arXiv:1305.4413v1 [stat.ME] 19 May 2013

A Penalized Multi-trait Mixed Model for Association Mapping in Pedigree-based GWAS

Jin Liu[†1], Can Yang[†1], Xingjie Shi[1,2], Cong Li[1], Jian Huang[3], Hongyu Zhao[*1], and Shuangge Ma[*1]

[1]School of Public Health, Yale University, New Haven, CT 06520, U.S.A.
[2]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China
[3]Department of Statistics & Actuarial Science, Department of Biostatistics, University of Iowa, Iowa City, IA 52242, U.S.A.

May 21, 2013

Abstract

In genome-wide association studies (GWAS), penalization is an important approach for identifying genetic markers associated with trait while mixed model is successful in accounting for a complicated dependence structure among samples. Therefore, penalized linear mixed model is a tool that combines the advantages of penalization approach and linear mixed model. In this study, a GWAS with multiple highly correlated traits is analyzed. For GWAS with multiple quantitative traits that are highly correlated, the analysis using traits marginally inevitably lose some essential information among multiple traits. We propose a penalized-MTMM, a penalized multivariate linear mixed model that allows both the within-trait and between-trait variance components simultaneously for multiple traits. The proposed penalized-MTMM estimates variance components using an AI-REML method and conducts variable selection and point estimation simultaneously using group MCP and sparse group MCP. Best linear unbiased predictor (BLUP) is used to find predictive values and the Pearson's correlations between predictive values and their corresponding observations are used to evaluate prediction performance. Both prediction and selection performance of the proposed approach and its comparison with the uni-trait penalized-LMM are evaluated through simulation studies. We apply the proposed approach to a GWAS data from Genetic Analysis Workshop (GAW) 18.

Keywords: Multivariate linear mixed model; Penalization approach; Feature selection;

GWAS.

1 Introduction

Genome-wide association studies (GWAS) help us better understand the genetic basis of many complex traits [McCarthy et al., 2008]. A better understanding of the relationship between phenotypic trait and genetic variation for these quantitative and complex traits will yield insights that are essential to predict disease risk and develop personalized therapeutic treatments for human-beings. At the beginning stage of GWAS, researchers mainly focused on a single trait analysis [Burton et al., 2007]. Although GWAS have identified some of the genetic risk variants [Visscher et al., 2012], those identified variants can only explain a small fraction of phenotypic variance, which is known as the "missing heritability" problem [Manolio et al., 2009]. Recent analysis suggested that a substantial proportion of heritability was not missing but hidden in the common variants with small or moderate effects [Yang et al., 2010, Makowsky et al., 2011].

On one hand, these results suggest that recruiting a large sample size will help to identify genetic risk variants but it could be very expensive. On the other hand, researchers start to be interested in simultaneously analyzing multiple correlated traits recently to improve the statistical power [Korte et al., 2012]. This is because the correlated traits may share common genetic factors, which is known as pleiotropy [Sivakumaran et al., 2011]. For example, a "pleiotropic enrichment" method was applied to analyze the GWAS data sets of schizophrenia and cardiovascular disease. The power to detect schizophrenia-associated common variants was shown to be improved by exploiting the pleiotropy between these two phenotypes. More recently, a study on genome-wide SNP data for five psychiatric disorders in 33,332 cases and 27,888 controls identified four significant loci ($P < 5 \times 10^{-8}$) affecting

[†]Co-first authors. * To whom correspondence should be addressed.

multiple disorders [Consortium, 2013]. It is expected that successfully taking account for the pleiotropy structure will be helpful for identification of risk variants.

In this study, a GWAS from GAW 18 with multiple traits that are highly correlated is analyzed. This type of data exposes an opportunity to integratively analyze multiple traits from a GWAS. In this paper, we focus on GAW 18 data with two highly correlated traits – systolic blood pressure (SBP) and diastolic blood pressure (DBP). We propose a unified framework to simultaneously analyze multiple traits, in which we introduce a variance component to account for sample relatedness or the confounding effects of population stratification, and introduce some sparse penalties to detect risk variants. Our approach bridges the advantages of multi-trait linear mixed models with penalized regression techniques. For the choice of sparse penalties, we have two types of models — homogeneity and heterogeneity models. Homogeneity model assumes that all trait-associated markers/variants are consistent across all traits, while heterogeneity model assumes that a marker/variant may be associated with some traits but not others. Depending on the assumption of homogeneity and heterogeneity, group MCP and sparse group MCP can be used to conduct variable selection.

The rest of the article is organized as follows. In Section 2, we show the data structure and review variance components model in genetics. The estimation of variance components, predictions, penalized selection and method to collapse SNPs are described in Section 3. Numerical studies, including simulation in Section 4 and data analysis in Section 5, are conducted to investigate finite sample performance. The article concludes with discussion in Section 6.

2 Data and Model

2.1 GAW 18 Data

The genetic analysis workshop (GAW) 18 provides the type 2 diabetes genetic exploration by next–generation sequencing in ethnic samples (T2D–GENES) consortium data set that consists of 1,043 individuals from 20 Mexican American pedigrees enriched for type 2 diabetes from San Antonio, TX. The study included subjects in two different groups, including the San Antonio family heart study (SAFHS) and the San Antonio family diabetes/gallbladder study (SAFDGS), which are together referred to as the San Antonio family studies (SAFS). Whole genome sequence is being performed commercially at Complete Genomics, Inc and the GAW 18 data set is based on the sequence data for the first 483 T2D–GENES. GWAS data for 472,049 SNPs on odd numbered autosomes are provided for these 959 family members (464 directly sequenced and the rest imputed [Howie et al., 2012]). A variety of different phenotypic traits were measured at examination, e.g. systolic blood pressure (SBP), diastolic blood pressure (DBP). Clearly, SBP and DBP are highly correlated traits. GAW 18 data set brings a good opportunity to develop statistical models to handle multiple correlated traits in the pedigree-based samples. We aim at identifying risk variants while accounting for the correlation among multiple traits and the relatedness among the samples.

