)	84(48.84)
Withdrawal from cocaine	0(0.00)	1122(82.62)	172(100.00)
Using cocaine in larger amounts or over longer period than intended	191(86.04)	1161(85.49)	139(80.81)
Persistent desiring or unsuccessful cutting down cocaine use	214(96.40)	1245(91.68)	162(94.19)
Great amount of time spent in activities related to cocaine	181(81.53)	1132(83.36)	122(70.93)
Gave up or reduced important activities because of cocaine use	222(100.00)	1122(82.62)	0(0.00)
Cocaine use despite knowledge of problems likely caused by cocaine	197(88.74)	1181(86.97)	133(77.33)
Number of cocaine symptoms endorsed	5.03(0.97)	5.82(1.99)	4.72(1.06)
Age when first used cocaine	22.31(6.59)	21.64(5.97)	23.33(8.96)
Age onset of DSM-IV cocaine dependence	28.52(7.99)	27.13(7.09)	28.23(8.37)
Transition time in years from first cocaine use to first CD diagnosis	7.48(11.38)	12.22(20.93)	7.78(14.54)

$N(\%)$ is shown for the first seven binary variables, where N is the number of subjects who are positive on the corresponding variable within a group and % is the percentage of N in the group.
$\mu(\sigma^2)$ is shown for the last four continuous variable, where μ is the group mean and σ^2 the standard deviation.

5.4 Summary

We have developed an approach that identifies composite traits from multivariate phenotypes that are highly heritable as estimated based on genome-wide SNPs. The trait we derived is in the form of linear combination of variables related to the phenotype, that is $\mathbf{y} = \mathbf{X}\mathbf{w}$. A quadratic optimization problem has been formulated, in which optimal $\mathbf{w}$ is sought to optimize the log likelihood for estimating variance components in REML. In this formulation, variance components are set to their ideal values with the additive genetic variance component σ_g^2 equal to 1 and other components equal to 0. To overcome the issue of overfitting, we incorporate a regularization term in our formulation. An efficient algorithm based on the sequential quadratic programming framework has been developed to solve the proposed optimization problem. We have evaluated the proposed approach on both synthetic and real world data. The empirical results demonstrate the effectiveness of our approach as it identifies traits with much higher chip h^2 than commonly-used disease phenotypes.

In this study, the pairwise genetic relationship among subjects was estimated from genome-wide SNPs. However, it can certainly be estimated from SNPs restricted to a specific region, such as a particular chromosome or genes related to a pathway, to explore the genetic architecture of a trait. When SNPs within a specific region are used, the trait resulted from the proposed approach will achieve the maximized genetic variance component corresponding to this region. In an application, such as substance dependence, there are known pathways involved in the biological mechanism of the disorder, it may be interested to find out whether there is a composite trait, the variance of which can be largely explained by the variants within these pathways, which will be a future application of our approach.

Bibliography

[1] *Diagnostic and Statistical Manual of Mental Disorders: Fourth Edition (DSM-IV)*, 4th ed., Amer Psychiatric Pub, July 1994.

[2] G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean, *An integrated map of genetic variation from 1,092 human genomes*, Nature **491** (2012), no. 7422, 56–65.

[3] L. Almasy and J. Blangero, *Multipoint quantitative-trait linkage analysis in general pedigrees*, American Journal of Human Genetics **62** (1998), no. 5, 1198–211.

[4] D. J. Balding, M. J. Bishop, and C. Cannings, *Handbook of statistical genetics*, 3rd ed., John Wiley & Sons, Chichester, England ; Hoboken, NJ, 2007.

[5] D. Basu, S.A. Ball, R. Feinn, J. Gelernter, and H.R. Kranzler, *Typologies of drug dependence: comparative validity of a multivariate and four univariate models*, Drug and Alcohol Dependence **73** (2004), no. 3, 289–300.

[6] Karen F. S. Bell, Francesc X. Soriano, Sofia Papadia, and Giles E. Hardingham, *Role of histone acetylation in the activity-dependent regulation of sulfiredoxin and sestrin 2*, Epigenetics **4** (2009), no. 3, 152–158 (English).

[7] D. P. Berry and J. J. Crowley, *Residual intake and body weight gain: a new measure of efficiency in growing cattle*, Journal of animal science **90** (2012), no. 1, 109–115 (English).

