Incorporating Group Correlations in Genome-Wide Association Studies Using Smoothed Group Lasso

Jin Liu[1 *], Jian Huang[2,3], Shuangge Ma[1], and Kai Wang[3]

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

October 13, 2011

The University of Iowa

Department of Statistics and Actuarial Science

Technical Report No. 411

*To whom correspondence should be addressed. jin.liu.23292@yale.edu

Incorporating Group Correlations in Genome-Wide Association studies Using Smoothed Group Lasso

Jin Liu[1], Jian Huang[2,3], Shuange Ma[1], and Kai Wang[2]

[1]Division of Biostatistics, School of Public Health, Yale University

[2]Department of Biostatistics, University of Iowa

[2,3]Department of Statistics & Actuarial Science,University of Iowa

Abstract

In genome-wide association studies, penalization is becoming an important approach for identifying genetic markers associated with disease. Motivated by the fact that there exists natural grouping structure in SNPs and more importantly such groups are correlated, we propose a new penalization method for group variable selection which can properly accommodate the correlation between adjacent groups. This method is based on a combination of the group Lasso penalty and a quadratic penalty on difference of regression coefficients of adjacent groups. The new method is referred to as Smoothed Group Lasso, or SGL. It encourages group sparsity and smoothes regression coefficients for adjacent groups. Canonical correlations are applied to the weights between groups in the quadratic difference penalty. We derive a group coordinate descent algorithm for computing the solution path. This algorithm takes the solution of a closed form of SGL for a single group model and is efficient and stable in high-dimensional settings. The SGL method is further extended to logistic regression for binary response. With the assistance of MM algorithm, the logistic regression model with SGL penalty turns out to be an iteratively penalized least-square problem. Principal components are used to reduce dimensionality locally within groups. Simulation studies are used to evaluate the finite sample performance. Comparison with group Lasso shows that SGL is more effective in selecting true groups. We also analyze a rheumatoid arthritis data by applying the SGL method under logistic regression model.

Keywords: Group selection; Linkage Disequilibrium; Smoothing; Regularization; SNP.

1 Introduction

In genome-wide association studies (GWAS), hundreds of thousands of single nucleotide polymorphisms (SNPs) are genotyped using array-based technologies for a large number of

individuals, typically ranging from several hundred to several thousand. Standard GWAS methods are single SNP-based, in the sense that they analyze one SNP at a time. Single-SNP approaches may not be appropriate when we investigate a complex polygenic trait, since they fail to take into account the accumulated and/or joint effects of multiple genetic markers on the trait. In contrast, multivariate analysis, which describes the joint effects of multiple SNPs in a single statistical model, may be more appropriate.

In GWAS, it is expected that only a subset of SNPs are associated with the response variables. Thus to analyze SNP data in a multivariate model, variable selection is needed along with regularized estimation. Penalization methods have been adopted for such a purpose. SNPs are naturally ordered along the genome with respect to their physical positions. They can be highly correlated due to tight linkage and linkage disequilibrium. Therefore, it is sensible to group SNPs based on their physical locations and correlation patterns among them. Commonly adopted penalization approaches, such as Lasso, Bridge, SCAD and MCP [Tibshirani, 1996, Frank and Friedman, 1993, Fan and Li, 2001, Zou and Hastie, 2005], assume interchangeable effects and cannot effectively accommodate grouping structure. The "group versions" of Lasso, elastic net, SCAD and MCP have been developed to analyze data that have the grouping structure [Bakin, 1999, Yuan and Lin, 2006, Wang et al., 2007, Friedman et al., 2010b, Huang et al., 2011b].

In addition to the grouping structure in SNP data, there is also possible strong correlation among adjacent groups. For the dataset described in Section 5, we find that even after grouping SNPs based on their physical locations and correlations, there still exist strong correlations among groups (see Fig. 1). In a recent study [Huang et al., 2011a], it is shown that a sparse Laplacian shrinkage estimator, which "smoothes" over regression coefficients for highly correlated covariates, may have superior estimation and variable selection properties. This approach can effectively accommodate the correlation among covariates but not the

grouping structure.

[Figure 1 about here.]

In this article, we develop a novel penalization method for estimation and variable selection with GWAS data. Our goal is to identify markers associated with response variables, while properly accommodating *the unique characteristics of high-dimensionality, grouping structure and correlation between groups* of GWAS data. The proposed approach is referred to as smoothed group Lasso, or SGL. Its penalty is the sum of the group Lasso penalty and the quadratic penalty on difference of regression coefficients of adjacent groups. The group Lasso penalty promotes sparsity and can select groups of SNPs associated with responses. The second penalty term, the quadratic difference penalty, takes into account the natural ordering of groups and adaptively accommodates the correlation between adjacent groups. Here, the correlations between groups are measured with canonical correlations. We derive a group coordinate descent algorithm for computing the SGL estimator. It is efficient and stable even in high-dimensional settings. Beyond developing the new penalty, we also investigate several related practical problems. The first is an extension of the proposed approach to incorporate negative log-likelihood as a loss function for case-control studies. In practical data analysis, high correlations within groups lead to high colinearity among variables, which can have adverse effects on selection and estimation results. We propose applying principal components analysis (PCA) within each group to locally reduce dimensionality and colinearity. In addition, a modified multi-split method is used to evaluate the statistical significance of selected groups.

In Section 2, we introduce the SGL penalty and develop a group coordinate descent (GCD) algorithm for quadratic loss functions. Tuning parameter selection is also discussed. In Section 3, we investigate several related practical issues, including accommodating case-control data, reducing dimensionality within groups using PCA and evaluating significance

level. Simulation studies are conducted in Section 4. We analyze a case-control study on rheumatoid arthritis in Section 5. The article concludes with discussion in Section 6.

2 Smoothed Group LASSO

2.1 Data and Model Setting

Here we first consider quadratic loss functions, which naturally arise from linear regression with continuous responses. Extension to binary trait with logistic regression model is discussed in Section 3.

Suppose that the data consists of n subjects. Let y_i be the continuous response variable for subject i. The genotype at a SNP is scored as 0, 1, or 2 depending on the number of copies of a reference allele in a subject. The SNPs are divided into J groups, each with size $d_j, j = 1, \ldots, J$, according to their physical locations and correlation patterns. Our approach for grouping SNPs is discussed in more detail in Section 5 below. Let x_{ij} be the $d_j \times 1$ covariates vector corresponding to the jth group of the SNPs for the ith subject. Denote β_j as the $d_j \times 1$ vector of regression coefficients for x_{ij}. It measures the effect size for predictors in the jth group. Let $\boldsymbol{\beta} = (\beta_1', \ldots, \beta_J')'$. Assume the linear regression model $y_i = \beta_0 + \sum_{j=1}^{J} x_{ij}'\beta_j + \epsilon_i$, where β_0 is the intercept and ϵ_i is the random error. With centered response variables and standardized covariates, we can assume $\beta_0 = 0$. We consider the quadratic loss function

$$\ell(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{J} x_{ij}'\beta_j)^2 = \frac{1}{2n}||\, Y - \sum_{j=1}^{J} X_j\beta_j||^2,$$

where $Y = (y_1, \ldots, y_n)'$, X_j is an $n \times d_j$ matrix corresponding to the jth group, and $||\cdot||$ denotes the l_2 norm.