2.2 Variance Components Model in Genetics

Recently, mixed model has been extensively studied for correcting the genetic relatedness in association mapping in genome-wide association studies (GWAS). The genetic relatedness from population mixture and inbred strains can cause the problem of inflated false positive rates. However, most of the existing methods fail to consider the mixed models with multiple traits. Denote that in a GWAS, we have n subjects and p genes of genetic scores with m traits. Assume that we have two traits—trait 1 and trait 2. Note that it can be relaxed to

more than two traits. For each trait, the relatedness matrix (K) can be used to describe the genetic relatedness. For multiple traits, we vectorize the multiple traits. When it comes to the analysis of multiple traits, a natural extension is to use diag(K, K). Here, we still miss a variance component describing the relatedness between/among multiple traits. Lee et al. [2012] used the covariance components among random effects across multiple traits to describe this relatedness. We go further on this direction including covariance components among residuals across multiple traits. The variance covariance matrix for vectorized two traits is given:

$$VC = \begin{pmatrix} K\sigma^2_{g^{(1)}} + \mathrm{I}_n\sigma^2_{e^{(1)}} & K\sigma_{g^{(12)}} + \mathrm{I}_n\sigma_{e^{(12)}} \\ K\sigma_{g^{(12)}} + \mathrm{I}_n\sigma_{e^{(12)}} & K\sigma^2_{g^{(2)}} + \mathrm{I}_n\sigma^2_{e^{(2)}} \end{pmatrix}, \tag{1}$$

where $\sigma^2_{g^{(1)}}$ and $\sigma^2_{g^{(2)}}$ are variance components for random effects on trait 1 and trait 2, $\sigma_{g^{(12)}}$ is the covariance of random effects between trait 1 and trait 2, $\sigma^2_{e^{(1)}}$ and $\sigma^2_{e^{(2)}}$ are variance components for residuals on trait 1 and trait 2, and $\sigma_{e^{(12)}}$ is the covariance of residuals between trait 1 and trait 2. We implement "Average information - restricted maximal likelihood" method (AI-REML) [Gilmour et al., 1995] to estimate variance components. With the variance components fixed, we may implement penalization methods to conduct variable selection and point estimation simultaneously. More details are discussed in Section 3.

3 Variance Components and Penalized Regression

Let us first consider the linear mixed model (LMM) which is widely used in single-trait analysis [Zhang et al., 2010, Kang et al., 2010, Lippert et al., 2011, Zhou and Stephens, 2012, Rakitsch et al., 2013] and then extend it to handle multiple correlated traits. Let n be

the sample size, the LMM can be written as:

$$\begin{aligned} y_o &= Wv + Xb + g + e, \\ g &\sim \mathrm{N}(0, \sigma_g^2 \mathrm{K}), \\ e &\sim \mathrm{N}(0, \sigma_e^2 \mathrm{I}), \end{aligned} \tag{2}$$

where $y_o \in \mathbb{R}^{n\times 1}$ is the response vector representing the trait, $W \in \mathbb{R}^{n\times q}$ is the matrix of covariates (fixed effect) including the intercept and other covariates such as age and gender, b is the vector for regression coefficients of the covariates, $X \in \mathbb{R}^{n\times p}$ is the genotype matrix and b is the vector for the effect sizes of p SNPs (fixed effects), g is the random effect from $\mathrm{N}(\mathbf{0}, \sigma_g^2 \mathrm{K})$, and e is the residual error with variance σ_e^2. Here the covariance matrix K is the genetic relatedness matrix which describes the pedigree structure among the individuals and σ_g^2 is the variance component of g. The covariance matrix K can be constructed according to the known pedigree information or estimated from genome-wide SNP information. This model can be interpreted as follows: The random effect g can be considered as a global average of signals from the genetic background and the shared environmental influence and we call it "average polygenic effect". For those SNPs with large effects which are different from the genetic background, they are put into the design matrix X and considered as fixed effects. In this way, the markers with larger effects can be treated locally.

3.1 Computation of Variance Components

Now we extend single-trait model (2) to a multiple-trait model. Let y_o be an $n \times m$ response matrix with each row representing subject and each column representing a trait. Let $\mathbf{g}(= (g^{(1)}, \ldots, g^{(m)}))$ and $\mathbf{e}(= (e^{(1)}, \ldots, e^{(m)}))$ be $n\times m$ matrix of unobserved polygenic and random residual effects, respectively. Denote W be $n \times q$ non-genetic covariates and X be $n \times p$ genetic scores of candidate genes. Correspondingly, we denote $V(= (v^{(1)}, \ldots, v^{(m)}))$ and $B(= (b^{(1)}, \ldots, b^{(m)}))$ the corresponding $q \times m$ coefficient matrix for q non-genetic covariates

and $p \times m$ coefficient matrix for p genetic scores in m traits. We denote $y_o^{(l)}$ the lth trait and further denote that $v^{(l)}$, $b^{(l)}$, $g^{(l)}$ and $e^{(l)}$ are the corresponding vectors of coefficient for non-genetic effects, genetic effects, average polygenic effects and vector of random residual, respectively, for $l(= 1, \ldots, m)$. First, consider linear mixed model for lth trait:

$$y_o^{(l)} = Wv^{(l)} + Xb^{(l)} + g^{(l)} + e^{(l)},$$

where $g^{(l)} \sim \mathrm{N}(\mathbf{0}, \mathrm{K}\sigma^2_{g^{(l)}})$ and $e^{(l)} \sim \mathrm{N}(\mathbf{0}, \mathrm{I}_n\sigma^2_{e^{(l)}})$. $\sigma^2_{g^{(l)}}$ and $\sigma^2_{e^{(l)}}$ are variance components describing relations among subjects. In order to account for the genetic correlation and residual correlation for multiple traits, we introduce $\sigma_{g^{(l,k)}}$ and $\sigma_{e^{(l,k)}}$ to describe the covariance of average polygenic effects and residual for lth and kth trait, respectively. Consider the multivariate linear mixed model:

$$\mathbf{y}_o = WV + XB + \mathbf{g} + \mathbf{e}, \tag{3}$$

where we assume that

1. $\mathrm{vech}(\mathbf{e}) \sim \mathrm{N}(\mathbf{0}, \mathrm{I}_n \otimes \Sigma_e)$ where Σ_e is a $m \times m$ matrix describing covariance structure among multiple traits.

2. $\mathrm{vech}(\mathbf{g}) \sim \mathrm{N}(\mathbf{0}, \mathrm{K} \otimes \Sigma_g)$ where Σ_g is $m \times m$ matrix describing covariance structure among multiple traits. K is an $n \times n$ genetic relatedness matrix (twice of the kinship matrix) and it can be calculated using genome-wide genetic markers [Yang et al., 2010] or directly obtained from unknown pedigree information.

