VARIABLE SELECTION, SPARSE META-ANALYSIS AND GENETIC RISK PREDICTION FOR GENOME-WIDE ASSOCIATION STUDIES

Qianchuan He

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2012

Approved by:

Dr. Danyu Lin
Dr. Hao Helen Zhang
Dr. Donglin Zeng
Dr. Michael Wu
Dr. Christy L. Avery

Abstract

QIANCHUAN HE: Variable Selection, Sparse Meta-Analysis and Genetic Risk Prediction for Genome-Wide Association Studies
(Under the direction of Dr. Danyu Lin and Dr. Hao Helen Zhang)

Genome-wide association studies (GWAS) usually involve more than half a million single nucleotide polymorphisms (SNPs). The common practice of analyzing one SNP at a time does not fully realize the potential of GWAS to identify multiple causal variants and to predict risk of disease. Recently developed variable selection methods allow the joint analysis for GWAS data, but they tend to miss causal SNPs that are marginally uncorrelated with disease and have high false discovery rates (FDRs). Genetic risk prediction becomes highly challenging when the number of causal variants is large and many of the effects are weak. Existing methods mostly rely on marginal regression estimates, and their prediction power is quite limited. In meta-analysis, the involvement of multiple studies adds one more layer of complexity to variable selection. While existing variable selection methods can be potentially applied to meta-analysis, they require direct access to raw data, which are often difficult to be obtained.

In the first part of this dissertation, we introduce GWASelect, a statistically powerful and computationally efficient variable selection method for analyzing GWAS data. This method searches iteratively over the potential SNPs conditional on previously selected SNPs and is thus capable of capturing causal SNPs that are marginally correlated with disease as well as those that are marginally uncorrelated with disease. A special resampling mechanism is built into the method to reduce false-positive findings.

Simulation studies demonstrate that the GWASelect performs well under a wide spectrum of linkage disequilibrium patterns and can be substantially more powerful than existing methods in capturing causal variants while having a lower FDR.

In the second part, we propose a new approach, Sparse Meta-Analysis (SMA), which performs variable selection for meta-analysis based solely on summary statistics and allows the effect sizes of each covariate to vary among studies. We show that the SMA enjoys the oracle property if the estimated covariance matrix of the parameter estimators from each study is available. We also consider the situations in which the summary statistics include only the variances or no variance/covariance information at all. Simulation studies and real data analysis demonstrate that the proposed methods perform well. Since summary statistics are far more accessible than raw data, our methods have broader applications in high-dimensional meta-analysis than existing ones.

In the third part, we investigate the issue of genetic risk prediction when the number of true causal SNPs is large and many of the effect sizes are small. We show that the estimators obtained from marginal logistic regression can be severely biased and that using these estimators for prediction can lead to highly inaccurate results. To construct a joint-effects model, we propose a new method based on the smoothly clipped absolute deviation-supporting vector machine (SCAD-SVM). We conduct a series of simulation studies to show that our method outperforms the methods based on marginal estimators. We further assess the performance of our method by applying it to real GWAS studies.

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Danyu Lin, for his patience, caring and guidance. To me, he has been a role model in many ways. It is truly a privilege for me to have the opportunity to work with him and to learn from him. Without him, I would never have the opportunity to conduct the research presented in this dissertation.

I am extremely grateful to my co-advisor, Hao Helen Zhang, for her great guidance and tremendous help on my research projects. I am always impressed by her scope of knowledge and creative thinking, and I thank her for allowing me to work on these highly interesting projects that she has co-developed.

I am grateful as well to my committee members, Dr. Donglin Zeng, Dr. Michael Wu and Dr. Christy L. Avery, who have helped me to significantly improve my work. I thank them for generously giving their time and expertise on my dissertation projects. I particularly thank Dr. Avery for generously sharing with me many large data sets for my research.

I thank all my friends at Chapel Hill who have helped me in the past five and half years. I cherish the friendship with them for ever.

Finally, I would like to thank my wife, Ying Du, for her love, support and sacrifice. I also wish to thank my daughter, Angela, and my son, William, for the joy they have brought to me.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ARIC	Atherosclerosis Risk in Communities
ATT	Armitage trend test
AUC	Area Under Curve
CARDIA	Coronary Artery Risk Development in Young Adults
CCD	Cyclic coordinate descent
FDR	False discovery rate
GWAS	Genome-wide association study
HWE	Hardy-Weinberg equilibrium
LASSO	least absolute shrinkage and selection operator
LD	Linkage disequilibrium
LLA	Local linear approximation
MESA	Multi-Ethnic Study of Atherosclerosis
OLS	Ordinary Least Square
ROC	Receiver Operating Curve
SCAD	Smoothly clipped absolute deviation
SIS	Sure Independence Screening
SMA	Sparse Meta-Analysis
SNP	Single nucleotide polymorphism
SVM	Support Vector Machine
WTCCC	Wellcome Trust Case-Control Consortium

Chapter 1

Introduction

Genome-wide association studies (GWAS) have become one of the most important tools to study the genetics of human diseases. By typing a large number of single nucleotide polymorphisms (SNPs) in thousands of subjects, researchers are empowered to search the whole genome for potential targets that predispose a person to disease. At the same time, GWAS also pave the way for predicting personal genetic risk.

While many SNPs have been identified to be associated with diseases, current methods do not realize the full potential of GWAS. The most common way to analyze GWAS data is to examine each SNP one at a time. Although simple and convenient, this strategy ignores the correlations between SNPs and can yield highly biased results. A more appropriate way to analyze GWAS is to conduct joint analysis for all the SNPs (or at least a large subset of them) so that better estimation and more accurate prediction can be achieved.

Unfortunately, traditional statistical methods for joint analysis (such as multivariate linear regression and logistic regression) are not amenable to high dimensional data, such as GWAS data. An effective strategy to deal with high dimensions is to conduct variable selection. However, most of the existing variable methods are designed for a moderate number of features, and cannot handle the ultra-high dimensions associated

with GWAS. A few recently developed variable selection methods are designed for GWAS data, but they tend to miss causal SNPs that are marginally uncorrelated with disease and have high false discovery rates (FDRs). Indeed, how to select important variables and accurately estimate their joint effects under ultra-high dimensions is one of the most challenging topics in the current research of statistics.

Meta-analysis plays an important role in summarizing and synthesizing evidence from multiple GWAS studies. Because the dimensions of GWAS are high, it is desirable to incorporate variable selection into meta-analysis for better model interpretation and prediction. Existing variable selection methods require direct access to raw data, but in practice, it can be extremely difficult to collect GWAS data from multiple resources.

Genetic risk prediction represents another important application of GWAS. For many complex diseases, the number of true predictors tends to be high and the effects of genetic variants are usually weak. How to choose the most informative set of SNPs for prediction and how to construct an effective prediction model are still elusive. Currently adopted methods are primarily based on marginal regression coefficient estimates and often yield low prediction power. Constructing prediction models that are based on the joint effects of SNPs deserves more research efforts.

In this dissertation, we first conduct a literature review on the aforementioned issues in the remaining of this Chapter. In Chapter 2, we introduce a new method that is able to conduct variable selection at the genome-wide level. In Chapter 3, we propose a novel method for sparse meta-analysis, i.e., variable selection for meta-analysis. In Chapter 4, we investigate a strategy for building prediction models based on the joint effects of SNPs. In Chapter 5, we outline a future research project on variable selection for multivariate-outcome data.

1.1 Variable Selection for Genome-Wide Association Studies

Genome-wide association studies provide an important tool to study the genetic components of human diseases (McCarthy et al., 2008). In GWAS, researchers are able to examine more than half a million of SNPs in the human genome in search for potential associations between genetic variants and diseases. Many disease-associated SNPs have been identified (Ku et al., 2010), and the number of publications on GWAS has reached nearly one thousand in the year of 2011 according to the National Human Genome Research Institute's Catalog of GWAS (http://www.genome.gov/gwastudies/).

A number of methods have been used to analyze GWAS data, such as the Fisher's exact test, logistic regression, and the Armitage trend test (ATT). Regardless of which method is chosen, common practice is to analyze each SNP one at a time. However, the marginal effects of SNPs can deviate from their joint effects significantly due to correlations among the SNPs. For example, a group of SNPs may have weak marginal effects but strong joint effects. Conversely, a SNP that has a strong marginal effect may simply be an unimportant SNP that happens to correlate with a causal SNP. As Wu et al. (2009) have pointed out, marginal analysis "goes against the grain of most statisticians, who are trained to consider predictors in concert". A more appropriate way to analyze SNPs is to estimate their joint effects.

Traditional joint-analysis methods are paralyzed by the enormous dimension of GWAS. This is because the sample covariance matrix is no longer invertible when the dimension p is greater than the sample size n. To reduce p, there are two common strategies: one is Principle Component Analysis (PCA), and the other is variable selection. PCA is less appealing for GWAS, because the main goal of GWAS is to identify potential causal SNPs rather than to find the best linear combination of all the SNPs for explaining the observed variance.

Early variable selection methods require to search the model space exhaustively and

evaluate each model by a pre-specified criterion, such as the AIC (Akaike, 1973) and the BIC (Schwarz, 1978). When p is large, it is computationally infeasible to search the entire model space. This spurred the development of penalized regression methods, including the bridge regression (Frank and Friedman, 1993) and the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996). Let $\mathbf{y}$ denote the $n \times 1$ response vector, and $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^{\mathrm{T}}$ denote the $n \times p$ covariates matrix. Let $\boldsymbol{\beta}$ be the regression coefficients vector. The bridge regression methods take the form of

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|^{\gamma},$$

where λ is a tuning parameter, β_j is the jth component of $\boldsymbol{\beta}$, and γ is a number greater than zero. In contrast, the LASSO takes the form of

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Clearly, the LASSO is a special case of the bridge regression in that γ is fixed at 1. When $\gamma = 2$, bridge regression reduces to the well-known ridge regression.

The LASSO quickly becomes a popular tool for variable selection because it can estimate many covariates exactly as zero, and hence naturally yields a sparse model. For other bridge regression methods with $\gamma \neq 1$, the estimated models are either non-sparse or discontinuous (Fan and Li, 2001). The magic of the LASSO hinges on the absolute-value operator on the β_j's. When taken the first derivative, this operator essentially translates into a thresholding function that forces some regression coefficients to zero if those coefficients are below a certain threshold. The theoretical property of the LASSO was studied by Knight and Fu (2000). Surprisingly, they found that the limiting distribution of the LASSO can have a positive probability mass at 0 when the true parameter is equal to 0, which indicates that the LASSO does not have the

model-selection consistency. In light of this issue, Zou (2006) developed the adaptive LASSO that is consistent for model selection. The adaptive LASSO, in the form shown below,

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} w_j |\beta_j|,$$

replaces the penalty terms of the LASSO by $\lambda w_j|\beta_j|$, where w_j is a feature-specific penalty weight pre-specified by the user. Zou (2006) proved that, if w_j's are chosen to be sufficiently small for the true features and sufficiently large for the null features, then adaptive LASSO can select models consistently. In practice, one can choose $1/\hat{\beta}_j$ for w_j, where $\hat{\beta}_j$ is the least-squares estimate. However, if $p > n$, then $\hat{\beta}_j$ is not available and hence adaptive LASSO is no longer applicable.

On a different line, the inconsistency of the LASSO also triggered the innovation of other penalty terms. For example, Fan and Li (2001) realized that the convexity of the LASSO penalty is primarily responsible for the inconsistency of the LASSO, and introduced the smoothly clipped absolute deviation (SCAD) penalty which is concave. They showed that if the tuning parameter is chosen properly, then SCAD is able to select the correct model with probability tending to 1, and estimate the covariance matrix as efficiently as if the true features were known beforehand. They call this type of property as the oracle property.

Another variable selection method, the elastic net (Zou and Hastie, 2005), has also received wide attention in the past several years. The elastic net penalty is a weighted sum of the L_1 and the L_2 penalty, and was conjectured to behave somewhat between the LASSO and the ridge regression. Empirical evidence suggests that the elastic net can significantly improve the prediction accuracy when features are highly correlated. The elastic net was later extended to adaptive elastic net (Zou and Zhang, 2011), which was shown to have the oracle property too.

A method that is quite different from the penalized regression methods is the dantzig selector (Candes and Tao, 2007), which solves the following problem,

$$\min_{\boldsymbol{\beta}} ||\boldsymbol{\beta}||_1 \text{ subject to } ||\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})||_\infty \leq t.$$

Here $||.||_1$ denotes the L_1 norm, and $||.||_\infty$ denotes the L_∞ norm, i.e., the maximum absolute value of the components of the vector. Candes and Tao (2007) proved that the dantzig estimator can estimate the true $\boldsymbol{\beta}$ quite accurately even under $p > n$, with a loss that is within a logarithmic factor of the ideal mean squared error. However, it was found that the dantzig selector sometimes has erratic operating behaviors in variable selection (Hastie et al., 2009).

Most of the aforementioned methods were designed for a moderate number of predictors, and become nonfunctional when p is greater than n. For example, the adaptive LASSO can not be applied when $p > n$, because the penalty weights, usually borrowed from the least-squares estimates, are no longer available due to the non-invertibility of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Besides the non-invertibility issue, other challenges exist for high-dimensional variable selection, such as the spurious correlations between the null features and the true features, and the decay of the true signals (Fan and Lv, 2008). For example, when $p > n$, even if all the predictors are independent, the maximum absolute sample correlation coefficient between features can be unusually large (Figure 1.1). As a matter of fact, how to conduct variable selection under ultra-high dimensions has become one of the most challenging problems in the current statistical research.

One solution to deal with high-dimensionality is to conduct a pre-screening step to reduce the dimension, and this idea lies in the heart of the Sure Independence Screening (SIS) proposed by Fan and Lv (2008). Precisely, they suggest to shrink the number of features from a very large scale to a moderate scale that is below sample size by univariate correlation learning, and then select important predictors by a moderate-scale variable selection method, such as the LASSO or SCAD. Subsequent work on SIS

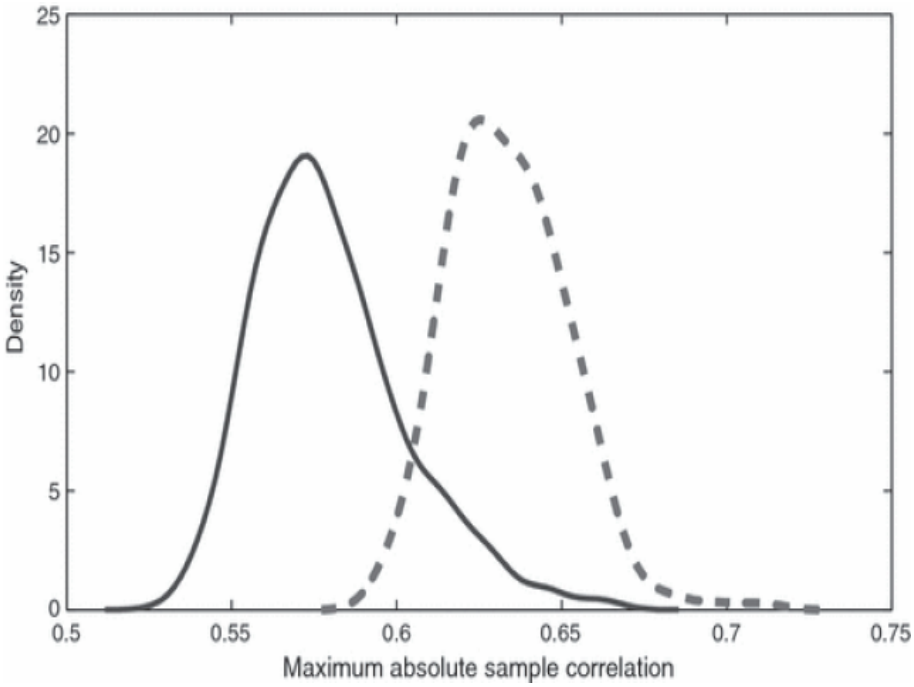

Figure 1.1: Distributions of the maximum absolute sample correlation coefficient when n=60 and p=1000 (inline image) and n=60 and p=5000 (- - - - -). (Fan and Lv, 2008.)

(Fan and Song, 2010) follows the same assumption that marginal screening utilities have a high probability of preserving all the important features.

In a similar spirit to the SIS, Wu et al. (2009) reduced the dimension of GWAS to several hundreds using a simple score criterion and applied the LASSO to the reduced set of SNPs. A drawback of this approach is that important features that are marginally uncorrelated with response are bound to be missed because the univariate screening step is based entirely on marginal correlations. Fan and Lv (2008) suggested the iterative sure independence screening (ISIS) procedure, which iterates the SIS procedure conditional on the previously selected features so as to capture important features that are marginally uncorrelated with response. Fan and Lv's work is confined to linear regression of a continuous response, and the number of features they considered is merely thousands.

It is important to control the false discovery rate (FDR) of a variable selection method. The FDR associated with ISIS, and indeed with any existing variable selection method, tends to be high. For GWAS, false discoveries can be easily made because the linkage disequilibrium (LD) between SNPs can be extremely high. Recently, Meinshausen and B$\ddot{u}$hlmann (2010) proposed the stability selection strategy to reduce the FDR. The procedure is to repeatedly subsample the original data and perform variable selection on each subsample. The rationale behind the stability selection is that, the features selected frequently among the subsamples tend to be truly associated with outcome and thus should be included in the final model. Fan et al. (2009) suggested a simpler way to reduce FDR. They divide the data into two halves, and then conduct variable selection for each half separately. Subsequently, the intersection of the two obtained models is designated as the final model. Zhao and Li (2010) provided some theoretic guidance on the control of FDR, but how practically useful their formula is remains to be examined.

Our review so far is mainly focused on the pivotal properties of variable selection methods, such as the model-selection consistency, oracle property and FDR. Next, we touch issues on to how to efficiently implement penalized regression methods and how to properly choose the tuning parameter. We first review some publications on the first issue.

