Quadratic Optimization to Identify Highly Heritable Quantitative Traits from Complex Phenotypic Features

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

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

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

ABSTRACT

Identifying genetic variation underlying a complex disease is important. Many complex diseases have heterogeneous phenotypes and are products of a variety of genetic and environmental factors acting in concert. Deriving highly heritable quantitative traits of a complex disease can improve the identification of genetic risk of the disease. The most sophisticated methods so far perform unsupervised cluster analysis on phenotypic features; and then a quantitative trait is derived based on each resultant cluster. Heritability is estimated to assess the validity of the derived quantitative traits. However, none of these methods explicitly maximize the heritability of the derived traits. We propose a quadratic optimization approach that directly utilizes heritability as an objective during the derivation of quantitative traits of a disease. This method maximizes an objective function that is formulated by decomposing the traditional maximum likelihood method for estimating heritability of a quantitative trait. We demonstrate the effectiveness of the proposed method on both synthetic data and real-world problems. We apply our algorithm to identify highly heritable traits of complex human-behavior disorders including opioid and cocaine use disorders, and highly heritable traits of dairy cattle that are economically important. Our approach outperforms standard cluster analysis and several previous methods.

Categories and Subject Descriptors

G.1.6 [**Numerical**]: Optimization—*Quadratic programming methods*; H.2.8 [**Database management**]: Database Application—*Data mining*

General Terms

Algorithms, Performance, Experimentation

*Correspondence to: Jinbo Bi, jinbo@engr.uconn.edu

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

Keywords

Quadratic optimization, Quantitative trait, Heritability

1. INTRODUCTION

Identifying genetic variants that underlie complex phenotypes is important in genetics. Genetic correlation analysis can help to uncover the underlying biological processes moderating or regulating a complex disease, such as cancer, heart disease, and substance dependence disorders, which facilitates the development of more effective treatments. Complex phenotypes, however, exhibit great heterogeneity, and are often products of a variety of genetic and environmental factors acting in concert. Refinement of a complex phenotype to reduce phenotypic heterogeneity that is aimed at separating genetic and environmental effects, and dissecting its genetic heterogeneity is a challenging problem.

The success of genetic correlation with a complex trait depends on the heritability of the trait. Heritability of a phenotype measures the proportion of the total phenotypic variance due to additive genetic effects [10]. It is a key population parameter that helps to understand the genetic architecture of traits. Higher heritability of a phenotype implies that the phenotype is more genetically influenced. Thus, there is chance to detect its causal genetic variants. Methods to obtain unbiased estimates of heritability from family pedigree data are well developed for quantitative phenotypes [6]. For example, SOLAR [1], a genetic linkage analysis software package, can estimate the heritability (both broad-sense and narrow-sense) of a quantitative trait.

In the context of genetic analysis, phenotypic heterogeneity means that there are diverse forms of a particular trait. Complex disease phenotypes are often characterized by a variety of clinical variables and symptoms. For example, to diagnose whether a patient has a lifetime drug dependence, clinicians interview the patient to understand many aspects of the patient's drug use and related behaviors, the negative consequences of drug use and treatment history, together with an assessment of other co-occuring medical conditions. All of these variables are used to make a diagnosis of dependence on certain drug. Hence the phenotype of *dependence on a drug*, such as opioid or cocaine, is characterized by all of these parameters.

Although a large number of variables are involved in the clinical diagnosis, multivariate data mining of these clinical features has seldom been utilized. In genetic analysis of a complex disease, the diagnosis itself is often regarded as the

phenotype, which is a binary trait, partitioning the population into two groups, one with the drug dependence disorder, and the other without. This binary trait simply cannot differentiate the heterogeneous clinical manifestations of the disorder, and is likely attributable to heterogeneous genetic causes. Hence, insights into the genetic etiology of drug dependence are limited. In general, diagnosis-induced binary traits often have low heritability and are not optimal for genetic association studies [12].

In the effort to identify clinical traits that are suitable for genetic analysis, researchers can perform a simple phenome scan which assesses the heritability of each collected clinical feature used in the diagnosis. This univariate analysis approach, however, cannot evaluate the interplay between different clinical features. It cannot answer the question of whether a combination of several clinical features will form a trait with higher heritability. In the few studies that have used multivariate analysis of clinical symptoms and features [8, 12, 5, 21], the most sophisticated approach is unsupervised cluster analysis, which is used to find sub-groups of a population that are homogeneous in their clinical features. Quantitative traits can be derived for each resultant cluster by calculating the membership likelihood in a cluster for each subject. Then heritability of the derived quantitative trait is estimated and used to assess the validity of the clusters. Since cluster analysis is completely unsupervised, the resultant clusters are not guaranteed to achieve high heritability. There is currently no empirically derived and statistically rigorous method to identify the optimal trait for a complex disease such as psychiatric illness [9].

In this paper, we propose to make explicit use of the heritability estimate during the derivation of quantitative traits of a complex phenotype. The problem of deriving a highly heritable quantitative trait based on a collected sample differs from traditional supervised or unsupervised machine learning problems where a human expert can either label each sample subject with a precise label, e.g. the membership of a subtype, or give no guidance at all. Mathematically, we are given data $\mathbf{X}_{n\times d}$ on a set of d phenotypic features $\mathbf{x}$ for a total of n subjects from multiple families, and our objective is to project $\mathbf{X}$ into some dimensions so that the empirical heritability of the projected traits is high.

We propose to construct a quantitative trait y in the linear form of $y = \mathbf{x}^\top \mathbf{w}$, and find the weight vector $\mathbf{w}$ that maximizes the empirical heritability of y. Non-linear projections, if desirable, can be formulated using kernel machines [20] following the proposed approach here. A quadratic optimization problem is formulated by decomposing the traditional maximum likelihood method for estimating heritability of a quantitative trait. We develop an efficient solver to optimize the proposed quadratic optimization problem, and evaluate the proposed algorithms on both synthetic and real world data. The computational results demonstrate the effectiveness and efficiency of our approach in deriving quantitative trait with high heritability. In particular, we apply the approach to analyze clinical features related to opioid dependence and cocaine dependence and pedigrees of small nuclear families. To further show the general applicability of our approach, we apply it to a set of dairy cattle traits collected for animal improvement together with extended pedigrees. Our approach can define quantitative traits of drug use with much higher heritability than those obtained by cluster analysis, and derive highly heritable quantitative traits for dairy cattle.