2.2 Penalized estimation

As discussed in Section 1, the goals of the SGL approach are two-fold. The first is to select groups of SNPs associated with response variables. The second is to smooth regression coefficients between adjacent groups with strong correlations.

To achieve the first goal, we use the group Lasso penalty [Bakin, 1999, Yuan and Lin, 2006, Meier et al., 2008], which is defined as

$$\rho(||\beta_j||_{\Sigma_j}; \sqrt{d_j}\lambda_1) = \lambda_1\sqrt{d_j}||\beta_j||_{\Sigma_j}, \tag{1}$$

where $||\beta_j||_{\Sigma_j} = (\beta_j'\Sigma_j\beta_j)^{1/2}$, $\Sigma_j = X_j'X_j/n$ is the empirical covariance matrix for the jth group, and $\lambda_1 > 0$ is a data-dependent tuning parameter. In expression (1), the rescaling factor $\sqrt{d_j}$ makes the penalty level proportional to group size. It ensures that small groups are not overwhelmed by large groups in selection. The group Lasso penalty has been investigated in multiple studies [Bakin, 1999, Yuan and Lin, 2006, Huang et al., 2009]. A blockwise standardization method has been proposed by Kim et al. [2006]. Meier et al. [2008] develop a block coordinate descent algorithm.

To achieve the second goal, we propose a new penalty that can adaptively incorporate possible correlations between adjacent groups. Specifically, consider

$$\frac{\lambda_2}{2}\sum_{j=1}^{J-1}\zeta_j d\Big(\frac{||\beta_j||_{\Sigma_j}}{\sqrt{d_j}} - \frac{||\beta_{j+1}||_{\Sigma_{j+1}}}{\sqrt{d_{j+1}}}\Big)^2,$$

where λ_2 is a data-dependent tuning parameter, the weight ζ_j is a measure of correlation between the jth and $(j+1)$th groups and $d = \max\{d_j : j = 1, \ldots, J\}$ is the largest group size. Here d is used to scale the squared difference of the two norms so that λ_2 can be on the same scale as λ_1. In this study, we set ζ_j as the canonical correlation between two groups. More details on this measure are provided in Appendix. This penalty has been motivated by the following considerations. When $\zeta_j = 0$, the two groups are unrelated, and there should

be no relationship between the regression coefficients. Hence the penalty reduces to zero. When ζ_j gets larger, the two groups are more highly correlated and hence the corresponding regression coefficients should be "more similar". A penalty on the difference of norm may shrink the difference. Note that we only penalize the difference between adjacent groups. Such groups are physically next to each other and hence are more likely to have similar regression coefficients if they are highly correlated. In addition, our empirical investigation shows that groups far away from each other tend to have $\zeta \sim 0$. Introducing a large number of penalties with $\zeta \sim 0$ may increase computational cost and reduce stability. Even though it is possible to extend the proposed penalty and consider all possible pairs of groups, we choose to focus on the adjacent pairs because of the above considerations.

In summary, the proposed SGL penalty function is

$$P_{\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \sum_{j=1}^{J} \lambda_1 \sqrt{d_j} ||\beta_j||_{\Sigma_j} + \frac{\lambda_2}{2} \sum_{j=1}^{J-1} \zeta_j d\Big(\frac{||\beta_j||_{\Sigma_j}}{\sqrt{d_j}} - \frac{||\beta_{j+1}||_{\Sigma_{j+1}}}{\sqrt{d_{j+1}}} \Big)^2.$$

Given a loss function $\ell(\beta_0, \boldsymbol{\beta})$, the SGL estimate $\hat{\boldsymbol{\beta}}$ is defined as the minimizer of

$$L_n(\boldsymbol{\beta}, \lambda_1, \lambda_2) = \ell(\boldsymbol{\beta}) + P_{\lambda_1,\lambda_2}(\boldsymbol{\beta}).$$

2.3 Group coordinate descent algorithm

The group coordinate descent(GCD) algorithm is originally proposed for group Lasso [Yuan and Lin, 2006] and has also been used for computing the group MCP solutions [Huang et al., 2011b]. It is a natural extension of the coordinate descent algorithm [Fu, 1998, Wu and Lange, 2007, Friedman et al., 2010a]. The GCD algorithm 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 which has a simple closed-form solution for a single group but lacks one with multiple groups.

First for each group, we orthogonalize the design matrix so that the empirical covariance matrix is equal to identity. We can write $\Sigma_j = R_j' R_j$ for a $d_j \times d_j$ upper triangular matrix R_j with positive diagonal entries via the Cholesky decomposition. Let $\tilde{X}_j = X_j R_j^{-1}$ and $b_j = R_j \beta_j$. With the transformation, the objective function is

$$L_n(\boldsymbol{b}, \lambda_1, \lambda_2) = \frac{1}{2n} || Y - \sum_{j=1}^{J} \tilde{X}_j b_j ||^2 + \sum_{j=1}^{J} \lambda_1 \sqrt{d_j} ||b_j|| + \frac{\lambda_2}{2} \sum_{j=1}^{J-1} \zeta_j d(\frac{||b_j||}{\sqrt{d_j}} - \frac{||b_{j+1}||}{\sqrt{d_{j+1}}})^2,$$

where $\boldsymbol{b} = (b_1', \ldots, b_J')'$. Note that $n^{-1} \tilde{X}_j \tilde{X}_j = R_j^{-1\prime} (n^{-1} X_j' X_j) R_j^{-1}$. Thus using the $|| \cdot ||_{\Sigma_j}$ norm amounts to standardizing the design matrices. Therefore, without loss of generality, we assume that X_j's are orthonormalized with $n^{-1} X_j' X_j = I_{d_j}$.

Given the group parameter vectors β_k $(k \neq j)$ fixed at their current estimates $\tilde{\beta}_k^{(s)}$, we seek to minimize the objective function $L_n(\boldsymbol{\beta}, \lambda_1, \lambda_2)$ with respect to the jth group parameter β_j, Here only the terms involving β_j in $L_n(\boldsymbol{\beta}, \lambda_1, \lambda_2)$ matter. Some algebra shows that this problem is equivalent to minimizing $R(\beta_j)$ defined as

$$R(\beta_j) = C(\tilde{\boldsymbol{\beta}}) + \frac{1}{2} a_j \beta_j' \beta_j - b_j' \beta_j + c_j ||\beta_j||, \quad j = 1, \ldots, J, \tag{2}$$

where $a_j = 1 + \frac{\lambda_2 d}{d_j} (\zeta_{j-1} + \zeta_j)$, $b_j = n^{-1} X_j' r + \tilde{\beta}_j^{(s)}$, $c_j = \lambda_1 \sqrt{d_j} - \frac{\lambda_2 d}{\sqrt{d_j}} (\zeta_{j-1} \frac{||\tilde{\beta}_{j-1}||}{\sqrt{d_{j-1}}} + \zeta_j \frac{||\tilde{\beta}_{j+1}||}{\sqrt{d_{j+1}}})$, and $C(\tilde{\boldsymbol{\beta}})$ is a constant free of $\boldsymbol{\beta}$.