The LASSO was initially implemented by a two-step procedure as follows. One first reexpresses the LASSO as least-squares problems with a number of inequality constraints, and then solves these problems by standard quadratic programming (Tibshirani, 1996). This algorithm was later replaced by the Least Angle Regression (LARS) algorithm, which is faster and only requires the same order of computation as a full least-squares problem (Efron et al., 2004). The Cyclic Coordinate Descent algorithm (CCD, also called the shooting algorithm) was hinted to solve the LASSO by Knight

and Fu (2000), but did not receive full attention until Friedman (2007) and Wu and Lange (2008) discovered that it is extremely efficient in solving the LASSO. This algorithm cycles through each feature one by one, and thus one only needs to optimize with respect to a single variable at each time. Since many features are estimated to be null features, the computation task quickly reduces to the optimization of a small subset of features, and hence the algorithm converges very fast.

It is more difficult to implement the SCAD because of the concavity of its penalty. The SCAD was initially implemented by a local quadratic approximation algorithm (Fan and Li, 2001), but this algorithm has a problem similar to the backward selection. That is, once a feature is estimated to be a null feature, it can never be selected back into the set of important features. Zou and Li (2008) proposed the local linear approximation algorithm, which covers all the penalized regression methods that have a concave penalty. With this algorithm, the SCAD can be casted as a series of LASSO problems, which can be quickly solved by the aforementioned CCD algorithm.

Another critical issue is how to choose the tuning parameter. A common strategy is to run a grid of tuning parameters, and then determine the best value of the tuning parameter by some evaluation criteria, such as the AIC and BIC. When p is large, BIC is no longer appropriate because it assigns a much higher probability to models with larger sizes than to models with smaller sizes. To correct this defect, Chen and Chen (2008) proposed the extended BIC and showed that it is consistent for model selection. Zou et al. (2007) provided some guidance on how to calculate the degree of freedom along the variable selection path. When AIC and BIC do not perform very well in practice, researchers often resort to the k-fold cross-validation or generalized cross-validation, which are often computation-intensive and sometimes require an independent validation data set.

1.2 Variable Selection for Meta-Analysis

Genetic studies of complex diseases often suffer from the problem of irreproducibility (Ioannidis, 2005). The leading factors responsible for this problem include small sample sizes, weak genetic effects and genetic heterogeneity (Burton et al., 2009; McCarthy et al., 2008; Ioannidis et al., 2007). Meta-analysis provides a way to obtain more robust and more reproducible results by combining the information from multiple studies. As a matter of fact, many highly influential findings in GWAS were discovered through meta-analysis (see for example, Scott et al., 2007; Zeggini et al., 2008; Lindgren et al., 2009).

Meta-analysis has been widely used in many quantitative research areas. By pooling multiple data sets together, one usually achieves higher statistical power, more accurate estimates, and improved reproducibility (Noble, 2006). Meta-analysis can be broadly classified into two classes, the one that requires access to the raw data, and the other one that only needs the summary statistics. The former one is sometimes called the *Integrative-analysis*. Let $\boldsymbol{\beta}_k = (\beta_{1k}, ..., \beta_{pk})^{\mathrm{T}}$ denote the vector of regression parameters in the kth study for $k = 1, ..., K$. Integrative-analysis can be conducted under two models: the first assumes that $\boldsymbol{\beta}_k$ are identical across all k studies, i.e., the fixed-effect model, while the second allows $\boldsymbol{\beta}_k$ to vary among different studies, i.e., the random effects model.

It is easier to collect summary statistics than the raw data, and many methods have been developed to analyze summary statistics. The frequently used one in GWAS is the weighted inverse variance method. Let $\hat{\boldsymbol{\beta}}_k$ be the estimator of $\boldsymbol{\beta}_k$, and $\hat{\boldsymbol{V}}_k$ the variance estimator of $\hat{\boldsymbol{\beta}}_k$. The inverse-variance estimate of the overall regression coefficients is

$$\left\{ \sum_{k=1}^{K} \hat{\boldsymbol{V}}_k^{-1} \right\}^{-1} \sum_{k=1}^{K} \hat{\boldsymbol{V}}_k^{-1} \hat{\boldsymbol{\beta}}_k,$$

and the inverse-variance estimate of the overall variance is

$$\left\{ \sum_{k=1}^{K} \hat{\boldsymbol{V}}_k^{-1} \right\}^{-1}.$$

Once these quantities are obtained, one can make statistical inference accordingly.

Traditional meta-analysis methods were mainly designed for low dimensional data sets. When the number of features becomes very large, such as that in the gene expression analysis or in the genome-wide association studies, it is desirable to incorporate variable selection into meta-analysis for better model interpretation and higher prediction accuracy.

A variety of variable selection techniques have been developed, including the LASSO (Tibshirani, 1996), the SCAD (Fan and Li, 2001) and the adaptive LASSO (Zou, 2006). If all the original data can be pooled together, a simple strategy is to apply these variable selection methods directly to the pooled data. By doing so, one essentially assumes that the effect sizes of a feature are identical across different studies (i.e., the fixed-effect model). For example, given a SNP, one needs to assume that its odds ratios are identical among all the studies. However, in real situations, the effect sizes of a feature often vary in different studies, i.e., following the random effects model. To avoid the 'fixed-effect' assumption, one can conduct variable selection for each study individually and then combine the results together. For example, let K be the total number of studies and $\widehat{\mathcal{M}}^{(k)}$ be the important set selected for the kth study, then $\cup_{k=1}^{K} \widehat{\mathcal{M}}^{(k)}$ may be considered as an estimate for the final set of important variables across multiple studies. The disadvantage of this strategy, though, is that each study is analyzed separately and hence it tends to be less efficient. It is also against the principle of meta-analysis, which emphasizes the importance of analyzing multiple studies together.

Very recently, Ma et al. (2011) proposed a more efficient approach that is able to overcome the weaknesses of the above methods. Their approach was motivated by the

gene expression analysis, where data collected from multiple resources are incompatible (due to different laboratory platforms) and are difficult to be combined. Their idea is explained as follows. Let $R_1(\boldsymbol{\beta}_1), ..., R_K(\boldsymbol{\beta}_K)$ be the likelihood for the K studies respectively, and $R \equiv R_1(\boldsymbol{\beta}_1) + \cdots + R_K(\boldsymbol{\beta}_K)$ be the likelihood summed over the K studies. They seek to solve

$$\text{argmax}_{\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_K} \left\{ R - \lambda_n \sum_{j=1}^{p} (\beta_{j1}^2 + \cdots + \beta_{jK}^2)^{1/2} \right\},$$

where λ_n is a tuning parameter. The penalty in the above expression is called the group bridge penalty. Under this formulation, the p features are treated as p groups, with each group consisting of its corresponding regression parameters from the K studies.

The group bridge penalty actually is a special case of the group penalty proposed by Yuan and Lin (2006), which has the form of

$$\lambda_n \sum_{j=1}^{p} \left\{ (\beta_{j1}, ..., \beta_{jK_j}) \times \Gamma_j \times (\beta_{j1}, ..., \beta_{jK_j})^{\mathrm{T}} \right\}^{1/2},$$

where Γ_j is a $K_j \times K_j$ weight matrix specified by the user. To see why the group penalty can be reduced to the group bridge penalty, simply let all $K_j = K$ and replace all Γ_j by the identity matrix. The motivation behind the group penalty is to produce sparsity at the group level, that is, some groups will be completely dropped during the variable selection process. By assigning different penalty weights to the regression parameters, the group-level sparsity allows one to select features based on some prior information, such as biological pathways or networks. Because of this attractive property, the group penalty has been adopted by a number of methods in the last few years (Ma et al., 2007; Wang et al., 2009; Meier et al., 2008; Zhou et al., 2010).

While Ma et al.'s approach (2011) is interesting, it requires direct access to the raw data. Unfortunately, it is not always possible to have all the raw data at hands due

to various reasons, such as IRB restriction, prohibition of data transfer, or unwillingness of the investigators to share data. Instead, most often we can only collect the summary statistics from each study, such as the least-squares estimates and their estimated variances. Surprisingly, some recent work by Lin and Zeng (2010a) and Lin and Zeng (2010b) indicated that summary statistics can provide almost an equal amount of information as the complete raw data when the sample sizes are sufficiently large. Hence, a natural question arises as to whether it is still possible to conduct effective variable selection solely based on summary statistics. This question turned out to have a positive answer, which is relegated to Chapter 3.

During the last decade, many progresses have been made on elucidating the asymptotic behaviors of the penalized regression methods. Knight and Fu (2000) obtained the limiting distribution of the LASSO estimator by deriving the limit of its penalized function. Fan and Li (2001) studied the oracle property for the SCAD in a completely different way. Fan and Li's proof essentially contains three steps: first, they showed that if the tuning parameter $\lambda_n \to 0$, then the estimator of the regression coefficients is $\sqrt{n}$ consistent; second, they proved that if $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, then with probability tending to 1, all the null features will be estimated as null features; third, based on the second result, they established the estimation efficiency for the true features. Zou (2006) followed the idea of Knight and Fu, and proved that the adaptive LASSO enjoys the oracle property as well. A by-product of Zou's article is that the nonnegative garrote (Breiman, 1995) was found to be a special case of the adaptive LASSO with additional sign constraints, agreeing with the results obtained by Yuan and Lin (2007).

All the above results assume that the dimension p is fixed and smaller than n. However, this assumption can be easily violated in real situations, for example, in the analysis of microarray data or GWAS. In the following discussion, we add a subscript n to the dimension p to emphasize that the dimension p_n is no longer a fixed number.

Let $\boldsymbol{\gamma}^0$ denote the true regression parameter, and $\hat{\boldsymbol{\gamma}}$ denote its estimator. A number of researchers have attempted to relax the assumption of '$p < n$'. Fan and Peng (2004) proved that, as long as $p_n^5/n \to 0$, even if p_n diverges possibly to infinity, one can still achieve $||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0|| = O(\sqrt{p_n/n})$ and the oracle property still holds. Huang et al. (2008) showed similar results for the bridge estimator when p_n is diverging. They further showed that, under the partial orthogonality condition (explained in the sequel), one can still achieve the oracle property even if $p_n > n$. Let $\mathcal{I}$ denote the set of important features, and $\mathcal{N}$ denote the set of unimportant features. Let x_{ij} be the jth covariate of the ith person. The partial orthogonality condition stipulates that, there exists a constant $c_0 > 0$ such that

$$\left| n^{-1/2} \sum_{i=1}^{n} x_{ij} x_{ik} \right| \leq c_0, \qquad j \in \mathcal{I}, k \in \mathcal{N},$$

for all large n. This condition essentially says that the correlations between important features and unimportant features cannot be too high so that it is possible to conduct marginal regression to reduce the dimension to a number less than n. Zou and Zhang (2011) proved similar asymptotic properties for the adaptive elastic-net. How to relax the partial orthogonality condition and how to better deal with the '$p > n$' situation remain to be one of the most active research areas in high dimensional data analysis.

1.3 Genetic Risk Prediction

Genome-wide association studies provide unprecedent opportunities for genetic risk prediction for human diseases (Kraft and Hunter, 2009). By harnessing the prediction power of single nucleotide polymorphisms, it is anticipated that disease prevention and clinical practice will be revolutionized in the near future (Collins, 2010).

The prediction power of SNPs varies dramatically among different diseases. For a

few diseases, such as the age-related macular degeneration, a handful of large-effect SNPs can yield a high prediction accuracy with regard to the disease outcome (Seddon et al., 2009). For many other diseases, the prediction power garnered from SNPs is fairly low. In fact, the proportion of phenotypic variance that can be explained by genetic factors is called the *heritability*, and is typically estimated from family studies. Because current methods that exploit SNPs for risk prediction can only explain a small proportion of the heritability, the remaining part is commonly called the 'missing heritability' (Manolio et al., 2009).

At least part of the missing heritability can be ascribed to insufficient number of SNPs being included in the prediction model. Current literature suggests that many complex diseases are contributed by a large number of causal SNPs (Barrett et al., 2008; Barret et al., 2009). Thus, it is natural to consider including as many SNPs as possible into the prediction model. A typical GWAS encompasses more than half a million SNPs, and it is certainly not proper to include all of them. One popular strategy is to conduct the marginal logistic regression for each SNP individually, and then use SNPs that reach a pre-specified statistical significance level for model construction (Wray et al., 2007). The resulting prediction model is essentially an inner product between those 'significant' SNPs and their estimated marginal effects, and we call this method as the Marginal method. This strategy has been adopted in a recent article (The international schizophrenia consortium, 2009), where thousands of SNPs were included in the prediction model, with the hope to harvest more prediction power. Another commonly used strategy is based on the genotype score, which is defined as the total number of risk alleles a person carries with respect to some predetermined candidate SNPs (James et al., 2008; Kang et al., 2010). Once the genotype score is obtained, one simply fits a logistic regression model (with the genotype score as the covariate) to construct the prediction model. This model can be somehow seen as a special case

of the first approach, with the absolute values of the effect sizes of all the SNPs being equal. We name the second method as the Count method.

There are at least two problems associated with the aforementioned two approaches. First, both of them rely on an arbitrary threshold/criterion to select SNPs and thus are quite *ad hoc*; second, it is known that marginal effects can be quite different from the joint effects of SNPs, hence the prediction power can be severely compromised. In fact, a recent study (Machiela et al., 2010) shows that using thousands of SNPs is no better than simply using dozens of SNPs for risk prediction when the above two methods are employed. Compared to the Marginal method and the Count method, a better strategy to deal with the aforementioned problems is to conduct variable selection on all the SNPs so that a smaller subset of SNPs can be prioritized and the joint effects of SNPs can be estimated.

While existing variable selection methods are abundant, few of them are suitable for the task of genetic risk prediction under the GWAS setting. This is because the dimension of GWAS is extremely high, and the linkage disequilibrium among SNPs is extensive. Beside these issues, the fact that the number of causal SNPs is large and many of their effects are weak poses additional challenges for risk prediction. The Support Vector Machine (SVM) is known for its superior performance in classification of high-dimensional data (Vapnik, 1999). For example, two articles that describe the application of the SVM to microarray data analysis (Furey et al., 2000; Guyon et al., 2002) have been cited more than 3000 times in total as of 2011. The SVM has many subtypes, and we focus on its linear subtype, i.e., the linear SVM. Let z_i denote the outcome for the ith subject, with 1 for case and -1 for control. Let α denote the intercept, and $\mathbf{x}_i$ and $\boldsymbol{\beta}$ be defined as in Section 1.1. The linear SVM can be written as

$$\min_{\alpha,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} [1 - z_i(\alpha + \mathbf{x}_i^T \boldsymbol{\beta})]_+ + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $[.]_+$ is called the hinge loss, and $[s]_+$ equals to the positive part of s.

Since the penalty term in the above function is an L_2 penalty, the SVM does not provide sparse solution with respect to $\boldsymbol{\beta}$. On the other hand, it has been observed that removing noise features from prediction models (i.e., achieving model sparsity) can often improve the prediction accuracy (Zou and Hastie, 2005; Kooperberg, et al., 2010). To achieve model sparsity, Zhang et al. (2006) replaced the L_2 penalty in the SVM with the smoothly clipped deviation (SCAD) penalty, and named the new method as the SCAD-SVM. The SCAD penalty is a concave penalty, and has the form of

$$p_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j| & \text{if } 0 \le |\beta_j| < \lambda \\ \frac{(a^2-1)\lambda^2-(|\beta_j|-a\lambda)^2}{2(a-1)}, i.e., -\frac{(|\beta_j|^2-2a\lambda|\beta_j|+\lambda^2)}{2(a-1)} & \text{if } \lambda \le |\beta_j| < a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } a\lambda \le |\beta_j| \ , \end{cases}$$

where a is usually chosen to be 3.7. It has been shown that the SCAD penalty possesses better theoretical property than the L_1 penalty (Fan and Li, 2001). Zhang et al. (2006) found that the SCAD-SVM competes favorably with the SVM in both variable selection and prediction.

The SCAD-SVM was designed for the analysis of gene expression data, whose dimension is much lower than GWAS. To extend the SCAD-SVM to the analysis of GWAS data, at least two issues need to be addressed. First, the original implementation of the SCAD-SVM requires inversion of a large matrix, which may encounter difficulties when the matrix is ill-conditioned. Second, dimension reduction is needed to reduce the enormous dimension of GWAS to a more manageable number. The recently proposed local linear approximation (LLA) algorithm (Zou and Li, 2008) has a potential to handle the first issue, while the Sure Independence Screening theory (Fan and Lv, 2008) provides some clues for the second issue. The local linear approximation works

as follows,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + p'_\lambda(|\beta_j^{(0)}|)(|\beta_j| - |\beta_j^{(0)}|),$$

which is essentially a first-order Taylor expansion of the $p_\lambda(|\beta_j|)$ around the $p_\lambda(|\beta_j^{(0)}|)$. The advantage of this approximation is that it accommodates the CCD algorithm and hence is highly efficient. The LLA algorithm is considered to be a significant improvement over the local quadratic approximation,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + \frac{1}{2}\{p'_\lambda(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}(\beta_j^2 - \beta_j^{(0)2}),$$

which has been found to be often less stable.

An important topic we have not touched so far is how to evaluate the prediction accuracy of various prediction methods. In simulation studies, because the true liability of disease is known, one can directly compare the estimated liability and the true liability. In real practice, the true liability is, of course, unknown, and other criteria need to be considered. A traditional criterion is the 0/1 mis-classification error rate. Let $\tilde{z}_i$ be the predicted outcome for z_i, then the mis-classification error rate under the SVM setting is defined as

$$\frac{1}{n}\sum_{i=1}^{n} |I(\tilde{z}_i = 1) - I(z_i = 1)|,$$

where $I(.)$ is an indicator function. This criterion essentially measures the proportion of subjects who were mis-classified with respect to their disease outcomes. Another popular measure of prediction accuracy is the Area Under Curve (AUC) with respect to the Receiver Operating Curve (ROC) (Lusted, 1971). The ROC is closely related to the concept of sensitivity and specificity, which is explained in the sequel. Let TP, TN, FP and FN denote the true positive, true negative, false positive and false

negative, respectively. The sensitivity is defined as $\frac{TP}{TP+FN}$, and the specificity is defined as $\frac{TN}{TN+FP}$. Plotting the sensitivity against the (1 – specificity) yields the ROC. An important difference between the two measures is that, the mis-classification error rate is critically dependent upon a single cut-off value on the decision rule, while AUC is averaged across all possible cut-off values. In fact, ROC curve depends only on the ranks of the liability scores (Lu and Elston, 2008), and hence is a more robust measure. In the current literature of genetic risk prediction, it appears that AUC is far more widely used than the mis-classification error rate (Wei et al., 2009; Kang et al., 2010; Machiela et al., 2011).