[8] J. Bi, *Multi-objective programming in SVMs*, Proceedings of the 20th International Conference on Machine Learning, 2003, pp. 35–42.

[9] J. Bi, J. Gelernter, J. Sun, and H. R. Kranzler, *Comparing the utility of homogeneous subtypes of cocaine use and related behaviors with DSM-IV cocaine dependence as traits for genetic association analysis*, American Journal of Medical Genetics (Part B): Neuropsychiatric Genetics **165B** (2013), no. 2, 148–156.

[10] L. J. Bierut, J. R. Strickland, J. R. Thompson, S. E. Afful, and L. B. Cottler, *Drug use and dependence in cocaine dependent subjects, community-based individuals, and their siblings*, Drug and Alcohol Dependence **95** (2008), no. 1-2, 14–22.

[11] Avrim Blum and Tom Mitchell, *Combining labeled and unlabeled data with co-training*, Proceedings of the eleventh annual conference on Computational learning theory (New York, NY, USA), COLT' 98, ACM, 1998, pp. 92–100.

[12] AA Boligon, MEZ Mercadante, F. Baldi, RB Lbo, and LG Albuquerque, *Multi-trait and random regression mature weight heritability and breeding value estimates in nelore cattle*, South African Journal of Animal Science **40** (2011), no. 5 (English).

[13] C. Braet and W. Beyers, *Subtyping children and adolescents who are overweight: Different symptomatology and treatment outcomes*, Journal of Consulting and Clinical Psychology **77** (2009), no. 5, 814–824.

[14] P. R. Burgel, N. Roche, J. L. Paillasseur, D. Caillaud, I. Tillie-Leblond, T. Perez, P. Chanez, R. Escamilla, I. Court-Fortune, and P. Carre, *Clinical COPD phenotypes: A novel approach using principal component and cluster analyses*, European Respiratory Journal **36** (2010), no. 3, 531–539.

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

[16] Kamalika Chaudhuri, Sham M. Kakade, Karen Livescu, and Karthik Sridharan, *Multi-view clustering via canonical correlation analysis*, Proceedings of the 26th Annual International Conference on Machine Learning (New York, NY, USA), ICML '09, ACM, 2009, pp. 129–136.

[17] Peikai Chen, Y. S. Hung, Yubo Fan, and S. T. C. Wong, *An integrative bioinformatics approach for identifying subtypes and subtype-specific drivers in cancer*, IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, 2012, pp. 169–176.

[18] E.E. Connor, J.L. Hutchison, and H.D. Norman, *Estimating feed efficiency of lactating dairy cattle using residual feed intake*, (Chapter 11) In Feed Efficiency in the Beef Industry, Hill, R.A. (ed.), Wiley-Blackwell, NJ (2012).

[19] David L. Davies and Donald W. Bouldin, *A cluster separation measure*, IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-1** (1979), no. 2, 224–227.

[20] J. C. Dunn, *Well separated clusters and optimal fuzzy-partitions*, Journal of Cybernetics **4** (1974), 95–104.

[21] D. S. Falconer and T.F. C. Mackay, *Introduction to quantitative genetics, 4th edition*, Benjamin Cummings, 1996 (English).

[22] T. Fawcett, *An introduction to ROC analysis*, Pattern Recognition Letters **27** (2006), no. 8, 861–874.

[23] J. Gelernter, H. R. Kranzler, R. Sherva, R. Koesterer, L. Almasy, H. Zhao, and L. A. Farrer, *Genome-wide association study of opioid dependence: multiple associations mapped to calcium and potassium pathways*, Biol Psychiatry **76** (2014), no. 1, 66–74.

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

[25] J. Gelernter, R. Sherva, R. Koesterer, L. Almasy, H. Zhao, H. R. Kranzler, and L. Farrer, *Genome-wide association study of cocaine dependence and related traits: Fam53b identified as a risk gene*, Mol Psychiatry (2013).

[26] Joel Gelernter, Carolien Panhuysen, Roger Weiss, Kathleen Brady, Victor Hesselbrock, Bruce Rounsaville, James Poling, Marsha Wilcox, Lindsay Farrer, and Henry R. Kranzler, *Genomewide linkage scan for cocaine dependence and related traits: Significant linkages for a cocaine-related trait and cocaine-induced paranoia*, American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics **136** (2005), no. 1, 45.

