Published in final edited form as:

Genet Epidemiol. 2016 July ; 40(5): 382–393. doi:10.1002/gepi.21975.

Analyzing Association Mapping in Pedigree-Based GWAS Using a Penalized Multitrait Mixed Model

Jin Liu^{1,†}, Can Yang^{2,†}, Xingjie Shi³, Cong Li⁴, Jian Huang⁵, Hongyu Zhao^{6,*}, and Shuangge Ma^{6,*}

¹Center of Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore

²Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

³Department of Statistics, Nanjing University of Finance and Economics, Nanjing, Jiangsu, China

⁴Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America

⁵Department of Statistics & Actuarial Science, University of Iowa, Iowa City, Iowa, United States of America

⁶Department of Biostatistics, Yale University, New Haven, Connecticut, United States of America

Abstract

Genome-wide association studies (GWAS) have led to the identification of many genetic variants associated with complex diseases in the past 10 years. Penalization methods, with significant numerical and statistical advantages, have been extensively adopted in analyzing GWAS. This study has been partly motivated by the analysis of Genetic Analysis Workshop (GAW) 18 data, which have two notable characteristics. First, the subjects are from a small number of pedigrees and hence related. Second, for each subject, multiple correlated traits have been measured. Most of the existing penalization methods assume independence between subjects and traits and can be suboptimal. There are a few methods in the literature based on mixed modeling that can accommodate correlations. However, they cannot fully accommodate the two types of correlations while conducting effective marker selection. In this study, we develop a penalized multitrait mixed modeling approach. It accommodates the two different types of correlations and includes several existing methods as special cases. Effective penalization is adopted for marker selection. Simulation demonstrates its satisfactory performance. The GAW 18 data are analyzed using the proposed method.

Keywords

GWAS; multitrait analysis; pedigree; mixed modeling; penalization

*Correspondence to: Hongyu Zhao, Department of Biostatistics, Yale University, New Haven, CT. hongyu.zhao@yale.edu; Shuangge Ma, Department of Biostatistics, Yale University, New Haven, CT. shuangge.ma@yale.edu.

[†]These authors are joint first authors.

Introduction

Genome-wide association studies (GWAS), since first published more than 10 years ago, have been playing an important role in biomedical research in understanding the molecular mechanisms underlying many complex diseases. An important goal of GWAS is to identify markers for disease outcomes and phenotypes so as to facilitate better understanding disease biology and more effectively building clinically useful models. Multiple types of variable selection techniques have been adopted for marker selection in GWAS. Among them, penalization has attracted extensive interest [Ma and Huang, 2008; Liu et al., 2012] because of its satisfactory performance in data analyses and simulations.

This study has been partly motivated by the Genetic Analysis Workshop (GAW) 18 data. GAW 18 was held in 2013, but the data were fully released only very recently. It provides data from the Type 2 Diabetes (T2D) Genetic Exploration by Next-generation sequencing in Ethnic Samples (T2D-GENES) Consortium. The dataset contains 1,043 individuals from 20 large Mexican American pedigrees enriched for type 2 diabetes from San Antonio, TX. The samples are from the San Antonio Family Heart Study (SAFHS) and the San Antonio Family Diabetes/Gallbladder Study (SAFDGS), together of which are referred to as the San Antonio Family Studies (SAFS). The part of data analyzed in this study includes 959 subjects from the 20 pedigrees who have measurements on the traits of interest. With the pedigree structure, naturally, the subjects are expected to be related. The familial relatedness of data can also be seen from the empirical relatedness calculated from the genotype data (Fig. A1, Appendix), where we can easily identify the 20 families. On each subject, multiple traits were measured. Of interest to our analysis are the systolic blood pressure (SBP) and diastolic blood pressure (DBP), which are well known to be correlated. In the GAW 18 dataset, the sample correlation coefficient is 0.431. Whole genome sequencing was performed by the Complete Genomics, Inc. The analyzed genetic data are based on the sequencing for the first 483 T2D-GENES. Although data are available on both common and rare variants, in this study, we focus on the common variants. For each subject, the genotypes of 472,049 SNPs over 11 odd-numbered autosomes are available. More specifically, the SNP values for 464 subjects were directly sequenced, and the rest were imputed [Howie et al., 2012]. The analysis goal is to identify important SNPs associated with the traits while properly accommodating the pedigree and multitrait structures.

Studies have suggested that ignoring the relatedness caused by the pedigree structure leads to ineffective model estimation and marker selection [Kang et al., 2010]. One of the most commonly adopted and powerful methods that can account for the pedigree relatedness among subjects is the linear mixed model (LMM) method. Several effective approaches have been developed to apply LMM to the analysis of GWAS data [Kang et al., 2010; Lippert et al., 2011; Rakitsch et al., 2013; Zhang et al., 2010; Zhou and Stephens, 2012]. On the other hand, more recently, the LMM approach has also been applied to analyze multiple correlated traits simultaneously [Korte et al., 2012]. This is motivated by the observation that correlated traits usually share common genetic factors, a phenomenon known as pleiotropy [Sivakumaran et al., 2011]. For example, a pleiotropic enrichment method has been developed to analyze GWAS data on schizophrenia and cardiovascular diseases. The power

to detect schizophrenia-associated common variants has been shown to increase by exploiting the pleiotropy between these two phenotypes.

Pedigree data are not uncommon in GWAS. The advantage of jointly analyzing multiple correlated traits has been established in multiple recent studies [Liu et al., 2012]. A common limitation of the existing methods, including those aforementioned, is that only one type of correlation is accounted for. Motivated by the GAW 18 data and data alike, the goal of this study is to develop an effective analysis method, which can accommodate both types of correlations, and effectively identify important variants. For this purpose, we will take advantage of recent development in advanced penalization techniques.