The rest of this paper is organized as follows. We briefly review the literature that is most relevant to the proposed work in Section 2. We describe our formulation that derives highly-heritable quantitative traits based on phenotypic features in Sections 3 and 4, followed by an algorithm evaluation on both synthetic and real world data in Section 5. We conclude this paper with a discussion in Section 6.

2. RELATED WORK

To date, 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 [5, 12, 14, 7, 8]. Traditional clustering algorithms such as k-means [19] and hierarchical clustering [18] have been extensively applied to phenotype complex diseases [22, 23, 4, 3]. 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 clusters [8, 5].

Figure 1 shows the flowchart of a common approach [7, 8, 16] for phenotypic subtyping. First, a standard clustering strategy, relying either on a single clustering method or k-means combined consecutively with hierarchical clustering, was applied to the phenotypic data to partition the sample. It assigns each subject to a specific cluster. To form a quantitative trait, a classification approach, typically logistic regression, is used to separate subjects in different clusters with a probabilistic classifier that is a function of an individual's phenotypic features. The resulting classifier is expected to report a higher membership score for subjects in the 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 empirically estimated using software such as SOLAR [1]. This approach is limited by the fact that, although heritability is used to validate the clusters, it is not used in their creation.

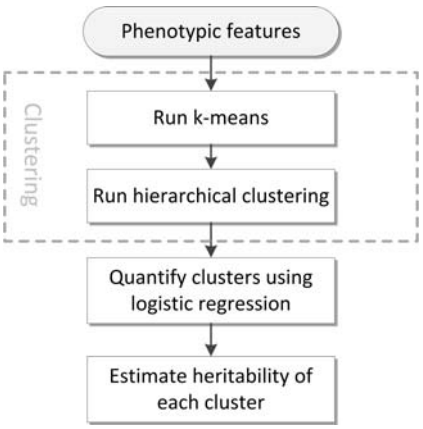

Figure 1: A common approach to phenotypic subtyping.

In very recent work [21], an approach was proposed to identify stable and heritable subtypes of opioid use and related behavior traits 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 [21]. 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 combining only highly heritable clinical features will necessarily lead to traits that are more heritable. Third, similar to the standard approach reviewed above, heritability was not used directly in the clustering process.

3. PROPOSED QUADRATIC OPTIMIZATION

In this paper, we propose a new approach to maximize a heritability-derived objective. Let $\mathbf{X}_{n\times d}$ be the data matrix on a set of d phenotypic features $\mathbf{x}$ for a total of n subjects from multiple families. The goal is to find a $y : y = \mathbf{x}^\top \mathbf{w}$ that yields a high heritability estimate. We limit our discussion to linear projections. The proposed approach can be extended to non-linear combinations, if desirable, by using kernel machines [20].

The heritability of a quantitative trait y can be estimated using a well-established maximum likelihood method based on linear mixture models [13, 2]. It 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. The reason to have separate means for males and females is that it aligns with the general observation that male and female subjects present differences in quantitative traits, such as height and weight. Each entry of $\mathbf{\Omega}_i$ is the phenotypic covariance of two family members j and k, given by (1).

$$cov(y^i_j, y^i_k) = 2\sigma^2_a \mathbf{\Phi}^i_{jk} + \sigma^2_d \mathbf{\Delta}^i_{jk} + \sigma^2_e \boldsymbol{\gamma}^i_{jk} \tag{1}$$

σ^2_a and σ^2_d are genetic components representing additive effects and dominant effects, respectively; σ^2_e denotes the variance due to environmental factors. Here another genetic component: epistatic effects σ^2_I is not considered, but the approach can be extended to include this component or effects from any other source. The quantity $\mathbf{\Phi}^i_{jk}$ 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 are identical by descent (IBD), which means that these two alleles are identical copies of the same ancestral allele. An allele is one of the alternative forms of a gene or a genetic locus. As the human genome is diploid, each human subject has two copies of an allele that can be in different forms at a specific genetic locus. The quantity $\mathbf{\Delta}^i_{jk}$ is the probability of members j and k sharing both alleles at a genetic locus. Both matrices $\mathbf{\Phi}_i$ and $\mathbf{\Delta}_i$ can be calculated from the family pedigrees. Readers can consult Chapter 32 in [2] for details. Exemplar entries of $\mathbf{\Phi}$ and $\mathbf{\Delta}$ between selected family members are illustrated in Table 1 where random mating is assumed. The parameter $\boldsymbol{\gamma}^i_{jk}$ is an environmental indicator that encodes whether j and k live together ($\boldsymbol{\gamma}^i_{jk} = 1$) or apart ($\boldsymbol{\gamma}^i_{jk} = 0$).

Table 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

The five parameters: μ_m, μ_f, σ^2_a, σ^2_d and σ^2_e, are estimated by maximizing the log likelihood of pedigrees over all sample families [13]. The log likelihood is computed by the following equation (2)

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

where μ^i denotes a vector of the respective means μ_m, μ_f for male or female members in the family i. Both gradient and Hessian of the equation (2) with respect to μ_m, μ_f, σ^2_a, σ^2_d and σ^2_e can be calculated, and a Newton-Raphson algorithm or a scoring method [13] can be applied to maximize the log likelihood (2).

The narrow sense heritability is defined by $h^2 = \sigma^2_a/\sigma^2_p$ where σ^2_p given by (3) is the total variance in y.

$$\sigma^2_p = \sigma^2_a + \sigma^2_d + \sigma^2_e \tag{3}$$

The broad sense heritability is defined by the portion of total variance due to all genetic variation: $H^2 = (\sigma^2_a + \sigma^2_d)/\sigma^2_p$. In this paper, we target quantitative traits with higher narrow sense heritability, which we call heritability through the remainder of this paper. If higher broad sense heritability is desirable, our formulation can be easily adapted to derive a quantitative trait of that sort.

