Identifying Heritable Composite Traits from Multivariate Phenotypes and Genome-wide SNPs

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

Henry R. Kranzler
Treatment Research Center
University of Pennsylvania Perelman School of Medicine
and VISN 4 MIRECC , Philadelphia VAMC
Philadelphia, PA, USA
kranzler@mail.med.upenn.edu

***Abstract*—An important approach to reducing missing heritability and enhancing success of genome-wide association studies (GWAS) for complex diseases is the identification of traits that are highly heritable and homogeneous in their etiology. Many approaches have been proposed to define such traits based on either cluster analysis or pedigree-based heritable component analysis. None of the existing methods, however, exploit the dense genome-wide genotypic data that are now readily available from GWAS, and with exome and whole genome sequencing more data will be available in the future. Moreover, because a phenotype can vary with respect to a covariate, such as age or race. The fixed effect due to the covariates may lead to a spuriously elevated estimate of heritability. Existing heritable component analysis methods have not considered covariate effects. We propose an optimization approach to identify composite traits with high heritability as a function of multiple phenotypic variables where heritability is estimated from genome-wide single neucleotide polymorphisms (SNPs). Our approach can model the covariate effects within heritability analysis. The proposed optimization problem can be efficiently solved by a sequential quadratic programming algorithm. A case study demonstrates the effectiveness of the proposed approach for finding composite traits with high SNP-based heritability.**

***Keywords*-phenotype-genotype analysis, chip heritability, quadratic optimization**

I. Introduction

Genome-wide association studies (GWAS) have had limited success in elucidating genetic etiology of complex diseases, such as substance use disorders [1], [2], [3]. The heterogeneity in disease phenotypes, e.g., in the variables that are used to diagnose and characterize the disease, is considered to be one of the major obstacles to greater success in gene finding through GWAS [3]. Defining composite traits that characterize more homogeneous subtypes of a disease could help to remove this obstacle. Unsupervised cluster analysis has commonly been used to partition a study population into subgroups based on clinical variables [4], [5], [6], [7]. This approach can create subgroups that are homogeneous phenotypically. However, it has limited utility in association analysis, because genetic data is not used in the creation of the subgroups.

For a complex disease, clinical variables vary among subjects, resulting in large phenotypic variance in the disease population. These variables also show varying degrees of heritability, e.g., some are more genetically influenced than others. Identifying highly heritable components of disease phenotypes is important because it increases the likelihood of identifying genetic associations with these components. Pedigree-based methods have been developed to identify principal components of phenotypic data that are highly heritable [8], [9], [10], [11]. These methods can only be used in a family-based study. As is widely recognized, it is challenging to recruit multi-member families on a large scale. With the availability of dense genotypic data, heritability can now be estimated from unrelated individuals with genome-wide genetic markers, using what is called chip heritability [12].

In this paper, we propose an optimization approach to identify a trait of high chip heritability by solving the inverse problem of (chip) heritability estimation. To estimate the chip heritability of a given trait, recently published methods use the restricted maximum likelihood (REML) method if the trait follows a mixed effect model with random genetic effects and fixed effects due to covariates [13], [12]. We propose to identify a linearly combined trait with high chip heritability when estimated using the REML method. 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 a real world dataset from a multi-site cocaine dependence study. Our experimental results show the effectiveness of the proposed approach.

II. Method

Given a set of n subjects, we use a vector $\mathbf{y}$ of length n to denote trait values for a quantitative trait y, a matrix $\mathbf{Z}_{n\times m}$ to represent standardized genotypic data at m genetic markers, and $\mathbf{C}_{n\times p}$ to represent data on p covariates. The matrix $\mathbf{Z}$ is calculated from the genotypic data as follows: let f_j be the frequency of the j-th genetic variant, r_{ij} be the number of copies of a reference allele at the j-th genetic

978-1-4799-5669-2/14/$31.00 ©2014 IEEE

variant of the i-th subject, and the standardized genotype z_{ij} is calculated as $(r_{ij} - f_j)/\sqrt{2f_j(1-f_j)}$. Following the model in the chip heritability estimation method, we consider the following mixed linear model that characterizes how a phenotype is related to genotypes and covariates:

$$\mathbf{y} = \mathbf{C}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\varepsilon}$ is a vector of length n, specifying residual effects. In Eq.(1), all covariates create 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 $\mathbf{V}$, can be calculated as:

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

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

$$\mathbf{V} = \mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_e^2. \tag{3}$$

where σ_g^2 and σ_e^2 can be estimated by the REML method [14], [15]; and the chip heritability estimated on the m genetic 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 methods propose to use genome-wide SNPs to estimate a GRM [13], [12].