Penalized Multitrait Mixed Modeling (MTMM)

LMM for a Single Trait

First consider the LMM approach for the analysis of a single trait [Lippert et al., 2011; Rakitsch et al., 2013]. This approach can be applied to analyze multiple traits separately, although ignoring the correlation among traits may result in inefficiency. Assume n subjects. For a single trait, the LMM is written as:

$$y_o = Wv + Xb + g + e, \quad g \sim N(0, \sigma_g^2 K), \quad e \sim N(0, \sigma_e^2 I), \quad (1)$$

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 effects) including the intercept and other (nongenetic) covariates such as age and gender, v is the $q \times 1$ vector of regression coefficients for the covariates, $X \in \mathbb{R}^{n \times p}$ is the genotype matrix, b is the vector for the effect sizes of p SNPs (fixed effects), g is the random effect vector with distribution $N(0, \sigma_g^2 K)$, and e is the residual with variance σ_e^2 . K is the genetic relatedness matrix and accommodates the pedigree structure among subjects.

Normalization conditions are imposed on K . As a result, σ_g^2 , the variance component of g , is needed. The covariance matrix K can be constructed using known pedigree information or estimated from genome-wide SNP information (which may lead to different results). 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 we refer to it as the “average polygenic effect.” SNPs with large effects that are different from the genetic background are put into the design matrix X and considered as fixed effects. In this way, markers with larger effects can be treated locally.

MultiTrait Mixed Modeling

We now extend the strategy described in the previous section to accommodate multiple correlated traits. Let y_o be the $n \times m$ response matrix with each row representing a subject and each column representing a trait. Let $\mathbf{g} = (g^{(1)}, \dots, g^{(m)})$ and $\mathbf{e} = (e^{(1)}, \dots, e^{(m)})$ be the $n \times m$ matrices of unobserved polygenic and random residual effects, respectively. Denote W as the $n \times q$ matrix of nongenetic covariates and X as the $n \times p$ matrix of the genetic scores of genes. Correspondingly, for the m traits, denote $V = (v^{(1)}, \dots, v^{(m)})$ and $B = (b^{(1)}, \dots,$

$b^{(m)}$) as the $q \times m$ coefficient matrix for the q nongenetic covariates and the $p \times m$ coefficient matrix for the p genetic scores. For $l = 1, \dots, m$, use the superscript “ (l) ” to denote measurements and parameters for the l th trait.

First, consider the LMM for the l th trait:

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

where notations have the same interpretations as those in the previous section. To accommodate the genetic correlation and residual correlation among multiple traits, we introduce $\sigma_{g^{(l,k)}}$ and $\sigma_{e^{(l,k)}}$ to describe the covariances of average polygenic effects and residuals for the l th and k th traits, respectively. Consider the MTMM:

$$y_o = WV + XB + g + e, \quad (2)$$

where (a) the vector $e \sim N(\mathbf{0}, I_n \otimes \Sigma_e)$, and Σ_e is the $m \times m$ matrix describing the covariance structure for the random residual effects among multiple traits; and (b) $g \sim N(\mathbf{0}, K \otimes \Sigma_g)$, and Σ_g is the $m \times m$ matrix describing the covariance structure for the unobserved polygenic effects among multiple traits. K is the genetic relatedness matrix and can be calculated using genome-wide genetic markers or directly obtained from known pedigree information.

To simplify notation, consider the scenario with two traits ($m = 2$), which matches the GAW 18 data. Accommodating more traits is straightforward. When $m = 2$, we have

$\Sigma_e = \begin{pmatrix} \sigma_{e^{(1)}}^2 & \sigma_{e^{(12)}} \\ \sigma_{e^{(12)}} & \sigma_{e^{(2)}}^2 \end{pmatrix}$ and $\Sigma_g = \begin{pmatrix} \sigma_{g^{(1)}}^2 & \sigma_{g^{(12)}} \\ \sigma_{g^{(12)}} & \sigma_{g^{(2)}}^2 \end{pmatrix}$. Denote $S = \text{diag}(W, W)$, $v = \text{vech}(V)$, $T = \text{diag}(X, X)$, $b = \text{vech}(B)$, $y = \text{vech}(y_o)$, $g = \text{vech}(g)$, and $e = \text{vech}(e)$. Here $\text{diag}(\cdot, \cdot)$ represents matrix construction by filling in the diagonal elements, and $\text{vech}(\cdot)$ represents vector construction. Then model (2) can be rewritten as

$$y = Sv + Tb + g + e,$$

$$\text{with } g \sim N(\mathbf{0}, K \otimes \Sigma_g), \quad e \sim N(\mathbf{0}, I_n \otimes \Sigma_e). \quad (3)$$

Integrating out g and e , we have $y \sim N(Sv + Tb, K \otimes \Sigma_g + I_n \otimes \Sigma_e)$. The log-likelihood function can be written as

$$L(v, b, \Sigma_g, \Sigma_e) = -\frac{1}{2n} [2n \log(2\pi) + \log(|H|) + (y - Sv - Tb)^T H^{(-1)} (y - Sv - Tb)], \quad (4)$$

where $H = K \otimes \Sigma_g + I_n \otimes \Sigma_e$.