To derive a trait y that has the highest possible heritability, i.e., 1, the variance or covariance of y, $cov(y^i_j, y^i_k)$ should be due to the additive effect σ^2_a only, and $\sigma^2_d = \sigma^2_e = 0$. Motivated by this fact, we propose to search for the optimal $\mathbf{w}$, μ_m and μ_f such that the resulting trait $y = \mathbf{x}^\top \mathbf{w}$ achieves the maximal log likelihood LL in (2) with $\mathbf{\Omega}_i$ fixed to:

$$(\mathbf{\Omega}_i)_{jk} = cov(y^i_j, y^i_k) = 2\sigma^2_a \mathbf{\Phi}^i_{jk}.$$

Given that a scaling factor will not change the results, we scale $\sigma_a = 1$. Then, maximizing the log likelihood LL in (2) is equivalent to finding the optimal solution of the following objective (4) after constants are eliminated:

$$\min_{\mathbf{w},\mu_m,\mu_f} \sum_i (\mathbf{X}_i\mathbf{w} - \mu^i)^\top \mathbf{\Phi}_i^{-1}(\mathbf{X}_i\mathbf{w} - \mu^i). \tag{4}$$

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

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

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 Problem (4) can be simplified to the following optimization problem (5):

$$\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^\top (\sum_i \mathbf{H}_i^\top \boldsymbol{\Phi}_i^{-1} \mathbf{H}_i) \boldsymbol{\beta} \tag{5}$$

This objective function can be scaled with the magnitude of $\boldsymbol{\beta}$. We control the magnitude of $\boldsymbol{\beta}$ by fixing the sample variance of the resulting trait to 1, which corresponds to a constraint $\boldsymbol{\beta}^\top \mathbf{H}^\top \mathbf{H} \boldsymbol{\beta} - n = 0$. Clearly, μ_m and μ_f are related to the sample means of clinical features $\mathbf{x}$ for male and female respectively. They are not true free parameters as they are determined once $\mathbf{w}$ is determined. Actually, μ_m and μ_f are equal to the sample phenotypic means of male and female, respectively when the optimal $\boldsymbol{\beta}$ is found. Let $\boldsymbol{\mu}_m$, $\boldsymbol{\mu}_f$ be respectively the two vectors of male and female means on features $\mathbf{x}$. Both $\boldsymbol{\mu}_m$ and $\boldsymbol{\mu}_f$ have a length of d. Let

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

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

Imposing all of these constraints yields an optimization problem with a quadratic objective subject to both quadratic and linear equality constraints as shown in (6).

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \boldsymbol{\beta}^\top (\sum_i \mathbf{H}_i^\top \boldsymbol{\Phi}_i^{-1} \mathbf{H}_i) \boldsymbol{\beta} \\ \text{subject to} \quad & \boldsymbol{\beta}^\top \mathbf{H}^\top \mathbf{H} \boldsymbol{\beta} - n = 0 \\ & \mathbf{a}_m \boldsymbol{\beta} = 0, \mathbf{a}_f \boldsymbol{\beta} = 0 \end{aligned} \tag{6}$$

In many applications, sparsity on the clinical features may be a desirable target. In other words, we expect to use few clinical features in the projection. In such a case, the objective can be regularized by a regularization term on $\mathbf{w}$, $R(\mathbf{w})$. Then the overall optimization problem becomes:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \boldsymbol{\beta}^\top (\sum_i \mathbf{H}_i^\top \boldsymbol{\Phi}_i^{-1} \mathbf{H}_i) \boldsymbol{\beta} + \lambda R(\mathbf{w}) \\ \text{subject to} \quad & \boldsymbol{\beta}^\top \mathbf{H}^\top \mathbf{H} \boldsymbol{\beta} - n = 0 \\ & \mathbf{a}_m \boldsymbol{\beta} = 0, \mathbf{a}_f \boldsymbol{\beta} = 0 \end{aligned} \tag{7}$$

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 according to specific requirement of an application. For example, if less features should be included in the projection, $R(\mathbf{w})$ can be implemented with the ℓ_1 vector norm: $||\mathbf{w}||_1$ which is defined by $\sum_i |w_i|$, where w_i is the i-th entry of $\mathbf{w}$. When features in $\mathbf{x}$ are clustered in different groups, $R(\mathbf{w})$ can be implemented by the $\ell_{1,2}$ vector norm defined as

$$||\mathbf{w}||_{2,1} = \sum_{\ell=1}^{L} \sqrt{\sum_{i \in \mathcal{G}_\ell} \mathbf{w}_i^2}. \tag{8}$$

where $\mathcal{G}_\ell$ contains the feature indices of a group ℓ. More specifically, if we focus on the implementation of (7) with the ℓ_1 norm, i.e., we solve the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \boldsymbol{\beta}^\top (\sum_i \mathbf{H}_i^\top \boldsymbol{\Phi}_i^{-1} \mathbf{H}_i) \boldsymbol{\beta} + \lambda ||\mathbf{w}||_1 \\ \text{subject to} \quad & \boldsymbol{\beta}^\top \mathbf{H}^\top \mathbf{H} \boldsymbol{\beta} - n = 0 \\ & \mathbf{a}_m \boldsymbol{\beta} = 0, \mathbf{a}_f \boldsymbol{\beta} = 0. \end{aligned} \tag{9}$$

We next introduce an algorithm to solve Problem (9) in the following section. Note that Problem (6) can be treated as a special case of Problem (9) at $\lambda = 0$. Hence, the solver for Problem (7) with $\lambda = 0$ serves a solver for Problem (6).