There is a difficulty in applying the model when $q + p + m(m + 1) > n$, when the number of parameters exceeds the number of samples (d is the number of covariates, p is the number of SNPs treated as fixed effects, $m(m + 1)$ is the number of variance components). In order to overcome this difficulty, we introduce sparse constraints on b to perform variable

selection, such that model (3) is well defined. To incorporate the feature of homogeneity and heterogeneity structure among traits, we propose to use group MCP and sparse group MCP. We call this approach as "Penalized Multi-Trait Mixed Models (Penalized-MTMM)".

For simplicity, here we only consider two traits ($m = 2$) but the framework for more than two traits remains the same. For $m = 2$, we have $\Sigma_e = \begin{pmatrix} \sigma^2_{e^{(1)}} & \sigma_{e^{(12)}} \\ \sigma_{e^{(12)}} & \sigma^2_{e^{(2)}} \end{pmatrix}$ and $\Sigma_g = \begin{pmatrix} \sigma^2_{g^{(1)}} & \sigma_{g^{(12)}} \\ \sigma_{g^{(12)}} & \sigma^2_{g^{(2)}} \end{pmatrix}$. Denote that $S = \text{diag}(W, W)$, $v = \text{vech}(V)$, $T = \text{diag}(X, X)$, $b = \text{vech}(B)$, $y = \text{vech}(\mathbf{y_o})$, $g = \text{vech}(\mathbf{g})$ and $e = \text{vech}(\mathbf{e})$. Model (3) becomes

$$\begin{aligned} y &= Sv + Tb + g + e, \\ g &\sim \text{N}(\mathbf{0}, \text{K} \otimes \Sigma_g), \\ e &\sim \text{N}(\mathbf{0}, \text{I}_n \otimes \Sigma_e), \end{aligned} \tag{4}$$

Integrating out g and e and we have $y \sim \text{N}(Sv + Tb, \text{K} \otimes \Sigma_g + \text{I}_n \otimes \Sigma_e)$. The log-likelihood can be analytically written as

$$L(v, b, \Sigma_g, \Sigma_e) = -\frac{1}{2}\left[2n\log(2\pi) + \log(|H|) + (y - Sv - Tb)^T H^{(-1)}(y - Sv - Tb)\right]. \tag{5}$$

where $H = \text{K} \otimes \Sigma_g + \text{I}_n \otimes \Sigma_e$. Now we introduce sparse penalties on the coefficient b and the penalized log-likelihood can be written as

$$L(v, b, \Sigma_g, \Sigma_e) = -\frac{1}{2}\left[2n\log(2\pi) + \log(|H|) + (y - Sv - Tb)^T H^{(-1)}(y - Sv - Tb)\right] - P_\lambda(b). \tag{6}$$

where λ is the regularization parameter.

Clearly, when b is fixed, the optimization of penalized log-likelihood function (6) can be solved by the standard "Average information - restricted maximal likelihood" method (AI-REML) [Gilmour et al., 1995]. When (v, Σ_g, Σ_e) are all known, we will show that maximization of penalized log-likelihood becomes a penalized least square problem. We will carefully discuss different penalties in next section.

3.2 Computation of Penalized-MTMM

We begin with $b = 0$ and solve (6) by AI-REML. After obtaining (v, Σ_g, Σ_e), we can transform the log-likelihood function (6) to penalized least square problem as follows. Let $\hat{H}$ and $\hat{v}$ be the estimate of H and v, given by AI-REML, respectively. Denote that $\tilde{y} = \hat{H}^{-1/2}(y - S\hat{v})$, and $\tilde{T} = \hat{H}^{-1/2}T$. Ignoring some constants, the unpenalized log-likelihood (6) becomes $(L(b)=)\|\tilde{y} - \tilde{T}b\|^2/2$. Hence, the maximization of the regularized penalized log-likelihood (6) is equivalent to the following optimization problem

$$\min_b \frac{1}{n} L(b) + \mathrm{P}_\lambda(b), \tag{7}$$

where $\mathrm{P}_\lambda(b)$ is a penalty function on the effects of genetic variants.

Similar to integrative analysis [Ma et al., In press], we can assume homogeneous or heterogeneous structure across multiple traits. Homogeneity model assumes that both traits share the same set of trait-associated covariates while heterogeneity model assumes that a covariate can be associated with some of traits but not others. For homogeneity model, group MCP has been demonstrated to conduct variable selection effectively, while sparse group MCP can be used to conduct variable selection between- and within-groups for heterogeneity model. Here, we choose to use minimax concave penalization (MCP) as basic penalty for the variant selection, since comparing with its alternative, e.g. Lasso [Tibshirani, 1996] and smooth clipped absolute deviation (SCAD) [Fan and Li, 2001], MCP belongs to the family of quadratic spline penalties and leads to oracle selection results requiring weaker conditions [Zhang, 2010]. We refer to Zhang [2010] and Mazumder et al. [2011] for detailed discussion.

The MCP is defined as

$$\rho(t; \lambda, \gamma) = \lambda_1 \int_0^{|t|} (1 - x/(\gamma\lambda))_+ dx.$$

Here λ is a penalty parameter, γ is a regularization parameter that controls the concavity of ρ and $x_+ = x1_{\{x\geq 0\}}$. The MCP can be easily understood by considering its derivative, which is

$$\dot{\rho}(t;\lambda,\gamma) = \lambda_1\big(1 - |t|/(\gamma\lambda)\big)_+ \text{sgn}(t),$$

where $\text{sgn}(t) = -1, 0$, or 1 if $t < 0, = 0$, or > 0, respectively. As $|t|$ increases from 0, MCP begins by applying the same rate of penalization as Lasso, but continuously relaxes that penalization until $|t| > \gamma\lambda$, a condition under which the rate of penalization drops to 0. It provides a continuum of penalties where the Lasso penalty corresponds to $\gamma = \infty$ and the hard-thresholding penalty corresponds to $\gamma \to 1+$. We note that other penalties, such as Lasso or SCAD, can also be used to replace MCP. We choose MCP because it possesses all the desirable properties of a penalty function and is computationally simple [Mazumder et al., 2011, Zhang, 2010].