It can be shown that the minimizer of $R(\beta_j)$ in expression (2) is

$$\tilde{\beta}_j = \frac{1}{a_j} \left(1 - \frac{c_j}{||b_j||} \right)_+ b_j. \tag{3}$$

This explicit solution greatly facilitates the implementation of the GCD algorithm described below.

Let $\tilde{\boldsymbol{\beta}}^{(0)} = (\tilde{\beta}_1^{(0)}, \ldots, \tilde{\beta}_J^{(0)})'$ be the initial value. A convenient choice for the initial value is zero (component wise). With fixed λ_1 and λ_2, the GCD algorithm proceeds as follows:

1. Set $s = 0$. Initialize the vector of residuals $r = Y - \sum_{j=1}^{J} X_j \tilde{\beta}_j^{(0)}$.

2. For $j = 1, \ldots, J$,

   (a) Calculate a_j, b_j and c_j in expression (2);

   (b) Update $\tilde{\beta}_j^{(s+1)} = \frac{1}{a_j}\left(1 - \frac{c_j}{||b_j||}\right)_+ b_j$ using expression (3);

   (c) Update $r \leftarrow r - X_j(\tilde{\beta}_j^{(s+1)} - \tilde{\beta}_j^{(s)})$ and $j \leftarrow j + 1$;

3. Update $s \leftarrow s + 1$;

4. Repeat Steps 2 and 3 until convergence.

Of note, in the above algorithm, we take the convention that $\zeta_0 = \zeta_J = 0$.

In step 2b, the SGL takes a form similar to group soft-thresholding operator for the Lasso estimates. The biggest difference lies in the support of c_j. With group Lasso, $c_j = \lambda_1\sqrt{d_j}$ and is always positive. However, with SGL, c_j as defined in step 2a can be negative or positive depending on the choice of λ_2 and the weight ζ. Under the simplified scenario with only one group, the group Lasso and SGL estimates are the same. With multiple groups, consider for example the jth group. If its adjacent groups are selected with nonzero regression coefficients, then c_{j-1} and c_{j+1} from the adjacent groups get smaller. Group j is then more likely to be selected, which is intuitively reasonable. Furthermore, in step 2a, λ_2 is reweighted with additional $1/\sqrt{d_j}$, which accounts for the size of the jth group. This implies that groups with smaller sizes are affected more by the same change of the adjacent groups. The solution path for a simulated dataset is provided in Fig. 2 for group Lasso and SGL with $\eta = 0.1$, $\eta = 0.2$ and $\eta = 0.5$, respectively, where $\eta = \lambda_1 + \lambda_2$ is the sum of the penalty parameters.

[Figure 2 about here.]

The convergence of this algorithm follows from Theorem 4.1(c) of Tseng [2001]. This can be shown as follows. The objective function of SGL can be written as $f(\boldsymbol{\beta}) = f_0(\boldsymbol{\beta}) + \sum_{j=1}^{J} f_j(\beta_j)$ where

$$f_0(\boldsymbol{\beta}) = \frac{1}{2n}||Y - \sum_{j=1}^{J} X_j\beta_j||^2 + \frac{\lambda_2}{2}\sum_{j=1}^{J-1}\zeta_j d\Big(\frac{||\beta_j||}{\sqrt{d_j}} - \frac{||\beta_{j+1}||}{\sqrt{d_{j+1}}}\Big)^2,$$

and $f_j(\beta_j) = \lambda_1\sqrt{d_j}||\beta_j||$. Since f is regular in the sense of (5) in Tseng (2001) and $\sum_{j=1}^{J} f_j(\beta_j)$ is separable (group-wise), the GCD solutions converge to a coordinatewise minimum point of f, which is also a stationary point of f.

2.4 Selection of tuning parameters

There are various methods that can be applied, which include AIC, BIC, cross-validation and generalized cross-validation. However, they are all based upon prediction error. In GWAS, disease markers may not be in the set of SNP markers. Practically it is rare that disease markers are a part of SNP data, which consequently results in non-true model for SNP data. Hence, the methods mentioned above may be inadequate in GWAS.

We adopt the approach proposed in Wu et al. [2009], which sets a predetermined number of selected SNPs based on the unique nature of GWAS. We implement a combination of bracketing and bisection to search for the tuning parameter that yields a predetermined number of selected SNPs. For this purpose, tuning parameters λ_1 and λ_2 are reparameterized as τ=λ_1+λ_2 and η=λ_1/τ. The value of η is fixed beforehand. We use a bisection approach to find the τ value such that $r(\tau)$, the number of selected markers, is equal to s. Let τ_{max} be the smallest value for which all estimated coefficients are 0. From the update steps 2a and 2b, $\tau_{max} = \max_j ||n^{-1}X_j'\mathbf{Y}||/(\eta\sqrt{d_j})$. We select ϵ (usually =0.01 in numerical studies) and let τ_{min}=$\epsilon\tau_{max}$. Initially, we set $\tau_l = \tau_{min}$ and $\tau_u = \tau_{max}$. If $r(\tau_u) < s < r(\tau_l)$, then we employ bisection. This involves testing the midpoint $\tau_m = \frac{1}{2}(\tau_l + \tau_u)$. There are three

possibilities. If $r(\tau_m) < s$, replace τ_u by τ_m. If $r(\tau_m) > s$, replace τ_l by τ_m. If $r(\tau_m) = s$, the calculation is terminated. In either of the first two cases, we bisect again and continue the loop until $r(\tau_m) = s$.