In REML, the log likelihood (after removing constants) used to estimate σ_g^2 and σ_e^2 can be written as:

$$\ell(\sigma_g^2, \sigma_e^2; \mathbf{y}, \mathbf{C}, \mathbf{Z}) = -\frac{1}{2}(\ln|\mathbf{V}| + \ln|\mathbf{C}^T\mathbf{V}^{-1}\mathbf{C}| + \mathbf{y}^T\mathbf{P}\mathbf{y}), \tag{4}$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{C}(\mathbf{C}^T\mathbf{V}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{V}^{-1}$. Given data: $\mathbf{y}$, $\mathbf{C}$ and $\mathbf{Z}$, σ_g^2 and σ_e^2 are optimized to maximize $\ell(\sigma_g^2, \sigma_e^2; \mathbf{y}, \mathbf{C}, \mathbf{Z})$ [15]. The chip heritability of a trait y contributed by genetic causal variants is computed using the optimal $\hat{\sigma_g}^2$ and $\hat{\sigma_e}^2$.

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 $\mathbf{Z}$ and 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 as to form a trait $\mathbf{y}$ 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 ease of 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 [16].

Note 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 in the log likelihood and removing constants yield the following objective function:

$$\min_{\mathbf{w}} \quad \mathbf{w}^T(\mathbf{X}^T\mathbf{P}\mathbf{X})\mathbf{w} \tag{5}$$

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

$$\mathbf{P} = \mathbf{G}^{-1} - \mathbf{G}^{-1}\mathbf{C}(\mathbf{C}^T\mathbf{G}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{G}^{-1}. \tag{6}$$

Since $\sigma_g^2 = 1$ and $\sigma_e^2 = 0$, the phenotypic covariance matrix $\mathbf{V} = \mathbf{G}$ (based on Eq.(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 so imposes a constraint in the optimization problem that the total phenotypic variance is scaled to 1. As the true distribution of the trait is not known, σ_p^2 cannot be estimated. It is commonly approximated by the sample variance of the trait (denoted by s^2).

$$\begin{aligned} s^2 &= \mathbf{w}^T\left(\frac{1}{n}\mathbf{X}^T\mathbf{X} - \frac{1}{n^2}\mathbf{X}^T\mathbf{1}\mathbf{1}^T\mathbf{X}\right)\mathbf{w} \\ &= \mathbf{w}^T\left(\mathbf{X}^T\left(\frac{1}{n}\mathbf{I} - \frac{1}{n^2}\mathbf{1}\mathbf{1}^T\right)\mathbf{X}\right)\mathbf{w}. \end{aligned}$$

where $\mathbf{I}$ is an identity matrix of $n \times n$ and $\mathbf{1}$ is an n-entry vector of all ones. Let

$$\mathbf{Q} = (1/n)\mathbf{I} - (1/n^2)\mathbf{1}\mathbf{1}^T. \tag{7}$$

Then, s^2 can be simplified to $\mathbf{w}^T(\mathbf{X}^T\mathbf{Q}\mathbf{X})\mathbf{w}$. Combining the objective function and the constraint together, the proposed optimization problem is formulated as:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^T(\mathbf{X}^T\mathbf{P}\mathbf{X})\mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^T(\mathbf{X}^T\mathbf{Q}\mathbf{X})\mathbf{w} = 1 \end{aligned} \tag{8}$$