Most of the existing methods for genetic risk prediction are focused on SNPs data alone. Ruderfer et al. (2010) proposed to incorporate family information into disease prediction, but their method can only handle low-dimensional data. Developing methods that can handle high-dimensional data as well as family structures is likely to become an interesting research topic in future.

Chapter 2

A Variable Selection Method for Genome-wide Association Studies

2.1 Introduction

In this chapter, we propose a new variable selection method, GWASelect, for genome-wide association studies. Our method is motivated by the Iterative Sure Independence Screening (ISIS) of Fan and Lv (2008). The ISIS essentially consists of two major components in its concept: the first one is to conduct dimension reduction through the utility of marginal regression, while the second one is to iteratively search for conditionally important predictors. The former component provides a powerful tool to reduce the dimension from ultra-high to a more manageable number, while the latter reminds us that marginal regression alone is not sufficient to capture the complexity of correlations among predictors.

Our extension of ISIS to GWASelect is not trivial. First, the original ISIS proposed by Fan and Lv was designed specifically for linear regression model, where the conditional screening can be readily performed based on the residuals. In contrast, we are dealing with the logistic regression model, where residuals cannot be used as response variables for conditional screening. Second, prediction errors tend to be much higher

for binary outcomes than continuous outcomes, because the former outcomes contain less information than the latter. Third, Fan and Lv were considering microarray data analysis, where the number of predictors is at most several thousands, whereas the number of SNPs we are dealing with can be extremely large, typically more than half a million. Fourth, the effects of causal SNPs on complex diseases tend to be small to modest, so the signal-to-noise ratio in GWAS data is low. Fifth, the LD among SNPs is extensive and can be extremely high in certain regions.

Another distinct feature of our method is that we incorporate a subsampling procedure into GWASelect to reduce the FDR. The subsampling procedure is based on the theory of stability selection by Meinshausen and B*ü*hlmann (2010), and has been proved to be consistent for variable selection.

We describe our approach in the next section. In Section 2.3, we demonstrate through simulation studies that GWASelect has robust performance under a variety of LD structures and can substantially increase the power and reduce the FDR compared to existing methods. In addition, the regression models generated by GWASelect significantly improve prediction accuracy. In Section 2.4, we apply GWASelect to the GWAS data from the Wellcome Trust Case-Control Consortium (WTCCC) (2007) and show that it yields several novel discoveries and improves prediction accuracy.

2.2 Methods

Our ISIS method consists of one marginal SIS and two rounds of conditional SIS. We first describe the marginal SIS procedure. The data contain n subjects and p SNPs. The genotypes of each SNP are standardized by its sample standard derivation. The SIS theory suggests to reduce the original set of features to a small subset whose dimension is in the order of $n/\log n$. Since binary outcomes generally contain less information than continuous outcomes, we shrink the dimension of SNPs from p to $n/(4\log n)$. The

SIS theory also suggests to use a large proportion of the subset for the marginal SIS; therefore, we choose to use t SNPs, where t is the integer part of $0.9n/(4\log n)$. That is, we perform the ATT (under the additive model) on each SNP and select the t most significant SNPs to form a set $\mathcal{S}_1$. Then we apply the LASSO to $\mathcal{S}_1$ as follows.

For $i = 1, \ldots, n$, let Y_i denote the disease status (1=case, 0=control), and X_i denote the $(t+1)$-vector consisting of 1 and the genotypes of the t SNPs in $\mathcal{S}_1$. The genotype of each SNP is represented by the number of minor alleles. It is natural to assume the logistic regression model

$$\Pr(Y_i = 1|X_i) = \frac{\exp(\beta^{\mathrm{T}} X_i)}{1 + \exp(\beta^{\mathrm{T}} X_i)},$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_t)^{\mathrm{T}}$ denotes the vector of unknown regression coefficients. The penalized log-likelihood function takes the form

$$\widetilde{l}(\beta) = \sum_{i=1}^{n} \left[Y_i \beta^{\mathrm{T}} X_i - \log\{1 + \exp(\beta^{\mathrm{T}} X_i)\} \right] - \lambda \sum_{j=1}^{t} |\beta_j|,$$

where λ is the tuning parameter.

We adopt the cyclic coordinate decent algorithm (CCD) (Genkin et al., 2007; Friedman et al., 2010), which is tantamount to maximizing $\widetilde{l}(\beta)$ in a component-wise manner. Cross-validation can be used to determine the tuning parameter (and consequently the model size), but for now, we set the model size to a user-specified number, say d.(We will show later how to determine the model size adaptively.) That is, we run the LASSO on a dense grid of λ until it generates a model containing d predictors. If the exact number of d cannot be achieved, we choose the model whose size is right below d. This model is labeled $\mathcal{M}_1$.

To reduce potential collinearity, we prune $\mathcal{M}_1$ using pairwise correlations. Our analysis revealed that 99.9% of the pairwise correlations among the Illumina300K SNPs

have absolute values less than 0.8 (corresponding to r^2 of 0.64). Thus, we set the pruning threshold for r^2 to 0.64 so as to minimize the loss of information due to pruning. The pruned model is labelled $\mathcal{M}_1^*$. This marks the end of the marginal SIS.

Assuming that $\mathcal{M}_1^*$ contains t_1 SNPs, we label the set of the remaining $(p - t_1)$ SNPs as $\overline{\mathcal{M}}_1^*$. We use the conditional SIS described below to capture important SNPs in $\overline{\mathcal{M}}_1^*$ that are marginally uncorrelated with disease. The first step is to screen all the SNPs in $\overline{\mathcal{M}}_1^*$ to identify a small set of candidate SNPs that are correlated with Y conditional on $\mathcal{M}_1^*$. This step is computationally challenging because the cardinality of $\overline{\mathcal{M}}_1^*$ is close to p, which can be 1 million. We develop the following conditional score test to accomplish this task in a very efficient manner.

For the ith subject, let W_i be the $(t_1 + 1)$-vector consisting of 1 and the genotypes of the t_1 SNPs in $\mathcal{M}_1^*$. Let Z_j be the jth SNP in $\overline{\mathcal{M}}_1^*$, and Z_{ji} be the value of Z_j on the ith subject, where $j = 1, \ldots, p - t_1$. We assume the logistic regression model:

$$\Pr(Y_i = 1 | Z_{ji}, W_i) = \frac{\exp(\gamma Z_{ji} + \eta^{\mathrm{T}} W_i)}{1 + \exp(\gamma Z_{ji} + \eta^{\mathrm{T}} W_i)},$$

where γ and η are unknown regression coefficients. We are interested in testing the null hypothesis $H_0 : \gamma = 0$. It is computationally intensive to fit the above model for each of the $(p - t_1)$ SNPs. To bypass this difficulty, we perform the conditional score test.

Specifically, we calculate $S = U/V^{1/2}$, where

$$
\begin{aligned}
U &= \sum_{i=1}^{n} \left\{ Y_i - \frac{\exp(\widehat{\eta}^{\mathrm{T}} W_i)}{1 + \exp(\widehat{\eta}^{\mathrm{T}} W_i)} \right\} Z_{ji}, \\
V &= I_{\gamma\gamma} - I_{\gamma\eta} I_{\eta\eta}^{-1} I_{\gamma\eta}^{\mathrm{T}}, \\
I_{\gamma\gamma} &= \sum_{i=1}^{n} \frac{\exp(\widehat{\eta}^{\mathrm{T}} W_i)}{\{1 + \exp(\widehat{\eta}^{\mathrm{T}} W_i)\}^2} Z_{ji}^2, \\
I_{\gamma\eta} &= \sum_{i=1}^{n} \frac{\exp(\widehat{\eta}^{\mathrm{T}} W_i)}{\{1 + \exp(\widehat{\eta}^{\mathrm{T}} W_i)\}^2} Z_{ji} W_i{}^{\mathrm{T}}, \\
I_{\eta\eta} &= \sum_{i=1}^{n} \frac{\exp(\widehat{\eta}^{\mathrm{T}} W_i)}{\{1 + \exp(\widehat{\eta}^{\mathrm{T}} W_i)\}^2} W_i W_i{}^{\mathrm{T}},
\end{aligned}
$$

and $\widehat{\eta}$ is the maximum likelihood estimator of η under H_0. Note that $\widehat{\eta}$ and $I_{\eta\eta}$ do not involve any data in $\overline{\mathcal{M}}_1^*$ and thus need to be calculated only once at the outset of the conditional SIS. Given $\widehat{\eta}$ and $I_{\eta\eta}^{-1}$, we calculate the test statistic S for each of the $(p - t_1)$ SNPs in $\overline{\mathcal{M}}_1^*$. In vein with the SIS theory, we choose the most significant q SNPs, where q is the integer part of $0.05n/(4 \log n)$, and call this set of SNPs $\mathcal{S}_2$.(We use 0.05 since $0.9 + 0.05 + 0.05 = 1$, where 0.9 pertains to the marginal SIS, and $(0.05 + 0.05)$ to the two rounds of conditional SIS.)

The first step of the conditional SIS is aimed at identifying important SNPs that are marginally uncorrelated (but conditionally correlated) with disease while weakening the priority of those unimportant SNPs that are highly associated with disease through their correlations with the SNPs in $\mathcal{M}_1^*$. In the second step, we combine $\mathcal{S}_2$ with $\mathcal{M}_1^*$ and run the LASSO to select a model $\mathcal{M}_2$ with d SNPs. During this process, new SNPs may be selected, and previously selected SNPs have a chance to be removed from the model. We prune $\mathcal{M}_2$ to form a new model $\mathcal{M}_2^*$. This completes the conditional SIS.

To increase the opportunities of capturing important SNPs, we repeat the conditional SIS once and call the final model $\mathcal{M}_3^*$. We refer to $\mathcal{M}_3^*$ as the ISIS model.

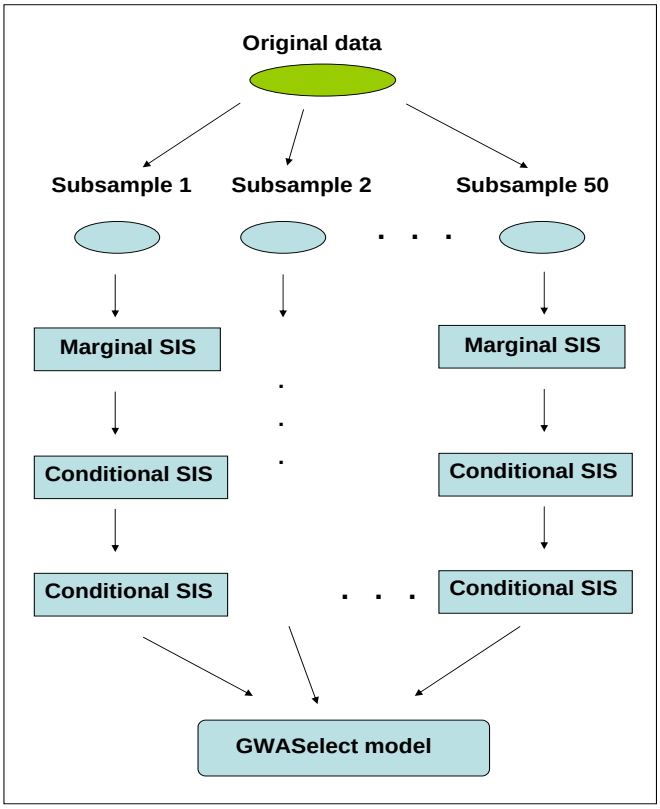

Figure 2.1: Flowchart of the proposed GWASelect method.

To reduce the FDR, we combine the extended ISIS procedure with the stability selection strategy (Meinshausen and Bühlmann, 2010) to create the GWASelect method, as illustrated in Figure 2.1. Specifically, we randomly obtain half of the cases and half of the controls from the GWAS data to form a subsample and then run the ISIS on this subsample. The resulting model is named $\mathcal{T}_1$. Repeating this subsampling concatenated with the ISIS 50 times, we obtain $\mathcal{T}_1, \ldots, \mathcal{T}_{50}$. Let $\mathcal{T} = \cup_{j=1}^{50} \mathcal{T}_j$, and denote $\mathcal{T} = \{v_1, \ldots, v_L\}$. We then calculate the selection probabilities for the L SNPs in $\mathcal{T}$

$$\pi_l = \sum_{j=1}^{50} I(v_l \in \mathcal{T}_j)/50, \quad l = 1, \ldots, L,$$

where $I(\cdot)$ is the indicator function. We choose the d SNPs with the highest selection probabilities from $\mathcal{T}$ to form the GWASelect model.

It is sometimes desirable to determine the model size adaptively from the data. To this end, we develop dynamic-GWASelect (d-GWASelect), which contains two modifications to the GWASelect. The first modification is that cross-validation is used to determine the tuning parameter for the LASSO embedded in the ISIS. Specifically, we divide the data randomly into 5 equal parts, with the kth ($k = 1, \ldots, 5$) part being the testing data and the remaining 4 parts being the training data. For a given tuning parameter λ, we apply the LASSO to the training data and select the SNPs that have nonzero regression coefficients. We calculate the liability score (i.e., the linear predictor) for each testing subject. Let $\mathcal{J}_1$ denote the set of subjects with the highest $\delta \times 100\%$ liability scores, and $\mathcal{J}_2$ the set with the lowest $\delta \times 100\%$, where δ is a user-specified number between 0 and 0.5. We then calculate the δ-error-rate, defined as $(\sum_{i \in \mathcal{J}_1} |Y_i - 1| + \sum_{i \in \mathcal{J}_2} |Y_i - 0|)/(2\delta\widetilde{n})$, where $\widetilde{n}$ is the number of subjects in the testing data. We choose the value of λ that minimizes the δ-error-rate averaged over the 5 testing data sets for $\delta = 0.1$.

The second modification is that, instead of fixing the model size at d, we specify a selection threshold ξ and select all SNPs with selection probabilities $\geq \xi$. As shown in the next section, the influence of ξ on the final model is typically small.

2.3 Simulation Studies

Each simulated data set contained 2,000 cases and 2,000 controls. For each subject, we simulated 20 chromosomes, each containing 3,000 SNPs. The disease status was generated from the logistic regression model containing 10 causal variants, $G_1, \ldots, G_{10}$, with the vector of log odds ratios β^*.

We considered three simulation schemes for the causal SNPs. In the first scheme, we simulated 10 independent causal SNPs that are located on 10 different chromosomes, with minor allele frequencies (MAFs) of 0.3. We set $\beta^* = (-0.35, -0.35, 0.35, 0.35, 0.35, 0.35, 0.35, -0.35, -0.35, -0.35)^{\mathrm{T}}$.

In the second scheme, we let $\{G_1, \ldots, G_{10}\}$ reside on one chromosome and have a special correlation structure such that the correlation between any two causal variants is nearly 0.6. We set $\beta^* = (0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3)^{\mathrm{T}}$.

In the third scheme, multiple causal SNPs were generated to be marginally uncorrelated with Y. We let the first causal SNP be independent of the other 9 causal SNPs. The latter were simulated to form three clusters, $\{G_2, G_3, G_4\}$, $\{G_5, G_6, G_7\}$ and $\{G_8, G_9, G_{10}\}$, each cluster residing on one chromosome. The three clusters are independent of each other, but within each cluster, SNPs have a compound symmetry correlation structure with correlation 0.5. We set $\beta^* = (0.5, -0.5, -0.5, 0.5, 0.5, -0.5, 0.5, -0.5, 0.5, -0.5)^{\mathrm{T}}$. Under this scheme, $\mathrm{corr}(Y, G_4)$, $\mathrm{corr}(Y, G_6)$ and $\mathrm{corr}(Y, G_9)$ are equal to 0.

For all three schemes, the positions of the causal variants on the chromosomes were randomly chosen. Thus, our simulation results would not be affected by any local LD patterns.

It is not trivial to simulate non-causal SNPs as they are desired to mimic the actual LD structure of human population. There exist several genome simulators based on the coalescent approach (Hudson, 2002), for which the users have to arbitrarily specify a number of parameters. As an alternative, the GWAsimulator of Li and Li (2008) employs a moving-window mechanism and can simulate genotypes based on the Illumina HumanHap300 chip data. We adopted the latter approach. Because the average distance between two SNPs for the Illumina HumanHap300 chip data is roughly 10 kb, the total length we simulated is approximately 600 Mb, which accounts for 1/5 of the whole genome. The LD was well preserved and no trimming was done for the simulated data.

Hoggart et al. (2008) explored variable selection from a Bayesian point of view by imposing the Laplace prior or the normal exponential gamma prior on each SNP. The former prior yields the LASSO procedure, while the latter generates a more sparse model and is called hyper-LASSO (HLASSO). We included the HLASSO in our simulation studies. Thus, we analyzed the simulated data by five methods: 1) the ATT method, for which the threshold for declaring significance was set to $0.05/60,000$ (i.e., Bonferroni correction); 2) the method by Wu et al. (2009); 3) the (extended) ISIS method; 4) the HLASSO; and 5) the GWASelect method. For the HLASSO, we applied a tuning parameter that yields an average model size of 15; for the other methods except the ATT, we also set the model sizes to 15. We chose 15 because most biology labs are likely to restrict their resources to a small number of top SNPs.

There are different criteria to evaluate a variable selection method. We chose to use the true discovery rate (TDR) and false discovery rate (FDR) (Benjamini and Hochberg, 1995) because the main goal of GWAS is to identify causal variants. For genetic studies, how to define the true discovery and false discovery is a delicate issue. This is because once a SNP is declared to be significant, all SNPs that are close to

and in LD with that SNP will be followed up. We defined the true positive and false positive as follows. If a captured SNP was no more than 50 SNPs away from a true causal SNP and had $r^2 > 0.05$ with that same causal SNP, then we classified it as a true positive. (Our experiments revealed that replacing 50 with 20 yielded similar results; Hoggart et al. (2008) provided a rationale for choosing 0.05 for r^2.) If more than one SNP satisfied these conditions, we counted them only as one true positive cluster. The remaining captured SNPs were classified as false positives. If two false positive SNPs were no more than 10 SNPs apart (i.e., within 100 kb in distance), we counted them as only 1 false positive cluster. The calculations of the TDR and FDR were based on clusters, rather than on individual SNPs. For each simulation scheme, the number of replications was set to 200. The results are shown in Table 2.1.