3 Practical Considerations

3.1 Accommodating case-control data with logistic regression

Consider a case-control study with n subjects. For the ith subject, let $y_i \in \{0, 1\}$ denote the response variable and $x_i = (x_{i1}', \ldots, x_{iJ}')'$. The logistic regression model assumes that $p(x_i) = \Pr(y_i = 1|x_i) = 1/(1 + \exp(-(\beta_0 + \sum_{j=1}^{J} x_{ij}'\beta_j)))$. The SGL estimate is defined as the minimizer of the penalized negative log-likelihood, that is,

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) = \underset{(\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left[\ell(\beta_0, \boldsymbol{\beta}) + P_{\lambda_1, \lambda_2}(\boldsymbol{\beta}) \right]. \tag{4}$$

The negative log-likelihood function in expression (4) is

$$\ell(\beta_0, \boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \cdot (\beta_0 + \sum_{j=1}^{J} x_{ij}'\beta_j) - \log(1 + e^{(\beta_0 + \sum_{j=1}^{J} x_{ij}'\beta_j)}) \right]. \tag{5}$$

When implementing the GCD algorithm, there is no simple, closed-form solution for penalized estimation with a single group. To tackle this problem, we propose using an MM approach [Ortega and Rheinboldt, 2000]. Note that negative log-likelihood (5) is a convex function. With the MM approach, we majorize the negative log-likelihood by a quadratic loss given by

$$\ell_Q(\beta_0, \boldsymbol{\beta}|\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}) = \frac{1}{8n} \sum_{i=1}^{n} (z_i - \beta_0 - x_i'\boldsymbol{\beta})^2 + C(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}),$$

where $z_i = \tilde{\beta}_0 + x_i^T\tilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)(1-\tilde{p}(x_i))}$ and $\tilde{p}(x_i) = \frac{1}{1+e^{-(\tilde{\beta}_0 + x_i^T\tilde{\boldsymbol{\beta}})}}$ are evaluated at the current estimate $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}})$, and the last term is free of $(\beta_0, \boldsymbol{\beta})$.

With fixed (λ_1, λ_2), our computational algorithm consists of a sequence of nested loops:

Outer loop: Update the majorized quadratic function l_Q using the current estimates $(\tilde{\beta}_0, \tilde{\boldsymbol{\beta}})$.
Inner loop: Run the GCD algorithm developed for the penalized least-squares problem (Section 2.3) and solve for

$$argmin_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{p+1}} \left\{ \ell_Q(\beta_0, \boldsymbol{\beta}|\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}) + P_{\lambda_1,\lambda_2}(\boldsymbol{\beta}) \right\}.$$

We note that in the penalized group least-squares problem (2.3), we do not estimate β_0. In logistic regression with the SGL penalty, we can estimate it after estimating all other β_js for each majorized function as $\hat{\beta}_0 = \sum_i^n (z_i - x_i)'\hat{\boldsymbol{\beta}}/n$. In addition, τ_{max} is not explicitly defined as in linear models. We evaluate the quadratic approximation for the negative log-likelihood at all coefficients $\beta_j, j = 1, \ldots, J$, equal to zero. Then τ_{max} can be calculated in a similar way.

3.2 Reducing within-group colinearity and dimensionality

Because SNPs are densely located in many regions, there may exist high correlations within a group of SNPs due to high linkage disequilibrium. This may cause instability problem in Cholesky decomposition when some eigenvalues of the correlation matrices are too small. In our group selection, we are more interested in the group effects as opposed to specific covariates within groups. To reduce the dimensionality within groups and to tackle the colinearity problem, in data analysis when there is evidence for a lack of stability, we propose to first conduct principal component analysis (PCA) within groups. Specifically, we conduct PCA for each group. We choose the number of PCs such that 90% of the total variation is explained. Then PCs, as opposed to the original covariates, are used for downstream analysis. Our empirical study suggests that this simple step may effectively guarantee that the smallest eigenvalues of the covariance matrices are not too small and that the Cholesky decomposition is stable.

3.3 Significance level for selected SNPs

With penalization methods, the "importance" of a covariate usually is determined by whether its regression coefficient is nonzero. In GWAS, the p-value is also of interest as a significance measure. Computing p-value with penalization methods is challenging. Wu et al. [2009] proposed a leave-one-out approach for the computing p-values of the selected SNPs in the reduced model. 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 that is a modification of the method proposed by Meinshausen et al. [2009] to obtain the p-values. With linear regression, we use the F-test for each group to evaluate whether there are elements in this group with significant effects. With logistic regression, we use the likelihood ratio statistic. This procedure will put us in a position to produce p-values at the group level. It is simulation-based and automatically 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}. In case-control studies, we split the samples in a way that maintains the case-control ratio.

2. Fit data in D_{in} with the SGL 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. In case-control studies, the likelihood ratio test is evaluated at this step.

   (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. [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 Study

We conduct simulation to better gauge performance of SGL. For comparison, we also consider group Lasso, which has a statistical framework closest to that of SGL. Four simulation examples are considered. The first three are linear models with normal residuals. The fourth one has binary responses. SNPs in the first two models are generated with a two-stage procedure, which has been adopted from Wu et al. [2009]. First, we draw the predictor vector x_i from a p-dimensional multivariate normal distribution. Then, with the assumption that SNPs have equal allele frequencies, the genotype of the ith SNP is set to be 0, 1 or 2 according to whether $x_{ij} < -c, -c < x_{ij} < c$, or $x_{ij} > c$. The cutoff point -c is the first quartile of a standard normal distribution. For the third and fourth examples, the genotype data is excerpted from a real Rheumatoid Arthritis (RA) study (details provided in Section 5). In all examples, we set $n = 400$ and $p = 5000$.

Example 1. In this example, there are 1253 groups for 5000 SNPs. For phenotypic irrelevant groups i and j, $\text{cov}(x_i, x_j) = 0.6^{|i-j|}$ and $\text{cov}(x_i, x_i) = 1$. For phenotypic relevant groups k and l, $\text{cov}(x_k, x_l) = 0.8$ if k and l are in the same group, $\text{cov}(x_k, x_l) = 0.6$ if k and l are

not in the same group, and $\text{cov}(x_k, x_k) = 1$. There are no correlations between irrelevant and relevant groups. The size for all irrelevant groups is four. The non-zero groups have regression coefficients as follows: $\beta_{25} = \beta_{28} = (0.1, 0.1, 0.1, 0.1)'$, $\beta_{26} = 1$, $\beta_{27} = (-1, 1, -1)'$, $\beta_{1002} = -\beta_{1006} = (0.2, 0.2, 0.2, 0.2)'$, $\beta_{1003} = -0.8$, $\beta_{1004} = (-0.8, -0.8)'$ and $\beta_{1005} = -0.8$. The response variable Y is generated from a linear regression model with normal residuals with mean 0 and standard deviation 1.5. In this example, the relevant groups are independent of irrelevant groups. Within all groups, SNPs are highly correlated.

Example 2. In this example, we use the same regression coefficients and grouping structure as with Example 1. A different, particularly auto-regressive correlation structure is adopted. For SNP i and j, $\text{cov}(x_i, x_j) = 0.7^{|i-j|}$.

Example 3. In this example, the genotype data is excerpted from a real data set. To make the LD structure as realistic as possible, genotypes are obtained from the rheumatoid arthritis (RA) study provided by the Genetic Analysis Workshop (GAW) 16. This study involves 2062 individuals, among which 400 are randomly chosen. Five thousand SNPs are selected from chromosome 6. For individual i, its phenotype y_i is generated from a linear regression model. The regression coefficients have elements all equal to zero except that $(\beta'_{705}, \ldots, \beta'_{707}) = (0.1, 0.2, -0.1, 0.2, 1, -0.1, -1, 0.1, -1, 0.1, -0.6, 0.2)$ and $(\beta'_{709}, \ldots, \beta'_{714}) = (0.1, -0.6, 0.2, 0.3, -0.1, 0.3, 0.4, -1.2, 0.1, 0.3, -0.7, 0.1, 1, 0.2, -0.4, 0.1, 0.5, -0.2, 0.1)$. SNPs are grouped if the value of absolute lag-1 autocorrelation is larger than a certain value, which is 0.2 in Examples 3 and 4, and the number of groups is 1432.

