Penalized Methods for Multiple Outcome Data in Genome-Wide Association Studies

Jin Liu[1], Shuangge Ma[1], and Jian Huang[2] *

[1]Division of Biostatistics, School of Public Health, Yale University
[2]Department of Statistics & Actuarial Science, and Department of Biostatistics, University of Iowa

The University of Iowa

Department of Statistics and Actuarial Science

Technical Report No. 413

Abstract

Genome-wide association studies have been extensively conducted, searching for markers for complex diseases. Penalization methods have been adopted for the analysis of joint effects of a large number of SNPs (single nucleotide polymorphisms) and marker identification. This study has been motivated by the analysis of Ocular Hypertension Treatment Study (OHTS), which for each subject, measures two highly correlated response variables and a large number of SNPs. Existing penalization methods have been designed to analyze a single response variable, cannot effectively accommodate the correlation structure among multiple response variables, and hence can be inefficient. With multiple response variables sharing the same set of markers, we first apply joint modeling to accommodate the correlation structure. A group penalization approach is adopted to select markers associated with all outcomes. An efficient computational algorithm is developed. Simulation study and analysis of the OHTS data show that the proposed method can outperform existing penalization analysis methods.

Keywords:GWAS; Multiple Outcomes; Penalization.

1 Introduction

This study has been partly motivated by the analysis of Ocular Hypertension Treatment Study (OHTS) on glaucoma. Glaucoma is a group of diseases that can lead to damage to

*To whom correspondence should be addressed. jian-huang@uiowa.edu

the eye's optic nerve and result in blindness. Open-angle glaucoma is the most common form of glaucoma. At first, it has no symptoms, but will gradually deteriorate patients' eyes until there is no vision left. In the literature, several clinical risk factors for glaucoma have been suggested, including age, race, eye pressure, certain characteristics in the anatomy of the optic nerve, and thinness of the cornea NIH [2002]. An important cause of open-angle glaucoma is the increased pressure inside the eye(s). OHTS is the first large-scale study to demonstrate that reducing elevated eye pressure can safely and effectively delay and possibly prevent the disease Kass et al. [2002]. Particularly, this study shows that, at 60 months, the cumulative probability of developing glaucoma was 4.4% for study participants who received the eye drops (which could reduce eye pressure). In contrast, the cumulative probability was 9.5% for study participants who did not receive the eye drops. The difference is statistically significant.

As with many complex human diseases, clinical risk factors and environmental exposures have failed to provide a comprehensive description of glaucoma development. Recently, Ramdas et al. Ramdas et al. [2010] performed a meta-analysis of GWAS (genome-wide association studies) to identify genetic variants associated with optic disc area and vertical cup-disc ratio. CCT (central corneal thickness) is an important predictive factor for the development of glaucoma among OHTS participants. There are two measurements per subject, namely right eye CCT and left eye CCT. CCT measurements on the same subjects tend to be highly correlated. In the OHTS dataset we analyze, there are 1,058 subjects. Records with missing CCT values are removed from analysis, leading to an effective sample size of 1,054. On those subjects, the correlation coefficient between CCT OD (right eye) and CCT OS (left eye) measurements is 0.958. The five-number summary statistics for CCT OT and CCT OS are $(440.0, 547.2, 573.0, 572.9, 599.0, 692.0)$ and $(444.0, 546.0, 574.0, 573.7, 600.0, 704.0)$. 11217 SNPs on chromosome 22 were profiled, searching for markers associated with CCT

OD and CCT OS.

GWAS data has extremely high dimensionality. To tackle this problem, some statistical approaches analyze one SNP at a time and then adjust for multiple comparisons. Such approaches are easy to implement, however, may contradict the fact that the development and progression of complex diseases are caused by the aggregated effects of multiple SNPs. They may miss SNPs with weak marginal but strong joint effects. In the analysis of joint effects of a large number of SNPs, regularized estimation is needed. In addition, it is expected that only a subset of profiled SNPs are associated with the response variables. Thus, marker selection is needed along with estimation.