We now regularize the linear model by including a regularizer on $\mathbf{w}$ that aims to avoid the overfitting problem. If overfitting occurs, the optimal $\mathbf{w}$ of Problem (8) may correspond to a trait that has high heritability using the data that are used to train the linear model, but when the model is applied to a new sample, the resultant trait has low heritability. 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:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^T(\mathbf{X}^T\mathbf{P}\mathbf{X})\mathbf{w} + \lambda R(\mathbf{w}) \\ \text{subject to} \quad & \mathbf{w}^T(\mathbf{X}^T\mathbf{Q}\mathbf{X})\mathbf{w} = 1, \end{aligned} \tag{9}$$

where λ is a hyper-parameter that has 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 statistical learning theory [16], minimizing $\mathbf{w}^T(\mathbf{X}^T\mathbf{P}\mathbf{X})\mathbf{w}$ corresponds to empirical risk minimization, whereas minimizing the objective in Eq.(9) corresponds to structural risk minimization that improves the generalizability of the resultant model. There are many different ways to realize $R(\mathbf{w})$. A common choice is $\|\mathbf{w}\|_2^2 = \sum_i w_i^2$. Another regularizer defined by $\|\mathbf{w}\|_1 = \sum_i |w_i|$ is a better choice when model sparsity is required to select fewer variables for use in the model. In more complicated applications, where variables may be grouped and 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.

Unlike our earlier work [11] that used related individuals to derive highly-heritable traits based on pedigrees, this study employs genome-wide SNPs to compute a GRM to replace the kinship matrix derived from pedigrees. However, similar to [11], an efficient sequential-quadratic-optimization (SQP) algorithm can be developed to solve Problem (9), while the regularization term is realized by $\|\mathbf{w}\|_1$.

III. Computational Results

We validated the proposed approach in the analysis of a real-world data set aggregated from a multi-site genetic study of cocaine dependence (CD) involving the University of Connecticut Health Center, Yale University School of Medicine, the University of Pennsylvania Perelman School of Medicine, McLean Hospital and the Medical University of South Carolina. All subjects underwent phenotypic assessment and provided a blood or saliva sample for genotyping according to procedures approved by the institutional review board at each participating site. There were 6,621 subjects genotyped at a total of 1,140,420 SNPs genome-wide. Among them, 2,674 were African American, and only these subjects were involved in our experiments to avoid spurious findings due to population structure. We removed 537 subjects with other family members in the data so that GRM could be computed for unrelated individuals.

A series of data cleaning steps were performed to ensure the quality of genotypic markers. Markers that met any of the following conditions were excluded: low call rate ($< 98\%$), G/C and A/T markers (to avoid strand issues), deviation from Hardy-Weinberg equilibrium ($p < 1e-8$), a significant cohort calling discrepancy and being monomorphic. We also removed non-autosomal markers, so that only markers from the 22 autosomal chromosomes were used in the analysis. After data cleaning, 690,864 SNPs remained. Genetic relationship was estimated for each pair of subjects using the Genome-wide Complex Trait Analysis (GCTA) software [15] and all 690,864 SNPs. We subsequently excluded 385 subjects whose relatedness with other subjects was greater than 0.025 (corresponding to the relatedness of second cousins). Ultimate, 1,752 subjects were used in our analysis.

All subjects were interviewed with a computer-assisted survey instrument, the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA) [17], which consists of questions designed to elicit key features of cocaine use and related behaviors. The questions include those that evaluate the presence of seven criteria for the DSM-IV diagnosis of CD. These seven criteria are as follows:

- *F1* - tolerance to the effects of cocaine;
- *F2* - withdrawal from cocaine with abstinence;
- *F3* - using cocaine in larger amounts or over a 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* - having given 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.

In our experiments, the seven variables are coded as 0 (absent) or 1 (present). In a previous study [1], the number of CD criteria (0-7) was shown to be a better trait for use in genetic association studies than the binary trait represented by the presence and absence diagnosis of CD. The cocaine criterion count is defined as the number of positive responses to the seven variables, a composite trait resulting 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 yielded a trait with a higher heritability (h^2) estimate than that of the cocaine criterion count. Because all seven variables are binary, their h^2s were not estimated here.