[27] Joel Gelernter, Carolien Panhuysen, Marsha Wilcox, Victor Hesselbrock, Bruce Rounsaville, James Poling, Roger Weiss, Susan Sonne, Hongyu Zhao, Lindsay Farrer, and Henry R. Kranzler, *Genomewide linkage scan for opioid dependence*

and related traits, American Journal of Human Genetics. **78** (2006), no. 5, 759–769.

[28] Santhosh Girirajan, Wendy S. Meschino, Marjan M. Nezarati, Alexander Asamoah, Kelly E. Jackson, Gordon C. Gowans, Judith A. Martin, Erin P. Carmany, David W. Stockton, Rhonda E. Schnur, Lynette S. Penney, Jill A. Rosenfeld, Donna M. Martin, Salmo Raskin, Kathleen Leppig, Heidi Thiese, Rosemarie Smith, Erika Aberg, Dmitriy M. Niyazov, Luis F. Escobar, Dima El-Khechen, Kisha D. Johnson, Bradley P. Coe, Robert R. Lebel, Kiana Siefkas, Susie Ball, Natasha Shur, Marianne McGuire, Campbell K. Brasington, J. E. Spence, Laura S. Martin, Carol Clericuzio, Blake C. Ballif, Sumit Parikh, Lisa G. Shaffer, Evan E. Eichler, Neil Friedman, Amy Goldstein, Robyn A. Filipink, Juliann S. McConnell, and Brad Angle, *Phenotypic heterogeneity of genomic disorders and rare copy-number variants*, The New England Journal of Medicine **367** (2012), no. 14, 1321–1331 (English).

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

[30] L. Hagen and A. B. Kahng, *New spectral methods for ratio cut partitioning and clustering*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. **11** (1992), no. 9, 1074–1085.

[31] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference and prediction*, New York: Springer, 2001.

[32] William G. Hill and Naomi R. Wray, *Heritability in the genomics era - concepts and misconceptions*, Nature Reviews Genetics **9** (2008), no. 4, 255–266 (English).

[33] 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 **43** (2008), no. 5, 505–15.

[34] C. A. Hodgkinson, Q. Yuan, K. Xu, P. H. Shen, E. Heinz, E. A. Lobos, E. B. Binder, J. Cubells, C. L. Ehlers, J. Gelernter, J. Mann, B. Riley, A. Roy, B. Tabakoff, R. D. Todd, Z. Zhou, and D. Goldman, *Addictions biology: haplotype-based analysis for 130 candidate genes on a single array*, Alcohol and Alcoholism **43** (2008), no. 5, 505–15.

[35] B. N. Howie, P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies*, PLoS Genet **5** (2009), no. 6, e1000529.

[36] Valerie W. Hu, Anjene Addington, and Alexander Hyman, *Novel autism subtype-dependent genetic variants are revealed by quantitative trait and subphenotype association analyses of published gwas data*, PloS ONE **6** (2011), no. 4, e19067 (English).

[37] Lester Ingber, *Very fast simulated re-annealing*, Mathematical and Computer Modelling **12** (1989), no. 8, 967–973.

[38] Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, and Sobel EM, *Mendel: The swiss army knife of genetic analysis programs*, Bioinformatics **29** (2013), 1568–1570.

[39] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Optimization by simulated annealing*, Science, Number 4598, 13 May 1983 **220, 4598** (1983), 671–680.

[40] L. Klei, D. Luca, B. Devlin, and K. Roeder, *Pleiotropy and principal components of heritability combine to increase power for association analysis*, Genet Epidemiol **32** (2008), no. 1, 9–19.

[41] 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 **33** (2008), no. 1, 41–53.

[42] Abhishek Kumar and Hal Daume III, *A co-training approach for multi-view spectral clustering*, Proceedings of the 28th International Conference on Machine Learning (New York, NY, USA) (Lise Getoor and Tobias Scheffer, eds.), ACM, 2011, pp. 393–400.