3.2.1 Group MCP

For the homogeneity model, group penalization methods can be implemented, e.g. group Lasso, group bridge, group MCP. Here, we choose to use group MCP [Huang et al., 2012] since comparing with its alternative, it possesses oracle properties with less conditions. We have $b = \text{vech}(B)$ where B_j is the jth row of B corresponding to the regression coefficients of the jth gene on multiple traits, $\|B_j\|_{\Sigma_j} = (B_j'\Sigma_j B_j)^{1/2}$ and $\Sigma_j = \tilde{T}_j'\tilde{T}_j/n$ is the empirical covariance matrix for the jth group. We can write $\Sigma_j = R_j'R_j$ for an $m \times m$ upper triangular matrix R_j with positive diagonal entries via the Cholesky decomposition. Let $V_j = \tilde{T}_j R_j^{-1}$ and $\beta_j = R_j B_j$. With the transformation, the penalty function of group MCP is $\text{P}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^p \rho\left(\|\beta_j\|; \sqrt{m}\lambda, \gamma\right)$ and the objective function corresponding to group MCP [Huang et al., 2012] can be written as:

$$L_{GM}(\boldsymbol{\beta},\lambda) = \frac{1}{2n}\|\tilde{y} - \sum_{j=1}^{p} V_j\beta_j\|^2 + \text{P}_\lambda(\boldsymbol{\beta}), \tag{8}$$

where $\boldsymbol{\beta} = (\beta_1', \ldots, \beta_p')'$.

The rationale behind the group MCP can also be understood by its univariate solution with the jth group. Consider the linear regression of y upon x_j (covariates on the jth group), with unpenalized least squares solution $z_j = n^{-1}x_j'y$ (recall that x_j has been standardized so that $x_j'x_j/n = \mathrm{I}_n$). For this linear regression problem, the group MCP estimator has the following closed form:

$$\hat{\beta}_j = \begin{cases} \frac{\gamma}{\gamma - 1}\mathrm{S}_1(z_j, \sqrt{m}\lambda), & \text{if } \|z_j\|_2 \leq a\sqrt{m}\lambda \\ z_j, & \text{if } \|z_j\|_2 > \gamma\sqrt{m}\lambda \end{cases},$$

where $\mathrm{S}_1(z, \lambda) = (1 - \lambda/\|z\|_2)_+ z$. Then one can implement group coordinate descent (GCD) algorithm to solve for the optimizer of objective function (8) [Huang et al., 2012].

3.2.2 Sparse Group MCP

For heterogeneity structure among traits, sparse group MCP can be applied to conduct variable selection. We orthogonalize covariates within groups in the same fashion as the group MCP. Denote $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$. Then, the penalty function of sparse group MCP is $\mathrm{P}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \rho\left(\|\beta_j\|; \sqrt{m}\lambda_1, \gamma\right) + \sum_{j=1}^{p}\sum_{k=1}^{m} \rho\left(|\beta_{jk}|; \lambda_2, \gamma\right)$ and the objective function corresponding to sparse group MCP [Liu et al., 2013] can be written as:

$$L_{SGM}(\boldsymbol{\beta}, \lambda) = \frac{1}{2n}\|\tilde{y} - \sum_{j=1}^{p} V_j\beta_j\|^2 + \mathrm{P}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}). \tag{9}$$

Similar to Zou and Li [2008], Breheny and Huang [2009], one can use local linear approximation (LLA) for the penalty function and iteratively solve the problem using the optimizer in Friedman et al. [2010]. In Liu et al. [2013], they used a two-step strategy to solve for the optimizer of objective function (9). In this way, GCD algorithm can be implemented to solve (9) instead of solving objective function with LLA penalty function. Breheny and Huang [2011] argued that GCD algorithm, alternative to LLA can be implemented with more efficiency. Consider univariate group solution on the linear regression of y upon x_j

(covariates on the jth group), with unpenalized least squares solution $z_j = n^{-1}x_j'y$ (recall that x_j has been standardized so that $x_j'x_j/n = \mathrm{I}_n$).

By setting first order derivative of objective function be zero, we have:

$$-z_j + g(\beta_j)\beta_j + \mathbf{t} = 0, \tag{10}$$

where $z_j = (z_j^1, \ldots, z_j^M)'$, $g(\beta_j) = \left(1 + \frac{1}{||\beta_j||_2}\right)\begin{cases}\sqrt{m}\lambda_1 - \frac{\|\beta_j\|_2}{\gamma}, & \text{if } \|\beta_j\|_2 \le \gamma\sqrt{m}\lambda_1 \\ 0, & \text{if } \|\beta_j\|_2 > \gamma\sqrt{m}\lambda_1\end{cases}$. Denote z_j^k as the kth element of z_j. First, fix $g(\beta_j)$ at the current estimate $\tilde{\beta}_j$, we use g short for $g(\tilde{\beta}_j)$. The kth element in equation (10) can be rewritten as:

$$-\frac{z_{jk}}{g} + \beta_{jk} + \mathrm{sgn}(\beta_{jk})\begin{cases}\frac{\lambda_2}{g} - \frac{|\beta_{jk}|}{\gamma g}, & \text{if } |\beta_{jk}| \le \gamma\lambda_2 \\ 0, & \text{if } |\beta_{jk}| > \gamma\lambda_2\end{cases} = 0. \tag{11}$$

The solution to equation (11) is

$$\widehat{g\beta_{jk}} = \begin{cases}\frac{\mathrm{S}_2(z_{jk},\lambda_2)}{1-\frac{1}{\gamma g}}, & \text{if } |z_{jk}| \le \gamma\lambda_2 g \\ z_{jk}, & \text{if } |z_{jk}| > \gamma\lambda_2 g.\end{cases}$$

Here $\mathrm{S}_2(z,\lambda) = \mathrm{sgn}(z)(|z| - \lambda)_+$. For $k = 1, \ldots, m$, set $u_k = \widehat{g\beta_{jk}}$ and $\mathbf{u} = (u_1, \ldots, u_m)'$. Taking $\mathbf{u}$ back into its definition,

$$\beta_j + \frac{\beta_j}{\|\beta_j\|_2}\begin{cases}\sqrt{m}\lambda_1 - \frac{\|\beta_j\|_2}{a}, & \text{if } \|\beta_j\|_2 \le a\sqrt{m}\lambda_1 \\ 0, & \text{if } \|\beta_j\|_2 > a\sqrt{m}\lambda_1\end{cases} = \mathbf{u}. \tag{12}$$

Expression (12) can be solved in a similar manner as with the gMCP, leading to

$$\hat{\beta}_j = \begin{cases}\frac{a}{a-1}\mathrm{S}_1(\mathbf{u}, \sqrt{m}\lambda_1), & \text{if } \|\mathbf{u}\|_2 \le a\sqrt{m}\lambda_1 \\ \mathbf{u}, & \text{if } \|\mathbf{u}\|_2 > a\sqrt{m}\lambda_1\end{cases}. \tag{13}$$

To optimize the group MCP or sparse group MCP objective function, group coordinate descent algorithm (GCD) can be implemented. Breheny and Huang [2011] explored coordinate descent algorithms (CDA) for nonconvex penalized regression, including MCP and SCAD. The extension of CDA to group level is natural, their details can be found in Huang et al. [2012], Liu et al. [2013].