Scheme 1 was designed to compare the five methods under a scenario where all causal variants are independent and their effects are moderate. Under this scheme, all five methods yield high TDRs (>95%), but the FDRs are highly variable. Despite a large model size, the ATT method has the lowest FDR. This seemingly paradoxical phenomenon is explained by the fact that most of the SNPs in the ATT model are highly clustered due to strong LD. The GWASelect model has an elevated FDR, but far lower than the ISIS and the HLASSO, and slightly lower than the Wu et al. model. This demonstrates that, by repeated subsampling and variable selection, GWASelect is able to remove many noise features from the model. Overall, the ATT method appears to be a good option when causal variants are independent with moderate effects, but if one wishes to achieve higher power without too many false discoveries, the GWASelect method would be a reasonable choice.

In scheme 2, all ten causal variants are correlated with each other, which makes variable selection more challenging. It can be shown that under this scheme, the marginal

effects of the causal SNPs are much higher than their joint effects. For variable selection, this has the undesired effect of including unimportant SNPs that are in proximity of the causal SNPs. Reflecting this fact, the ATT, the ISIS, and the HLASSO all have FDR above 30%. The GWASelect is able to keep the FDR at a low level and preserve most of the power because of the stability selection. The Wu et al. method has high power and a relatively low FDR, suggesting that this method is particularly capable of distinguishing causal SNPs from unimportant SNPs that are in LD with them.

Scheme 3 represents a more complex correlation structure in which the three causal SNPs (i.e., the fourth, sixth and ninth SNPs) are marginally uncorrelated with Y. As expected, methods that are strongly driven by marginal correlations, such as the ATT and the Wu et al. method, almost completely missed G_4, which drives down their power to 70%. Both the ISIS and the HLASSO methods achieved higher power, but at the price of high FDR (around 30%). Overall, the GWASelect model offers a more balanced solution in terms of the TDR and FDR.

In summary, only the HLASSO and the GWASelect were able to keep their power above 90% under all three schemes, and the latter appears to have a much lower FDR. The other three methods either lack power under some schemes or entail high FDRs in others.

Next, we investigated the prediction accuracy of the five methods. For each scheme, we further simulated 2,000 testing subjects under the prospective sampling. To avoid numerical instabilities, we pruned the obtained models and used the pruned models for prediction. We calculated the true liability score and the estimated liability score for each subject and used the correlation between the two scores as a measure of prediction accuracy. We also calculated the absolute difference between the model-predicted and true disease probabilities, termed as p-diff, to measure the prediction error. The results are shown in Table 2.2.

Table 2.1: True and false discoveries of variable selection methods when the model sizes are fixed at 15 (except for the ATT method)

	ATT	Wu et al.	ISIS	HLASSO	GWASelect
Scheme 1					
Model size	26	15	15	15	15
TPC[a]	9.59	9.92	9.98	9.99	9.93
FPC[b]	0.06	2.02	4.04	3.84	1.52
TDR[c] (%)	95.9	99.2	99.8	99.9	99.3
FDR[d] (%)	0.6	15.8	28.5	26.4	12.4
Scheme 2					
Model size	103	15	15	15	15
TPC[a]	9.90	9.68	7.99	9.27	9.07
FPC[b]	4.8	1.05	3.88	4.89	0.03
TDR[c] (%)	99.0	96.8	79.9	92.7	90.7
FDR[d] (%)	31.5	8.6	31.0	32.8	0.3
Scheme 3					
Model size	41	15	15	15	15
TPC[a]	7.07	6.97	8.80	9.99	9.29
FPC[b]	0.08	4.97	4.88	4.47	1.92
TDR[c] (%)	70.7	69.7	88.0	99.9	92.9
FDR[d] (%)	1.0	40.3	35.3	29.4	16.0
G4[e] (%)	0	1	89	100	96

a. Number of true positive clusters

b. Number of false positive clusters

c. True discovery rate

d. False discovery rate

e. The rate of capturing the fourth causal SNP, which is marginally uncorrelated with disease under schemes 3.

Table 2.2: Prediction accuracy of variable selection methods when the model sizes are fixed at 15 (except for the ATT method)

	ATT	Wu et al.	ISIS	HLASSO	GWASelect
Scheme 1					
p-diff[a]	0.023	0.023	0.028	0.028	0.021
liab-correl[b]	0.931	0.943	0.919	0.920	0.948
log-likelihood	-760.0	-759.3	-763.4	-763.1	-758.6
Scheme 2					
p-diff[a]	0.067	0.028	0.073	0.048	0.053
liab-correl[b]	0.912	0.986	0.912	0.961	0.955
log-likelihood	-976.0	-938.9	-983.8	-955.7	-957.7
Scheme 3					
p-diff[a]	0.045	0.050	0.037	0.028	0.027
liab-correl[b]	0.801	0.771	0.874	0.937	0.926
log-likelihood	-720.0	-725.9	-710.9	-701.9	-701.7

a. The absolute difference between the model-predicted and true disease probabilities

b. liability correlation

The Wu et al. method excels under scheme 2, consistent with its high TDR and low FDR under this scheme. However, both the Wu et al. and the ATT are less accurate than the other methods under scheme 3 because they missed those marginally uncorrelated SNPs. The HLASSO performs well under scheme 2 and 3, suggesting that high prediction power can be achieved even if some noise features are included in the model. Overall, only the HLASSO and the GWASelect have prediction accuracy above 90% under all three schemes.

To assess data-adaptive choice of model size, we repeated the above simulation studies but now incorporated a 5-fold cross-validation into all the methods (except the ATT) by using the 10%-error-rate as the evaluation criterion (see Methods) For d-GWASelect, we set the selection threshold ξ to 0.3. All effect sizes were set to be moderate. For both schemes 1 and 3, $\beta^* = (0.4,\ -0.4, -0.4, 0.4, 0.5, -0.5, 0.5, -0.6, 0.6, -0.6)^{\mathrm{T}}$. For scheme 2, $\beta^* = (0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2)^{\mathrm{T}}$. The results are shown in Tables 2.3 and 2.4.

The d-GWASelect remains to be a robust variable selection method under all three schemes and indeed appears to have a better performance than the version with a fixed model size. The Wu et al., the ISIS and the HLASSO now entail extremely high FDR and poor prediction accuracy. The reason is that cross-validation often favors a large model size for logistic regression, especially when the signal-noise ratio is low. The d-GWASelect method, however, has a well-controlled model size because stability selection sifts away many noise features. In all, the d-GWASelect enjoys low FDR, high TDR and excellent prediction performance. Replacing the selection threshold with 0.4 yielded highly similar results (data not shown). We also explored the cross-validation by using the deviance (instead of the 10%-error-rate) as the evaluation criterion, and the d-GWASelect remains more favorable than the other methods (Tables 2.8 and 2.9 in Supplementary Materials).

Table 2.3: True and false discovery rates when cross validation is incorporated into variable selection (except for the ATT method)

	ATT	Wu et al.	ISIS	HLASSO	d-GWASelect
Scheme 1					
Model size	32	102	75	42	20
TPC[a]	9.95	10.00	10.00	10.00	10.00
FPC[b]	0.04	63.61	47.24	30.77	0.85
TDR[c] (%)	99.5	100.0	100.0	100.0	100.0
FDR[d] (%)	0.4	86.1	82.0	40.8	7.2
Scheme 2					
Model size	77	49	50	14	20
TPC[a]	9.73	9.88	9.40	7.96	9.09
FPC[b]	1.26	16.95	21.63	5.97	0.21
TDR[c] (%)	97.3	98.8	94.0	79.6	90.9
FDR[d] (%)	10.8	54.8	67.7	17.1	2.0
Scheme 3					
Model size	39	101	68	59	22
TPC[a]	7.01	7.13	9.99	9.87	9.85
FPC[b]	0.04	62.82	39.20	47.41	0.65
TDR[c] (%)	70.1	71.3	99.9	98.7	98.5
FDR[d] (%)	0.4	89.6	78.8	57.3	5.7
G4[e] (%)	0	1	100	89	87

a. Number of true positive clusters

b. Number of false positive clusters

c. True discovery rate

d. False discovery rate

e. The rate of capturing the fourth causal SNP, which is marginally uncorrelated with the disease outcome under scheme 3.

Table 2.4: Prediction accuracy of variable selection methods with cross-validation incorporated (except for the ATT method)

	ATT	Wu et al.	ISIS	HLASSO	d-GWASelect
Scheme 1					
p-diff[a]	0.018	0.076	0.073	0.046	0.016
liab-correl[b]	0.951	0.671	0.702	0.849	0.968
log-likelihood	-661.4	-745.9	-739.3	-719.6	-659.7
Scheme 2					
p-diff[a]	0.033	0.046	0.063	0.027	0.027
liab-correl[b]	0.912	0.873	0.802	0.942	0.946
log-likelihood	-754.3	-771.4	-795.4	-756.8	-749.4
Scheme 3					
p-diff[a]	0.039	0.082	0.061	0.062	0.017
liab-correl[b]	0.786	0.519	0.751	0.787	0.964
log-likelihood	-645.1	-725.5	-681.7	-713.7	-624.8

a. The absolute difference between the model-predicted and true disease probabilities

b. liability correlation

2.4 Analysis of WTCCC Data

The WTCCC study examined approximately 2,000 subjects for each of seven common diseases and a shared set of approximately 3,000 controls. Each subject was genotyped on the Affymetrix GeneChip 500K Mapping Array Set. We provide detailed analysis for the data on Type II diabetes (T2D [MIM 125853, http://www.ncbi.nlm.nih.gov/omim]) and Type I diabetes (T1D [MIM 222100]), and the analysis for some of the other five diseases is presented in Supplementary Tables 2.10-2.11.

We excluded a small number of subjects according to the sample exclusion lists provided by the WTCCC. In addition, we excluded a SNP if 1) it is on the SNP exclusion list provided by the WTCCC; 2) it has a poor cluster plot as defined by the WTCCC; 3) its MAF<0.01 in both cases and controls; or 4) it has extreme departure from Hardy-Weinberg equilibrium (p-value$< 10^{-4}$). Approximately 390,000 SNPs were used in the analysis, and there were 2,938 controls, 1,924 T2D cases and 1,963 T1D cases.

Figure 2.2 indicates the SNPs selected by the ATT, Wu et al., HLASSO and GWASelect for T2D; the details are shown in Table 2.5 and Supplementary Table 2.13. Under the ATT method, 15 SNPs reach the genomewide significance of p-value $< 10^{-7}$. The most significant one is rs4506565 (p-value$= 7.5 \times 10^{-13}$), which is located in gene *TCF7L2*. The other 14 SNPs are clustered within either *TCF7L2* or *FTO*. These results are consistent with the WTCCC's findings. The HLASSO model is essentially identical to the ATT model, albeit with a smaller model size.

For the Wu et al. and GWASelect methods, we set the model sizes to 20. Both methods successfully detected *TCF7L2* and *FTO*. They also identified a locus that spans *TSPAN8/LGR5*, which was one of the most significant loci reported in a recent meta-analysis of 10,128 subjects (Zeggini et al., 2008). This finding demonstrates empirically that regression-based variable selection methods can be more powerful than

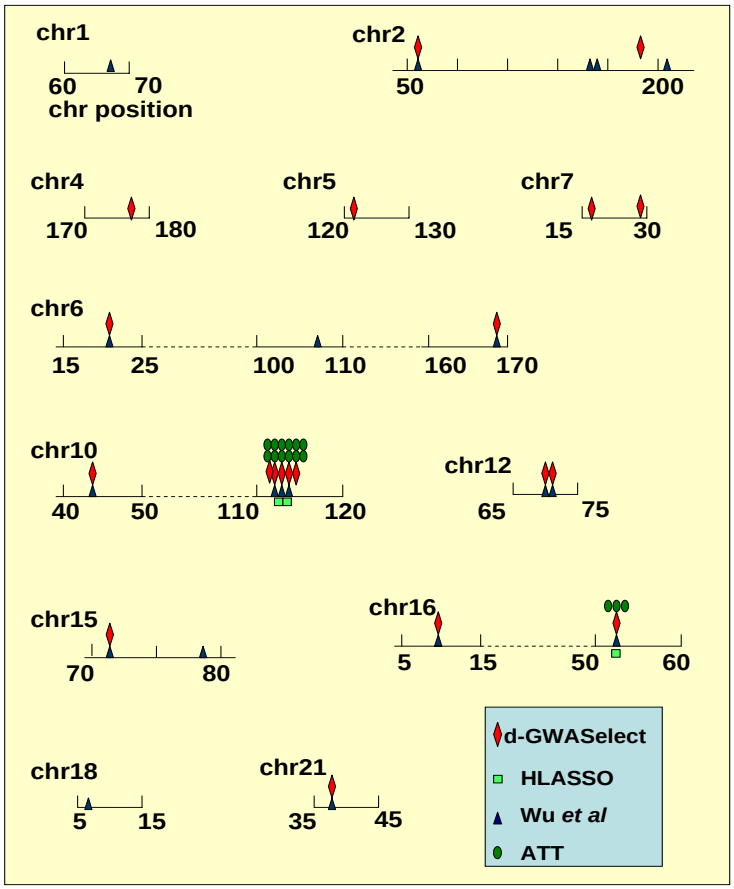

Figure 2.2: The T2D models selected by four different methods.

the ATT method.

It is interesting to compare the GWASelect and Wu et al. models. Five SNPs, rs11688935, rs6846031, rs6872465, rs2389591 and rs10435018, show up only in the GWASelect model. Among these SNPs, rs6846031 was selected partly due to its conditional correlation with T2D, underscoring the importance of conditional screening in variable selection. This finding also indicates that genetic factors underlying T2D are not simply in parallel with each other, but rather form a complex structure that needs to be carefully dissected.

Several SNPs in our GWASelect model have not been reported in the literature on

T2D. Some of them are plausibly related to T2D. For example, GULP1 is an adaptor protein that binds and directs the trafficking of LRP1 (Su et al., 2002), a protein that has been shown to play a critical role in adipocyte energy homeostasis and insulin sensitivity (Hofmann et al., 2007). Thus, genetic variants in *GULP1* may potentially influence the amount of LRP1 in adipocyte cells and thereby modulate a person's risk to T2D. As another example, the CREB5 was recently found to be down-regulated along with other members of the insulin signaling cascade when stimulated by a ligand of PPARγ, which is known to be associated with T2D (Herrmann et al., 2009). This suggests that CREB5 is closely related to PPARγ and the insulin pathway. The other SNPs do not have known connections with T2D, but further investigation of those loci may reveal novel mechanisms or pathways related to T2D.

For prediction of T2D, the δ-error-rates (with δ=0.1) of all four models are over 40%, suggesting that T2D is greatly influenced by other types of genetic variations and environmental factors. Since it is not very meaningful to compare prediction errors at such high level, we turned our attention to the T1D data because it is well-known that T1D is genetically more homogeneous than T2D.

For the T1D data, we used cross-validation to choose the tuning parameter for the d-GWASelect method and set the selection threshold ξ to 0.20. For the Wu et al. method, we set the model size to 15. The results are shown in Figure 2.3, Table 2.6 and Supplementary Table 2.12. The d-GWASelect model contains 14 SNPs, among which *ADAM29*, *SYNGAP1*, *CUX2* and *ALDH2* do not appear in any of the other three models. The gene *SYNGAP1* was observed to have strong conditional correlation with T1D, demonstrating again that selection solely based on marginal correlation is insufficient. Searching the T1DBase (http://www.t1dbase.org) revealed that all 4 genes have expressions in pancreas, although none has been previously considered as strong candidates for T1D. Interestingly, the *CUX2* has been shown to directly regulate the

Table 2.5: List of SNPs selected by the GWASelect for the WTCCC-T2D

SNP[a]	Chromosome	Gene[b]
rs11688935	2	*GULP1*
rs903228	2	*ASB3/LOC129656*
rs6846031	4	*VEGFC/NEIL3*
rs6872465	5	*PRDM6*
rs10806665	6	*THBS2/SMOC2*
rs9465871	6	*CDKAL1*
rs2389591	7	*TMEM195/LOC729920*
rs10435018	7	*CREB5*
rs7917983	10	*TCF7L2*
rs7901695	10	*TCF7L2*
rs4506565	10	*TCF7L2*
rs4132670	10	*TCF7L2*
rs7077039	10	*TCF7L2*
rs9326506	10	*ZNF239*
rs1495377	12	*TSPAN8/LGR5*
rs7961581	12	*TSPAN8/LGR5*
rs2930291	15	*CCDC33*
rs8050136	16	*FTO*
rs2099106	16	*C16orf72/GRIN2A*
rs6517434	21	*KCNJ6*

a. rs number identified from dbSNP

b. Gene symbol from Entrez Gene (http://www.ncbi.nlm.nih.gov/gene/)

expression of *NeuroD* (Lulianella et al., 2008), a gene that can cause T1D if mutated.

Finally, we compared the prediction accuracy of the four methods. We randomly divided the data into three parts, two as the training data and one as the testing data. Since the training data set contains only 2/3 of the original data, we reduced ξ from 0.20 to 0.10 to ensure that a similar number of loci are included in the d-GWASelect model. Since the true liability scores and disease probabilities are unknown in real data, we measured the prediction errors by the δ-error-rates for $\delta = 0.1$, 0.15 and 0.25 (see Methods for detail). Considering that pruning was done before each model was used for prediction, we report the actual (i.e., effective) number of SNPs used by each model for prediction. Under default settings, the effective model sizes of the Wu et al., the HLASSO and the d-GWASelect are 14, 4 and 21, respectively. Since the former two models are much smaller, we also evaluated the prediction accuracy of the former two under 21 effective SNPs. (We were not able to evaluate the ATT under 21 effective SNPs due to numerical instabilities.) The results are reported in Table 2.7. Clearly, the d-GWASelect performs the best or nearly the best for all three δ-error-rates. We have also calculated the area under the ROC curve for the four methods, and GWASelect achieves the highest value (Supplementary Table 2.15).