With high-dimensional data, penalization has been extensively applied for regularized estimation and variable selection. Commonly used penalization methods include LASSO Tibshirani [1996], elastic net Zou and Hastie [2005], bridge Frank and Friedman [1993], Fu [1998], SCAD Fan and Li [2001], MCP Zhang [2010] and others. Such methods can effectively analyze data with a single response variable with interchangeable covariate effects. When there exists hierarchical structure among covariates, for example the "pathway, SNP-within-pathway" two-level structure, the "group" version of the aforementioned penalization methods have been proposed. The group penalty is usually a composite penalty. For example with group SCAD Wang et al. [2007], the outer penalty is the SCAD penalty, and the inner penalty is the ridge penalty. We note that, such group penalization methods are still mainly used for the analysis of data with a single response variable. For the aforementioned penalization methods, computational algorithms based on coordinate descent or boosting have been developed.

In this study, our goal is to analyze data with multiple correlated response variables and conduct marker selection. In "classic" statistical analysis with a small number of covariates, data with multiple response variables can be accommodated under the framework

of multivariate analysis of variance (MANOVA) Stevens [2002] and multivariate analysis of covariance (MANCOVA). However, such methods cannot accommodate high dimensional covariates. It is possible to first apply existing penalization methods, for example LASSO, analyze each response variable separately, and then combine the analysis results using meta-analysis methods. However, such an approach ignores the correlation among response variables and hence can be less informative.

In OHTS, there are two continuously distributed, highly correlated response variables. Under a joint modeling framework, we propose first transforming multi-response data into uni-response data following the same distribution. Then a group LASSO approach is applied to the transformed uni-response data. With two responses, the effect of one SNP needs to be represented by two regression coefficients, which naturally form a "group". We emphasize that, unlike other group penalization studies in which one group usually corresponds to multiple covariates, here one group corresponds to a single covariate for multiple responses.

The rest of the paper is organized as follows. In Section 2, we describe the data and model setup. In Section 3, we describe marker selection using the group LASSO method, propose a coordinate descent algorithm for estimating model parameters and discuss tuning parameter selection. Section 4 evaluates the proposed method using simulated data. Section 5 applies the proposed method to the OHTS data. The article concludes with discussions in Section 6.

2 Analysis of Multi-response Data

Consider data with multiple correlated response variables. With data like OHTS, it is reasonable to assume that multiple responses share a certain common ground, particularly the same set of susceptible SNPs. However, we note that although the response variables are correlated, they are not identical. With the inherent heterogeneity, it is not sensible to reinforce the same model with the same regression coefficients for different response variables.

Let $M(>1)$ be the number of response variables and n be the number of subjects. Denote $y^1, \ldots, y^M$ as the response variables and $\mathbf{x}$ as the $n \times p$ covariate matrix. For $m = 1, \ldots, M$, assume that y^m is associated with $\mathbf{x}$ via the model $y^m \sim \phi(\mathbf{x}\beta^m)$, where $\beta^m (= (\beta_1^m, \ldots, \beta_p^m)')$ is the regression coefficient corresponding to the mth response variable, and ϕ is the link function. We first transform the original data frame. For simplicity of notation, we use the same symbol y but with different subscripts for the new response variable. Although the proposed method can accommodate different covariates for distinct response variables, we assume that the same set of covariates are measured for all responses. Let y_i be the length-M vector of response variables, and $\mathbf{y} = (y_1', \ldots, y_n')'$. Covariates for the ith subject with a length-M vector of response variables has the form $X_i = (U_{i1}, \ldots, U_{ip})$ where $U_{ij} = x_{ij} I_m$. The regression coefficient vector is $B = (\beta_1', \ldots, \beta_p')'$ where $\beta_j = (\beta_j^1, \ldots, \beta_j^M)'$.

To better illustrate the basic features of the model settings here, consider a dataset with M=2 response variables and p SNPs. Assume that only the first 4 SNPs are associated with responses. Then the coefficients may look like

$$\begin{aligned} \beta^1 &= (0.1, 0.4, 0.3, 0.8, 0, \ldots, 0)' \\ \beta^2 &= (0.03, 0.2, 0.5, 0.6, 0, \ldots, 0)', \end{aligned}$$

and correspondingly,

$$B' = (0.1, 0.03, 0.4, 0.2, 0.3, 0.5, 0.8, 0.6, 0, \ldots, 0).$$

The regression coefficient B and corresponding model have the following features. First, only the first 4 disease-associated SNPs have nonzero regression coefficients (i.e. the model is sparse). Thus, marker identification amounts to discriminate SNPs with non-zero coefficients from those with zero coefficients. This strategy has been commonly used in regularized marker selection. Second, as the two response variables share the same susceptible SNPs, there is natural grouping structure with the transformed covariates. For example, the first

two regression coefficients/covariates correspond to the first SNP. Thus, they form a *group* of size two and should be selected at the same time.