Penalized Estimation and Selection

With whole genome data similar to the GAW 18, the number of unknown parameters is much larger than the sample size. In addition, among all of the SNPs profiled, only a very small subset is expected to be associated with the traits. Thus, regularized estimation and marker selection are needed. In this study, we adopt penalization, which has been the choice of a large number of studies [Liu et al., 2012; Ma and Huang, 2008]. Consider the penalized log-likelihood function

$$-\frac{1}{2}[2n \log(2\pi) + \log(|H|) + (y - S\mathbf{v} - T\mathbf{b})^T H^{(-1)}(y - S\mathbf{v} - T\mathbf{b})] - P_\lambda(b). \quad (5)$$

$P_\lambda(b)$ is the penalty function on the effects of genetic variants, and λ is the data-dependent tuning parameter. The choice of penalty is discussed in the next section.

When b is fixed, optimization of the unpenalized log-likelihood function (4) can be solved by the “Average information-restricted maximal likelihood” method (AI-REML) [Gilmour et al., 1995]. When $(\mathbf{v}, \Sigma_g, \Sigma_e)$ are all known, optimization of the penalized log-likelihood becomes a penalized least squares problem. Computationally, we start with $b = 0$ and solve (4) by AI-REML. After obtaining the estimates for $(\mathbf{v}, \Sigma_g, \Sigma_e)$, we can transform the problem into a penalized least squares one as follows. Let $\hat{H}$ and $\hat{\mathbf{v}}$ be the estimates of H and $\mathbf{v}$, respectively. Denote $\tilde{\mathbf{y}} = \hat{H}^{-1/2}(y - S\hat{\mathbf{v}})$ and $\tilde{T} = \hat{H}^{-1/2}T$. Then optimizing the penalized log-likelihood function is equivalent to

$$\min_b \left\{ \frac{1}{n} L(b) + P_\lambda(b) \right\}, \quad (6)$$

where $L(b) = \|\tilde{\mathbf{y}} - \tilde{T}\mathbf{b}\|^2/2$. Optimizing with respect to b and $(\mathbf{v}, \Sigma_g, \Sigma_e)$ can be iterated until convergence.

Group Minimax Concave Penalization

The proposed penalty is built on the minimax concave penalization [MCP, Mazumder et al., 2011; Zhang, 2010]. The MCP is defined as

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

Here λ is the tuning parameter and controls the degree of sparsity. γ is the regularization parameter and controls the concavity of ρ . $x_+ = x1_{\{x > 0\}}$. Numerically and statistically, MCP has been shown to have satisfactory performance. It has been the choice for penalty in quite a few recent studies. It is interesting to note that when $\gamma \rightarrow \infty$, MCP becomes Lasso, the most popular penalty.

Following the development of recent integrative analysis [Ma et al., 2011; Zhao et al., 2015], for marker selection, two structures can be assumed, namely the homogeneity and

heterogeneity structures. Under the homogeneity structure, multiple traits share the same set of important markers, whereas under the heterogeneity structure, the sets of important markers for different traits can differ. It has been noted in the literature that the heterogeneity structure can be more flexible. However, when the traits are highly related and so their marker sets are expected to be similar, selection and estimation that assume the homogeneity structure can be more accurate [Zhao et al., 2015]. With the GAW 18 data, SBP and DBP have highly related biological mechanisms, and their measurements are significantly correlated. Thus it is sensible to conduct marker selection under the homogeneity structure. Extension to the heterogeneity structure can follow the literature [Zhao et al., 2015].

Building on the MCP, we adopt the group MCP (GMCP) for marker selection. Denote B_j as the $m \times 1$ vector composed of the elements in the j th row of B . It contains the regression

coefficients of the j th SNP on the m traits. Define $\|B_j\|_{\sum_j} = (B_j' \sum_j B_j)^{1/2}$, where $\sum_j = \tilde{T}_j' \tilde{T}_j / n$, and $\tilde{T}_j$ consists of the components of $\tilde{T}$ corresponding to the j th SNP. Write $\sum_j = R_j' R_j$ 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 $\beta_j = R_j B_j$. Denote $\beta = (\beta_1', \dots, \beta_p')'$. The GMCP is defined as

$$P_\lambda(\beta) = \sum_{j=1}^p \rho(\|\beta_j\|; \sqrt{m}\lambda, \gamma).$$

The overall penalized objective function is

$$L(v, b, \sum_g, \sum_e) - P_\lambda(\beta).$$

Similar to MCP, when $\gamma \rightarrow \infty$, GMCP becomes group Lasso (GLasso), which has also been popular in the literature. The rationale of the GMCP is as follows. With multiple traits, the effects of one SNP are represented by m coefficients, with one for each trait. Those m coefficients naturally form a group. When the homogeneity structure is assumed, multiple traits share the same set of important SNPs. Thus in selection, it is desired to have the “group in, group out” property, which can be achieved using a group penalty. We adopt the $\|\cdot\|_{\sum_j}$ norm—as opposed to the simple ℓ_2 norm—to simplify computation.

With the notations described above, simple calculation shows that the proposed optimization can be rewritten as

$$\min_{\beta} \left\{ L_{\text{GM}}(\beta) = \frac{1}{2n} \|\tilde{y} - \sum_{j=1}^p V_j \beta_j\|^2 + P_\lambda(\beta) \right\},$$

which can be realized using a group coordinate descent (GCD) approach [Huang et al., 2012]. In the GCD, consider optimization with respect to the j th group (SNP). When there

is no penalty, regress $\tilde{y}$ upon V_j (covariates of the j th group), and the least squares solution is $z_j = n^{-1} V_j' \tilde{y}$. With the GMCP, the penalized estimate has the following closed form

$$\hat{\beta}_j = \begin{cases} \frac{\gamma}{\gamma-1} S_1(z_j, \sqrt{m}\lambda), & \text{if } \|z_j\|_2 \leq \gamma \sqrt{m}\lambda \\ z_j, & \text{if } \|z_j\|_2 > \gamma \sqrt{m}\lambda \end{cases},$$

where $S_1(z, \lambda) = (1 - \lambda/\|z\|_2)_+ z$.