[43] Abhishek Kumar, Piyush Rai, and Hal Daume III, *Co-regularized multi-view spectral clustering*, Advances in Neural Information Processing Systems 24 (J. Shawe-Taylor, R.S. Zemel, P. Bartlett, F.C.N. Pereira, and K.Q. Weinberger, eds.), 2011, pp. 1413–1421.

[44] Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, Klemm AK, and Manolio TA, *A catalog of published genome-wide association studies*, Available at: www.genome.gov/gwastudies (Accessed July. 2014).

[45] K. Lange, J. Westlake, and M. A. Spence, *Extensions to pedigree analysis. iii. variance components by the scoring method*, Ann Hum Genet **39** (1976), no. 4, 485–91, Lange, K Westlake, J Spence, M A Research Support, U.S. Gov't, P.H.S. England Annals of human genetics Ann Hum Genet. 1976 May;39(4):485-91.

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

[47] Elin Lehrmann, Carlo Colantuoni, Amy Deep-Soboslay, Kevin G. Becker, Ross Lowe, Marilyn A. Huestis, Thomas M. Hyde, Joel E. Kleinman, and William J. Freed, *Transcriptional changes common to human cocaine cannabis and phencyclidine abuse*, PLoS ONE **1** (2006), no. 1, e114 (English).

[48] Ulrike Luxburg, *A tutorial on spectral clustering*, Statistics and Computing **17** (2007), 395–416.

[49] Lukas Meier, Sara Van De Geer, and Peter Bühlmann, *The group lasso for logistic regression*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70** (2008), no. 1, 53–71 (English).

[50] H.B. Moss, C.M. Chen, and H.Y. Yi, *Subtypes of alcohol dependence in a nationally representative sample*, Drug and Alcohol Dependence **91** (2007), no. 2-3, 149–158.

[51] Fionn Murtagh, *Multiple correspondence analysis and related methods*, Psychometrika **72** (2007), no. 2, 275–277.

[52] National Institute of Health, *Study of addiction: Genetics and environment (SAGE)*, NIH Project Website **http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1** (2009).

[53] Mark J. Niciu, Grace Chan, Joel Gelernter, Albert J. Arias, Kara Douglas, Roger Weiss, Raymond F. Anton, Lindsay Farrer, Joseph F. Cubells, and Henry R.

Kranzler, *Subtypes of major depression in substance dependence*, Addiction **104** (2009), no. 10, 1700–1709.

[54] J. Nocedal and S. J. Wright, *Numerical optimization*, 2nd ed., Springer, New York, 2006.

[55] J. Ott and D. Rabinowitz, *A principal-components approach based on heritability for combining phenotype information*, Hum Hered **49** (1999), no. 2, 106–11.

[56] K. Oualkacha, A. Labbe, A. Ciampi, M. A. Roy, and M. Maziade, *Principal components of heritability for high dimension quantitative traits and general pedigrees*, Statistical Applications in Genetics and Molecular Biology **11** (2012), no. 2.

[57] C. I. Panhuysen, Y. Yu, L. A. Farrer, H. R. Kranzler, R. D. Weiss, K. Brady, J. Gelernter, and J. Poling, *Confirmation and generalization of an alcohol-dependence locus on chromosome 10q*, Neuropsychopharmacology **35** (2010), no. 6, 1325–1332.

[58] H. D. Patterson and R. Thompson, *Recovery of inter-block information when block sizes are unequal*, Biometrika **58** (1971), no. 3, pp. 545–554 (English).

[59] 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 **91** (2007), no. 1, 85–90.

[60] A. P. Reynolds, G. Richards, B. De la Iglesia, and V. J. Rayward-Smith, *Clustering rules: A comparison of partitioning and hierarchical clustering algorithms*, Journal of Mathematical Modelling and Algorithms **5** (1992), 475–504.

[61] A. P. Reynolds, G. Richards, and V. J. Rayward-Smith, *The application of k-medoids and pam to the clustering of rules*, Intelligent Data Engineering and Automated Learning, Lecture Notes in Computer Science, vol. 3177, 2004, pp. 173–178.

[62] Peter Rousseeuw, *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, J. Comput. Appl. Math. **20** (1987), no. 1, 53–65.

[63] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*, Cambridge Univeristy Press, 2004.

[64] Jianbo Shi and Jitendra Malik, *Normalized cuts and image segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000), no. 8, 888–905.