4. OPTIMIZATION

The objective function in (9) is not continuously differentiable because of the one norm regularization term. In order to convert it to a canonical optimization problem and gradient-based methods can be applied, we introduce two sets of variables $\mathbf{u} \geq 0$ and $\mathbf{v} \geq 0$ both with equal length of $\mathbf{w}$. We replace $\mathbf{w}$ by $\mathbf{u} - \mathbf{v}$. Then $||\mathbf{w}||_1 = \sum_{i=1}^{d} u_i + v_i$. Correspondingly, we re-organize the variables as follows

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

Let

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

and

$$\begin{aligned} \mathbf{a}_m^\top = [\boldsymbol{\mu}_m, -\boldsymbol{\mu}_m, -1, 0], \mathbf{a}_f^\top = [\boldsymbol{\mu}_f, -\boldsymbol{\mu}_f, 0, -1], \\ \mathbf{b} = [\ _{2d}, 0, 0], \end{aligned}$$

where $_{2d}$ is a vector with length of $2d$ consisting of all ones. By change of variables, Problem (9) can be proved to be equivalent to Problem (10):

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \quad & f : \boldsymbol{\gamma}^\top (\sum_i \mathbf{C}_i^\top \boldsymbol{\Phi}_i^{-1} \mathbf{C}_i) \boldsymbol{\gamma} + \lambda \sum_{i=1}^{2d} \gamma_i \\ \text{subject to} \quad & g_1 : \boldsymbol{\gamma}^\top \mathbf{C}^\top \mathbf{C} \boldsymbol{\gamma} - n = 0 \\ & g_2 : \mathbf{a}_m^\top \boldsymbol{\gamma} = 0 \\ & g_3 : \mathbf{a}_f^\top \boldsymbol{\gamma} = 0 \\ & g_{4:e} : \mathbf{b} \cdot \boldsymbol{\gamma} \succeq 0 \end{aligned} \tag{10}$$

where $e = 2d + 5$ is the total number of constraints in the problem. As we have an equality constraint in a quadratic form, Problem (10) is not a convex optimization problem. However, considering its special structure, we can solve it efficiently in the framework of sequential quadratic programming (SQP) [15], as the gradient of both the objective function f and the constraints $g_{i:i=1:e}$ can be calculated as follows:

$$\begin{aligned} \triangledown f = (\sum_i \mathbf{C}_i^\top \boldsymbol{\Phi}_i^{-1} \mathbf{C}_i) \boldsymbol{\gamma} + \lambda \mathbf{b} \\ \triangledown g_1 = \mathbf{C}^\top \mathbf{C} \boldsymbol{\gamma}, \triangledown g_2 = \mathbf{a}_m', \triangledown g_3 = \mathbf{a}_f' \\ \triangledown g_{4:e} = \begin{bmatrix} I_{2d \times 2d} & 0 \\ 0 & 0 \end{bmatrix} \end{aligned}$$

where $I_{2d \times 2d}$ is the identity matrix of dimension $2d \times 2d$. The Lagrange function for this problem is,

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\alpha}) = f(\boldsymbol{\gamma}) + \sum_i \alpha_i g_i(\gamma). \tag{11}$$

where $\boldsymbol{\alpha}$ contains all Lagrangian multipliers. The Hessian of $\mathcal{L}$ with respect to $\boldsymbol{\gamma}$ can be calculated as,

$$\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathcal{L} = \sum_i \mathbf{C}_i^\top \boldsymbol{\Phi}_i^{-1}\mathbf{C}_i + \alpha_1 \mathbf{C}^\top\mathbf{C} \tag{12}$$

As other SQP methods, we solve the proposed problem iteratively. At each iteration $t+1$, we solve a quadratic programming subproblem given in (13) to find the direction to move towards one optimal solution.

$$\begin{aligned} \min_{\mathbf{p}} \quad & f(\boldsymbol{\gamma}_t) + \nabla f(\boldsymbol{\gamma}_t)^\top \mathbf{p} + \frac{1}{2}\mathbf{p}^\top \nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathcal{L}(\boldsymbol{\gamma}_t, \boldsymbol{\alpha}_t)\mathbf{p} \\ \text{subject to} \quad & \nabla g_i(\boldsymbol{\gamma}_t)^\top \mathbf{p} + g_i(\boldsymbol{\gamma}_t) = 0, i \in [1:3] \\ & \nabla g_i(\boldsymbol{\gamma}_t)^\top \mathbf{p} + g_i(\boldsymbol{\gamma}_t) \succeq 0, i \in [4:e] \end{aligned} \tag{13}$$

where $\mathbf{p}$ is a direction to be determined along with the objective function can be reduced. The solution $\hat{\mathbf{p}}_t$ to this subproblem together with its corresponding optimal Lagrangian multipliers $\hat{\mathbf{q}}_t$ are used to update $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ as follows:

$$\boldsymbol{\gamma}_{t+1} = \boldsymbol{\gamma}_t + s\hat{\mathbf{p}}_t, \boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + s(\hat{\mathbf{q}}_t - \boldsymbol{\alpha}_t). \tag{14}$$

where s is the learning step size which is a scalar and can be found by general line search with ℓ_1 merit function defined as following:

$$\phi(\boldsymbol{\gamma}, z) = f(\boldsymbol{\gamma}) + z\sum_{i=1}^{3} |g_i(\gamma)|. \tag{15}$$

z is a multiplier that can be chosen in each step with the following constraint:

$$z \geq \frac{\nabla f(\boldsymbol{\gamma}_t)^\top \hat{\mathbf{p}}_t + (\sigma/2)\hat{\mathbf{p}}_t{}^\top \nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathcal{L}(\boldsymbol{\gamma}_t, \boldsymbol{\alpha}_t)\hat{\mathbf{p}}_t}{(1-\rho)\sum_{i=1}^{3}|g_i(\gamma_t)|}, \tag{16}$$

for some parameter $\rho \in (0,1)$. σ is a constant, either 1 when $\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathcal{L}(\boldsymbol{\gamma}_t, \boldsymbol{\alpha}_t)$ is positive definite, otherwise 0. The directional derivative of ϕ in the direction $\hat{\mathbf{p}}_t$ is given in (17).

$$D(\phi(\boldsymbol{\gamma}_t, z), \hat{\mathbf{p}}_t) = \nabla f(\gamma_t)^\top \hat{\mathbf{p}}_t - z\sum_{i=1}^{3}|g_i(\gamma_t)| \tag{17}$$

Algorithm 1 summarizes the algorithm that we used to solve Problem (10).