There is one tuning parameter λ and one regularization parameter γ . In general, smaller values of γ are better at retaining the unbiasedness of MCP for large coefficients, but 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 large enough to avoid this problem but not too large. Simulation studies in Breheny and Huang [2011] suggest that $\gamma = 3$ is a reasonable choice for GMCP. We search for the optimal λ using V -fold cross validation ($V = 5$ in our numerical study). More details regarding the choice of tuning parameter for GMCP can also be found in Huang et al. [2012] and Liu et al. [2012].

Practical Considerations

As suggested in some recent studies [Madsen and Browning, 2009; Wu et al., 2011], grouping SNPs at the gene level may improve the power of detecting susceptibility genes. This is especially useful when the minor allele frequencies (MAF) are low. Although grouping has been more extensively used with rare variants, it is also applicable to the analysis of common variants as in this study. We group SNPs using the weighted-sum method [Madsen and Browning, 2009]. The weights are calculated as:

$$\hat{w}_k^j = 1 / \sqrt{\text{MAF}_k^j (1 - \text{MAF}_k^j)},$$

where MAF_k^j is the MAF of the k th SNP in the j th gene. Then, the genetic score for subject i and gene j can be calculated as

$$X_{ij} = \sum_{k=1}^{p_j} \hat{w}_k^j G_{ik}^j,$$

where G_{ik}^j is the genotype at the k th SNP of the j th gene for the i th subject, and p_j is the number of SNPs in the j th gene.

The proposed method can be used to analyze a training dataset, identify important markers, and construct models. When there is a testing dataset with unobserved traits, prediction can be made. Specifically, using notations similar to those above, according to the best linear unbiased predictor (BLUP) principle, the predicted value y_p is given by

$$y_p = W\hat{v} + X\hat{b} + K_{tt}\hat{H}^{-1}(y_t - W\hat{v} - X\hat{b}), \quad (7)$$

where $\hat{v}$ and $\hat{b}$ are the estimates for the nongenetic effects and genetic effects, K_{tt} is the covariance matrix between the training set and testing sample, and $\hat{H}$ is the matrix of variance components from the training set.

Simulation Study

We conduct simulation to better gauge performance of the proposed methods. The genotype data are excerpted from the GAW 18. We randomly choose 400 subjects from 20 families and 5,000 genetic scores of genes collapsed from SNPs and generate two correlated traits. That is, $n = 400$, $p = 5,000$, and $m = 2$. We consider six correlation scenarios for residuals (Σ_e) and random effects (Σ_g). Details are provided in Table A1 (Appendix). Scenarios 1–3 represent the case that Σ_e and Σ_g are proportional with weak, moderate, and strong correlations, while scenarios 4–6 represent the case that Σ_e and Σ_g are not proportional with different combinations of values. More specifically, scenario 4 has correlation weak in Σ_e but strong in Σ_g . Scenario 5 has correlation weak in Σ_g but strong in Σ_e . Scenario 6 has strong correlations in both Σ_e and Σ_g , but the diagonal elements are different in Σ_e . Although the proposed method has been designed for the homogeneity structure, to be comprehensive, we also conduct simulation under the heterogeneity structure. More specifically, under the simulated homogeneity structure, the indices of important genes are 1–5 and 16–20 for both traits. Under the simulated heterogeneity structure, the indices of important genes are 1–5 and 16–20 for trait 1 and 1–5 and 21–25 for trait 2. Here under the heterogeneity structure, the two traits still share common important genes, which is reasonable with the level of biological and statistical correlations in the GAW 18 and data alike. All nonzero coefficients are set as 0.5. The genetic relatedness matrix K is calculated using the 5,000 genes. Here we do not consider the low-dimensional nongenetic fixed effects, which are relatively easy to handle. The traits y_0 are generated under model (2).

We analyze the simulated data using the proposed method. We also apply the proposed modeling with GLasso, which is a special case of GMCP with $\gamma = \infty$, for selection. As a “baseline,” for each trait separately, we adopt the LMM and apply both MCP and Lasso for marker selection. With a sequence of tuning parameter values, we construct the ROC (receiver operating characteristic) curves to compare the selection accuracy of different methods. Similar evaluation has been conducted in multiple published studies. The ROC curves under the homogeneity and heterogeneity structures are shown in Figures 1 and 2, respectively. It is noted that as usually only the region with small FPR (false-positive rate) is of interest, only partial curves are plotted. To more quantitatively compare different methods, we compute the scaled partial AUC [area under curve, Walter, 2005], which is defined as $AUC_{rs}^* = AUC_{rs} / (s - r)$, where AUC_{rs} is the partial AUC for $r \leq FPR \leq s$. The scaled partial AUC results are summarized in Table 1.

Simulation suggests that, across the whole simulation spectrum, the proposed MTMM with GMCP and GLasso significantly outperforms the LMM analysis with MCP and Lasso. For

example with correlation scenario 1, under the homogeneity structure, the mean scaled PAUCs are 0.537 (MTMM+GMCP), 0.654 (MTMM+GLasso), 0.355 (LMM+MCP), and 0.432 (LMM+Lasso), respectively. Under the heterogeneity structure, the mean scaled PAUCs are 0.540, 0.653, 0.351, and 0.429, respectively. It is interesting to observe the superior performance of the proposed method under the heterogeneity structure. The two simulated traits have different sets of markers. Thus, it may seem that the separate analysis, which allows such difference, should fit such data better. However, under our specific simulations settings, the two traits still share common markers. By borrowing information across traits, the proposed method can more accurately identify the shared markers and, as a result, also have better overall identification accuracy. More closely examining Figures 1 and 2 suggests that in the region with extremely small FPR, GMCP outperforms GLasso with larger TPR (true-positive rate) values. However, when FPR gets larger, GLasso can have better performance.