Example 4. In this example, the genotype data and the regression coefficients are the same as with Example 3. The linear predictors are generated in the same way as with Example 3. The binary response variables are generated from Bernoulli distributions with probability $\Pr(y_i = 1|x_i) = \frac{1}{1+e^{-(\beta_0 + x_i'\boldsymbol{\beta})}}$.

[Table 1 about here.]

We analyze Examples 1-4 using the SGL and group Lasso. PCA is used to reduce within-group collinearity for the simulated datasets. We are aware of other group selection methods, including for example group bridge and group MCP. We focus on comparison with group Lasso because of its similarity with the SGL, which may help us better understand the effects of the smooth penalty. Summary statistics based on 100 replicates are shown in Table 1. Simulation suggests that the SGL is computationally affordable, with analysis of one replicate taking about 5 minutes on a desktop PC. For each replicate in all four examples, we prefix the number of selected groups equal to 15, and use the method described in Section 2.4 for tuning parameter selection. With a total of 9 true positive groups, selecting 15 groups can ensure that the majority or all of the true positives can be selected. As shown in Table 1, we have experienced with different η values and found that $\eta = 0.3$ is an appropriate choice with linear regression and $\eta = 0.1$ is appropriate with logistic regression. Note that when $\eta = 1$, the SGL penalty becomes the group Lasso.

[Table 2 about here.]

Table 1 suggests that SGL is capable of selecting the majority of true positives. We do observe a few false positives, which is reasonable considering the extremely high dimensionality and noisy nature of data. Under all simulated scenarios, SGL outperforms group Lasso by identifying more true positives and/or less false positives, which supports the necessity of smoothing. We also examine the multi-split approach. For a representative dataset simulated under Examples 3 and 4 (Table 2 and Table 3), respectively, we show the selected group norms and their corresponding p-values. Note that the true trait-related groups are from 705 to 708 and from 710 to 714. We see that SGL models under both linear regression and logistic regression select a more clustered set of groups. Furthermore, the SGL models select more groups that are false negative and some of their p-values are significant. Hence the SGL models outperform the group Lasso in the case of strong correlations among groups.

[Table 3 about here.]

5 Analysis of Rheumatoid Arthritis Data

Rheumatoid arthritis (RA) is a long-term condition that leads to inflammation of the joints and surrounding tissues. It can also affect other organs. RA is a complex human disorder with a prevalence ranging from around 0.8% in Caucasians to 10% in some native American groups [Amos et al., 2009]. Several demographical and environmental risk factors have been suggested, including for example gender and smoking. In addition, there are solid evidences that multiple genetic risk factors contribute to the risk of RA. Genetic risk factors underlying RA have been mapped to the HLA region on region 6p21 [Newton et al., 2004], PTPN22 locus at 1p13 [Begovich et al., 2004],and the CTLA4 locus at 2q33 [Plenge et al., 2005]. There are some other loci with weaker effects reported, including loci at 6q (TNFAIP3), 9p13 (CCL21), 10p15 (PRKCQ) and 20q13 (CD40) [Amos et al., 2009].

The GAW 16 data is from the North American Rheumatoid Arthritis Consortium (NARAC). It was the initial batch of whole genome association data for the NARAC cases (N=868) and controls (N=1194) after removing duplicated and contaminated samples. SNP genotype data were generated using an Illumina 550k platform and available for 868 cases and 1194 controls. After quality control and removing SNPs with low minor allele frequencies, there are 31,670 SNP measurements on chromosome 6.

In Fig. 3 (Appendix), we provide the plot of ζ values. It is easy to see that some correlations are very high. Note that there are more groups having ζs smaller than 0.6, since the SNPs are grouped if the absolute lag-1 autocorrelations are larger than 0.6. The proportion of $\zeta_j > 0.6$ for 100 non-overlapping groups is also plotted (Fig. 3 (Appendix)).

[Figure 3 about here.]

With SNP data, one possible way of grouping SNPs is based on the distance to the closest genes. This can be done with the help of the annotation files. However, overlapping of genes happens frequently with SNP data. Thus, sometimes it can be difficult to identify which group a SNP belongs to. Here we use an alternative way to group SNPs. The lag-1 Pearson correlation coefficients are first calculated for all SNPs. Then we group SNPs using lag-1 correlations: if two adjacent SNPs have absolute correlation larger than 0.6, we put them in the same group. We choose the threshold to be 0.6 as it leads to a reasonable number of groups, neither too large nor to small. Different from simulation studies, we do not know the number of true groups beforehand. We choose the predetermined number of selected groups equal to 100. This choice has been motivated by several considerations. Fig. 3(a) and Fig. 3(b) suggest that 100 groups should be sufficient to catch the important groups. In addition, 100 is large enough so that true positives are likely to be caught; On the other hand, it is not too large so that there should be not many false positives. From simulation studies, we choose $\eta = 0.1$ for data analysis. The value of the optimal tuning parameter for SGL with $\eta = 0.1$ is 1.783, whereas the value of the optimal tuning parameter for group Lasso is 0.1384. As described in Section 3.2, we apply PCA to reduce dimensionality within groups. Therefore, a direct plot of point estimates cannot be produced. We use the group norms to plot against their original positions. The plots for SGL and group Lasso are provided in Fig. 3(a) and Fig. 3(b), respectively. The smaller dots in the plots are estimates with insignificant p-values. The larger dots stand for estimates with significant p-values. For comparison, we also conduct single SNP-based analysis. Results are provided in Fig. 3.

From Fig. 3, we see that single-SNP analysis produces estimates with too much noise while group Lasso and SGL are capable of conducting screening and yielding much sparser estimates. Comparing with group Lasso, group estimates from the SGL method are more clustered in the HLA region that has been found to be associated RA. Moreover, there are

more significant groups (larger dots) identified by the SGL method in the HLA region. It is consistent with the results from simulation. The analysis of RA data shows that the SGL approach is more efficient than the group Lasso.

6 Discussion

Penalization provides an effective way of analyzing the joint effects of multiple SNPs in GWAS. Because of the natural ordering of SNPs on the genome and possible high linkage disequilibrium among tightly linked SNPs, SNP data can be highly correlated and hence have the natural grouping structure. In addition, adjacent groups can still be highly correlated, which may give rise to similar association with the phenotype of interest. Existing penalized marker selection methods do not effectively accommodate all the aforementioned properties of SNP data. In this article, we propose a new penalized marker selection approach. It uses the group Lasso penalty for group marker selection and, more importantly, a new penalty to smooth the regression coefficients between adjacent groups. The proposed approach is intuitively reasonable. We also investigate several related issues, including computation, within-group local dimension reduction, and evaluation of significance. Our numerical studies, including simulation and data analysis, show that the proposed approach has satisfactory performance.