Motivated by the OHTS data, we describe the proposed approach for studies with quantitative traits under linear models. The proposed approach can be easily extended to other types of response variables and other statistical models, as long as the joint modeling of response variables can be conducted. In a study with M response variables, the least square loss function for transformed data can be written as

$$\sum_{i=1}^{n}(y_i - X_iB)'\Sigma^{-1}(y_i - X_iB),$$

where Σ is the covariance matrix among residuals.

3 Penalized Estimation and Marker Selection

3.1 Penalized estimation

From definition, $\beta_j = (\beta_j^1, \ldots, \beta_j^M)'$ is the coefficients for the M responses at the jth locus. We define $\hat{B}$ as the minimizer of the penalized least squares loss function:

$$\begin{aligned}\hat{B} &= \underset{B}{\operatorname{argmin}} \left\{ \frac{1}{2n}\sum_{i=1}^{n}(y_i - X_iB)'\Sigma^{-1}(y_i - X_iB) + \lambda\sum_{j=1}^{p}\sqrt{d_j}||\beta_j||_{\Sigma_j} \right\} \\ &= \underset{B}{\operatorname{argmin}} \left\{ \frac{1}{2n}||\tilde{y} - \sum_{j=1}^{p}\tilde{U}_j\beta_j||^2 + \lambda\sum_{j=1}^{p}\sqrt{d_j}||\beta_j||_{\Sigma_j} \right\}. \qquad (1)\end{aligned}$$

Here $\tilde{y}_i = \Sigma^{-\frac{1}{2}}y_i$, $\tilde{X}_i = \Sigma^{-\frac{1}{2}}X_i$, $\tilde{U}_{ij} = \Sigma^{-\frac{1}{2}}U_{ij}$, $\tilde{y}' = (\tilde{y}_1', \ldots, \tilde{y}_n')$, $\tilde{U}_j = (\tilde{U}_{1j}', \ldots, \tilde{U}_{nj}')'$, $\Sigma_j = n^{-1}\tilde{U}_j'\tilde{U}_j$ and d_j is the number of levels at the jth locus (usually equals M). Note that prior to the transformation, we assume that the response follows a multivariate normal distribution, which has been motivated by data similar to OHTS. In contrast, after transformation, the new response $\tilde{y}$ follows a univariate normal distribution. We center the response and make the grand mean equal to zero.

The proposed penalty has been motivated by the following considerations. For a given SNP locus, we treat its regression coefficients for M response variables as a "group", so that we can evaluate its overall effects. The within-group penalty has an L_2 norm, and the group-level penalty has an L_1 norm. Thus, the proposed penalty may have the following main properties. First, it can conduct group-level selection. Second, if a group is selected, then all members within that group are selected with non-zero estimates. But the magnitudes of regression coefficients may differ. On the other hand, if a group is not selected, all of its member are set to be zero. Such properties properly fit the goal of the proposed analysis.

As discussed in Huang et al. [2011], we need to orthogonalize the transformed covariates block-wise in order to achieve computational efficiency. Write $\Sigma_j = R_j' R_j$ for an upper triangular matrix R_j via Cholesky decomposition. Assume that Σ_j is invertible. Let $V_j = \tilde{U}_j R_j^{-1}$ and $b_j = R_j \beta_j$, then the penalized least-squares in expression (1) becomes

$$Q(\lambda, b) = \frac{1}{2n} ||\tilde{y} - \sum_{j=1}^{p} V_j b_j||^2 + \lambda \sum_{j=1}^{p} \sqrt{d_j} ||b_j||. \tag{2}$$

If we center $\tilde{y}$, there is no need to fit for intercept for (2).

3.2 Computational algorithm

We use the group cyclical coordinate descent (GCD) algorithm originally proposed for group LASSO and group MCP Yuan and Lin [2006], Huang et al. [2011]. The GCD algorithm is a natural extension of the coordinate descent algorithm Fu [1998], Wu and Lange [2007], Friedman et al. [2010]. It optimizes a target function with respect to a single group parameter at a time and iteratively cycles through all group parameters until convergence is reached. It is particularly suitable for problems such as the current one that has a simple closed-form solution with a single group but lacks one with multiple groups.