To complement the marker identification analysis, we also evaluate prediction performance. Specifically, a cross-validation based approach is adopted. A simulated replicate is split into fivefolds with equal sizes. Fourfolds are used as the training data, and the remaining fold is used as the testing data. Models are constructed using the training data, and prediction is conducted for the testing data as described above. Penalization is adopted in this study, which may lead to estimates shrunk toward zero. To eliminate the effect of this shrinkage, we compute the correlation between the predicted and observed values. Across the simulated scenarios and replicates, under the homogeneity structure, the MTMM+GMCP generates mean (SD) correlation 0.361 (0.049), compared to 0.306 (0.162) with the LMM+MCP. Under the heterogeneity structure, the mean (SD) correlations are 0.432 (0.072) and 0.300 (0.188), respectively. The GLasso- and Lasso-based methods generate smaller correlations. The MTMM+GMCP method has the best prediction performance.

Beyond those described above, additional simulations are also conducted and described in the Appendix, which again demonstrate the superiority of proposed method. We note that quite a few other penalties can be coupled with the proposed MTMM or LMM. The Lasso-based methods are compared because of the popularity of Lasso. Under simpler data settings, MCP has been shown to have performance better than or comparable to other penalties. Thus comparison with penalties other than Lasso is not pursued.

Analysis of the GAW 18 Data

Brief information on the data is provided in the first section. More detailed information is available at www.gaworkshop.org/gaw18/index.html and www.biomedcentral.com/1753-6561/8/S1/S1. Among the 472,049 genotyped SNPs, 292,355 are within the regions of genes (exon regions). We collapse SNPs within the same gene and generate genetic scores for 10,949 genes. After quality control (e.g., removal of subjects with certain covariates missing), 823 subjects with genetic scores for 10,549 genes are included in the downstream analysis. For those subjects, DBP ranges from 40.0 to 123.0 mm Hg (mean 77.0 mm Hg), and SBP ranges from 80.0 to 216.0 mm Hg (mean 129.0 mm Hg). For nongenetic variables, we include age and smoking status. Age ranges from 16.0 to 92.1 years (mean 49.3 years),

and 179 people are current smokers. For each subject, up to four measurements are available for each trait, and we focus on the first measurements.

In applying the proposed method, the genetic relatedness matrix is estimated from the genetic scores of the 10,949 genes. The variance components for the residuals (Σ_e) are $\begin{pmatrix} 0.799 & 0.322 \\ 0.322 & 0.565 \end{pmatrix}$, and those for the random effects (Σ_g) are $\begin{pmatrix} 0.168 & 0.148 \\ 0.148 & 0.157 \end{pmatrix}$. Therefore, the estimated heritability for DBP is $\sigma_{g(1)}^2 / (\sigma_{e(1)}^2 + \sigma_{g(1)}^2) = 0.174$, and that for SBP is $\sigma_{g(2)}^2 / (\sigma_{e(2)}^2 + \sigma_{g(2)}^2) = 0.217$. Here we note that the estimated heritability values may be slightly higher than the true values, as the common environmental factors shared by related individuals have not been accounted for.

In Table 2, we provide the marker identification and estimation results under the proposed MTMM. We provide results using the GMCP, as simulation suggests its superior performance when the FPR is low. For a direct comparison, we also apply the LMM+MCP. The MTMM+GMCP identifies 30 genes as associated with both SBP and DBP. For the dominating majority of these genes, the estimates have the same sign for the two traits. We note that there are a few very small values (possibly 0.000 due to truncation). It is possible that they correspond to no association between the genes and traits. In comparison, the LMM+MCP identifies 19 genes for DBP and five genes for SBP. There are only two overlapped genes. Considering the level of biological and statistical correlations between SBP and DBP, the lack of overlapped genes of the alternative method seems insensible. The results using the GLasso/Lasso-based methods are available from the authors.

Quick literature search suggests that the genes identified using the proposed MTMM +GMCP have sensible implications and replicate the findings of the previous studies. For example, Tayo et al. [2009] identified an intergenic SNP, which is located between the *RP43* and *GL CCI1* genes on chromosome 7, as associated with blood pressure in Nigerian families. Seymour and Chatham [1997] discovered that the *OGDH* gene is related to hypertrophy. Multiple studies [Kaess et al., 2011; Talmud et al., 2009] identified that the *ANKRD40* and *NCAM1* genes are associated with cardiovascular diseases. Kothapalli et al. [2007] found that the expression of gene *EXOSC2* is associated with docosahexaenoic acid (DHA).

To complement the identification analysis, we evaluate the stability of identified genes using a random sampling approach [Huang and Ma, 2010]. Specifically, we randomly sample from the original data and apply the “MTMM+GMCP” and “LMM+MCP” to identify genes associated with SBP and DBP. The process is repeated 1,000 times. With each gene, we calculate the probability that it is identified out of the 1,000 samplings. This probability has been referred to as the observed occurrence index (OOI) [Huang and Ma, 2010] and measures the relative stability. The OOI values under the two methods are also provided in Table 2. For most of the identified genes, the OOI values are high, suggesting satisfactory stability. In addition, we also evaluate prediction performance in a similar manner as in simulation. For “MTMM+GMCP” and “LMM+MCP”, the mean correlations (SD) between the predicted and observed values are 0.180 (0.035) and 0.130 (0.071), respectively. We note

that the correlations are not particularly high. This is reasonable as the levels of blood pressures are affected by a large number of factors, and genetic variations are expected to be able to explain only a proportion of the levels. The “MTMM+GMCP” outperforms “LMM+MCP” with better prediction.