2.5 Discussion

We have developed a new tool, GWASelect, for variable selection at the genome-wide level. This regression-based method has the ability to capture both marginally correlated and marginally uncorrelated causal SNPs and has low FDR. The advantages over the existing methods have been demonstrated through simulated and real data. Our method has two versions. The first version requires the specification of the model size d, for which we suggest to choose a number that is consistent with the current biological knowledge of the studied disease. The second version (d-GWASelect) does

Table 2.6: List of SNPs selected by the d-GWASelect for the WTCCC-T1D

SNP[a]	Chromosome	Gene[b]
rs6679677	1	*RSBN1/PTPN22*
rs41515647	1	*ST6GALNAC5*
rs17388568	4	*ADAD1*
rs330483	4	*ADAM29*
rs9273363	6	*HLA-DQB1*
rs9272346	6	*HLA-DQA1*
rs411136	6	*SYNGAP1*
rs1265566	12	*CUX2*
rs7398833	12	*CUX2*
rs10744777	12	*ALDH2*
rs17696736	12	*C12orf30*
rs11171739	12	*ERBB3*
rs12708716	16	*CLEC16A*
rs12924729	16	*CLEC16A*

a. rs number identified from dbSNP

b. Gene symbol from Entrez Gene

Table 2.7: Prediction errors for the WTCCC-T1D

model	effective size	δ-error-rate			log-likelihood
		0.1	0.15	0.25	
ATT	5	0.110	0.139	0.181	-2116.9
Wu et al.	14	0.119	0.139	0.179	-2075.1
	21	0.135	0.157	0.196	-2059.8
HLASSO	4	0.116	0.141	0.176	-2113.6
	21	0.126	0.151	0.191	-2073.5
d-GWASelect	21	0.107	0.131	0.178	-2058.6

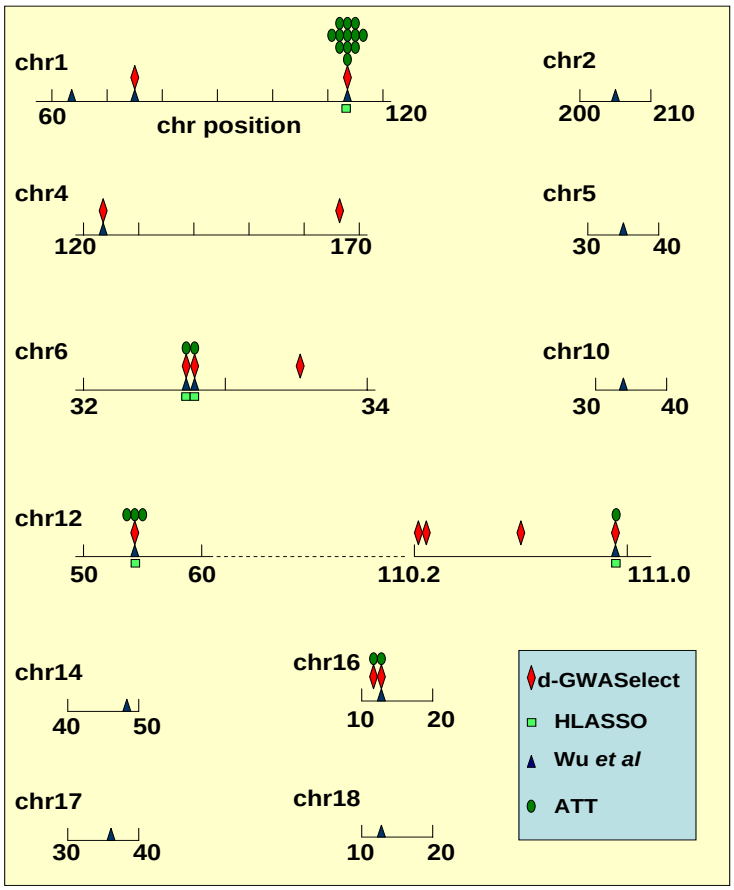

Figure 2.3: The T1D models selected by four different methods.

not require the specification of the model size, and this is the version we recommend for general use.

The correlation structures for causal variants used in our simulation studies have biological relevance. Scheme 2 mimics a scenario in which the causal variants form a gene cluster that contributes synergistically to the disease outcome, while scheme 3 reflects a scenario in which several biological pathways (or networks) affect the disease development.

We did not include Least Angle Regression (LARS) in our studies because it has been shown to have highly similar performance to LASSO (Hastie et al., 2009). Indeed, LASSO can be implemented by LARS with a small modification. Wu and Lange (2008) demonstrated that CCD is "considerably faster and more robust than LARS" and is "more successful than LARS in model selection".

The HLASSO adopts a concave penalty function, and it has been suggested that the CCD algorithm may not converge for nonconvex penalty (Wu et al., 2009; Friedman et al., 2010). A valid algorithm to implement concave penalty functions is the local linear approximation (Zou and Li, 2008), which amounts to multiple rounds of CCD and would make the HLASSO computation prohibitively expensive. For the WTCCC T1D data, running the CCD version of the HLASSO with 10 iterations on an Intel Quadcore Nehalem processor (2.4Ghz, 16GB memory) require 67.5 to 175 hours, depending on the value of the tuning parameter. In contrast, we have been running the GWASelect in a parallel computing environment, and the same analysis can be completed within several hours on 16 processors.

In an independent effort, Fan et al. (2009) developed an ISIS method for generalized linear models in the context of microarray data analysis. In their method, the conditional screening procedure requires fitting a separate regression model for each feature, which would create heavy computational burden for GWAS data. In addition, their

method tends to have high FDR. They observed that cross-validation tends to yield large models for logistic regression, resonating our findings in the Simulation Studies.

We can extend our methods to select interactions. Instead of considering all possible interaction terms, we may incorporate known biological network information (Franke et al., 2006) into our selection procedure. Another approach is to first extend the existing genetic network identification tools, such as the liquid association (Li, 2002) and bounded mode stochastic search (Dobra, 2007), to infer SNP interactions and then incorporate such information into our GWASelect procedure. In fact, Han et al. (2010) proposed a Markov blanket-based method to evaluate epistatic interactions for GWAS data. It will be interesting to compare to that method when we extend our work to interaction effects.

How to obtain p-values for high-dimensional variable selection is an active research area. The stochastic error introduced by the selection process makes it very difficult to assign p-values to the selected features. Meinshausen et al. (2009) offered one possible solution by "aggregating" p-values from stability selection, but our experiments indicated that this procedure is too conservative for SNP data, likely due to the ultra-high dimension and strong LD. We hope future progress will shed light on this important issue.

The prediction of genetic risk using GWAS data has drawn considerable attention in recent years. Wray et al. (2007) pioneered this area of research. Their approach selected genetic predictors by a univariate screening method. As shown in this chapter, our GWASelect method tends to provide more accurate prediction than univariate screening when the SNPs are in strong LD. Wei et al. (2009) explored genetic risk prediction through a Support Vector Machine (SVM) algorithm, but it is difficult to compare our results directly with theirs because 1) their analysis involved two other data sets besides the WTCCC-T1D data; 2) our testing samples are far smaller than

theirs; 3) interaction effects are not considered in our current work.

2.6 Supplementary Materials

2.6.1 Cross-validation using deviance

In the following two tables, we show that if the deviance is used as the evaluation criterion for cross-validation, the Wu et al. method, the ISIS method and the HLASSO method entail high FDRs, whereas the d-GWASelect method maintains high power, low FDR and good predictive capabilities.

2.6.2 Analysis of the WTCCC data

We analyzed the WTCCC data by four different methods, the ATT, the Wu et al., the HLASSO and the GWASelect. For the HLASSO method, we chose 0.1 as the shape parameter (as suggested by the User Manual of the HLASSO), and a tuning parameter that corresponds to 1×10^{-7} for the SNP-wise type-I-error-rate α. For the other methods, the details have been described in the main text of this paper.

In the following tables (Tables 2.10-2.13), we list SNPs that were identified by different variable selection methods for some of the seven WTCCC diseases (some SNPs were omitted if their loci have already been represented by other SNPs). We also examined whether the identified SNPs have been replicated by other GWAS as collected in the National Human Genome Research Institute's Catalog of Published Genome-Wide Association Studies as of September 1, 2010. In Table 2.14, the replication rates (i.e., the proportion of distinct loci replicated by other studies) for the four methods were calculated. However, we suggest caution be exercised when interpreting Table 2.14. First, although the ATT and the HLASSO have higher replication rates, their model

Table 2.8: True and false discovery rates when the deviance is used as the evaluation criterion for cross-validation (except for the ATT method)

	ATT	Wu et al.	ISIS	HLASSO	d-GWASelect
Scheme 1					
Model size	32	100	77	96	21
TPC[a]	9.95	10.00	10.00	10.00	9.99
FPC[b]	0.04	63.60	47.20	82.63	0.76
TDR[c] (%)	99.5	100.0	100.0	100.0	99.9
FDR[d] (%)	0.4	86.1	81.9	71.5	6.6
Scheme 2					
Model size	77	48	52	66	26
TPC[a]	9.73	9.97	9.73	7.96	9.56
FPC[b]	1.26	17.27	22.06	55.89	0.32
TDR[c] (%)	97.3	99.7	97.3	79.6	95.6
FDR[d] (%)	10.8	55.9	67.2	37.8	2.8
Scheme 3					
Model size	39	98	71	107	24
TPC[a]	7.01	7.13	9.99	9.96	9.88
FPC[b]	0.04	62.84	39.19	93.66	0.68
TDR[c] (%)	70.1	71.3	99.9	99.6	98.8
FDR[d] (%)	0.4	89.6	78.8	79.2	5.9
G4[e] (%)	0	1	100	97	91

a. Number of true positive clusters

b. Number of false positive clusters

c. True discovery rate

d. False discovery rate

e. The rate of capturing the fourth causal SNP, which is marginally uncorrelated with the disease outcome under scheme 3.

Table 2.9: Prediction accuracy when the deviance is used as the evaluation criterion for cross-validation (except for the ATT method)

	ATT	Wu et al.	ISIS	HLASSO	d-GWASelect
Scheme 1					
p-diff[a]	0.018	0.076	0.073	0.092	0.016
liability correlation	0.951	0.671	0.701	0.698	0.968
log-likelihood	-661.4	-746.1	-739.8	-837.7	-659.6
Scheme 2					
p-diff[a]	0.033	0.047	0.063	0.074	0.027
liability correlation	0.912	0.872	0.803	0.832	0.948
log-likelihood	-754.3	-771.7	-795.2	-898.2	-749.0
Scheme 3					
p-diff[a]	0.039	0.082	0.061	0.099	0.017
liability correlation	0.786	0.519	0.751	0.672	0.965
log-likelihood	-645.1	-725.7	-682.3	-819.4	-624.7

a. The absolute difference between the model-predicted and true disease probabilities

sizes are small and they essentially only contain loci with strong effects. Second, a number of SNPs captured by the GWASelect may not be able to be replicated by currently used methods due to the different selection mechanisms, and hence the replication rate of the GWASelect may be underestimated. Finally, although the replication provides a useful measure for the reliability of a variable selection method, it should not be regarded as the gold standard for making the final judgement. The ultimate criterion to judge whether a detected locus is a true or false discovery would be functional studies. Overall, these tables are shown mainly to provide potential candidates for future research, rather than to compare the performances of the illustrated methods.

We have systematically searched the Online Mendelian Inheritance in Man for functional evidence for the identified loci (http://www.ncbi.nlm.nih.gov/omim), but most of the evidence is circumstantial or weak, if it exists at all. Taking the T1D as an example, among all the loci identified by all the studied methods (including the Wu et al. method), only for the HLA-DQB1 and the HLA-DQA1 loci did we find functional

Table 2.10: List of SNPs selected by different methods for Bipolar Disorder (BD)

SNP	Chr	Gene	ATT	Wu et al.	HLASSO	d-GWASelect
rs41515647	1	*ST6GALNAC5*				•
rs7570682	2	*MRPS9*		•		
rs11123306	2	*DPP10*		•		
rs1375144	2	*DPP10*		•		
rs12472797	2	*LPIN1*		•		
rs1133353	2	*CAPN10*		•		
rs514636	3	*LAMP3*		•		
rs4276227	3	*CMTM8*		•		
rs4627791	3	*CMTM8*		•		•
rs715891	5	*PPP2R2B*				•
rs10993706	9	*SYK*				•
rs10982256	9	*DFNB31*		•		
rs11622475	14	*TDRD9*		•		•
rs10134944	14	*SLC35F4*		•		
rs1344484	16	*CHD9*		•		
rs7243929	18	*PTPRM*		•		•
rs12980129	19	*ZNF99*		•		•
rs2837588	21	*DSCAM*		•		•

Note: No SNPs were found to be replicated by other GWAS as collected in the National Human Genome Research Institute's Catalog of Published GWAS.

evidence that would directly link the two loci to the T1D, while for the other loci, we found them either poorly annotated or lack of direct functional evidence for T1D. More biochemical/genetics studies are merited.

Table 2.11: List of SNPs selected by different methods for Coronary artery disease (CAD)

SNP	Chr	Gene	ATT	Wu et al.	HLASSO	d-GWASelect
rs4846770	1	*MIA3*		•		
rs903228	2	*ASB3*		•		
rs906766	3	*MED12L*		•		•
rs2562544	5	*SLC1A3*		•		
rs383830	5	*TMEM157*		•		
rs449650	5	*TMEM157*		•		
rs6922269	**6**	*MTHFD1L*		•		
rs1333049	**9**	*CDKN2A/2B*	•	•	•	•
rs6490506	13	*ZMYM2*		•		
rs8055236	16	*CDH13*		•		
rs889595	16	*FOXF1*		•		
rs41537748	19	*IL28A*		•		•
rs688034	22	*SEZ6L*		•		

SNPs in bold were replicated by other studies as collected in the National Human Genome Research Institute's Catalog of Published GWAS.

Table 2.12: List of SNPs selected by different methods for T1D

SNP	Chr	Gene	ATT	Wu et al.	HLASSO	d-GWASelect
rs2269241	**1**	*PGM1*		•		
rs41515647	1	*ST6GALNAC5*		•		•
rs6679677	**1**	*RSBN1*	•	•	•	•
rs3087243	**2**	*CTLA4*		•		
rs17388568	4	*ADAD1*		•		•
rs330483	4	*ADAM29*				•
rs1025039	5	*SPEF2*		•		
rs9273363	**6**	*HLA-DQB1*	•	•	•	•
rs9272346	**6**	*HLA-DQA1*	•	•	•	•
rs411136	6	*SYNGAP1*				•
rs2666236	10	*NRP1*		•		
rs17696736	**12**	*C12orf30*	•	•	•	•
rs1265566	12	*CUX2*				•
rs7398833	12	*CUX2*				•
rs10744777	12	*ALDH2*				•
rs11171739	12	*ERBB3*	•	•	•	•
rs7157296	14	*C14orf138*		•		
rs12708716	**16**	*CLEC16A*	•			•
rs12924729	16	*CLEC16A*	•	•		•
rs7221109	**17**	*TNS4*		•		
rs2542151	**18**	*PTPN2*		•		

SNPs in bold were replicated by other studies as collected in the National Human Genome Research Institute's Catalog of Published GWAS.

Table 2.13: List of SNPs selected by different methods for T2D

SNP	Chr	Gene	ATT	Wu et al.	HLASSO	d-GWASelect
rs4655595	1	*PDE4B*		•		
rs903228	2	*ASB3/LOC129656*		•		•
rs7593730	**2**	*RBMS1*		•		
rs6718526	2	*RBMS1*		•		
rs11688935	2	*GULP1*				•
rs17248501	2	*PARD3B*		•		
rs6846031	4	*VEGFC/NEIL3*				•
rs6872465	5	*PRDM6*				•
rs10806665	6	*THBS2/SMOC2*		•		•
rs1665901	6	*KIAA1553/PDSS2*		•		
rs9465871	6	*CDKAL1*		•		•
rs2389591	7	*TMEM195*				•
rs10435018	7	*CREB5*				•
rs9326506	10	*ZNF239*		•		•
rs7917983	10	*TCF7L2*	•			•
rs7901695	**10**	*TCF7L2*	•	•		•
rs4506565	10	*TCF7L2*	•	•	•	•
rs4132670	10	*TCF7L2*	•			•
rs7077039	10	*TCF7L2*	•	•	•	•
rs1495377	12	*TSPAN8/LGR5*		•		•
rs7961581	**12**	*TSPAN8/LGR5*		•		•
rs2930291	15	*CCDC33*		•		•
rs2903265	15	*ZFAND6*		•		
rs2099106	16	*C16orf72*		•		•
rs8050136	**16**	*FTO*	•	•	•	•
rs543759	18	*PTPRM*		•		
rs6517434	21	*KCNJ6*		•		•

SNPs in bold were replicated by other studies as collected by NHGRI.

Table 2.14: The replication rates for different methods on the WTCCC data

	ATT	Wu et al.	HLASSO	d-GWASelect
total # of distinct loci	19	94	18	60
loci replicated by other studies	13	20	11	13
replication rate	68%	21%	61%	22%

Table 2.15: The AUC achieved by different methods for the WTCCC-T1D data on genetic prediction (two different model sizes were evaluated for the Wu et al. model and the HLASSO model)

	ATT	Wu et al.		HLASSO		d-GWASelect
effective size	5	14	21	4	21	21
AUC	0.7890	0.7869	0.7808	0.7901	0.7829	0.7942

Chapter 3

Sparse Meta-Analysis With High-Dimensional Data

3.1 Introduction

Meta-analysis is commonly used in many scientific areas. By combining multiple data sources, one can achieve higher statistical power, more accurate estimation, and improved reproducibility (Noble, 2006). Traditional meta-analysis methods were designed mainly for low-dimensional data sets. When the number of covariates becomes very large, as in gene expression studies and genome-wide association studies, it is desirable to incorporate variable selection into meta-analysis to improve model interpretation and prediction accuracy.