The GCD algorithm proceeds as follows. For a given λ,

1. Let $\tilde{b}^{(0)}$ be the initial estimate. A sensible initial estimate is zero (component-wise). Initialize the vector of residuals $r = \tilde{y} - \sum_{j=1}^{p} V_j \tilde{b}_j^{(0)}$ and $s = 0$.

2. For $j = 1, \ldots, p$, repeat the following steps:

   (a) Calculate the least-square estimates with respect to b_j

   $$\begin{aligned} \tilde{v}_j &= n^{-1} V_j' (\tilde{y} - \sum_{j=1}^{p} V_j \tilde{b}_j^{(s)} + V_j \tilde{b}_j^{(s)}) \\ &= n^{-1} V_j' r + \tilde{b}_j^{(s)}. \end{aligned}$$

   (b) Compute

   $$\tilde{b}_j^{(s+1)} = S(\tilde{v}_j; \sqrt{d_j}\lambda) = \left(1 - \frac{\sqrt{M}\lambda}{||\tilde{v}_j||}\right)_+ \tilde{v}_j. \tag{3}$$

   (c) Update $r \leftarrow r - V_j(\tilde{b}_j^{(s+1)} - \tilde{b}_j^{(s)})$. $s \leftarrow s + 1$.

3. Iterate Step 2 until convergence.

Breheny and Huang Breheny and Huang [2011] discussed the convergence of coordinate descent algorithms for SCAD and MCP. We now consider the GCD for group LASSO. For any given λ, starting from an initial $b^{(0)}$, the GCD algorithm generates a sequence of updates $b^{(s)} = (b_1^{(s)\prime}, \ldots, b_p^{(s)\prime})$, $s = 1, 2, \ldots$, where

$$b_j^{(s)} = argmin_{b_j} Q(b_1^{(s)}, \ldots, b_{j-1}^{(s)}, b_j, b_{j+1}^{(s-1)}, \ldots, b_p^{(s-1)}; \lambda), 1 \leq j \leq p.$$

Since the sequence $\{Q(b^s; \lambda) : s \geq 1\}$ is non-increasing and bounded below by 0, it always converges. The following proposition is concerned about the convergence of $\{b^{(s)} : s \geq 1\}$.

Proposition 1. *For any fixed λ, the GCD updates $\{b^{(s)} : s \geq 1\}$ converge to a local minimizer of the group LASSO criterion $Q(\lambda)$ and satisfy the inequality*

$$Q(b^{(s-1)}; \lambda) - Q(b^{(s)}; \lambda) \geq \frac{1}{2}||b^{(s-1)} - b^{(s)}||^2.$$

Huang et al.Huang et al. [2011] showed that the GCD algorithm for group MCP converges. When $\gamma = +\infty$, group MCP becomes the group Lasso. The group Lasso should be at least no worse than the the convergence of group MCP.

3.3 Choice of tuning parameter

There are various methods that can be applied, which include AIC Akaike [1974], BIC Schwarz [1978], cross-validation Hastie et al. [2009] and generalized cross-validation Wahba [1985]. Chen and Chen Chen and Chen [2008] developed a family of extended Bayes information criteria (EBIC) to overcome the overly liberal selection problem caused by the small-n-large-p situation. Furthermore, Chen and Chen [2010] established the consistency the EBIC under the generalized linear models in the small-n-large-p situation. For group LASSO, Yuan and Lin Yuan and Lin [2006] propose an approximation to the degree freedom (DF). Here, we apply EBIC with approximated DF to select the tuning parameter λ. The BIC is defined as:

$$\text{BIC} = n\text{log}RSS_\lambda/n + \tilde{\text{df}}\left\{\hat{\mu}(\lambda) \equiv X\hat{\beta}\right\}(\text{log}n + 2\gamma\text{log}p), \gamma \geq 0$$

where d is the number of predictors. The DF for group LASSO Yuan and Lin [2006] is defined as:

$$\tilde{\text{df}}\left\{\hat{\mu}(\lambda) \equiv X\hat{\beta}\right\} = \sum_j I(||\hat{\beta}_j|| > 0) + \sum_j \frac{||\hat{\beta}_j||}{||\hat{\beta}_j^{\text{LS}}||}(p_j - 1) \tag{4}$$

where p_j is the number of predictors in the jth group and $\hat{\beta}_j^{\text{LS}}$ is the least-square estimates.