We ran 10 times three-fold cross validation (CV) to determine a proper value of λ. Once λ was determined, we ran the proposed method on the entire sample to identify the final trait. All the reported values of h^2 were estimated using GCTA with a GRM computed using the 690,864 SNPs. In all of these expriments, we used age, sex and the first three PCs of the GRM as covariates.

The test values of h^2s for all traits derived in the CV are plotted in Figure 1. When $\lambda = 4$, the derived traits had the highest heritability estimates, with a mean of 0.29. We then set $\lambda = 4$ and applied our method, developing a trait from the entire sample, which had an estimated chip $h^2 = 0.3$ (s.e. 0.27). We also estimated the chip h^2 of the cocaine criterion count using the same sample set, GRM and covariates. It had an h^2 estimate near zero. These results demonstrate the effectiveness of our approach in identifying a heritable composite trait from a complex multivariate phenotype.

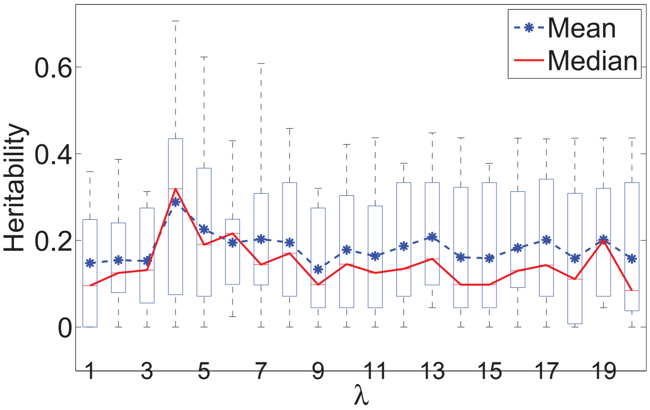

Figure 1. Case study on cocaine dependence dataset: testing values of h^2 of the composite traits derived in three-fold cross validation with varying λ.

IV. Conclusion

We developed an approach to identify composite traits from multivariate phenotypes that are highly heritable when estimated using genome-wide SNPs. The trait we derived is in the form of a linear combination of variables related to the phenotype. A quadratic optimization problem was formulated, which searches for the combination weights 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. Our empirical results on a case study demonstrate the effectiveness of our approach as it identifies traits with much higher chip h^2 than commonly-used disease phenotypes.

In this paper, the pairwise genetic relationships among subjects were estimated from genome-wide SNPs. However, it could also be estimated using 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 resulting from the proposed approach will achieve the maximized genetic variance component corresponding to this region. In an application, such as substance dependence, where there are known pathways underlying the biology of the disorder, it may be of interest to determine whether a composite trait exists, the variance of which can be largely explained by variants within these pathways, which will be a future application of our approach.

Acknowledgment

JB and her student JS were supported by NSF IIS-1320586 and DBI-1356655. Disclosure: HRK has served as a consultant or advisory board member for the following companies: Alkermes, Lilly, Lundbeck, Otsuka, Pfizer, and Roche. He is a member of the Alcohol Clinical Trials Group of the American Society of Clinical Psychopharmacology, which is supported by Abbvie, Alkermes, Ethypharm, Lilly, Lundbeck, and Pfizer.

References

[1] 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.

[2] 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*, vol. 76, no. 1, pp. 66–74, 2014.

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

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

[5] 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," *Am J Med Genet B Neuropsychiatr Genet*, vol. 165B, no. 2, pp. 148–56, 2014.

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

[7] 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.

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

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

[10] 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*, vol. 11, no. 2, 2012.

[11] J. Sun, J. Bi, and H. R. Kranzler, "Quadratic optimization to identify highly heritable quantitative traits from complex phenotypic features," in *SIGKDD'13*. New York, NY, USA: ACM, 2013, pp. 811–819.

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

[13] J. Yang, B. Benyamin, ..., and P. M. Visscher, "Common snps explain a large proportion of the heritability for human height," *Nat Genet*, vol. 42, no. 7, pp. 565–9, 2010.

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

[15] 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*, vol. 88, no. 1, pp. 76–82, 2011.

[16] V. N. Vapnik, "An overview of statistical learning theory," *Ieee Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[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*, vol. 91, no. 1, pp. 85–90, 2007.