To be comprehensive, we also analyze data under the heterogeneity structure. The MTMM is adopted. The MCP is applied to allow for different sets of markers for different traits. The results are provided in the Appendix (Table A2). Some of the findings replicate those described in Table 2. As expected, the two structures lead to different results. It can be of interest to data adaptively determine which structure is more suitable. However, this will involve high-dimensional hypothesis testing and has not been conducted in the literature. As it is highly nontrivial and “deviates” from the proposed methodological development, it is not pursued.

Discussion

Pedigree data have been shown to be powerful in genetic studies. Recently, more and more studies have collected and analyzed data on multiple traits. Complicated correlations arise with pedigree and multitrait data. Such correlations, coupled with the need for marker identification, make analysis challenging. LMM is a powerful tool for accommodating correlations. In this study, we have extended the existing methods and effectively accommodated the two types of correlations. We adopt a GMCP approach for marker identification. The proposed method has an intuitive formulation and can be realized using an effective coordinate descent approach. Simulation and the analysis of the GAW 18 data demonstrate satisfactory performance of the proposed method and advantage over the alternative analysis.

With the GAW 18 data, the traits are continuously distributed, and it is sensible to adopt linear models. Pedigree and multitrait data may also have other types of distributions. The proposed modeling strategy and marker identification method can be potentially extended to accommodate other types of response distributions. Under the specific GAW 18 settings, it is reasonable to assume the homogeneity structure. When the heterogeneity structure is desired, the GMCP needs to be replaced by techniques (e.g., the sparse group penalties or composite penalties) that allow for trait-specific markers [Zhao et al., 2015]. In this study, we have mostly focused on methodological development and implementation. When the complex correlations are not present, a recent study [Zhang et al., 2015] establishes the consistency properties of the GMCP approach. Based on that study, it may be reasonable to conjecture that the proposed method also has the consistency properties. However, Zhang et al. [2015] also suggests that the proof will be highly challenging. Thus we postpone the theoretical investigations to future study. Our simulation covers a wide spectrum of scenarios. More extensive simulations are omitted. In data analysis, the proposed method identifies a consistent set of markers with satisfactory stability and better prediction, which support the proposed analysis to a large extent. Bioinformatics analysis is needed to fully support the findings.

Acknowledgments

We thank the editor and reviewers for their careful review and insightful comments, which have led to a significant improvement of this article. This study has been supported by the Duke-NUS Medical School WBS: R-913-200-098-263, grant NO. 61501389 from National Natural Science Foundation of China (NSFC), grants HKBU 22302815 and HKBU 12202114 from Hong Kong Research Grant Council, and grants FRG2/14-15/069 and FRG2/14-15/077 from Hong Kong Baptist University, and CA142774 and CA016359 from NIH.

References

- Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann Appl Stat.* 2011; 5:232–253. [PubMed: 22081779]
- Gilmour A, Thompson R, Cullis B. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics.* 1995; 51:1440–1450.
- Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis G. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet.* 2012
- Huang J, Ma S. Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Anal.* 2010; 16:176–195. [PubMed: 20013308]
- Huang J, Wei F, Ma S. Semiparametric reregression pursuit. *Stat Sinica.* 2012; 22:1403–1426.
- Kaess BM, Tomaszewski M, Braund PS, Stark K, Rafelt S, Fischer M, Hardwick R, Nelson CP, Debiec R, Huber F, et al. Large-scale candidate gene analysis of HDL particle features. *PLoS One.* 2011; 6:e14529. [PubMed: 21283740]
- Kang H, Sul J, Service S, Zaitlen N, Kong S, Freimer N, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet.* 2010; 42:348–354. [PubMed: 20208533]
- Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet.* 2012; 44:1066–1071. [PubMed: 22902788]
- Kothapalli K, Anthony JC, Pan BS, Hsieh AT, Nathanielsz PW, Brenna JT. Differential cerebral cortex transcriptomes of baboon neonates consuming moderate and high docosahexaenoic acid formulas. *PLoS One.* 2007; 2:e370. [PubMed: 17426818]
- Lippert C, Listgarten J, Liu Y, Kadie C, Davidson R, Heckerman D. FaST linear mixed models for genome-wide association studies. *Nat Methods.* 2011; 8:833–835. [PubMed: 21892150]
- Liu J, Huang J, Ma S. Analysis of genome-wide association studies with multiple outcomes using penalization. *PLoS One.* 2012; 7:e51198. [PubMed: 23272092]
- Liu J, Huang J, Xie Y, Ma S. Integrative analysis of cancer prognosis data with sparse group penalization. 2012
- Ma S, Huang J. Penalized feature selection and classification in bioinformatics. *Brief Bioinform.* 2008; 9:392–403. [PubMed: 18562478]
- Ma S, Huang J, Song X. Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics.* 2011; 12:763–775. [PubMed: 21415015]
- Madsen B, Browning S. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009; 5:e1000384. [PubMed: 19214210]
- Mazumder R, Friedman J, Hastie T. SparseNet: coordinate descent with non-convex penalties. *J Am Stat Assoc.* 2011; doi: 10.1198/jasa.2011.tm09738
- Rakitsch B, Lippert C, Stegle O, Borgwardt K. A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics.* 2013; 29:206–214. [PubMed: 23175758]
- Seymour A-ML, Chatham JC. The effects of hypertrophy and diabetes on cardiac pyruvate dehydrogenase activity. *J Mol Cell Cardiol.* 1997; 29:2771–2778. [PubMed: 9344771]
- Sivakumaran S, Agakov F, Theodoratou E, Prendergast J, Zgaga L, Manolio T, Rudan I, McKeigue P, Wilson J, Campbell H. Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet.* 2011; 89:607–618. [PubMed: 22077970]