Algorithm 1 A sequential quadratic programming approach for solving problem (10)

Input: $\mathbf{C}_i$, $\boldsymbol{\Phi}_i$, $\mathbf{a}'_m$, $\mathbf{a}'_f$, λ
Output: $\boldsymbol{\gamma}$
1. Initialize $\boldsymbol{\gamma}$ with $\mathbf{u} = 1$, $\mathbf{v} = 0$, and μ_m, μ_f equal to the sample male and female means of the obtained trait when $\mathbf{w} = 1$ applied.
2. Initialize $\boldsymbol{\alpha}$ with all ones.
3. Evaluate f, ∇f, ∇g_i and $\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathcal{L}$ with $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$.
4. Solve problem (13) to obtain $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$.
5. Choose z satisfy Eq. (16).
6. Do line search to find a step size s satisfying Eq.(18) with $\eta \in (0, 0.5)$.

$$\phi(\boldsymbol{\gamma} + s\hat{\mathbf{p}}, z) \leq \phi(\boldsymbol{\gamma}, z) + \eta s D(\phi(\boldsymbol{\gamma}, z), \hat{\mathbf{p}}) \tag{18}$$

7. Update $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ as in Eq.(14).
Repeat 3-7 until $\boldsymbol{\gamma}$ reaches a fixed point.

5. EVALUATION

In our experiments, we first evaluated the effectiveness of our approach using synthetic data for which we knew the ground truth to test against. We then compared our approach on real-world data sets with several clustering approaches [12, 5, 21]. To the best of our knowledge, these methods [12, 5, 21] are most suitable for comparison with the proposed one. Heritability for the derived traits was estimated using the *polygenic* program in the SOLAR software package [1]. The *polygenic* function estimates the narrow sense heritability of a quantitative trait.

5.1 Synthetic data

We synthesized data by creating the pedigrees for 250 nuclear families, each of which had 4 family members: father, mother and two children. Hence, we had 1000 subjects in total in the pedigree. The gender of all of the children were randomly assigned. Both the kinship matrix $\boldsymbol{\Phi}$ and the delta matrix $\boldsymbol{\Delta}$ were calculated for each family. Since all simulated families had the exact same structure, they had the exact same kinship matrix and delta matrix as follows:

$$\boldsymbol{\Phi} = \begin{bmatrix} 0.5 & 0 & 0.25 & 0.25 \\ 0 & 0.5 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.5 \end{bmatrix},$$

$$\boldsymbol{\Delta} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.25 \\ 0 & 0 & 0.25 & 1 \end{bmatrix},$$

with family members aligned in such order: father, mother, the first child and the second child in both columns and rows.

We then simulated a quantitative trait y which corresponded to a vector where each entry was the phenotype of a subject in the pedigree. The simulation procedure was designed according to how the heritability of a quantitative trait was typically estimated. We randomly drew points from a 4-dimensional multivariate Gaussian distribution: $N(\boldsymbol{\mu}, \boldsymbol{\Omega})$ for each family. The 4 dimensions corresponded to the 4 family members in each simulated family. Notice that the μ used in the simulation of each family may vary between families according to the gender of family members. More precisely, if a family member is male, μ was set to μ_m; or otherwise it was set to μ_f. The covariance matrix Ω was given as follows:

$$\boldsymbol{\Omega} = 2\sigma_a^2\boldsymbol{\Phi} + \sigma_d^2\boldsymbol{\Delta} + \sigma_e^2\mathbf{I}. \tag{19}$$

Without loss of generality, for simplicity, here we use identity matrix $\mathbf{I}$ as the matrix $\boldsymbol{\gamma}$ in (1). The quantitative trait was simulated with the following choice of parameters

$$[\sigma_a^2, \sigma_d^2, \sigma_e^2, \mu_m, \mu_f] = [0.8,\ 0.1,\ 0.1,\ 0.5,\ 1.1]. \tag{20}$$

We next simulated 5 phenotypic features by randomly drawing from a Gaussian distribution and used them to approximate the simulated quantitative trait by a linear function of $y_i = \mathbf{x}_i^\top \mathbf{w}$ for each simulated subject i. We regressed on the simulated $\mathbf{x}_i$ and the quantitative trait y_i and the best $\mathbf{w} = [1.3,\ 1.5,\ 0.7,\ 0,\ 1.1]$ in the simulation. The fourth variable had no effect on this simulated trait.

Before we ran the proposed algorithm, we estimated the heritability of the simulated quantitative trait. The heritability reported by SOLAR was 0.75, which was close to the parameters we set and was accurate enough taking into account the random nature of the simulation. The heritability we implanted in the data was 0.8 according to (20). This means that there is at least one specific combination of the 5 simulated features that gives us a trait of around 0.75 heritability. Hence, if our approach works, it should yield quantitative traits with a heritability estimate of at least 0.75.

We conducted a set of experiments with the proposed algorithm where we chose $\lambda = 0,\ 1,\ 2,\ 3,\ 4$ in succession. For all of the choices of λ tested, the proposed method can recover the true $\mathbf{w}$ with good accuracy. And as expected, $\mathbf{w}$ was shrunk when we increased λ. The proposed algorithm could completely rule out the effect of the fourth variable when $\lambda = 4$. The estimated heritability of the five derived quantitative traits (corresponding to $\lambda = 0,\ 1,\ 2,\ 3,\ 4$) are as following: 0.797, 0.796, 0.796, 0.796, 0.796 and 0.789, all of which are higher than the implanted one. This result shows that our approach can successfully derive quantitative traits of high heritability. When λ increases, heritability tends to decrease but without significant changes for all tested λs.

5.2 Two data sets of drug use and related behaviors

We evaluated the proposed approach on two real-world problems in genetic studies of opioid use and cocaine use separately. 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. All of the subjects identified themselves as either African American (AA) or European American (EA). Opioid use or cocaine use and related behaviors were assessed with two separate sections of the interview, each of which was dedicated to the diagnosis of opioid dependence or cocaine dependence. The computer-assisted interview is called the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA) [17].