Note that when $p_j = 1$ for all $j = 1, \ldots, p$, group LASSO becomes LASSO and its DF is the number of non-zero parameters selected. Therefore, one can take the LASSO as a special case of group LASSO and so does the DF in expression(4).

3.4 Significance level for selected SNPs

With penalization methods, the "importance" of a covariate usually is determined by whether its regression coefficient is nonzero. In some studies, p-value may also be of interest as a significance measure. Computing p-value with penalization methods is challenging. Wu et al. Wu et al. [2009] proposed a leave-one-out approach for computing p-value by assessing the correlations among selected SNPs in the reduced model. Wu et al. Wu et al. [2009] also commented that this approach may be invalid because it neglects the complex selection procedure for defining the reduced model in the first place.

Here, we use a multi-split method modified from the method proposed by Meinshausen et al. Meinshausen et al. [2009] to obtain p-values. With linear regression, we use F-test for each group to evaluate whether there are elements in this group with significant effects. This procedure will put us in a position to produce p-values at the group level. It is simulation-based and can adjust for multiple comparisons. Multi-split method proceeds as follows:

1. Randomly split data into two disjoint sets of equal size: D_{in} and D_{out}.
2. Fit data in D_{in} with the proposed method. Denote the set of selected groups by S.
3. Compute $\tilde{P}_j$, p-value for group j, as follows:
   (a) If group j is in set S, set $\tilde{P}_j$ equal to the p-value from the F-test in the regular linear regression where group j is the only group.
   (b) If group j is not in set S, set $\tilde{P}_j = 1$.
4. Define the adjusted p-value as $P_j = \min\{\tilde{P}_j|S|, 1\}, j = 1, \ldots, J$, where $|S|$ is the size of set S.

This procedure is repeated B times for each group. Let $P_j^{(b)}$ denote the adjusted p-value for group j in the bth iteration. For $\pi \in (0,1)$, let q_π be the π-quartile of $\{P_j^{(b)}/\pi; b = 1,\ldots,B\}$. Define $\tilde{Q}_j(\pi) = \min\{1, q_\pi\}$. Meinshausen et al. Meinshausen et al. [2009] shows that $\tilde{Q}_j(\pi)$ is an asymptotically correct p-value, adjusted for multiplicity. They also propose an adaptive version that selects a suitable value of quartile based on data:

$$Q_j = \min\left\{1, (1-\log\pi_0)\inf_{\pi\in(\pi_0,1)}\tilde{Q}_j(\pi)\right\},$$

where π_0 is chosen to be 0.05. It is shown that $Q_j, j = 1,\ldots,J$, can be used for both FWER (family-wise error rate) and FDR (false discovery rate) control Meinshausen et al. [2009].

4 Simulation Studies

We consider six different simulation scenarios, each with 500 subjects and 5,000 or 10,000 SNPs. We simulate two distinct response variables for each subject. The correlation between the two responses are set to be 0.1, 0.5 and 0.9. For each response variable, there are 12 SNPs with nonzero effects. We further consider the scenario where the 12 SNPs can be clustered into 3 clusters. The correlation among each cluster is 0.2. The correlation among SNPs not associated with response is set to be 0.2 but independent of those associated clusters. The genotypes are first generated from multivariate normal distribution and then categorized into 0,1 or 2. To mimic a SNP with equal allele frequency, we categorize genotype in a way similar to Wu et al. [2009]. The genotype is set to be 0, 1 or 2 depending on whether $x_{ij} < -c$, $-c \le x_{ij} \le c$ or $x_{ij} > c$, where c is the $3rd$-quartile of x. For the first response variable, the regression coefficient is

$$(\underbrace{0,\ldots,0}_{24},2,2,2,2,\underbrace{0,\ldots,0}_{12},1,-1,1,-1,\underbrace{0,\ldots,0}_{12},2,-1,2,-1,\underbrace{0,\ldots,0}_{p-60}).$$

For the second response variable, the regression coefficient is

$$(\underbrace{0,\ldots,0}_{24},0.5,0.5,0.5,0.5,\underbrace{0,\ldots,0}_{12},-2,2,-2,1,\underbrace{0,\ldots,0}_{12},1,-2,1,-2,\underbrace{0,\ldots,0}_{p-60}).$$

Both response variables depend on the same genotypic data and are correlated through the residuals. Clustering structure exists in this simulation.