3.2.3 Choice of tuning parameter

With MCP, there is one tuning parameter λ and one regularization parameter γ. Generally speaking, smaller values of γ are better at retaining the unbiasedness of the MCP penalty for large coefficients, but they also have the risk of creating objective functions with a nonconvexity problem that are difficult to optimize and yield solutions that are discontinuous with respect to λ. It is therefore advisable to choose a γ value that is big enough to avoid this problem but not too big. Simulation studies in Breheny and Huang [2011] and Liu et al. [2012] show that $\gamma = 3$ is a reasonable choice for group MCP and $\gamma = 6$ is a reasonable choice for sparse group MCP, respectively. For group MCP, we search for tuning parameters λ using V-fold cross validation ($V = 5$ in our numerical study). For sparse group MCP, we fix the ratio of λ_1 and λ_2 to be 1. Then λ_1 and λ_2 can be searched through V-fold cross validation. It is expected that tuning parameter cannot go down to very small values which correspond to regions not locally convex. The cross validation criteria over non-locally convex regions may not be monotone. More details regarding the choice of tuning parameter for group MCP and sparse group MCP can be found in [Huang et al., 2012] and Liu et al. [2012], respectively.

3.3 Genetic Scores on Collapsed SNPs

Recent GWAS have shown that common variants can only account for small proportion of heritability. Among all potential explanations to this missing heritability, the large number of variants of small effects and rare variants (possibly with large effects) can be partially remedied using a weighted-sum method [Madsen and Browning, 2009]. We group SNPs at gene level using this weighted-sum method. This process puts the analysis for genetic markers at gene level and is capable of dealing with rare variants together with common variants. The proposed approach can be easily implemented in GWAS with common variants only or

longitudinal measurements on traits with independent samples.

3.4 Trait Prediction

Given a training sample of genetic variants and traits, we can predict the unobserved traits using a testing sample. According to best linear unbiased predictor (BLUP), the predictive value of lth trait $y_p^{(l)}$ is given by

$$y_p^{(l)} = W\hat{v}^{(l)} + X\hat{b}^{(l)} + K_{tt}\hat{H}^{(l)^{-1}}(y_t^{(l)} - W\hat{v}^{(l)} - X\hat{b}^{(l)}) \tag{14}$$

where $y_t^{(l)}$ is the lth trait of the training set, K_{tt} is the covariance matrix between the training sample and the testing sample and $\hat{H}^{(l)}$ is matrix of variance components corresponding to lth trait. To evaluate estimates from the penalized-MTMM, we first extract variance components and genetic effects corresponding to each trait. Then, we marginally evaluate the prediction on each trait and calculate the Pearson's correlation between the predictive values and their corresponding observations. This procedure puts the comparison between penalized-MTMM and uni-trait penalized-LMM methods on the same page.

4 Simulation Study

We conduct simulation to better gauge performance of the proposed methods. The genotype data is excerpted from a T2D–GENES study with twenty pedigree families (Section 2). We consider six scenarios of correlations. For all scenarios, we set $n = 400$, $p = 5000$ and $m = 2$. We consider two traits in this study. The covariance among residual (Σ_e) and random effects (Σ_d) in six scenarios are listed in Table 1. Scenario 1–3 represent the cases that Σ_e and Σ_d is proportional with weak, moderate and strong correlation while scenario 4–6 represent the cases that Σ_e and Σ_d is not proportional with multiple combination on Σ_e and Σ_d. We also consider homogeneity and heterogeneity structure between two traits in this simulation

study. In homogeneity structure, the index of important variants are (1–5, 16–20) for both traits while the index of important variants are (1–5, 16–20) for trait 1 and (1–5, 21–25) for trait 2 in heterogeneity structure.

We analyze simulated data using the proposed penalized-MTMM approach on multiple traits. For comparison, we also consider penalized-LMM considering one trait at a time and linear model on each trait without consideration of variance components adjusting for relatedness among samples. For all scenarios on covariance components on both unobserved random residual and polygenic effects, ROC curves for homogeneity and heterogeneity are shown in Figure 1 and 2, respectively. We also calculate the partial area under the curve (P-AUC) for each methods under each scenario (Table 2). One can observe that under homogeneity structure, the area under the curve for group MCP and sparse group MCP using penalized-MTMM is larger than that from uni-trait penalized-LMM using MCP and uni-trait linear model. Under heterogeneity model, the P-AUC using penalized-MTMM is larger for sparse group MCP in four scenarios but also comparable in other two scenarios. The main reason for this phenomenon is that only two traits in the study. We postulate if the number of traits is going up to three or four, the improvement of sparse gMCP over gMCP in heterogeneity becomes more obvious. Furthermore, one can observe that uni-trait penalized-LMM using MCP is consistently better than univariate linear model, since uni-trait LMM-Pen takes into account the confounding relatedness in samples that is better in identifying genetic variants. To better compare prediction, we compare the multi-trait and uni-trait methods through simulation studies.

5 Analysis of GAW 18 data

We analyze GAW 18 data described in Section 2. GAW 18 T2D–GENES study provides a GWAS data consisting of SNPs from odd autosomes. Totally, there are 472,049 SNPs over

11 autosomes for 959 samples from 20 large pedigrees. Among those SNPs, 292,355 SNPs are within the scope of gene. Using the collapsing techniques described in Section 3.3, SNPs within gene scope are collapsed into genetic scores for 10,949 genes. After quality control, 849 samples with genetic scores for 10,549 genes are used for further analysis.