- Talmud PJ, Drenos F, Shah S, Shah T, Palmen J, Verzilli C, Gaunt TR, Pallas J, Lovering R, Li K, et al. Gene-centric association signals for lipids and apolipoproteins identified via the Human CVD Beadchip. *Am J Hum Genet.* 2009; 85:628–642. [PubMed: 19913121]
- Tayo BO, Luke A, Zhu X, Adeyemo A, Cooper RS. Association of regions on chromosomes 6 and 7 with blood pressure in Nigerian families. *Circ Cardiovasc Genet.* 2009; 2:38–45. [PubMed: 20031566]
- Walter S. The partial area under the summary ROC curve. *Stat Med.* 2005; 24:2025–2040. [PubMed: 15900606]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 2011; 89:82–93. [PubMed: 21737059]
- Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat.* 2010; 38:894–942.
- Zhang Q, Zhang S, Liu J, Huang J, Ma S. Penalized integrative analysis under the accelerated failure time model. *Stat Sinica.* 2015 In press.
- Zhang Z, Ersoz E, Lai C, Todhunter R, Tiwari H, Gore M, Bradbury P, Yu J, Arnett D, Ordovas J, et al. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet.* 2010; 42:355–360. [PubMed: 20208535]
- Zhao Q, Shi X, Huang J, Liu J, Li Y, Ma S. Integrative analysis of “-omics” data using penalty functions. *Wiley Interdiscip Rev Comput Stat.* 2015; 7:99–108. [PubMed: 25691921]
- Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet.* 2012; 44:821–824. [PubMed: 22706312]

Appendix

This file contains additional numerical results.

Additional Simulation I

We conduct additional simulation and compare the proposed method against those that consider only one type of correlation. More specifically, we consider both the homogeneity and heterogeneity structures and set

$$\sum_g = \begin{pmatrix} 0.20 & 0.10 \\ 0.10 & 0.20 \end{pmatrix}, \quad \sum_e = \begin{pmatrix} 0.40 & 0.20 \\ 0.20 & 0.40 \end{pmatrix}.$$

Figure A1.
Pedigree relatedness observed in the GAW 18 data.

Table A1

Six simulation scenarios for covariances of both unobserved random residuals and polygenic effects

	Σ_e	Σ_g
Scenario 1	$\begin{pmatrix} 0.20 & 0.04 \\ 0.04 & 0.20 \end{pmatrix}$	$\begin{pmatrix} 0.40 & 0.08 \\ 0.08 & 0.40 \end{pmatrix}$

	Σ_e	Σ_g
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	$\begin{pmatrix} 0.20 & 0.16 \\ 0.16 & 0.24 \end{pmatrix}$	$\begin{pmatrix} 0.40 & 0.24 \\ 0.24 & 0.40 \end{pmatrix}$

In model fitting, we consider (Case 1) $\hat{\Sigma}_g$ is constrained to be diagonal, and (Case 2) $\hat{\Sigma}_e$ is constrained to be diagonal. The ROC plots are shown in the figure below. For clarity of view, we only provide results using the GMCP/MCP. Overall, we observe the superior performance of the MTMM+GMCP.

Additional Simulation II

In this set of simulation, we consider diagonal correlations. More specifically, for both the homogeneity and heterogeneity structures, we set

$$\Sigma_g = \begin{pmatrix} 0.20 & 0.00 \\ 0.00 & 0.20 \end{pmatrix}, \quad \Sigma_e = \begin{pmatrix} 0.40 & 0.00 \\ 0.00 & 0.40 \end{pmatrix}.$$

The ROC plots are shown below. The observed pattern is similar to that described in the main text.

Table A2

Analysis of the GAW 18 data under the heterogeneity structure

Gene	DBP		SBP	
	Estimate	OOI	Estimate	OOI
AIRE			-0.011	0.635
ANKRD40			0.034	0.965
C19orf38			0.038	0.937
CYP4Z1			0.009	0.288
ELMO1	-0.031	0.864		
EXOSC2			0.018	0.757
FBXL18	0.004	0.456		
HEATR2			0.014	0.829
HS3ST3B1	0.009	0.475		
IL27RA			0.003	0.292
LOC153469	-0.001	0.378		
MMP27			0.001	0.443
MOBK2C	-0.009	0.656		
NCAM1	-0.021	0.843		
NDUFA13	-0.014	0.448		

Gene	DBP		SBP	
	Estimate	OOI	Estimate	OOI
NEUROG1	0.006	0.353		
NOM1	−0.004	0.368		
OR8A1			−0.040	0.995
PLAA			−0.001	0.298
PYCR2	0.025	0.890		
RAMP2			−0.009	0.420
RPA3	−0.034	0.940		
SCYL1BP1	−0.009	0.545		
SLC16A4	−0.012	0.680		
SLC25A36			0.001	0.365
TLCD2	0.022	0.771		
ZNF3			−0.006	0.664
ZNF611			−0.007	0.391
ZNF826			0.036	0.959
ZRANB2			−0.003	0.353

Additional Simulation III

In this set of simulation, we consider data with three traits. More specifically, under both the homogeneity and heterogeneity structures, we generate data with three traits in a similar manner as Scenario 2 previously described. The ROC plots are shown below. The observation is similar to those described above.