In individual marker and group selections, it has been shown that some penalties can outperform Lasso-based penalties. It is possible to extend the proposed approach, for example, by replacing the group Lasso with group MCP or group SCAD. Such an extension may incur high computational cost and will not be pursued. There are multiple ways of defining difference between groups and hence smoothing. The proposed way is computationally simple and intuitively reasonable. The proposed approach can accommodate more subtle structure in SNPs. As a consequence, it inevitably demands new structure and tun-

ings. For example, there may be multiple ways of constructing groups. We are aware of "more automatic" approaches, for example, the hierarchical clustering plus Gap approach. However, such methods may rely on assumptions that SNP data clearly violate; and some methods cannot fully accommodate the spatial adjacency of SNPs on chromosome. The adopted tuning parameter selection approach has been developed in published studies. Our literature review suggests that there is a lack of consensus on tuning parameter selection with high-dimensional SNP data. However, a comprehensive investigation on tuning parameter selection is beyond the scope of the current paper. In our study, we use a predetermined number of groups to be selected. Usually, this number can be determined based on prior knowledge, limitation of resources for downstream analysis, and possible trial and error. We note that the proposed SGL can be easily coupled with other tuning parameter selection approaches such as cross validation. In this article, we focus on the development of the new methodology. Further work is needed to investigate the theoretical properties of the SGL method.

Acknowledgements

The rheumatoid arthritis data was made available through the Genetic Analysis Workshop 16 with support from NIH grant R01-GM031575. The data collection was supported by grants from the National Institutes of Health (N01-AR-2-2263 and R01-AR-44422), and the National Arthritis Foundation. This study has been supported by awards CA120988 and CA142774 from NIH.

References

C. Amos, W. Chen, M. Seldin, E. Remmers, K. Taylor, L. Criswell, A. Lee, R. Plenge, D. Kastner, and P. Gregersen. Data for genetic analysis workshop 16 problem 1, association analysis of rheumatoid arthritis data. *BMC Proceedings*, 3:S2, 2009.

S. Bakin. *Adaptive regression and model selection in data mining problems.* PhD thesis, Australian National University, Canberra, Australia, 1999.

A. Begovich, V. Carlton, L. Honigberg, S. Schrodi, A. Chokkalingam, H. Alexander, K. Ardlie, Q. Huang, A. Smith, J. Spoerke, M. Conn, M. Chang, S. Chang, R. Saiki, J. Catanese, D. Leong, V. Garcia, L. Mcallister, D. Jeffery, A. Lee, F. Batliwalla, E. Remmers, L. Criswell, M. Seldin, D. Kastner, C. Amos, J. Sninsky, and P. Gregersen. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.*, 75:330–337, 2004.

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

I.E. Frank and J.H. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35:109–148, 1993.

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

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

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

J. Huang, S. Ma, H. Xie, and C.-H. Zhang. A group bridge approach for variable selection. *Biometrika*, 96:339–355, 2009.

J. Huang, S. Ma, H. Li, and C.H. Zhang. The sparse Laplacian shrinkage estimator for high-dimensional regression. *Ann. Statist.*, 39:2021–2046, 2011a.

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

Y. Kim, J. Kim, and Y. Kim. The blockwise sparse regression. *Statist. Sinica*, 16:375–390, 2006.

L. Meier, S. van de Geer, and P. Bühlmann. Group Lasso for logistic regression. *J. R. Statist. Soc. B*, 70:53–71, 2008.

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

J. Newton, S. Harney, B. Wordsworth, and M. Brown. A review of the MHC genetics of rheumatoid arthritis. *Genes Immun.*, 5:151–157, 2004.

J. Ortega and W. Rheinboldt. *Iterative solution of nonlinear equations in several variables.* Classics in Applied Mathematics. SIAM, Philadelphia, PA, 4th edition, 2000.

R. Plenge, L. Padyukov, E. Remmers, S. Purcell, A. Lee, E. Karlson, F. Wolfe, D. Kastner, L. Alfredsson, D. Altshulder, P. Gregersen, L. Klareskog, and J. Rioux. Replication of putative candidate gene associations with rheumatoid arthritis in over 4,000 samples from north america and sweden: association of susceptibility with PTPN22, CTLA4 and PADI4. *Am. J. Hum. Genet.*, 77:1044–1060, 2005.

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

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109:475–494, 2001.

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.*, 2:224–244, 2007.

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

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

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

Appendix

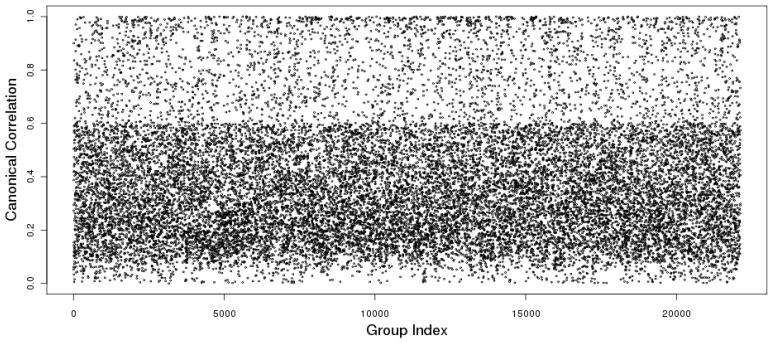
Canonical correlation

With the SGL penalty, the canonical correlation is suitable to measure the associations between adjacent groups. Canonical correlation analyzes the correlation between a linear combination of variables in one set and a linear combination of variables in another set. It searches for coefficient vectors $\mathbf{a}$ and $\mathbf{b}$ such that

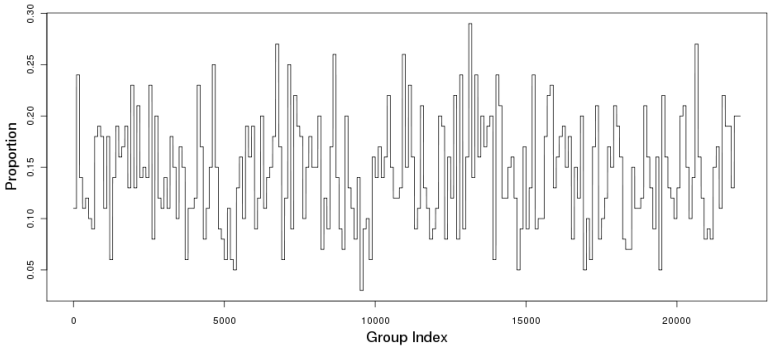
$$\text{Corr}(U, V) = \frac{\mathbf{a}'\Sigma_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\Sigma_{11}\mathbf{a}}\sqrt{\mathbf{b}'\Sigma_{22}\mathbf{b}}}$$

is maximized. Here $U = \mathbf{a}'X^{(1)}$, $V = \mathbf{b}'X^{(2)}$, $\Sigma_{11} = \mathrm{Cov}(X^{(1)}, X^{(1)})$, $\Sigma_{22} = \mathrm{Cov}(X^{(2)}, X^{(2)})$ and $\Sigma_{12} = \mathrm{Cov}(X^{(1)}, X^{(2)})$. By the change of basis and Cauchy-Schwartz inequality, it can be shown that $\max_{\mathbf{a},\mathbf{b}} \mathrm{Corr}(\mathbf{a}'X^{(1)}, \mathbf{b}'X^{(2)}) = \sqrt{\pi_1}$, where π_1 is the largest eigenvalue of $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$.