The variance components for residuals are $\begin{pmatrix} 0.900 & 0.490 \\ 0.490 & 0.903 \end{pmatrix}$ and the proportion between Σ_d and Σ_e is (θ=) 0.100. One can deduce heritability for this data set that is $\theta/(1+\theta)$ (=0.091). The estimates and corresponding observed occurrence index (OOI) of the proposed approach using group MCP and sparse group MCP are shown in Table 3 and 4. For comparison, we conduct uni-trait analysis using MCP, the estimates and corresponding observed occurrence index (OOI) are shown in Table 5. To evaluate prediction performance, we calculate the correlation between the predictive values using BLUP in Section 3.4 and their corresponding observations. We carry out this procedure via V-fold cross-validation. The mean (sd) correlation is 0.152(0.057) and 0.186(0.084) for SBP and DBP, respectively using the proposed method on group MCP. The mean correlation is 0.139(0.070) and 0.192(0.112) for SBP and DBP, respectively using the proposed method on sparse group MCP. For comparison, The mean correlation is 0.125(0.078) and 0.146(0.074) for SBP and DBP, respectively using uni-trait method on MCP.

[Table 1 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

6 Discussion

We have presented a penalized multi-trait mixed model (Penalized-MTMM) for detecting pleiotropic genetic associations among multiple traits in the presence of pedigree structure. The approach combines the advantages of mixed models that allow for elegant correction for pedigree-based family data, integrative analysis of multiple traits that borrow strengths across traits and joint multi-variant models that take the joint effects of sets of genetic variants into account rather than one single variant at a time. In the joint multi-variant models, we consider both homogeneity and heterogeneity structure using group MCP and sparse group MCP, respectively. We use ROC to evaluate selection performance for penalized-MTMM comparing with penalized-LMM considering one trait at a time and a linear model. To evaluate prediction performance, we use BLUP to find the predictive values and the correlations of the predictive values and their corresponding observations are calculated subsequently. Our numerical studies show that the proposed approach has satisfactory performance.

Confounder effects and population structure induce spurious correlations between genotype and phenotype, complicating the genetic analysis. Mixed models accounting for the presence of such structure are well studied and have been shown to greatly reduce the impact of this confounding source of variability. For instance, EIGENSTRAT was built upon the idea of extracting the major axes of population differentiation using a PCA decomposition of the genotype data and subsequently including them into the model as additional covariates [Price et al., 2006]. The penalized-MTMM can consider both of confounder effects and population structure depending the choice of random effects. On the other hand, mixed models also show its strength in coping with repeated measures in longitudinal studies. The

penalized-MTMM can handle data from longitudinal studies with multiple traits.

In the similar fashion, our method can be applied to conduct integratively analysis of multiple GWAS with correlated traits.

References

P. Breheny and J. Huang. Penalized methods for bi-level variable selection. *Statistics and Its Interface*, 2:369–380, 2009.

P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Statist.*, 5:232–253, 2011.

Paul R Burton, David G Clayton, Lon R Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem H Ouwehand, Nilesh J Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*, 2013.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, 96:1348–1360, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group Lasso and a sparse group Lasso. *arXiv:1001.0736*, 2010.

A.R. Gilmour, R. Thompson, and B.R. Cullis. Average information REML:an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 51:1440–1450, 1995.

B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, and G.R. Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics*, 2012.

J. Huang, F. Wei, and S. Ma. Semiparametric reregression pursuit. *Statistica Sinica*, 22: 1403–1426, 2012.

H. Kang, J. Sul, S. Service, N. Zaitlen, S. Kong, N. Freimer, , C. Sabatti, and E. Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nat Genetics*, 42:348C354, 2010.

Arthur Korte, Bjarni J Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature genetics*, 44(9):1066–1071, 2012.

S. Lee, J. Yang, M.E. Goddard, P.M. Visscher, and N.R. Wray. Estimation of pleiotropy between complex diseases using snp-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, page doi: 10.1093/bioinformatics/bts474, 2012.

C. Lippert, J. Listgarten, Y. Liu, C.M. Kadie, R.I. Davidson, and D. Heckerman. FaST linear mixed models for genome-wide association studies. *Nat Methods*, 8(10):833–835, 2011.

J. Liu, J. Huang, Y. Xie, and S. Ma. Integrative analysis of cancer prognosis data with sparse group penalization. 2012.

J. Liu, J. Huang, and S. Ma. Integrative analysis of multiple-typed cancer genomic datasets under heterogeneity model. *Accepted: Stat. Med.*, 2013.

S. Ma, J. Huang, and X. Song. Integrative analysis and variable selection with multiple high-dimensional datasets. *Biostatistics*, In press.

B. Madsen and S. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2):e1000384, 2009.

Robert Makowsky, Nicholas M Pajewski, Yann C Klimentidis, Ana I Vazquez, Christine W Duarte, David B Allison, and Gustavo de los Campos. Beyond missing heritability: prediction of complex traits. *PLoS genetics*, 7(4):e1002051, 2011.

Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

R. Mazumder, J. Friedman, and T. Hastie. SparseNet:Coordinate descent with non-convex penalties. *J. Am. Stat. Assoc.*, page doi:10.1198/jasa.2011.tm09738., 2011.

M. McCarthy, G. Abecasis, L. Cardon, D. Goldstein, J. Little, J. Ioannidis, and J. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5):356–369, 2008.

A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

B. Rakitsch, C. Lippert, O. Stegle, and K. Borgwardt. A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, 29(2):206–214, 2013.

Shanya Sivakumaran, Felix Agakov, Evropi Theodoratou, James G Prendergast, Lina Zgaga, Teri Manolio, Igor Rudan, Paul McKeigue, James F Wilson, and Harry Campbell. Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*, 89(5):607–618, 2011.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B*, 58: 267–288, 1996.

Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

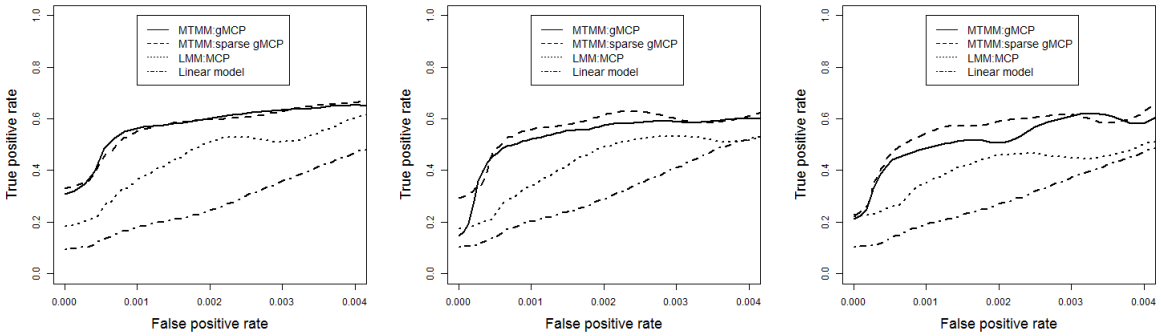
Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38:894–942, 2010.