5.2.1 Opioid use and related behaviors

A total of 4964 subjects with 1888 from small nuclear families and 3076 unrelated individuals were aggregated for the opioid use study. We included unrelated individuals in the analysis, because they contribute to phenotypic variance estimation even though they have no effect on covariance estimation. There are 23 questions in the opioid dependence section of the SSADDA, resulting in 220 clinical variables. These variables represent features of opioid use and related behaviors, including age of onset and frequency of opioid use, the occurrence of psychosocial and medical consequences of opioid use, etc. Of the 220 variables, 69 were identified and used as key features for the purpose of subtyping opioid use and related behaviors in a previous study [21]. We used these selected features in the current analysis.

We ran the proposed algorithm on this dataset and we set λ from 0 to 10 with a step size of 1. Estimated heritability of the derived quantitative traits is reported in Figure 2, which also shows the percentage of clinical variables that were selected by the ℓ_1 norm sparse regularization in our analysis. As was observed in the synthetic data, the heritability dropped when as features were removed from the model. We obtained the highest heritability, approximately 1, at $\lambda = 0$, where all features were used in the model. The lowest heritability of 0.9543 was obtained at $\lambda = 10$, where less than 44% of the features were selected for use in the model.

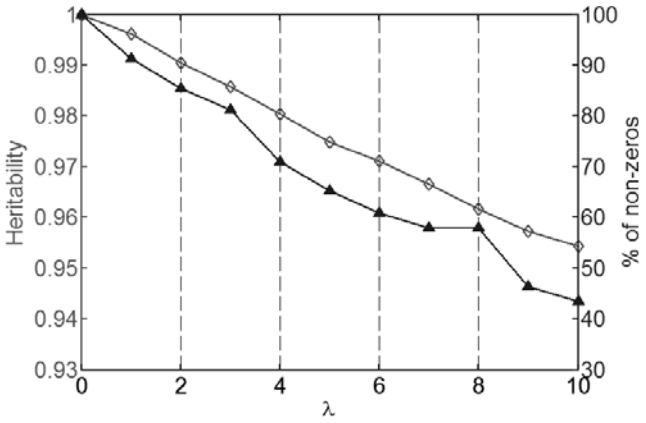

Figure 2: Heritability estimates of derived quantitative traits for opioid use and related behaviors when λ varies (marked by diamonds) and the percentage of clinical variables included in the corresponding models (marked by triangles).

We also examined the 10 clinical features that received the highest weights in the model. Figure 3 shows the coefficients w of these 10 clinical features (i.e., when $\lambda = 10$). The clinical questions corresponding to these 10 variables are listed below. These features may be worth investigating in future genetic studies of opioid dependence.

- **A1:** When you stopped, cut down, or went without (OPIATE, which reflects the opioid drug used most commonly by the respondent), did you have nausea, or did you vomit and have other withdraw symptoms for most of the day for 2 days or longer?
- **A2:** Please think about the period when you were using (OPIATE) the most. During that period, how many days per month did you use (OPIATE)?
- **A3:** Have you given up or greatly reduced important activities like sports, work, or associating with friends or relatives while using opiate?
- **A4:** Because of your (OPIATE) use, did you ever experience having trouble concentrating or having such trouble thinking clearly for more than 24 hours that it interfered with your functioning?
- **A5:** Did you ever bring up any problems you might have had with (OPIATE) with any professional?
- **A6:** Did using (OPIATE) cause you to have an overdose?
- **A7:** Because of your (OPIATE) use, did you ever experience feeling jumpy or easily startled or nervous for more than 24 hours to the point that it interfered with your functioning?

- **A8:** Because of your (OPIATE) use, did you ever experience feeling depressed or uninterested in things for more than 24 hours to the point that it interfered with your functioning?
- **A9:** How many times you have ever injected an opiate drug?
- **A10:** When you stopped, cut down, or went without (OPIATE), did you yawn and have other withdraw symptoms for most of the day for 2 days or longer?

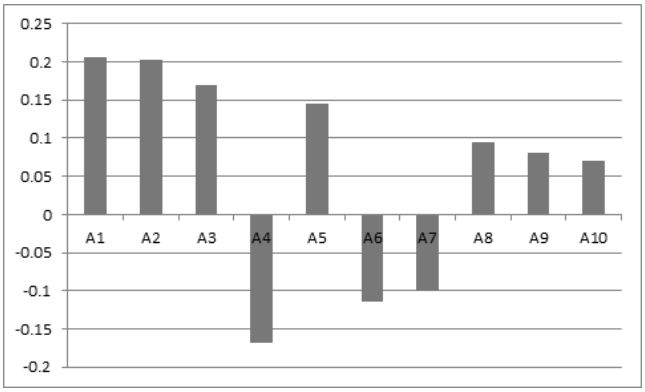

Figure 3: Weights of the top 10 variables used in the model that characterized a quantitative trait of opioid use and related behaviors.

We compared the proposed approach with a cluster analysis method published recently [21], which created a quantitative trait of estimated heritability of 0.76 using the same data. In the original study using this approach [21], it was applied to a larger sample, which included the current data as a subset and yielded five clusters, with the highest estimated heritability of the clusters being 0.76. The heritability was estimated using the same function in the SOLAR.

Another cluster analysis method [5] was also applied to the same sample used in our study. The highest heritability that the method in [5] achieved in its derived quantitative traits was 0.66. Using the proposed method, we were able to derive quantitative traits that approximated the highest possible heritability, which demonstrated a clear benefit of our algorithm.

5.2.2 Cocaine use and related behaviors

Phenotype data were collected for a set of 9436 subjects for the studies of cocaine use and related behavior. Of these 9436 subjects, 2268 were from small nuclear families; and 7268 of them were unrelated individuals. The sample overlapped with the one used for opioid dependence, but the variables were derived from a different section of the SSADDA interview. The cocaine use section of the SSADDA contains 25 questions on (1) age of onset, frequency, and intensity of cocaine use; (2) route of cocaine administration; (3) occurrence of psychosocial and medical consequences of cocaine use; (4) attempts to quit cocaine use; and (5) cocaine abuse treatment sought and received, resulting in 160 variables. There were 68 variables identified as key variables in the derivation of highly heritable quantitative traits for cocaine use and related behaviors [12]. We used these variables as inputs in the proposed method.