There exist many variable selection methods, such as LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001) and adaptive-LASSO (Zou, 2006). When raw data are available, these methods can be applied to each study and the selection results can be combined. However, this strategy fails to borrow the information shared among the studies and therefore may be inefficient. There are two approaches to better use of multiple data sets: *integrative analysis* and *meta-analysis*. Integrative analysis pools raw data from multiple studies, while meta-analysis combines summary statistics from

multiple studies. In the context of integrative analysis, Ma et al. (2011) studied variable selection by employing a penalized likelihood method with the group bridge penalty.

In practice, it is rarely possible to obtain the original data due to high cost, IRB restrictions, or unwillingness of investigators to share data. A natural question arises as to whether it is possible to conduct effective variable selection using only summary statistics. In addition, it is unclear how to extract common information shared by different studies while allowing heterogeneity among studies. Furthermore, the high-dimensional nature of -omics studies makes information extraction and model building difficult.

In this article, we propose a new approach, *sparse meta-analysis* (SMA), for variable selection in meta-analysis based solely on summary statistics. To our knowledge, no such method exists in the literature. We show that the SMA estimator can achieve selection consistency and can be as efficient as if the raw data were available. A key feature of the SMA is its flexibility in handling both the homogeneous structure (which assumes that the effects of each covariate are either all zero or all non-zero across studies) and the heterogeneous structure (which allows the effects of each covariate to be partly zero among studies). These two structures are shown in Figure 3.1 in Supplemental Materials. The heterogeneous structure is useful in many settings. For example, there is biological evidence that genetic variants may exhibit on/off effects due to genetic modifiers, environmental exposure or epigenetic mechanisms (Zeisel, 2007).

The rest of the chapter is organized as follows. In Section 3.2, we describe the SMA methods for various forms of summary statistics. In Section 3.3, we study the selection consistency and asymptotic normality of the SMA estimators, considering both the situations that the dimension p is fixed and p is diverging. Section 3.4 contains numerical results. Section 3.5 provides an application to real GWAS studies.

3.2 Sparse Meta-analysis

3.2.1 Data and Models

Suppose that there are K independent studies, with n_k participants in the kth study. The raw data consist of $(y_{ik}, \boldsymbol{x}_{ik}), k = 1, \ldots, K; i = 1, \ldots, n_k$, where y_{ik} is the response for the ith subject in the kth study, and $\boldsymbol{x}_{ik}$ is the corresponding p-vector of covariates. We assume the following linear models

$$y_{ik} = \boldsymbol{x}_{ik}^{\mathrm{T}} \boldsymbol{\beta}_k^0 + \epsilon_{ik}, \quad k = 1, \ldots, K; i = 1, \ldots, n_k,$$

where $\boldsymbol{\beta}_k^0 \equiv (\beta_{1k}^0, \ldots, \beta_{pk}^0)^{\mathrm{T}}$ is a p-vector of regression coefficients for the kth study, $E(\epsilon_{ik}) = 0$ and $Var(\epsilon_{ik}) = \sigma_k^2$. We divide the covariates into two disjoint sets: important set $I = \{j = 1, \ldots, p : \beta_{jk}^0 \neq 0 \text{ for some } k\}$ and unimportant set $U = \{j = 1, \ldots, p : \beta_{jk}^0 = 0 \text{ for all } k = 1, \ldots, K\}$. Our goal is to identify the set I correctly and to estimate the effects of those covariates in I. There are two possible structures of I:

1) Homogeneous structure: for any $j \in I$, $\beta_{jk}^0 \neq 0$ for all $k = 1, \ldots, K$.

2) Heterogeneous structure: for any $j \in I$, $\beta_{jk}^0 \neq 0$ for at least one k.

The homogeneous structure requires each covariate in I to be completely active across all K studies, whereas the heterogeneous structure allows each covariate in I to be partly active among the K studies. The former structure is a special case of the latter. If we treat the regression coefficients associated with a covariate in the K studies as a group, then the homogeneous structure assumes sparsity only at the group level, whereas the heterogeneous structure assumes additional within-group sparsity.

In meta-analysis, the only information available pertains to the summary statistics, in the form of the ordinary least-squares (OLS) estimates $\widetilde{\boldsymbol{\beta}}_k (k = 1, \ldots, K)$. Often, the corresponding variance estimates for individual regression coefficients are also available. In prospectively designed meta-analysis, it is possible to obtain the estimated covariance

matrices $\widetilde{\boldsymbol{V}}_k \equiv \widehat{\mathrm{Cov}}(\widetilde{\boldsymbol{\beta}}_k)(k = 1, \ldots, K)$.

3.2.2 SMA Estimators

We now introduce the SMA approach to variable selection and effect estimation. We first focus on the heterogeneous structure. We propose to minimize the following objective function with respect to $\boldsymbol{\beta} \equiv (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_K^{\mathrm{T}})^{\mathrm{T}}$

$$Q_n(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K) \equiv \sum_{k=1}^{K} (\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)^{\mathrm{T}} \widetilde{\boldsymbol{V}}_k^{-1} (\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) + \lambda \sum_{j=1}^{p} \left(\sum_{k=1}^{K} w_{jk} |\beta_{jk}| \right)^{\frac{1}{2}}, \qquad (3.1)$$

where λ is a tuning parameter, and w_{jk} is a user-specified penalty weight for $|\beta_{jk}|$. The penalty term in (3.1) is different from the commonly used group penalty function. The group lasso penalty (Yuan and Lin, 2006) employs the L_2 norm for the coefficients of each covariate and thus a group is entirely selected or dropped. Our penalty term adopts the group L_1 norm, which allows the selection of individual covariates within groups. This penalty shares the spirit of the group bridge penalty (Huang et al., 2009) wherein $\gamma = \frac{1}{2}$. The main difference is that we assign a weight w_{jk} to each coefficient whereas Huang et al. (2009) used an un-weighted norm $\sum_{k=1}^{K} |\beta_{jk}|$. In the next section, we show that the presence of these weights is crucial to the oracle property of the new estimators. We suggest to set $w_{jk} = |\widetilde{\beta}_{jk}|^{-1}$. Let $\widehat{\boldsymbol{\beta}} \equiv (\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}, ..., \widehat{\boldsymbol{\beta}}_K^{\mathrm{T}})^{\mathrm{T}}$ be the minimizer of (3.1).

For the homogeneous structure, we propose to use a common penalty weight for all the coefficients associated with a given covariate. That is, we minimize

$$Q_n^*(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K) \equiv \sum_{k=1}^{K} (\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)^{\mathrm{T}} \widetilde{\boldsymbol{V}}_k^{-1} (\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) + \lambda \sum_{j=1}^{p} \left(\sum_{k=1}^{K} w_j |\beta_{jk}| \right)^{\frac{1}{2}}, \qquad (3.2)$$

where $w_j = (\sum_{k=1}^{K} |\widetilde{\beta}_{jk}|/K)^{-1}$ for $j = 1, \ldots, p$. Let $\widehat{\boldsymbol{\beta}}^*$ denote the minimizer of (3.2).

In practice, one may be provided only the diagonal elements of $\widetilde{\boldsymbol{V}}_k^{-1}$, i.e., $\widehat{\mathrm{Var}}(\widetilde{\beta}_{jk})$ $(j = 1, \ldots, p)$, for $k = 1, \ldots, K$. In that case, we construct a working covariance matrix $\widehat{\mathbf{C}}_k \equiv \mathrm{diag}\{\widehat{\mathrm{Var}}(\widetilde{\beta}_{1k}), ..., \widehat{\mathrm{Var}}(\widetilde{\beta}_{pk})\}$ and minimize

$$\sum_{k=1}^{K}(\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)^{\mathrm{T}}\widehat{\mathbf{C}}_k^{-1}(\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) + \lambda\sum_{j=1}^{p}\left(\sum_{k=1}^{K} w_{jk}|\beta_{jk}|\right)^{\frac{1}{2}}. \tag{3.3}$$

We call the solution to (3.3) the SMA-Diag estimator. In some applications, even $\widehat{\mathrm{Var}}(\widetilde{\beta}_{jk})$ may not be available. Then we replace $\widetilde{\boldsymbol{V}}_k$ with the identity matrix and minimize

$$\sum_{k=1}^{K}(\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) + \lambda\sum_{j=1}^{p}\left(\sum_{k=1}^{K} w_{jk}|\beta_{jk}|\right)^{\frac{1}{2}}. \tag{3.4}$$

The solution to (3.4) is called the SMA-Id estimator.

3.2.3 Algorithms

The minimizations of (3.1), (3.3), and (3.4) can be done through similar optimization algorithms. Thus, we focus on the minimization of (3.1). The objective function in (3.1) is not convex in $\boldsymbol{\beta}$ and has a complex nonlinear form. For implementation, we propose to solve the following equivalent problem

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \quad & \sum_{k=1}^{K}(\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)^{\mathrm{T}}\widetilde{\boldsymbol{V}}_k^{-1}(\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) + \lambda_1\sum_{j=1}^{p}\gamma_j + \sum_{j=1}^{p}\gamma_j^{-1}\left(\sum_{k=1}^{K} w_{jk}|\beta_{jk}|\right) \\ \text{subject to} \quad & \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p) \geq 0, \end{aligned} \tag{3.5}$$

where $\lambda_1 > 0$ is a tuning parameter. There is one-to-one correspondence between λ and λ_1, and the proof for the equivalence of (3.1) and (3.5) is given in the Supplemental Materials. Here, we propose an iterative algorithm which alternately minimizes (3.5)

with respect to $\boldsymbol{\beta}$ (or $\boldsymbol{\gamma}$), with $\boldsymbol{\gamma}$ (or $\boldsymbol{\beta}$) fixed at their current values. When $\boldsymbol{\beta}$ is fixed, we can get a closed form solution for $\boldsymbol{\gamma}$. When $\boldsymbol{\gamma}$ is fixed, the minimization can be transformed into an adaptive-LASSO problem and solved by the cyclic coordinate descent algorithm (Friedman et al., 2007).

- Step 1: Initialize $\widehat{\boldsymbol{\beta}}_k^{(0)}$ by the least-squares estimates $\widetilde{\boldsymbol{\beta}}_k$ for all k. Set m=1.
- Step 2: Fix $\widehat{\boldsymbol{\beta}}_k^{(m-1)}, k = 1, \ldots, K$, at their current values and minimize (3.5) with respect to $\boldsymbol{\gamma}$. The solution is $\widehat{\gamma}_j^{(m)} \equiv \left(\sum_{k=1}^K w_{jk}|\widehat{\beta}_{jk}^{(m-1)}|\right)^{\frac{1}{2}} \lambda_1^{-\frac{1}{2}}, j = 1, \ldots, p$.
- Step 3: Fix $\widehat{\gamma}_j^{(m)}, j = 1, \ldots, p$, at their current values and minimize (3.5) with respect to $\boldsymbol{\beta}$. Denote the solution as $\widehat{\boldsymbol{\beta}}_k^{(m)}$, $k = 1, \ldots, K$.
- Step 4: Let $m = m + 1$, and go to Step 2 until convergence.

The tuning parameter λ controls the trade-off between model sparsity and model fit. Motivated by the work of Wang and Leng (2007), we determine the tuning parameter by a modified BIC criterion. Define $SSE_\lambda = \sum_{k=1}^K (\widehat{\boldsymbol{\beta}}_{k,\lambda} - \widetilde{\boldsymbol{\beta}}_k)^{\mathrm{T}} (\widehat{\boldsymbol{\beta}}_{k,\lambda} - \widetilde{\boldsymbol{\beta}}_k)$, where $\widehat{\boldsymbol{\beta}}_{k,\lambda}$ is the estimate of $\boldsymbol{\beta}_k^0$ under λ. Let $q_{\lambda,k}$ be the number of nonzero components of $\widehat{\boldsymbol{\beta}}_{k,\lambda}$. Then the modified BIC is defined as:

$$BIC_\lambda = SSE_\lambda + \sum_{k=1}^K (q_{\lambda,k} \log n_k / n_k). \tag{3.6}$$

One may use other tuning criteria, such as the BIC of Wang and Leng (2007) and the general cross validation (GCV), but criterion (3.6) works well in both simulated and real data analysis. This newly proposed BIC can be shown to be consistent for model selection (see Supplemental Materials).

3.3 Asymptotic Properties

In this section, we study the asymptotic properties of the SMA estimators in terms of selection and estimation. Since the heterogeneous structure is more general than the homogeneous structure, we mainly focus on the former and treat the latter as a special case. Specifically, Theorems 1, 2, 4 and 5 below pertain to the heterogeneous structure only, while the other results pertain to both structures. The proofs are relegated to Appendix 1 and the Supplemental Materials.

Without loss of generality, we assume that the first p_0 covariates are active or partly active. That is, $I = \{1, \ldots, p_0\}$ and $U = \{p_0 + 1, \ldots, p\}$. For $j = 1, \ldots, p_0$, define $\mathcal{M}_j = \{k : \beta_{jk}^0 \neq 0, k = 1, \ldots, K\}$ and $\mathcal{M}_j^c = \{1, \ldots, K\} \backslash \mathcal{M}_j$. Also, define $\mathcal{N} = \{(j,k) : \beta_{jk}^0 = 0, j = 1, \ldots, p; k = 1, \ldots, K\}$. For the penalty weights, define $g_{1n} = \max_{1 \leq j \leq p, 1 \leq k \leq K} w_{jk}$, $g_{2n} = \min\{w_{jk} : (j,k) \in \mathcal{N}\}$. Let $\boldsymbol{X}_k = (\boldsymbol{x}_{1k}, \ldots, \boldsymbol{x}_{n_k k})^{\mathrm{T}}$, $n = \sum_{k=1}^K n_k$, and $\boldsymbol{\beta}^0$ be the vector stacked from $\boldsymbol{\beta}_1^0, \ldots, \boldsymbol{\beta}_K^0$. For any index set $\mathcal{B}$, we denote its cardinality by $|\mathcal{B}|$. For any square matrix $\boldsymbol{H} = \{h_{ij}\}$, denote its smallest and largest eigenvalues by $\tau_{\min}(\boldsymbol{H})$ and $\tau_{\max}(\boldsymbol{H})$, respectively.

We study the asymptotic properties of the SMA estimators under two scenarios: (1) p is fixed; (2) p is diverging. (Hereafter, we use p_n instead of p when we wish to emphasize that the dimension diverges with n.) The theoretical analysis requires regularity conditions on the covariate design and the data distribution. We state below the basic assumptions and give additional conditions for the diverging p_n in Section 3.3.2:

(A1) $n_k/n \to \nu_k \in (0,1)$ as $n \to \infty$.

(A2) $n_k^{-1} \boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{X}_k \to \{\boldsymbol{\Sigma}^{(k)}\}^{-1}$, where $\boldsymbol{\Sigma}^{(k)}$ is positive definite, for $k = 1, \ldots, K$.

(A3) There exist constants $0 < b < b' < \infty$ such that

$$b \leq \tau_{\min}(n_k^{-1} \boldsymbol{X}_k' \boldsymbol{X}_k) \leq \tau_{\max}(n_k^{-1} \boldsymbol{X}_k' \boldsymbol{X}_k) \leq b', \quad \text{for } k = 1, \ldots, K.$$

(A4) There exist constants $0 < r_1 < r_2 < \infty$ such that

$$r_1 \leq \min\{|\beta_{jk}^0|, 1 \leq j \leq p_0, k \in \mathcal{M}_j\} \leq \max\{|\beta_{jk}^0|, 1 \leq j \leq p_0; k \in \mathcal{M}_j\} \leq r_2.$$

(A5) For $k = 1, \ldots, K$, $\sigma_k^2 < \infty$.

3.3.1 Fixed Dimension

Theorem 1. *Consider the SMA estimator under the heterogeneous structure. Suppose that conditions (A1)∼(A5) hold. Let* $t_{1n} = \max\{w_{jk} : 1 \leq j \leq p_0, k \in \mathcal{M}_j\}$, *and* $t_{2n} = \min\{w_{jk} : 1 \leq j \leq p_0, k \in \mathcal{M}_j\}$. *Define* $a_n = t_{1n} t_{2n}^{-\frac{1}{2}}$, *and* $b_n = t_{1n}^2 t_{2n}^{-\frac{3}{2}}$. *If* $\lambda n^{-\frac{1}{2}} a_n = O_p(1)$, $\lambda n^{-1} b_n = o_p(1)$ *and* $\lambda n^{-1} = o_p(1)$, *then* $\widehat{\boldsymbol{\beta}}$ *is* $\sqrt{n}$*-consistent.*

Theorem 1 states that if λ and the weights are chosen properly, the SMA estimator $\widehat{\boldsymbol{\beta}}$ will converge to the true $\boldsymbol{\beta}^0$ at the rate of $n^{-\frac{1}{2}}$. It also implies that the penalty weights for the nonzero coefficients are critical to the consistency. The following corollary gives examples and theoretical justifications on constructing the weights based on the OLS estimators.

Corollary 1. *For the heterogeneous structure, let* $w_{jk} = |\widetilde{\beta}_{jk}|^{-1}$ *for* $j = 1, \ldots, p$ *and* $k = 1, \ldots, K$. *For the homogeneous structure, let* $w_j = (\sum_{k=1}^K |\widetilde{\beta}_{jk}|/K)^{-1}$ *for* $j = 1, \ldots, p$. *Suppose that conditions (A1)∼(A5) hold. If* $\lambda n^{-\frac{1}{2}} = O_p(1)$, *then both* $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{\boldsymbol{\beta}^*}$ *are* $\sqrt{n}$*-consistent.*

Theorem 2. *Consider the heterogeneous structure, and let* $\widehat{\beta}_{jk,\lambda}$ *denote the estimator of* β_{jk}^0 *under* λ. *Assume that the conditions of Theorem 1 hold. If* $n^{-\frac{1}{2}} \lambda g_{2n} g_{1n}^{-\frac{1}{2}} \to \infty$, *then* $P(\widehat{\beta}_{jk,\lambda} = 0) \to 1$ *for any* $(j, k) \in \mathcal{N}$.

Theorem 2 shows that the penalty weights for both the nonzero and zero coefficients contribute to the identification of the zero coefficients. It also says that the proposed method can asymptotically estimate all zero coefficients exactly at zero.