Note that canonical correlation is always positive. This can be guaranteed by choosing an appropriate sign for $\mathbf{b}$. This property is desirable as SGL uses the canonical correlation as weight to smooth estimates.

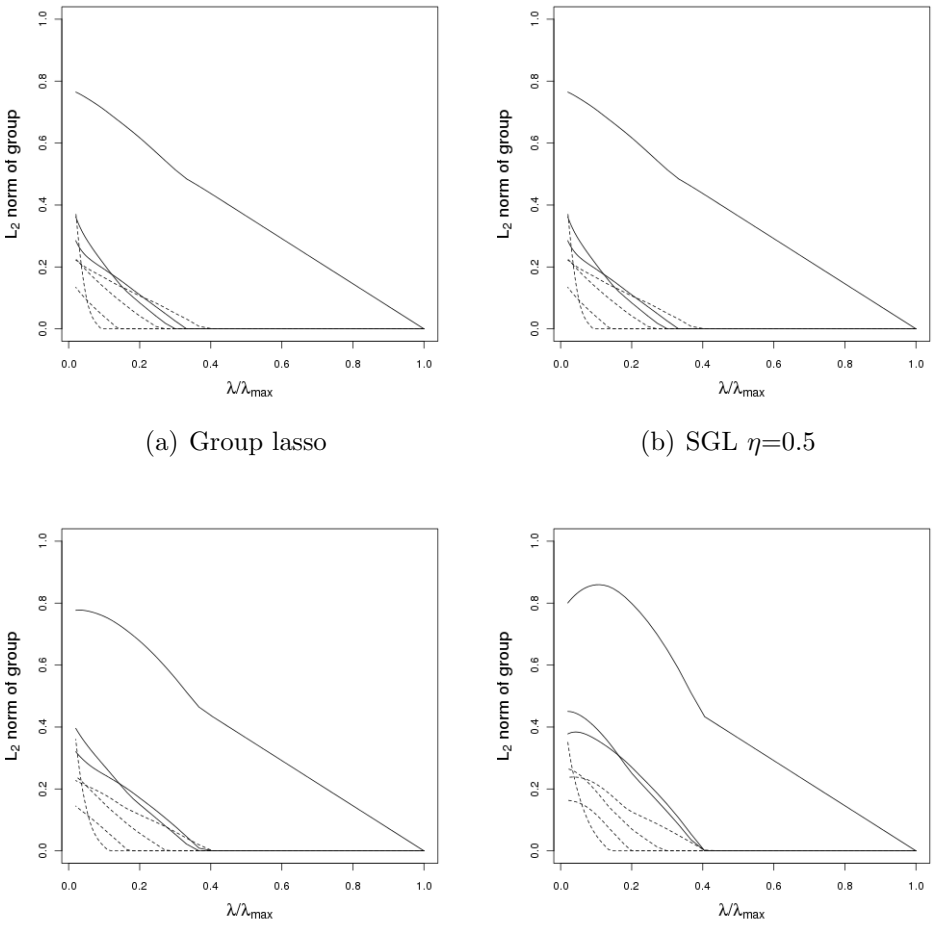


(a) Correlation coefficient ζ_j



(b) Correlation coefficients larger than 0.6 averaged within non-overlapping 100-SNPs windows.

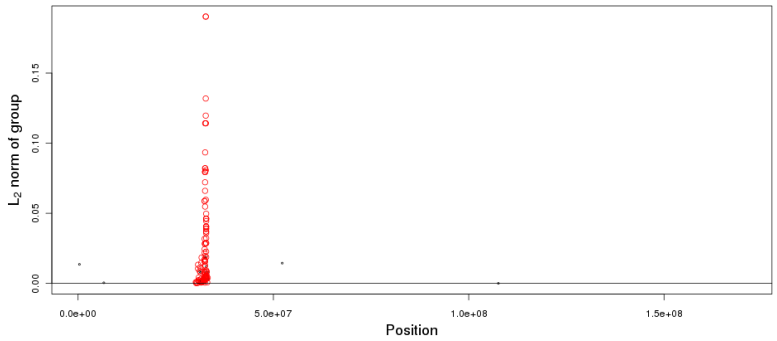
Figure 1: Plots of absolute lag-1 autocorrelation ζ_j on Chromosome 6 from Genetic Analysis Workshop 16 Rheumatoid Arthritis data.



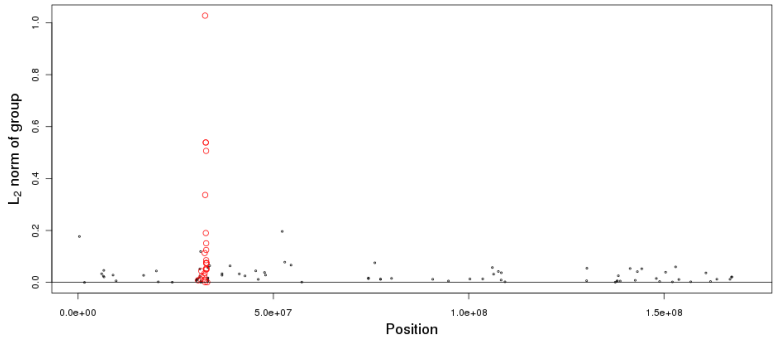
(a) Group lasso (b) SGL η=0.5

(c) SGL η=0.2 (d) SGL η=0.1

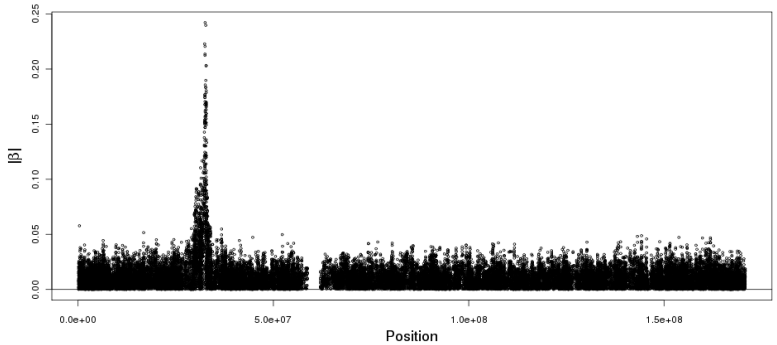
Figure 2: Solution path for a simulated data for (a) group Lasso, and SGL with (b) $\eta = 0.5$, (c) $\eta = 0.2$ and (d) $\eta = 0.1$, where $\eta = \lambda_1 + \lambda_2$. Black lines are paths of non-zero groups and grey lines are paths of irrelevant groups



(a) SGL $\eta = 0.1$



(b) group Lasso



(c) Single-SNP Regression

Figure 3: Plots of $||\beta||$ for SGL and group Lasso, and $|\beta|$ for single-SNP logistic regression.