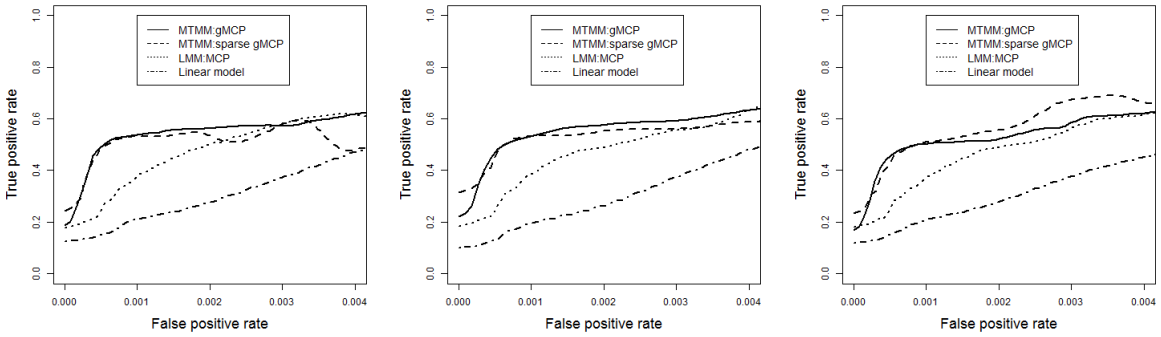
Z. Zhang, El Ersoz, C. Lai, R. Todhunter, H. Tiwari, M. Gore, P. Bradbury, J. Yu, D. Arnett, J. Ordovas, and E. Buckler. Mixed linear model approach adapted for genome-wide association studies. *Nat Genetics*, 42:355–360, 2010.

X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nat Genetics*, 44:821–824, 2012.

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, 36(4), 2008.

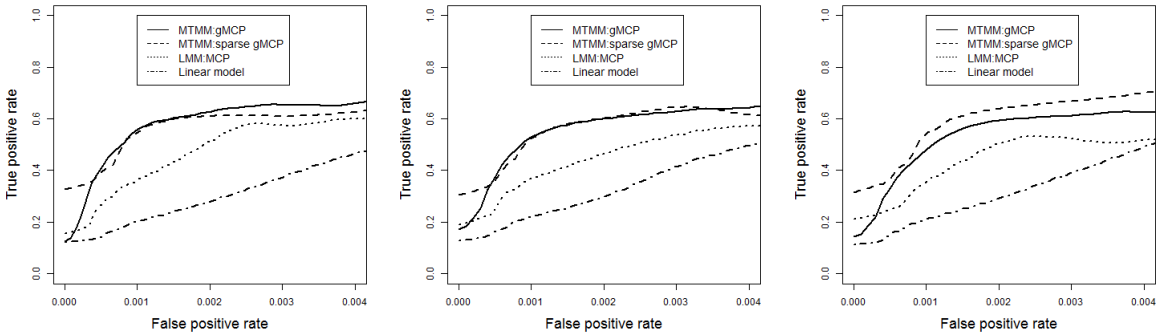


(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

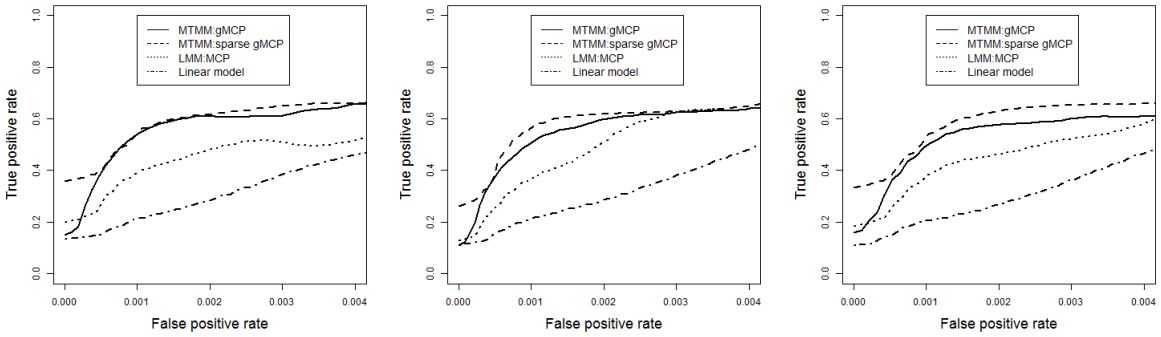


(d) Scenario 4 (e) Scenario 5 (f) Scenario 6

Figure 1: ROC plots for example 1–6 in homogeneity model.



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3



(d) Scenario 4 (e) Scenario 5 (f) Scenario 6

Figure 2: ROC plots for example 1–6 in heterogeneity model.

Table 1: Six scenarios for Covariance on both unobserved random residual and polygenic effects.

	Σ_e	Σ_d
Scenario 1	$\begin{pmatrix} 0.20 & 0.04 \\ 0.04 & 0.20 \end{pmatrix}$	$\begin{pmatrix} 0.40 & 0.08 \\ 0.08 & 0.40 \end{pmatrix}$
Scenario 2	$\begin{pmatrix} 0.20 & 0.10 \\ 0.10 & 0.20 \end{pmatrix}$	$\begin{pmatrix} 0.40 & 0.20 \\ 0.20 & 0.40 \end{pmatrix}$
Scenario 3	$\begin{pmatrix} 0.20 & 0.16 \\ 0.16 & 0.20 \end{pmatrix}$	$\begin{pmatrix} 0.40 & 0.32 \\ 0.32 & 0.40 \end{pmatrix}$
Scenario 4	$\begin{pmatrix} 0.20 & 0.04 \\ 0.04 & 0.24 \end{pmatrix}$	$\begin{pmatrix} 0.40 & 0.24 \\ 0.24 & 0.40 \end{pmatrix}$
Scenario 5	$\begin{pmatrix} 0.20 & 0.16 \\ 0.16 & 0.24 \end{pmatrix}$	$\begin{pmatrix} 0.40 & 0.04 \\ 0.04 & 0.40 \end{pmatrix}$
Scenario 6	$\begin{pmatrix} 0.20 & 0.16 \\ 0.16 & 0.24 \end{pmatrix}$	$\begin{pmatrix} 0.40 & 0.24 \\ 0.24 & 0.40 \end{pmatrix}$

Table 2: Partial AUC (standard deviation) under 6 scenarios for all methods in both homogeneity and heterogeneity models.