[65] D. Speed, G. Hemani, M. R. Johnson, and D. J. Balding, *Improved heritability estimation from genome-wide snps*, Am J Hum Genet **91** (2012), no. 6, 1011–21.

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

[67] J. Sun, J. Bi, and H. R. Kranzler, *A multi-objective program for quantitative subtyping of clinically-relevant phenotypes*, Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM2012), 2012, pp. 256–261.

[68] Jiangwen Sun, Jinbo Bi, and Henry R. Kranzler, *Quadratic optimization to identify highly heritable quantitative traits from complex phenotypic features*, Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA), KDD '13, ACM, 2013, pp. 811–819.

[69] S. T. Tay, S. H. Leong, K. Yu, A. Aggarwal, S. Y. Tan, C. H. Lee, K. Wong, J. Visvanathan, D. Lim, W. K. Wong, K. C. Soo, O. L. Kon, and P. Tan, *A combined comparative genomic hybridization and expression microarray analysis of gastric cancer reveals novel molecular subtypes*, Cancer research **63** (2003), no. 12, 3309–16.

[70] Robert Tibshirani, *Regression shrinkage and selection via the LASSO*, Journal of the Royal Statistical Society. Series B (Methodological) **58** (1996), no. 1, 267–288.

[71] J. Treutlein and M. Rietschel, *Genome-wide association studies of alcohol dependence and substance use disorders*, Curr Psychiatry Rep **13** (2011), no. 2, 147–55.

[72] Iven Van Mechelen, Hans-Hermann Bock, and Paul De Boeck, *Two-mode clustering methods: a structured overview*, Statistical Methods in Medical Research **13** (2004), no. 5, 363–394.

[73] V. Vapnik, *Statistical learning theory*, John Willey & Sons, Inc, New York, 1998.

[74] V. N. Vapnik, *An overview of statistical learning theory*, Ieee Transactions on Neural Networks **10** (1999), no. 5, 988–999.

[75] A.P. Verbyla, *A conditional derivation of residual maximum likelihood*, Australian Journal of Statistics **32** (1990), no. 2, 227–230.

[76] Dorothea Wagner and Frank Wagner, *Between min cut and graph bisection*, Mathematical Foundations of Computer Science, 1993, pp. 744–750.

[77] J. C. Wang, M. Kapoor, and A. M. Goate, *The genetics of substance dependence*, Annu Rev Genomics Hum Genet **13** (2012), 241–61.

[78] Y. Wang, Y. Fang, and M. Jin, *A ridge penalized principal-components approach based on heritability for high-dimensional data*, Hum Hered **64** (2007), no. 3, 182–91.

[79] M. Weatherall, J. Travers, P.M. Shirtcliffe, S.E. Marsh, M.V. Williams, M.R. Nowitz, S. Aldington, and R. Beasley, *Distinct clinical phenotypes of airways disease defined by cluster analysis*, European Respiratory Journal **35** (2010), no. 2, 459–460.

[80] R. West, *Theory of addiction*, Oxford: Blackwell Publishing (2006).

[81] C. J. Willer, Y. Li, and G. R. Abecasis, *Metal: fast and efficient meta-analysis of genomewide association scans*, Bioinformatics **26** (2010), no. 17, 2190–1.

[82] D.R. Williams, R. De Silva, D.C. Paviour, A. Pittman, H.C. Watt, L. Kilford, J.L. Holton, T. Revesz, and A.J. Lees, *Characteristics of two distinct clinical phenotypes in pathologically proven progressive supranuclear palsy: Richardson's syndrome and psp-parkinsonism*, Brain **128** (2005), no. 6, 1247–1258.

[83] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard,

and P. M. Visscher, *Common snps explain a large proportion of the heritability for human height*, Nat Genet **42** (2010), no. 7, 565–9.

[84] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, *Gcta: a tool for genome-wide complex trait analysis*, Am J Hum Genet **88** (2011), no. 1, 76–82.

[85] Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin, *An improved glmnet for l1-regularized logistic regression*, Journal of Machine Learning Research **13** (2012), 1999–2030 (English).

[86] Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani, *1-norm support vector machines*, Neural Information Processing Systems, MIT Press, 2003, p. 16.