To better gauge performance of the proposed approach, we also consider the following alternative approach. We first analyze each response variable separately using Lasso, and then combine the results by examining the overlapped SNPs. For both approaches, we apply the EBIC method described in Section 3.3 to select the tuning parameter λ. We evaluate the number of SNPs identified, the number of true positives, false discovery rate (FDR) and false negative rate (FNR).

The results based on 100 Monte Carlo replicates are summarized in Table 1. Note that the true trait-related SNPs are 25—28, 41—44 and 57—60 on both phenotypes. Totally, there are 12 SNPs associated with 2-level phenotypes. Simulation study suggests that the proposed method is more effective in selecting true positives. Comparing to the proposed method, analyzing each response separately has fewer true positives. A tradeoff is that the proposed method may also identify a few more false positives. However, it is still at a comparably acceptable range. Hence the proposed method outperforms the individual Lasso selection. p-values evaluated by the multi-split method for the selected groups are presented in Table 2. It can be shown that most true positives have significant p-values and all false positives have insignificant p-values.

[Table 1 about here.]

[Table 2 about here.]

5 Application to Ocular Hypertension Treatment Study

The OHTS data is described in Section 1. We refer to the original publication for more detailed descriptions. The data we analyze contains 1,054 subjects, two highly correlated response variables on each subject – CCT OD and CCT OS, and 11217 SNPs on chromosome 22. We analyze the data using three different approaches: the traditional one-SNP-at-a-time approach, analysis of individual response using LASSO, and the proposed approach. In Figure 1 the upper two panels, we show the absolute values of β estimates from the single-SNP analysis. When analyzing each response separately using LASSO, we use the method described in section 3.3 to select the tuning parameter λ. We find that the selected model model contains no SNP (hence results now plotted). The proposed approach identifies 4 SNPs. Information on the identified SNPs, corresponding estimates and p-values are shown in Table 3. One can see that 2 of them are highly significant and other 2 SNPs are moderately significant. With our limited knowledge on susceptibility SNPs for glaucoma, we are not able to objectively evaluate the biological implications of identified SNPs. As an alternative, we consider the following evaluation of prediction performance. (a) Randomly split the sample into given parts with equal sizes; (b) Analyze four parts using the proposed approach; (c) Use the obtained model and make prediction for subjects in the left-out part; (d) Repeat Steps (b) and (c) over all five parts. The proposed approach is a penalization approach, which shrinks estimates towards zero. Thus, the prediction sum of squared errors may not be the appropriate measure. Instead, we evaluate the correlation between predicted and observed values. The mean prediction correlation coefficient is 0.085 , which suggests that the proposed approach is able to provide a reasonable prediction using identified SNPs.

We note that the identified SNPs are only capable of explaining a small percentage of variation of CCT. Such a result is in fact not surprising. CCT is a complex trait and is

potentially related to multiple clinical risk factors, environmental exposures, and genetic risk factors. The predictors we analyze all come from chromosome 22, which is expected to cover only a small percentage of risk factors. As one can see from the results that the null models are identified by Lasso, the signals are weak in this OHTS dataset. Despite that the result of the proposed method is not dramatic, it does advance from individual Lasso analysis.

[Table 3 about here.]

[Figure 1 about here.]

6 Conclusions

In the study of complex diseases, it is not uncommon that a single trait cannot provide a comprehensive description of disease, and hence multiple traits need to be measured. In this article, we study the scenario where multiple traits share the same set of susceptibility SNPs. A penalization approach is proposed for marker selection. The proposed approach can accommodate the joint effects of multiple SNPs and is hence more informative than single-SNP based analysis. Compared with existing approaches that analyze different traits separately, it can more effectively accommodate the correlation among traits and hence be more effective in marker selection. Numerical studies, including simulation and analysis of OHTS data, show satisfactory performance of the proposed approach.

The OHTS data we analyze have two continuous response variables with marginally normal distributions. With other types of response variables, there is a rich literature on joint modeling, which can be adopted to couple with the proposed marker selection. The proposed approach is based on the group Lasso penalty. We expect that other "group-type" penalties, such as group SCAD or group bridge, can be applied. The group Lasso is selected

because of its computational simplicity. Analysis of chromosome 22 data shows that the proposed approach may identify SNPs missed by single-response analysis, and such SNPs may have reasonable prediction performance. The ultimate performance of the proposed approach and identified SNPs needs to be examined together with other risk factors and susceptibility SNPs on other chromosomes.