Similar to the analysis on opioid use data, we also ran our algorithm with λ ranging from 0 to 10 with a step size of 1. Figure 4 reports the heritability of the derived traits and the percentage of clinical variables retained in the model. Heritability decreased when fewer features were used in the quantitative trait model. We reached the highest heritability of 0.88 when $\lambda = 0$ and the corresponding quantitative trait used all clinical features. The lowest heritability was 0.87 when $\lambda = 10$, where more than 54 % of the clinical features were ruled out. Even when the number of features used in the formation of the traits was significantly reduced, the heritability of the resulting trait was more or less stable. The results suggested that the phenotypic and clinical features of cocaine use and related behaviors appeared to be less heritable than those of opioid use.

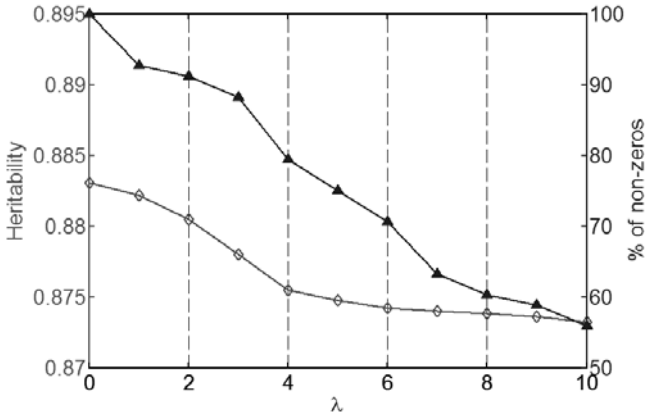

Figure 4: Heritability estimates of derived quantitative traits for cocaine use and related behaviors when λ varies (the diamond line) and the percentage of features included in the corresponding models (the triangle line).

The 10 clinical variables that received a large magnitude of weights when $\lambda = 10$ were examined. Their corresponding weights are shown in Figure 5. The questions corresponding to these 10 variables are listed as follows:

- **A1:** Did you ever use cocaine at least once a week for a month or more?
- **A2:** Have you often wanted to stop or cut down on cocaine?
- **A3:** Have you ever been under the effects of cocaine when it increased your chances of getting hurt, for instance, when driving a car or boat, using knives, machinery or guns, crossing against traffic, climbing or swimming?
- **A4:** How old were you the first time when you injected cocaine?
- **A5:** Has there ever been a period of a month or more when a great deal of your time was spent using cocaine, getting cocaine, or getting over its effects?
- **A6:** Have you ever stayed high from cocaine for a whole day or more?
- **A7:** How many times in your life have you used cocaine?
- **A8:** Did you ever use alcohol or any other drug to make yourself feel better when coming down from the effects of cocaine?

- **A9:** Have you ever used cocaine to keep from having any these problems (or to make them go away)?
- **A10:** Have you often used cocaine on more days or in larger amounts than you intended to?

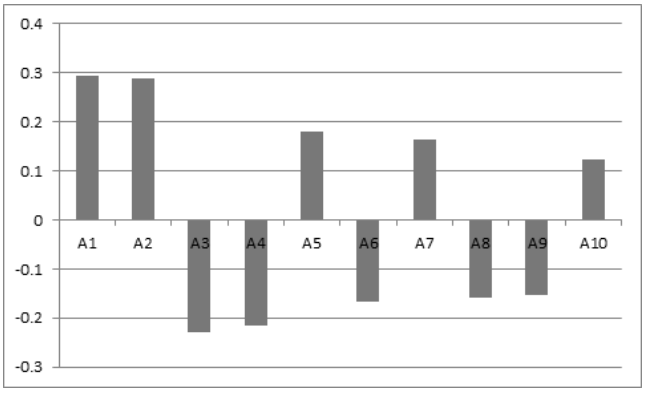

Figure 5: Weights of the top 10 variables in the model that characterizes the quantitative trait for cocaine use and related behaviors.

A subset of our cocaine-use data was employed previously in [12] to derive highly heritable quantitative traits for cocaine use and related behaviors. The highest heritability reported in [12] was 0.5. We applied that cluster analysis method to the same sample used in this study to search for highly heritable traits, and the highest heritability we could find was 0.644. Instead, the proposed method achieved a quantitative trait with a much higher heritability (i.e., 0.88).

5.3 Economically important traits of dairy cattle

A subset of pedigrees consisting of 2,544 extended families and 19,097 individuals selected from a large cattle pedigree data set (downloaded from *http://aipl.arsusda.gov/eval.htm*) were used in our analysis. All individual bulls are of Holstein breed and are from the United States. Among the 19,097 individuals, 10,216 bulls were evaluated on 38 quantitative traits, such as reliability of yield and average number of lactations per daughter. We first estimated the heritability for all 38 individual traits and found 11 traits with low-to-moderate heritability. Of these 11 traits, 8 had an estimated heritability less than 0.35, and the other three had an estimated heritabilities of 0.64, 0.65 and 0.73. We ran our algorithm on the 11 traits to derive highly heritable quantitative traits. Similar to the analysis for human subjects, we tested multiple choices of λ ranging from 0 to 10.

The heritability estimates of the derived traits when λ varies are plotted in Figure 6 together with the percentage of features remaining in the model. As expected, we obtained the highest heritability of 0.907 when $\lambda = 0$. The heritability decreased slightly when λ increased, with the lowest heritability of 0.898 when $\lambda = 10$. Two features are dropped by the model when $\lambda \geq 6$. All of the 11 derived traits had a higher heritability than that of any single input trait, which shows the effectiveness of our algorithm when heritability is estimated using extended pedigrees.

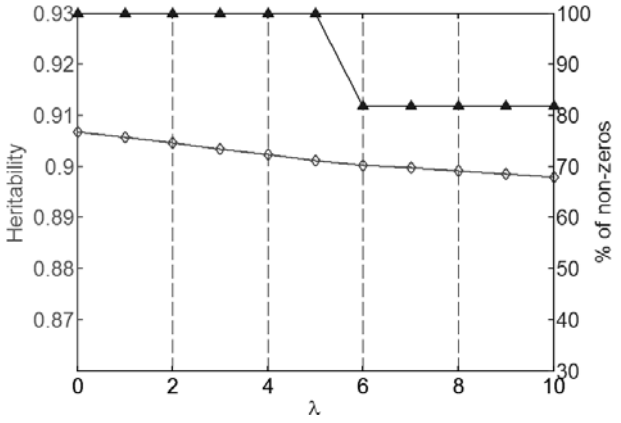

Figure 6: Heritability estimates of derived quantitative traits of dairy cattle when λ varies (the diamond line) and the percentage of features included in the corresponding models (the triangle line).