Additional Simulation IV

Under the settings described in the main text, all nonzero regression coefficients are equal. We now conduct simulation and examine the case with heterogeneous effect sizes. Specifically, under the homogeneity structure, the coefficients for the first and second traits are

$$(0.5, -0.5, 1, -1, 0.4, 0, \underbrace{\dots, 0}_{10}, \underbrace{0.5, \dots, 0.5}_5, \underbrace{0, \dots, 0}_{4980})$$

and

$$(\underbrace{0.5, \dots, 0.5}_5, \underbrace{0, \dots, 0}_{10}, \underbrace{-0.5, \dots, -0.5}_5, \underbrace{0, \dots, 0}_{4980}).$$

Under the heterogeneity structure, the coefficients for the first trait remain the same. The coefficients for the second trait are

$$(\underbrace{0.5, \dots, 0.5}_5, \underbrace{0, \dots, 0}_{15}, \underbrace{-0.5, \dots, -0.5}_5, \underbrace{0, \dots, 0}_{4975}).$$

The ROC plots are shown below, and the pattern is similar to the previously observed.

(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

(e) Scenario 5

(f) Scenario 6

Figure 1.
Simulation: ROC curves under the homogeneity structure.

(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

(e) Scenario 5

(f) Scenario 6

Figure 2.
Simulation: ROC curves under the heterogeneity structure.

Table 1

Simulation: scaled partial AUC (standard deviation)

		MTMM+GMCP	MTMM+GLasso	LM+MCP	LM+Lasso
Scenario 1	Homogeneity	0.537(0.134)	0.654(0.109)	0.355(0.084)	0.432(0.082)
Scenario 2	Homogeneity	0.486(0.128)	0.556(0.131)	0.340(0.087)	0.410(0.095)
Scenario 3	Homogeneity	0.411(0.131)	0.552(0.118)	0.359(0.092)	0.433(0.079)
Scenario 4	Homogeneity	0.517(0.105)	0.598(0.126)	0.332(0.108)	0.419(0.084)
Scenario 5	Homogeneity	0.520(0.089)	0.606(0.117)	0.337(0.094)	0.420(0.094)
Scenario 6	Homogeneity	0.449(0.128)	0.542(0.128)	0.324(0.100)	0.413(0.092)
Scenario 1	Heterogeneity	0.540(0.089)	0.653(0.086)	0.351(0.082)	0.429(0.079)
Scenario 2	Heterogeneity	0.531(0.101)	0.658(0.091)	0.352(0.083)	0.426(0.089)
Scenario 3	Heterogeneity	0.558(0.093)	0.691(0.080)	0.324(0.102)	0.420(0.082)
Scenario 4	Heterogeneity	0.510(0.092)	0.632(0.093)	0.333(0.091)	0.413(0.075)
Scenario 5	Heterogeneity	0.489(0.103)	0.455(0.114)	0.324(0.095)	0.420(0.087)
Scenario 6	Heterogeneity	0.518(0.093)	0.650(0.089)	0.327(0.084)	0.421(0.080)

Table 2

Analysis of the GAW 18 data using different methods

Gene	MTMM+GMCP			LMM+MCP		
	DBP Estimate	SBP Estimate	OOI	DBP Estimate	SBP Estimate	OOI
AIRE	-0.004	-0.010	0.582			
ANKHD1-EIF4EBP3	0.001	0.000	0.404	0.017	0.656	
ANKRD40	-0.001	0.021	0.881			
C19orf38	-0.012	0.027	0.956			
C1orf128	0.004	0.002	0.438	0.015	0.659	
C3orf20	0.004	0.002	0.440	0.007	0.474	
ELMO1	-0.025	-0.005	0.740			
EXOSC2	0.041	0.058	0.985		0.038	0.988
FNDC3A	0.002	0.004	0.504		0.015	0.713
HEATR2	0.000	0.003	0.680			
HS3ST3B1	0.015	0.006	0.667	0.022	0.795	
INOC1	0.042	0.038	0.952	0.019	0.731	0.004 0.546
LCE1A	0.024	0.023	0.842	0.007	0.489	
LOC390998	0.028	0.021	0.885	0.028	0.854	
MOBK12C	-0.009	-0.003	0.577			
NCAMI	-0.016	-0.004	0.758	-0.024	0.787	
NDUFA13	-0.012	0.007	0.635			
OGDH	0.026	0.025	0.908	0.022	0.761	0.011 0.665
OR8A1	-0.001	-0.032	0.978			
PAPOLB	0.006	0.007	0.611			
PIAS1	0.003	0.002	0.430	0.016	0.674	
PYCR2	0.021	0.006	0.856	0.008	0.597	
RPA3	-0.036	-0.010	0.933	-0.030	0.884	
SLC16A4	-0.000	0.000	0.417			
SYT3	0.025	0.018	0.833	0.024	0.807	
TLCD2	0.015	-0.003	0.673			

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Gene	MTMM+GMCP			LMM+MCP		
	DBP Estimate	SBP Estimate	OOI	DBP Estimate	SBP Estimate	OOI
ZCCHC10	0.022	0.012	0.738	0.027	0.783	
ZNF3	-0.004	-0.009	0.663			
ZNF611	-0.007	-0.014	0.698		-0.011	0.728
ZNF826	-0.006	0.027	0.937			
ARSB				-0.007	0.444	
DFFA				0.005	0.383	
MCM7				-0.001	0.382	
MKNK1				-0.002	0.309	
SFRS12				0.007	0.522	