Table 1: True positive (number of groups), false discovery rate (FDR) and false negative rate (FNR) for simulated data.

	Example 1			Example 2		
η	TP*	FDR	FNR	TP*	FDR	FNR
0.1	8.95(0.22)	0.40(0.02)	0.006(0.02)	8.59(0.53)	0.43(0.04)	0.05(0.06)
0.2	8.88(0.36)	0.41(0.02)	0.01(0.04)	8.48(0.50)	0.44(0.03)	0.06(0.06)
0.3	8.45(0.58)	0.44(0.04)	0.06(0.06)	8.13(0.56)	0.46(0.04)	0.10(0.06)
0.4	7.90(0.76)	0.47(0.05)	0.12(0.08)	7.77(0.65)	0.48(0.04)	0.14(0.07)
0.5	7.67(0.76)	0.49(0.05)	0.15(0.08)	7.58(0.61)	0.50(0.04)	0.16(0.07)
0.6	7.29(0.62)	0.51(0.04)	0.19(0.07)	7.57(0.66)	0.50(0.04)	0.16(0.07)
0.7	7.36(0.67)	0.51(0.05)	0.18(0.08)	7.44(0.65)	0.50(0.04)	0.17(0.07)
0.8	6.97(0.77)	0.54(0.05)	0.23(0.09)	7.16(0.72)	0.52(0.05)	0.20(0.08)
0.9	6.68(0.76)	0.56(0.05)	0.26(0.09)	7.06(0.75)	0.53(0.05)	0.22(0.08)
1	6.24(0.61)	0.58(0.04)	0.31(0.07)	6.72(0.83)	0.55(0.06)	0.25(0.09)
	Example 3			Example 4		
η	TP*	FDR	FNR	TP*	FDR	FNR
0.1	9.00(0.00)	0.40(0.00)	0.00(0.00)	8.41(0.81)	0.44(0.05)	0.07(0.09)
0.2	8.87(0.37)	0.41(0.03)	0.01(0.04)	7.61(0.87)	0.49(0.06)	0.15(0.10)
0.3	8.52(0.52)	0.43(0.04)	0.05(0.06)	7.41(0.74)	0.51(0.05)	0.18(0.08)
0.4	8.15(0.46)	0.46(0.03)	0.09(0.05)	7.33(0.77)	0.51(0.05)	0.19(0.09)
0.5	8.07(0.46)	0.46(0.03)	0.10(0.05)	7.14(0.75)	0.52(0.05)	0.21(0.08)
0.6	7.86(0.45)	0.48(0.03)	0.13(0.05)	6.91(0.74)	0.54(0.05)	0.23(0.08)
0.7	7.79(0.54)	0.48(0.04)	0.13(0.06)	6.63(0.68)	0.56(0.05)	0.26(0.08)
0.8	7.63(0.60)	0.49(0.04)	0.15(0.07)	6.34(0.81)	0.58(0.05)	0.30(0.09)
0.9	7.60(0.53)	0.49(0.04)	0.16(0.06)	5.55(0.87)	0.63(0.06)	0.38(0.10)
1	6.21(0.67)	0.59(0.05)	0.31(0.08)	4.20(0.84)	0.72(0.06)	0.53(0.09)

* True Positive.

** The optimal η for linear models (Example 1—3) is 0.3.

*** The optimal η for logistic regression model (Example 4) is 0.1.

**** When $\eta = 1$, the SGL becomes the group Lasso.

Table 2: Multi-split p-values for a simulated dataset in example 3.

	Group Info.		Group LASSO		SGL	
Group index	Start index	End index	$\|\|\hat{\beta}\|\|$	p-value	$\|\|\hat{\beta}\|\|$	p-value
150	462	463	0.008	1	0	1
179	551	551	0.043	1	0	1
634	2025	2025	0.016	1	0	1
654	2120	2123	0.011	1	0	1
664	2152	2155	0.002	1	0	1
695	2269	2270	0	1	0.002	1
703	2283	2285	0.009	1	0.0005	1
704	2286	2286	0.072	1	0.018	1
705	2287	2290	0.291	8.2e-08	0.046	5.3e-08
706	2291	2296	0	1	0.026	9.4e-05
708	2299	2299	0	1	0.058	1
709	2300	2303	0.006	2.9e-06	0.196	9.3e-09
710	2304	2306	0.578	2.5e-09	0.176	8.9e-10
711	2307	2307	0.078	0.060	0.139	0.003
712	2308	2310	0	1	0.254	6.9e-10
713	2311	2312	0.544	3.1e-07	0.191	5.9e-08
714	2313	2318	0	1	0.222	5.5e-08
715	2319	2319	0	1	0.058	1
716	2320	2321	0	1	0.014	4.9e-04
773	2528	2531	0.028	1	0	1
782	2558	2559	0.005	1	0	1
1038	3462	3462	0.043	1	0.001	1

Table 3: Multi-split p-values for a simulated dataset in example 4.

	Group Info.		Group LASSO		SGL	
Group index	Start index	End index	$\|\|\hat{\beta}\|\|$	p-value*	$\|\|\hat{\beta}\|\|$	p-value*
75	202	202	0.005	1	0	1
175	541	541	0.042	1	0	1
179	551	551	0.043	1	1.0e-04	1
184	567	567	0.008	1	0	1
272	883	883	0.073	1	0	1
296	963	963	0.011	1	0	1
425	1325	1326	0.018	1	0	1
469	1469	1469	0.007	1	0	1
580	1812	1812	0.011	1	7.0e-04	1
596	1866	1866	0	1	1.1e-04	1
683	2219	2220	0.127	1	0.003	1
705	2287	2290	0	1	0.001	0.002
707	2297	2298	0	1	5.3e-4	1
708	2299	2299	0	1	0.036	1
709	2300	2303	0	1	0.048	1
710	2304	2306	0.288	4.3e-09	0.059	1.1e-8
711	2307	2307	0.152	0.007	0.071	0.004
712	2308	2310	0	1	0.077	1.4e-07
713	2311	2312	0.328	1.4e-09	0.075	2.0e-09
714	2313	2318	0	1	0.067	1
715	2319	2319	0	1	0.051	1
716	2320	2321	0	1	0.009	1.9e-04
753	2464	2464	0.013	1	0	1
1361	4737	4737	0.020	1	0	1