Corollary 2. *Consider both the heterogenous and homogeneous structures. Let* $\widehat{\beta}^*_{jk,\lambda}$ *be the estimator of* β^0_{jk} *under* λ *for the homogeneous structure. Assume that* w_{jk} *and* w_j *are chosen as in Corollary 1. If* $\lambda n^{-\frac{1}{2}} = O_p(1)$ *and* $\lambda n^{-\frac{1}{4}} \to \infty$, *then* $P(\widehat{\beta}_{jk,\lambda} = 0) \to 1$ *and* $P(\widehat{\beta}^*_{jk,\lambda} = 0) \to 1$ *for all* $(j,k) \in \mathcal{N}$.

Corollary 2 suggests that λ must be at a rate n^δ with $1/4 < \delta \leq 1/2$ in order to achieve the selection consistency. Together with Theorem 1 and Corollary 1, it implies that the proposed method can effectively distinguish zero coefficients from nonzero coefficients and estimate the nonzero coefficients consistently.

Define $\mathcal{A}_k = \{j : \beta^0_{jk} \neq 0\}$ and $\mathcal{A}^c_k = \{j : \beta^0_{jk} = 0\}$ for $k = 1, \ldots, K$, and $\mathcal{A} = \bigcup_{k=1}^K \mathcal{A}_k$. Let $\boldsymbol{\beta}^0_{\mathcal{A}_k}$ be the subvector of $\boldsymbol{\beta}^0_k$ corresponding to $\mathcal{A}_k$, and $\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k}$ be the submatrix of $\boldsymbol{\Sigma}^{(k)}_k$ corresponding to $\mathcal{A}_k$. Let $\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k\mathcal{A}^c_k}$ denote the submatrix of $\boldsymbol{\Sigma}^{(k)}$ with its row indices corresponding to $\mathcal{A}_k$ and column indices corresponding to $\mathcal{A}^c_k$. Define $\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}^c_k\mathcal{A}_k} = \{\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k\mathcal{A}^c_k}\}^{\mathrm{T}}$. Other subvectors and submatrices are to be understood in the same fashion. The following theorem gives conditions under which the nonzero estimators behave as a multivariate random normal variable.

Theorem 3. *Consider the heterogeneous structure. Suppose that conditions (A1)*$\sim$*(A5) hold. If* $\lambda n^{-\frac{1}{2}} a_n = o_p(1)$, $\lambda n^{-1} b_n = o_p(1)$, $n^{-\frac{1}{2}} \lambda g_{2n} g_{1n}^{-\frac{1}{2}} \to \infty$, *and* $\lambda n^{-1} = o_p(1)$, *then* $\sqrt{n_k}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k} - \boldsymbol{\beta}^0_{\mathcal{A}_k}) \to N\left(0, \sigma^2_k \{\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k} - \boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k\mathcal{A}^c_k}[\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}^c_k}]^{-1}\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}^c_k\mathcal{A}_k}\}\right)$ *for* $k = 1, ..., K$. *For the homogeneous structure, a similar result holds.*

Remark 1. *If the penalty weights are chosen as in Corollary 1, then the conditions in Theorem 3 can be simplified to* $n^{-\frac{1}{2}}\lambda = o_p(1)$ *and* $n^{-\frac{1}{4}}\lambda \to \infty$. *If we set* $\lambda = O_p(n^\omega)$ *with* $\frac{1}{4} < \omega < \frac{1}{2}$, *then the asymptotic normality (for both structures) is guaranteed.*

3.3.2 Diverging Dimension

In modern scientific studies, the dimension can be very large. In this section, we allow $p_n \to \infty$ as $n \to \infty$. We also alow $p_0 \to \infty$. Recall that $g_{1n} = \max_{1 \leq j \leq p, 1 \leq k \leq K} w_{jk}$,

$g_{2n} = \min\{w_{jk} : (j,k) \in \mathcal{N}\}$, and $t_{1n} = \max\{w_{jk} : 1 \le j \le p_0, k \in \mathcal{M}_j\}$. We make the following assumptions:

(C1) $p_n^2/n \to 0$.

(C2) $\lambda\sqrt{t_{1n}} = O_p(p_n)$.

(C3) (i) $\lambda/\sqrt{n} \to 0$; (ii) $\lambda(np_n)^{-1/2} g_{2n} g_{1n}^{-\frac{1}{2}} \to \infty$.

(C4) For each k, $n_k^{-1/2} \max_{1 \le i \le n_k} \boldsymbol{x}_{ik}^{\mathrm{T}} \boldsymbol{x}_{ik} \to 0$.

Condition (C1) specifies the diverging rate of p_n. Under this condition, Yohai and Maronna (1979) established the root-$\sqrt{n/p_n}$ consistency of the OLS estimators. Condition (C2) is needed for the proof of consistency. Condition (C3) is needed for the selection consistency and asymptotic normality. Similar conditions were considered by Huang et al. (2009). Condition (C4) ensures that the predictor matrix has a reasonably good behavior and is needed for the proof of the asymptotic normality. Similar conditions were considered by Yohai and Maronna (1979) and Huang et al. (2009).

Theorem 4. *Consider the heterogeneous structure. Assume that conditions (A1)∼(A5) and (C1) and (C2) hold. Then* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(\sqrt{p_n/n})$.

Corollary 3. *Assume that conditions (C1) and (C2) hold. For the heterogeneous structure, let* $w_{jk} = |\widetilde{\beta}_{jk}|^{-(2+\vartheta)}$ *for some* $\vartheta > 0$, $j = 1, ..., p_n$ *and* $k = 1, ..., K$. *For the homogeneous structure, let* $w_j = (\sum_{k=1}^{K} |\widetilde{\beta}_{jk}|^{2+\vartheta}/K)^{-1}$ *for* $j = 1, ..., p_n$ *and* $k = 1, ..., K$. *Then* $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{\boldsymbol{\beta}}^*$ *are* $\sqrt{n/p_n}$*-consistent.*

The above corollary shows that it is possible to find proper penalty weights and tuning parameter λ that satisfy the requirements of Theorem 4. Next, we show that the SMA is consistent in model selection when p_n is diverging.

Theorem 5. *Consider the heterogeneous structure. Assume that conditions (A1)∼(A5) and (C1)∼(C3) hold. Then* $P(\widehat{\beta}_{jk,\lambda} = 0) \to 1$ *for any* $(j,k) \in \mathcal{N}$.

Corollary 4. *Consider both the heterogenous and homogeneous structures. Let* $\widehat{\beta}^*_{jk,\lambda}$ *be as defined in Corollary 2. Assume that* w_{jk} *and* w_j *are chosen as in Corollary 3. Then* $P(\widehat{\beta}_{jk,\lambda} = 0) \to 1$ *and* $P(\widehat{\beta}^*_{jk,\lambda} = 0) \to 1$ *for all* $(j,k) \in \mathcal{N}$. *In the special case of* $\vartheta = 1$, *condition (C3)(ii) is simplified to* $\lambda n^{\frac{1}{4}} p_n^{-\frac{5}{4}} \to \infty$.

The above theorem and corollary indicate that, even if p_n grows large, we can still estimate the zero coefficients exactly at zero. Finally, we investigate the asymptotic normality of the estimators.

Theorem 6. *Assume that conditions (A1)∼(A5) and (C1)∼(C4) hold. Let* $\boldsymbol{\gamma}_n$ *be a vector of length* p_0 *with norm 1. Define* $s^2_{n,k} = \sigma_k^2 \boldsymbol{\gamma}_n^{\mathrm{T}} (\boldsymbol{X}^{\mathrm{T}}_{\mathcal{A}_k} \boldsymbol{X}_{\mathcal{A}_k}/n_k)^{-1} \boldsymbol{\gamma}_n$ *for* $k = 1, \ldots, K$. *Then*

$$\sqrt{n_k} s_{n,k}^{-1} \boldsymbol{\gamma}_n^{\mathrm{T}} (\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k} - \boldsymbol{\beta}^0_{\mathcal{A}_k}) \to N(0,1).$$

Remark 2. *It can be easily verified that, if the penalty weights are chosen as in Corollary 3 and* $\lambda = O_p(p_n)$, *then all the conditions in Theorem 6 are satisfied.*

3.4 Numerical Studies

We conducted extensive simulation studies to compare the following methods: 1) the method that utilizes the raw data along with a group penalty, i.e.,

$$\min_{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K} \sum_{k=1}^{K} \left\{ {n_k}^{-1} \sum_{i=1}^{n_k} (y_{ik} - \boldsymbol{x}_{ik}^{\mathrm{T}} \boldsymbol{\beta}_k)^2 \right\} + \lambda \sum_{j=1}^{p} \left(\sum_{k=1}^{K} w_{jk} |\beta_{jk}| \right)^{\frac{1}{2}} \tag{3.7}$$

with $w_{jk} = |\widetilde{\beta}_{jk}|^{-1}$, which we call the Gold method; 2) the SMA; 3) the SMA-Diag; 4) the aLASSO-U method, which first applies the adaptive-LASSO to each of the K studies to obtain K models and then takes the union of the K models as the final model; 5) the aLASSO-I method, which is the same as the aLASSO-U except that it takes the intersection of the K models. Methods 4) and 5) represent two

natural ways of combining variable selection results obtained from individual studies. The SMA-Id method was considered only for the $p > n$ case. To measure the performance of each method, we calculated the estimated model sparsity, defined as $\sum_{k=1}^{K} |\widehat{\mathcal{M}}^{(k)}|$ for the heterogeneous structure and $K^{-1}\sum_{k=1}^{K} |\widehat{\mathcal{M}}^{(k)}|$ for the homogeneous structure, where $\widehat{\mathcal{M}}^{(k)}$ denotes the set of selected covariates for the kth study. We further calculated the correct_0 rate $M^{-1}\sum_{m=1}^{M} \widehat{\pi}_m$ and the incorrect_0 rate $M^{-1}\sum_{m=1}^{M} \widehat{\zeta}_m$, where $\widehat{\pi}_m = \sum_{k=1}^{K}\sum_{j=1}^{p} I(\widehat{\beta}_{jk} = 0)I(\beta_{jk}^0 = 0)/\sum_{k=1}^{K}\sum_{j=1}^{p} I(\beta_{jk}^0 = 0)$, $\widehat{\zeta}_m = \sum_{k=1}^{K}\sum_{j=1}^{p} I(\widehat{\beta}_{jk} = 0)I(\beta_{jk}^0 \neq 0)/\sum_{k=1}^{K}\sum_{j=1}^{p} I(\beta_{jk}^0 \neq 0)$ for the mth simulation, $I(\cdot)$ is the indicator function, and M is the total number of simulations. Similar criteria were used by Wang (2009). We set $M = 100$. To assess the prediction accuracy, we further simulated K data sets, $(\widetilde{y}_{ik}, \widetilde{\boldsymbol{x}}_{ik})$ for $k = 1, \ldots, K$ and $i = 1, \ldots, n_k$, and calculated the prediction error $K^{-1}\sum_{k=1}^{K}\left\{n_k^{-1}\sum_{i=1}^{n_k}(\widetilde{y}_{ik} - \widetilde{\boldsymbol{x}}_{ik}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_k)^2\right\}$. Our work was motivated mainly by GWAS, which typically involve thousands of subjects. Thus, we focused on large simulation studies. Our methods are also applicable to small-scale datasets and the corresponding results are given in Supplemental Materials 3.8.4.

3.4.1 Small Dimensions

We simulated 5 studies, each with 50 covariates. The sample sizes for the five studies are 2000, 1800, 1600, 1400 and 1200. The 50 covariates were evenly divided into 10 blocks, and each block followed a multivariate normal distribution with mean being the unit vector and covariance matrix following either the compound symmetry or the auto-regressive correlation structure. The variances of the 50 covariates followed a Uniform(1.0, 2.0) distribution.

We first considered the heterogeneous structure, in which 10 covariates were active or partly active among the studies. For $j = 1, \ldots, p$, the nonzero coefficients for the jth covariate were simulated under the random effects model $\beta_{jk}^0 = (-1)^j \times R_j$ for $k \in \mathcal{M}_j$,

where $R_j \sim N(\mu_j, \sigma_j^2)$, $\mu_j \sim \text{Uniform}(0.5, 1.0)$, and $\sigma_j = \mu_j/2$. The choice of (μ_j, σ_j) essentially allows β_{jk}^0 $(k = 1, ..., K)$ to have the same sign for a given j, and this is a reasonable assumption. The results are shown in Table 3.1.

The results suggest that the performance of the SMA is comparable to that of the Gold method. This is consistent with our theoretical results that the two methods are asymptotically equivalent. The SMA-Diag appears to have good performance, although it is less efficient than the SMA because the SMA-Diag omits part of the information in the estimated covariance matrix. The aLASSO-U clearly over-selects models, and its parameter estimation error is almost always higher than those of the SMA and SMA-Diag. The prediction errors for the first 4 methods are close, although the SMA tends to be slightly better than the aLASSO-U. The aLASSO-I method always under-selects models, resulting in grossly inflated estimation errors and prediction errors. Indeed, under the heterogeneous structure, the aLASSO-I is bound to fail by its design. The results for the homogeneous structure (Table 3.6 in Supplemental Materials) show similar patterns.

3.4.2 Large Dimensions

We increased the dimension to 200 but kept the other conditions the same as before. For the SMA and SMA-Diag, we let $w_{jk} = |\widetilde{\beta}_{jk}|^{-3}$ for the heterogeneous structure and $w_j = (\sum_{k=1}^K |\widetilde{\beta}_{jk}|^3/K)^{-1}$ for the homogeneous structure ($j = 1, \ldots, p$ and $k = 1, \ldots, K$). The results are shown in Table 3.2 and Table 3.7. The SMA method continues to perform similarly to the Gold method. The SMA-Diag is less efficient than the SMA but still possesses the desirable model sparsity. The aLASSO-U model includes a large number of noise covariates, while the aLASSO-I tends to miss important covariates. We further tested the five methods under smaller effect sizes with $\mu_j \sim \text{Uniform}(0.2, 0.5)$ for all j, and the SMA and SMA-Diag still perform well (see Tables

Table 3.1: Comparison of the SMA and other methods under the heterogeneous structure for $p = 50$

Method	$\sum_{k=1}^{5} \lvert\widehat{\mathcal{M}}^{(k)}\rvert$	Correct_0(%)	Incorrect_0(%)	$\lvert\lvert\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\rvert\rvert^2/5$	Pred_err
Correlation structure: Auto-regressive ($\rho = 0.3$)					
Gold	40.61	99.9	1.44	0.004	1.003
SMA	40.74	99.8	1.51	0.004	1.003
SMA-Diag	41.09	99.7	1.39	0.005	1.004
aLASSO-U	115.65	64.3	0	0.006	1.006
aLASSO-I	32.60	100.0	20.54	1.169	2.793
Correlation structure: Auto-regressive ($\rho = 0.6$)					
Gold	40.66	99.9	1.59	0.004	1.003
SMA	42.65	98.9	1.34	0.005	1.003
SMA-Diag	46.23	97.3	1.20	0.009	1.009
aLASSO-U	130.05	57.4	0	0.008	1.006
aLASSO-I	32.85	100.0	20.00	1.145	2.756
Correlation structure: Compound symmetry ($\rho = 0.3$)					
Gold	40.66	99.9	1.46	0.004	1.003
SMA	41.02	99.7	1.44	0.004	1.003
SMA-Diag	41.37	99.6	1.37	0.005	1.004
aLASSO-U	126.50	59.1	0	0.007	1.006
aLASSO-I	32.75	100.0	20.17	1.120	2.728
Correlation structure: Compound symmetry ($\rho = 0.5$)					
Gold	40.72	99.8	1.66	0.005	1.003
SMA	42.33	99.1	1.32	0.006	1.003
SMA-Diag	43.46	98.6	1.39	0.008	1.007
aLASSO-U	140.85	52.2	0	0.010	1.007
aLASSO-I	32.80	99.9	20.32	1.136	2.742

Table 3.2: Comparison of the SMA and other methods under the heterogeneous structure for $p = 200$

Method	$\sum_{k=1}^{5} \lvert\widehat{\mathcal{M}}^{(k)}\rvert$	Correct_0(%)	Incorrect_0(%)	$\lvert\lvert\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\rvert\rvert^2/5$	Pred_err
Correlation structure: Auto-regressive($\rho = 0.3$)					
Gold	41.01	99.9	1.59	0.004	1.009
SMA	41.26	99.9	2.32	0.005	1.010
SMA-Diag	42.37	99.8	2.44	0.007	1.013
aLASSO-U	436.80	58.7	0	0.014	1.022
aLASSO-I	32.85	100.0	20.07	1.097	2.640
Correlation structure: Auto-regressive ($\rho = 0.6$)					
Gold	40.91	99.9	1.56	0.004	1.009
SMA	43.13	99.7	2.51	0.007	1.012
SMA-Diag	48.05	99.1	3.12	0.018	1.027
aLASSO-U	442.05	58.2	0	0.016	1.022
aLASSO-I	32.70	100.0	20.24	1.123	2.668
Correlation structure: Compound symmetry ($\rho = 0.3$)					
Gold	40.83	100.0	1.54	0.005	1.010
SMA	41.22	99.9	2.46	0.006	1.012
SMA-Diag	42.43	99.7	2.54	0.009	1.015
aLASSO-U	474.30	54.8	0	0.016	1.025
aLASSO-I	32.85	100.0	19.98	1.086	2.625
Correlation structure: Compound symmetry ($\rho = 0.5$)					
Gold	40.71	100.0	1.63	0.005	1.009
SMA	42.58	99.7	2.63	0.007	1.011
SMA-Diag	46.79	99.3	2.78	0.012	1.019
aLASSO-U	477.85	54.4	0	0.018	1.024
aLASSO-I	32.85	100.0	20.17	1.094	2.633

3.9 and 3.10 in Supplemental Materials).

3.4.3 $p > n$

When p is greater than n, variable selection in meta-analysis becomes extremely challenging because the OLS estimators cannot be obtained. The LASSO estimators may sometimes be available for each study because the LASSO can handle '$p > n$' via the coordinate descent algorithm (Friedman et al., 2007). Although our theory does not cover the $p > n$ situation, we evaluated the empirical performance of our SMA-Id method in this challenging case.