	Data Model	MTMM		LMM	Linear Model
Penalty		gMCP	sparse gMCP	MCP	MCP
Scenario 1	Homogeneity	0.537(0.134)	0.484(0.140)	0.355(0.084)	0.344(0.062)
Scenario 2	Homogeneity	0.486(0.128)	0.468(0.115)	0.340(0.087)	0.372(0.067)
Scenario 3	Homogeneity	0.411(0.131)	0.452(0.125)	0.359(0.092)	0.355(0.071)
Scenario 4	Homogeneity	0.517(0.105)	0.471(0.130)	0.332(0.108)	0.359(0.063)
Scenario 5	Homogeneity	0.520(0.089)	0.487(0.102)	0.337(0.094)	0.367(0.063)
Scenario 6	Homogeneity	0.449(0.128)	0.460(0.128)	0.324(0.100)	0.357(0.055)
Scenario 1	Heterogeneity	0.540(0.089)	0.499(0.095)	0.351(0.082)	0.360(0.065)
Scenario 2	Heterogeneity	0.531(0.101)	0.501(0.106)	0.352(0.083)	0.375(0.062)
Scenario 3	Heterogeneity	0.558(0.093)	0.574(0.108)	0.324(0.102)	0.366(0.063)
Scenario 4	Heterogeneity	0.510(0.092)	0.521(0.104)	0.333(0.091)	0.350(0.072)
Scenario 5	Heterogeneity	0.489(0.103)	0.515(0.122)	0.324(0.095)	0.357(0.070)
Scenario 6	Heterogeneity	0.518(0.093)	0.558(0.090)	0.327(0.084)	0.342(0.058)

Table 3: Gene selected by penalized-MTMM using gMCP.

Gene	Trait1	Trait2	OOI	Gene	Trait1	Trait2	OOI
DFFA	0.001	0.000	0.400	NCAM1	-0.017	-0.005	0.690
LOC390998	0.031	0.024	0.900	OR8A1	-0.002	-0.030	0.970
MOBKL2C	-0.009	-0.002	0.440	FNDC3A	0.008	0.015	0.650
SLC16A4	-0.007	-0.000	0.550	INOC1	0.029	0.026	0.920
LCE1A	0.018	0.016	0.750	PIAS1	0.002	0.001	0.340
PYCR2	0.004	0.001	0.500	TLCD2	0.004	-0.001	0.520
ALCAM	-0.002	-0.002	0.440	HS3ST3B1	0.012	0.004	0.660
EIF2B5	-0.001	-0.002	0.460	FLII	-0.004	-0.005	0.360
ZCCHC10	0.019	0.011	0.680	RAMP2	-0.004	-0.011	0.540
ANKHD1	0.007	0.004	0.520	ANKRD40	-0.002	0.022	0.850
LOC100130230	-0.003	-0.005	0.540	C19orf38	-0.009	0.025	0.970
PAPOLB	0.015	0.016	0.820	NDUFA13	-0.002	0.001	0.370
RPA3	-0.032	-0.010	0.890	ZNF826	-0.006	0.020	0.900
ELMO1	-0.008	-0.001	0.430	SYT3	0.016	0.011	0.800
OGDH	0.040	0.037	0.990	ZNF611	-0.008	-0.017	0.770
EXOSC2	0.029	0.040	0.990	AIRE	-0.004	-0.011	0.570
LOC390084	-0.008	-0.007	0.510				

Table 4: Gene selected by penalized-MTMM using sparse gMCP.

Gene	Trait 1		Trait 2		Gene	Trait 1		Trait 2	
	Est.	OOI	Est.	OOI		Est.	OOI	Est.	OOI
AIRE	-0.004	0.020	-0.007	0.450	MOBKL2C	-0.006	0.510		
ANKRD40	0.001	0.580	0.005	0.580	NCAM1	-0.012	0.750		
C19orf38	-0.004	0.690	0.010	0.690	OGDH	0.024	0.810	0.031	0.980
CSF1	-0.005	0.030	-0.008	0.490	OR8A1	-0.004	0.770	-0.016	0.770
EIF2B5	-0.007	0.000	-0.010	0.490	PAPOLB	0.010	0.200	0.015	0.610
ELMO1	-0.009	0.610			PYCR2	0.004	0.510		
EXOSC2	0.027	0.000	0.042	1.000	RAMP2	-0.005	0.050	-0.009	0.570
FLII	-0.004	0.000	-0.006	0.550	RPA3	-0.018	0.810	-0.001	0.350
FNDC3A	0.017	0.000	0.026	0.710	SCYL1BP1	-0.001	0.530		
HS3ST3B1	0.004	0.540	0.000	0.290	SLC16A4	-0.016	0.760		
INOC1	0.009	0.560	0.010	0.600	TLCD2	0.007	0.650		
LCE1A	0.001	0.320	0.001	0.420	ZNF611	-0.014	0.020	-0.022	0.880
LOC100130230	-0.007	0.000	-0.011	0.700	ZNF826	-0.001	0.540	0.003	0.540
LOC390998	0.013	0.700	0.012	0.700					

Table 5: Gene selected by uni-trait LMM using MCP separately on each trait.

Gene	Trait 1		Trait 2		Gene	Trait 1		Trait 2	
	Est.	OOI	Est.	OOI		Est.	OOI	Est.	OOI
ANKHD1	0.017	0.710			NCAM1	-0.024	0.800		
ARSB	-0.007	0.490			OGDH	0.022	0.730	0.011	0.760
C1orf128	0.015	0.680			PIAS1	0.016	0.700		
C3orf20	0.007	0.500			PYCR2	0.008	0.660		
DFFA	0.005	0.450			RPA3	-0.030	0.900		
HS3ST3B1	0.022	0.810			SFRS12	0.007	0.580		
INOC1	0.019	0.760	0.004	0.540	SYT3	0.024	0.750		
LCE1A	0.007	0.430			ZCCHC10	0.027	0.780		
LOC390998	0.028	0.830			EXOSC2			0.038	0.980
MCM7	-0.001	0.410			FNDC3A			0.015	0.720
MKNK1	-0.002	0.270			ZNF611			-0.011	0.750