Acknowledgements

This study has been supported by awards CA120988 and CA142774 from NIH.

References

H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Control*, 19(6):716–723, 1974.

P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Statist.*, 5(1):232–253, 2011.

J. Chen and Z. Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

J. Chen and Z. Chen. Extended bic for small-n-large-p sparse GLM. 2010.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, 96(456):1348–1360, 2001.

I. Frank and J. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

J. Friedman, T. Hastie, and R. Tibshirani. Regularized paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1–22, 2010.

W. Fu. Penalized regressions:the bridge versus the LASSO. *J. Comp. Graph. Statist.*, 7(3): 397–416, 1998.

T. Hastie, R. Tishirani, and J.H. Friedman. *The elements of statistical learning:data mining, inference, and prediction.* Springer-Verlag New York,LLC, second edition, 2009.

J. Huang, F. Wei, and S. Ma. Semiparametric reregression pursuit. *Accepted for publication by Statistica Sinica*, 2011.

M. Kass, D. Heuer, E. Hiqqinbotham, C. Johnson, J. Keltner, J. Miller, R. Parrish, M. Wilson, and M. Gordon. The ocular hypertension treatment study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Arch Ophthalmol.*, 120(6):701–713, 2002.

N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *J. Am. Stat. Assoc.*, 104(488):1671–1681, 2009.

NIH. Eye drops delay onset of glaucoma in people at higher risk, June 2002. URL `http://www.nei.nih.gov/news/pressreleases/061302.asp`.

W. Ramdas, L. van Koolwijk, M. Ikram, N. Jansonius, P. de Jong, A. Bergen, A. Isaacs, N. Amin, Y. Aulchenko, R. Wolfs, A. Hofman, F. Rivadeneira, B. Oostra, A. Uitterlinden, P. Hysi, C. Hammond, H. Lemij, J. Vingerling, C. Klaver, and C. van Duijin. A genome-wide association study of optic disc parameters. *Plos Genetics*, 6(6), 2010.

G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

J.P. Stevens. *Applied multivariate statistics for the social sciences.* Mahwah, NJ: Lawrence Erblaum, 2002.

R. Tibshirani. Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B*, 58(1):267–288, 1996.

G. Wahba. A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothingl. *Ann. Statist.*, 13:1378–1402, 1985.

L. Wang, G. Chen, and H. Li. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494, 2007.

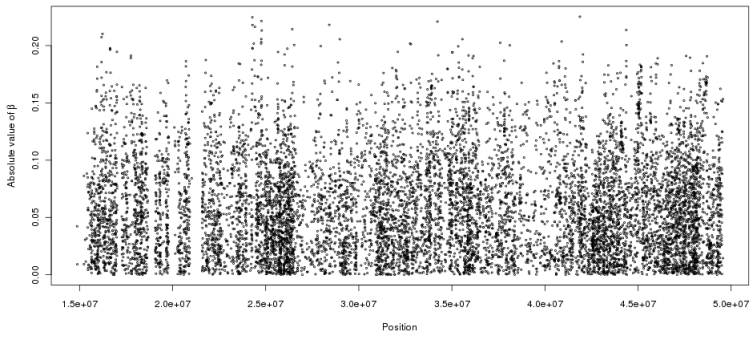
T. Wu and K. Lange. Coordinate descent procedures for LASSO penalized regression. *Ann. Appl. Statist.*, 2(1):224–244, 2007.

T. Wu, Y. Chen, T. Hastie, E. Sobel, and K. Lange. Genomewide association analysis by LASSO penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

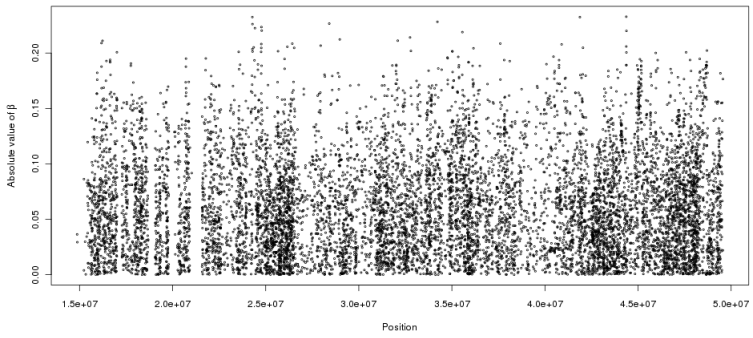
M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68:49–67, 2006.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 2010.