6. CONCLUSION

Discovering genetic risk factors for complex diseases is an important task in medicine and biology. The power of most genotype-phenotype association studies is positively associated with the heritability of the quantitative traits studied [2]. However, for many complex diseases, such as substance use disorders, currently no highly heritable quantitative traits are available, despite the fact that twin studies show them to be heritable diseases [11]. There is a lack of effective methods to derive such a trait, a problem that is exacerbated by the complicated structure of their clinical features. Researchers have been using cluster analysis to identify subgroups of a study population and then derive cluster-based quantitative traits to maximize heritability and homogeneity of the clinical features [12, 5, 21]. However, all previous approaches search for quantitative traits without explicitly maximizing heritability, using heritability only as the evaluation metric.

In this paper, we have proposed a quadratic optimization problem to derive quantitative traits of a linear form $y = \mathbf{x}^\top \mathbf{w}$ by explicitly maximizing heritability. We searched for the optimal $\mathbf{w}$ that maximizes the log likelihood in the heritability estimation. An optimization algorithm based on the framework of sequential quadratic programming was developed to efficiently solve the proposed formulation. The proposed approach was evaluated on synthetic data and three real-world problems. The empirical results demonstrate the effectiveness of the proposed approach to the identification of highly heritable quantitative traits. Comparing the results with those from existing cluster analysis methods on the two real-world substance dependence data sets clearly showed that the new approach was superior. Our future work will include a more thorough evaluation of the proposed method on multiple other datasets that represent difficult subtyping problems. We plan to implement the proposed method with a few more choices of other regularization terms, such as the $\ell_{1,2}$-norm discussed in an early section of this paper. These different regularization terms can help to deal with complex data structures in the clinical features, such as the different groups of clinical features in the opioid and cocaine use datasets.

7. ACKNOWLEDGMENTS

This work was supported by NIH grants DA12849, DA12690, AA03510, AA11330, and AA13736 and Philadelphia VA Mental Illness Research, Education, and Clinical Centers (MIRECCs).

8. REFERENCES

[1] L. Almasy and J. Blangero. Multipoint quantitative trait linkage analysis in general pedigrees. *American journal of human genetics*, 62(5):1198–211, 1998.

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

[3] C. Braet and W. Beyers. Subtyping children and adolescents who are overweight: Different symptomatology and treatment outcomes. *J. Consult. Clin. Psychol. Journal of Consulting and Clinical Psychology*, 77(5):814–824, 2009.

[4] 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. *Eur. Respir. J. European Respiratory Journal*, 36(3):531–539, 2010.

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

[6] D. S. Falconer and T. C. Mackay. *Introduction to quantitative genetics, 4th Edition*. Benjamin Cummings, 1996.

[7] J. Gelernter, C. Panhuysen, R. Weiss, K. Brady, V. Hesselbrock, B. Rounsaville, J. Poling, M. Wilcox, L. Farrer, and H. 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 : the official publication of the International Society of Psychiatric Genetics.*, 136(1):45, 2005.

[8] J. Gelernter, C. Panhuysen, M. Wilcox, V. Hesselbrock, B. Rounsaville, J. Poling, R. Weiss, S. Sonne, H. Zhao, L. Farrer, and H. R. Kranzler. Genomewide linkage scan for opioid dependence and related traits. *American Journal of Human Genetics.*, 78(5):759, 2006.

[9] D. C. Glahn, J. E. Curran, A. M. Winkler, M. A. Carless, J. W. Kent, J. C. Charlesworth, M. P. Johnson, H. H. H. Goring, S. A. Cole, T. D. Dyer, E. K. Moses, R. L. Olvera, P. Kochunov, R. Duggirala, P. T. Fox, L. Almasy, J. Blangero, and D. Molecular Substrates of Neuroplasticity in. High dimensional endophenotype ranking in the search for major depression risk genes. *Biological psychiatry*, 71(1):6–14, 2012.

[10] W. G. Hill and N. R. Wray. Heritability in the genomics era - concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266, 2008.

[11] K. S. Kendler, K. C. Jacobson, C. A. Prescott, and M. C. Neale. Specificity of genetic and environmental risk factors for use and abuse/dependence of cannabis, cocaine, hallucinogens, sedatives, stimulants, and opiates in male twins. *Am J Psychiatry*, 160(4):687–95, 2003.

[12] 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. *Addictive Behavior*, 33(1):41–53, 2008.

[13] K. Lange, J. Westlake, and M. A. Spence. Extensions to pedigree analysis. iii. variance components by the scoring method. *Ann Hum Genet*, 39(4):485–91, 1976. 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.

[14] M. J. Niciu, G. Chan, J. Gelernter, A. J. Arias, K. Douglas, R. Weiss, R. F. Anton, L. Farrer, J. F. Cubells, and H. R. Kranzler. Subtypes of major depression in substance dependence. *Addiction*, 104(10):1700–1709, 2009.

[15] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

[16] 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(6):1325–1332, 2010.

[17] 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(1):85–90, 2007.

[18] 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*, 5:475–504, 1992.

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

[20] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge Univeristy Press, 2004.

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

[22] M. Weatherall, J. Travers, P. Shirtcliffe, S. Marsh, M. Williams, M. Nowitz, S. Aldington, and R. Beasley. Distinct clinical phenotypes of airways disease defined by cluster analysis. *European Respiratory Journal*, 35(2):459–460, 2010.

[23] D. Williams, R. De Silva, D. Paviour, A. Pittman, H. Watt, L. Kilford, J. Holton, T. Revesz, and A. Lees. Characteristics of two distinct clinical phenotypes in pathologically proven progressive supranuclear palsy: Richardson's syndrome and psp-parkinsonism. *Brain*, 128(6):1247–1258, 2005.