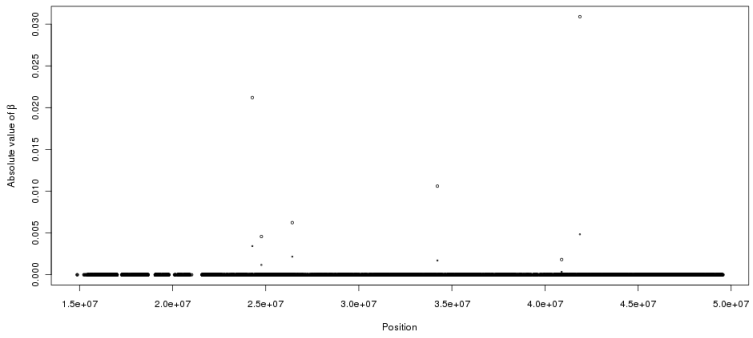
H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B*, 67(2):301–320, 2005.



(a) Phenotype 1



(b) Phenotype 2



(c) The proposed method

Figure 1: Absolute value of $|\beta|$ estimates from the simple linear regression on both levels and the proposed method.

Table 1: Simulation studies on 6 different settings. The numbers are mean (standard deviation) based on 100 replicates.

		Combined Individual			
p	ρ	True Positive	Model Size	FDR	FNR
5000	0.1	12.32(1.73)	14.34(2.85)	0.127(0.099)	0.487(0.072)
5000	0.5	11.83(2.21)	13.70(3.30)	0.120(0.099)	0.507(0.092)
5000	0.9	11.96(2.34)	13.52(3.13)	0.103(0.089)	0.502(0.097)
10000	0.1	11.18(2.28)	12.89(3.28)	0.115(0.107)	0.534(0.095)
10000	0.5	11.19(2.20)	12.59(3.02)	0.097(0.091)	0.534(0.092)
10000	0.9	10.90(2.45)	12.55(3.51)	0.113(0.100)	0.546(0.102)
		Proposed Approach			
p	ρ	True Positive	Model Size	FDR	FNR
5000	0.1	19.26(4.09)	23.18(6.89)	0.141(0.122)	0.198(0.170)
5000	0.5	21.18(3.24)	24.96(6.64)	0.122(0.123)	0.118(0.135)
5000	0.9	22.98(1.26)	31.32(6.56)	0.240(0.134)	0.043(0.052)
10000	0.1	17.47(4.17)	20.28(6.76)	0.107(0.111)	0.279(0.187)
10000	0.5	19.21(4.91)	21.47(6.81)	0.082(0.100)	0.199(0.204)
10000	0.9	22.70(1.28)	31.78(7.98)	0.249(0.156)	0.054(0.053)

Table 2: Multi-split p-values for simulated data.

	p=5000		p=10000	
SNP index	$\|\|\hat{\beta}\|\|$	p-value	$\|\|\hat{\beta}\|\|$	p-value
25	0.185	1.46e-07	0.185	1.39e-05
26	0.138	2.67e-05	0.175	7.54e-05
27	0.156	3.68e-07	0.176	6.07e-04
28	0.161	1.45e-06	0.213	4.23e-07
38	0.013	1.000		
41	0.371	6.45e-04	0.564	3.23e-04
42	0.472	0.126	0.569	2.02e-04
43	0.295	0.028	0.429	4.15e-04
44	0.435	0.012	0.826	1.88e-10
57	0.075	0.217	0.073	1.000
59	0.117	3.26e-03	0.147	0.784
60	0.100	0.204	0.042	1.000
2081	0.004	1.000		
2200	0.011	1.000		
4964			0.009	1.000
8482			0.008	1.000

* Empty cells stand for SNPs that are not identified from the model

Table 3: Multi-split p-values for simulated data.

SNP Name	Position	Band	$\|\|\hat{\beta}\|\|$	p-value
rs5752182	24281739	22q11.23	0.0168	6.36e-04
rs5762319	26426963	22q12.1	0.0060	0.022
rs5995107	34222746	22q12.3	0.0094	0.026
rs5996279	41876677	22q13.2	0.0281	5.34e-05