We set the sample sizes of the 5 studies to 600, 500, 600, 500 and 400. The dimension p is set to 1000. To create high variability, we let the variances of the p covariates follow the Uniform(1.0, 3.0) distribution. We first ran the LASSO for each study to obtain the estimates for all β_{jk}^0. Since the corresponding variance estimates are usually not provided, we applied the SMA-Id to the LASSO estimates to conduct variable selection. We compared the SMA-Id to the LASSO-U and LASSO-I, which are the same as the aLASSO-U and aLASSO-I except for the use of the LASSO instead of the aLASSO. The results are shown in Table 3.3 and Table 3.8. The LASSO-U and LASSO-I yield either extremely large models or highly skewed parameter estimates. The SMA-Id tends to select a model that is quite close to the true model and improve parameter estimation and prediction accuracy. These results suggest that, even though the 'feed-in' summary statistics are biased, the SMA-Id can transform them into a more informative model for both variable selection and prediction.

3.4.4 Sub-group Structures

In the above studies, we considered both the homogeneous and heterogeneous structures. Sometimes, one may encounter a structure that is a hybrid of the two, which

Table 3.3: Comparison under the heterogeneous structure for $p > n$

Method	$\sum_{k=1}^{5} \lvert\widehat{\mathcal{M}}^{(k)}\rvert$	Correct_0(%)	Incorrect_0(%)	$\lvert\lvert\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\rvert\rvert^2/5$	Pred_err
Correlation structure: Auto-regressive ($\rho = 0.3$)					
SMA-Id	38.81	100.0	5.34	0.060	1.109
LASSO-U	963.9	81.4	0.07	0.066	1.120
LASSO-I	30.85	100.0	24.76	1.360	3.794
Correlation structure: Auto-regressive ($\rho = 0.6$)					
SMA-Id	38.65	100.0	5.80	0.077	1.124
LASSO-U	1058.20	79.5	0.10	0.083	1.135
LASSO-I	30.9	100.0	24.63	1.399	3.832
Correlation structure: Compound symmetry ($\rho = 0.3$)					
SMA-Id	38.50	100.0	6.10	0.086	1.127
LASSO-U	1134.50	77.9	0.10	0.093	1.142
LASSO-I	30.60	100.0	25.37	1.400	3.795
Correlation structure: Compound symmetry ($\rho = 0.5$)					
SMA-Id	38.34	100.0	6.59	0.131	1.144
LASSO-U	1272.70	75.2	0.10	0.139	1.166
LASSO-I	29.40	100.0	28.29	1.532	3.912

NOTE: $p = 1000$ and the sample sizes range from 400–600.

we call the sub-group structure. More detail on this structure can be found in Section 3.8.3 of the Supplemental Materials. It is shown in Tables 3.11-3.13 that with a proper choice of the penalty weight w_{jk} our methods are applicable to this structure.

3.4.5 Small Sample Sizes

We further tested our methods under smaller sample sizes, such as $n_k = 100$ for $k = 1, ..., 5$. We varied the heterogeneous structure to make it more challenging for variable selection (see Figure 3.2 in Supplemental Materials). Our methods continue to perform better than the methods based on the LASSO or the adaptive-LASSO (see Tables 3.14 and 3.15 in Supplemental Materials).

3.5 Real Data Analysis

We consider the Multi-Ethnic Study of Atherosclerosis (MESA) study (Bild et al., 2002) and the Coronary Artery Risk Development in Young Adults (CARDIA) study (Friedman et al., 1988). A major goal of these two studies is to investigate genetic factors that influence the development of cardiovascular diseases.

The MESA study contains 1568 whites and 2249 blacks, and the CARDIA study contains 1261 whites and 1422 blacks. The phenotype of interest is *lipidpc1*, a continuous variable that was derived from the principal component analysis of several lipid-related measurements and represents a summary of lipid traits in a person. A previous study revealed a number of single nucleotide polymorphisms (SNPs) that are associated with lipidpc1 (Avery et al., 2011). Since a genetic study rarely reports more than 100 SNPs with summary statistics, we focus on the top 100 SNPs to conduct variable selection. We include environmental covariates, i.e., center, age, gender and the top 10 principle components for ancestry, in all of our analysis.

We analyze the data of black and white samples separately, which amounts to 4 studies, i.e., MESA-Black, MESA-White, CARDIA-Black and CARDIA-White. We adopt the heterogeneous structure. We compare the SMA to the Gold method, the SMA-Diag and the aLASSO-U method. The aLASSO-I method is omitted because of its inherent incompatibility with the heterogeneous structure. The results are shown in Tables 3.4 and 3.5. (Information on the Gold model can be found in Table 3.16 in Supplemental Materials.) As expected, both the SMA and the SMA-Diag yield sparse models that are largely consistent with the Gold model, and many of the selected covariates overlap in these three models. In contrast, the aLASSO-U selects a much larger model which is difficult to interpret. The estimated regression coefficients vary across different studies and sometimes shrink to zero among the 4 studies. This is consistent with the heterogeneous structure.

Table 3.4: Variable selection in the CARDIA and MESA followed by prediction in the ARIC study

	Model sizes				
Method	CARDIA-Blk.	MESA-Blk.	CARDIA-Whi.	MESA-Whi.	Pred_err
Gold	11	5	13	5	2.75
SMA	10	9	17	14	3.06
SMA-Diag	11	8	16	12	3.04
aLASSO-U	70	70	70	70	3.41

To investigate the prediction performance of these methods, we consider prediction in the Atherosclerosis Risk in Communities (ARIC) study (The ARIC Investigators, 1989). This study includes 8907 white and 2532 black participants. Since the ARIC cohort is demographically similar to the MESA cohort, we apply the regression coefficient estimates obtained from the MESA-Black to predict the ARIC-Black, and the MESA-White to predict the ARIC-White. The prediction errors are shown in Table 3.4. The Gold method achieves the highest prediction accuracy, and the SMA and SMA-Diag both beat the aLASSO-U method. The observed difference between the Gold method and the SMA can be explained by the extremely small effects of some of the selected SNPs and the relatively insufficient sample sizes of our training data sets compared to many existing GWAS. The prediction error of the SMA-Diag is slightly lower than that of the SMA, and this is likely due to two reasons: first, the correlations among the 100 SNPs are quite mild and thus the two methods should perform quite similarly; second, the underlying distribution of the ARIC data is somewhat different from that of the MESA data and thus some stochastic deviation is expected. The larger prediction error of the aLASSO-U method is likely due to its large model size, demonstrating the importance of variable selection for predicting genetic risks.

The above prediction uses external data sets as the testing data, which may not have the same distributions as the training data sets. To further investigate the prediction power, we divide the ARIC-white data into 3 subsets according to the 3 centers from

Table 3.5: The SMA model based on the CARDIA and MESA studies

	Estimates of regression coefficients			
SNP	CARDIA-Black	MESA-Black	CARDIA-White	MESA-White
rs4420638	0.102	0.029	-0.352	-0.088
rs660240		0.221	0.333	0.107
rs445925	-0.191		-0.364	-0.293
rs6511720	0.259	0.187	0.248	0.212
rs1713222		-0.155	-0.162	-0.142
rs1168132		0.095		
rs2954021			0.174	0.035
rs9302635	-0.128		-0.121	-0.033
rs9534262	-0.067	-0.107		0.018
rs2307039				0.059
rs1203576		-0.065	-0.103	
rs5752792	0.041		0.123	
rs3916027			-0.015	-0.075
rs9348432	0.098			
rs7552841				-0.040
rs12401642	0.326		0.208	
rs873870			-0.017	-0.031
rs816060		0.212	0.056	0.085
rs2479409	-0.178		-0.093	
rs4689667				0.095
rs2760537	0.068			
rs7493705		0.062	-0.260	
rs10445281			0.163	
rs10743370			0.042	

NOTE: Zero estimates are left blank.

which the data were collected and treat each subset as a separate study. We then conduct variable selection on the three studies. Each study is randomly divided into 10 folds, with one fold as the testing data and the remaining 9 folds as the training data. The resulting prediction error is commonly called the 10-fold cross-validation prediction error. We find that the 10-fold cross-validation prediction errors for the Gold, SMA, SMA-Diag and aLASSO-U are 2.57, 2.77, 2.78, 2.89, respectively. Overall, the SMA and SMA-Diag are preferable to the aLASSO-U method in genetic risk prediction.

3.6 Discussion

Variable selection in meta-analysis is important in improving model interpretability and prediction accuracy. We have developed a class of variable selection methods that can make use of either the raw data or the summary statistics and thus have greatly broadened the applicability of variable selection in meta-analysis. When all the raw data are available, the Gold method should be the first choice. When only summary statistics are available, either the SMA or the SMA-Diag is recommended, depending on whether the estimated covariance matrix is provided or not. The SMA-Id is most useful in the '$p > n$' case.

Our theoretical work and simulation studies demonstrate that, under proper conditions, summary statistics can almost replace raw data for variable selection in meta-analysis, at least in the asymptotic sense. In a similar spirit (although not in the context of variable selection), Lin and Zeng (2010) showed that summary statistics are asymptotically equivalent to the original data for meta-analysis under the fixed-effects model. We conjecture that if the first part of equation (3.1) is replaced by other suitable risk functions, the asymptotic equivalence between summary statistics and raw data will continue to hold. If the penalty term in equation (3.1) is replaced by other reasonable penalties, the asymptotic equivalence may hold as well. Which risk function

or penalty term to use should depend on the scientific nature of the problem and the modelling aims of the investigators.

We have focused on the continuous outcome. We are currently extending our methods to discrete and censored data. The results will be communicated in separate reports.

3.7 Supplemental Materials

3.7.1 Performance under the homogeneous structure

We investigated the performance of the SMA and other methods under the homogeneous structure. Data were simulated in a similar manner as described in Section 3.4.1-3.4.3 of the main text, except that we replaced the heterogeneous structure with the homogeneous structure. Specifically, we let all the five studies share 10 common active covariates, as shown in the upper panel of Figure 3.1 in Supplemental Materials. To accommodate the homogeneous structure, we adjusted the Gold method by solving the following problem:

$$\min_{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K} \sum_{k=1}^{K} \left\{ {n_k}^{-1} \sum_{i=1}^{n_k} (y_{ik} - \boldsymbol{x}_{ik}^{\mathrm{T}} \boldsymbol{\beta}_k)^2 \right\} + \lambda \sum_{j=1}^{p} \left(\sum_{k=1}^{K} w_j |\beta_{jk}| \right)^{\frac{1}{2}} \tag{3.8}$$

with $w_j = (\sum_{k=1}^{K} |\widetilde{\beta}_{jk}|/K)^{-1}$. The results in Tables S1a-S1c show that the SMA has a similar performance as the Gold method and performs better than the other methods.

3.7.2 Performance of the SMA under small effect sizes

We assessed the performance of the SMA and other methods under relatively small effect sizes. The true model is as described in Section 3.4.2 of the main text, but we let $\mu_j \sim$Uniform(0.2, 0.5). The results in Tables 3.9 and 3.10 show that the SMA and the SMA-Diag methods still perform well under small effect sizes.

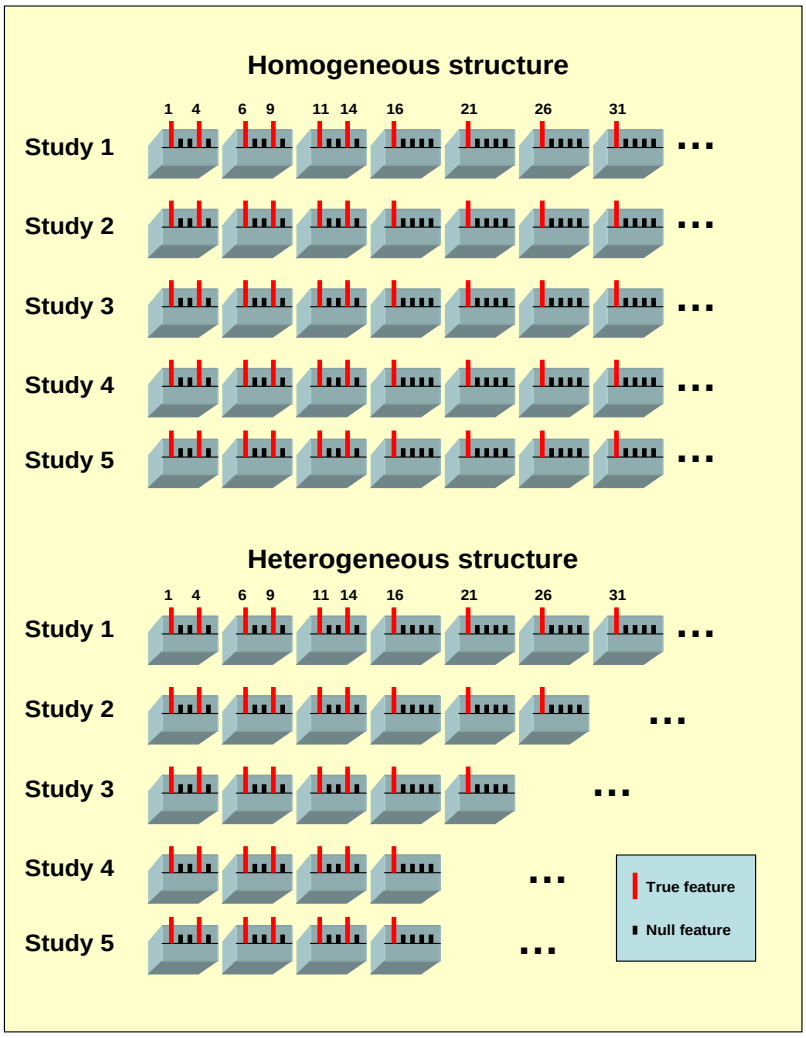

Figure 3.1: True models under homogeneous and heterogeneous structures. Only blocks harboring important covariates are shown.

Table 3.6: Comparisons of the SMA and other methods under the homogeneous structure for $p = 50$

Method	$\sum_{k=1}^{5} \lvert\widehat{\mathcal{M}}^{(k)}\rvert/5$	Correct_0(%)	Incorrect_0(%)	$\lvert\lvert\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\rvert\rvert^2/5$	Pred_err
Correlation structure: Auto-regressive ($\rho = 0.3$)					
Gold	9.96	100.0	0.40	0.005	1.005
SMA	10.01	99.9	0.42	0.005	1.005
SMA-Diag	10.04	99.8	0.36	0.005	1.006
aLASSO-U	24.45	63.9	0	0.008	1.009
aLASSO-I	9.26	100.0	7.40	0.438	1.647
Correlation structure: Auto-regressive ($\rho = 0.6$)					
Gold	9.97	100.0	0.40	0.005	1.005
SMA	10.17	99.5	0.28	0.005	1.005
SMA-Diag	10.41	98.9	0.18	0.008	1.010
aLASSO-U	26.05	59.9	0	0.009	1.010
aLASSO-I	9.24	100.0	7.60	0.455	1.674
Correlation structure: Compound symmetry ($\rho = 0.3$)					
Gold	9.98	100.0	0.24	0.005	1.006
SMA	10.02	99.9	0.20	0.005	1.006
SMA-Diag	10.04	99.8	0.22	0.006	1.007
aLASSO-U	25.65	60.9	0	0.008	1.010
aLASSO-I	9.38	100.0	6.20	0.371	1.531
Correlation structure: Compound symmetry ($\rho = 0.5$)					
Gold	9.98	100.0	0.20	0.005	1.006
SMA	10.11	99.7	0.18	0.006	1.006
SMA-Diag	10.23	99.4	0.26	0.008	1.010
aLASSO-U	27.96	55.1	0	0.010	1.010
aLASSO-I	9.27	100.0	7.30	0.414	1.588

Table 3.7: Comparisons of the SMA and other methods under the homogeneous structure for $p = 200$

Method	$\sum_{k=1}^{5} \lvert\widehat{\mathcal{M}}^{(k)}\rvert/5$	Correct_0(%)	Incorrect_0(%)	$\lvert\lvert\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\rvert\rvert^2/5$	Pred_err
Correlation structure: Auto-regressive ($\rho = 0.3$)					
Gold	10.01	100.0	0.28	0.005	1.011
SMA	10.03	100.0	0.28	0.005	1.011
SMA-Diag	10.25	99.9	0.24	0.006	1.013
aLASSO-U	90.13	57.83	0	0.016	1.026
aLASSO-I	9.34	100.0	6.60	0.330	1.486
Correlation structure: Auto-regressive ($\rho = 0.6$)					
Gold	10.01	100.0	0.28	0.005	1.011
SMA	10.11	99.9	0.26	0.005	1.011
SMA-Diag	12.74	98.6	0.04	0.015	1.024
aLASSO-U	86.65	59.7	0	0.017	1.024
aLASSO-I	9.34	100.0	6.60	0.350	1.504
Correlation structure: Compound symmetry ($\rho = 0.3$)					
Gold	10.02	100.0	0.30	0.005	1.011
SMA	10.04	100.0	0.26	0.005	1.011
SMA-Diag	10.36	99.8	0.16	0.007	1.014
aLASSO-U	93.93	55.8	0	0.017	1.026
aLASSO-I	9.39	100.0	6.20	0.329	1.474
Correlation structure: Compound symmetry ($\rho = 0.5$)					
Gold	10.00	100.0	0.30	0.006	1.011
SMA	10.06	100.0	0.22	0.006	1.011
SMA-Diag	11.32	99.3	0	0.011	1.019
aLASSO-U	97.19	54.1	0	0.020	1.027
aLASSO-I	9.35	100.0	6.60	0.359	1.532

Table 3.8: Comparisons of the SMA-Id and other methods under the homogeneous structure for $p > n$

Method	$\sum_{k=1}^{5} \lvert\widehat{\mathcal{M}}^{(k)}\rvert/5$	Correct_0(%)	Incorrect_0(%)	$\lvert\lvert\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\rvert\rvert^2/5$	Pred_err
Correlation structure: Auto-regressive ($\rho = 0.3$)					
SMA-Id	9.66	100.0	3.42	0.078	1.140
LASSO-U	226.09	78.2	0	0.080	1.141
LASSO-I	8.73	100.0	12.70	0.762	2.400
Correlation structure: Auto-regressive ($\rho = 0.6$)					
SMA-Id	9.65	100.0	3.54	0.094	1.148
LASSO-U	236.10	77.2	0	0.096	1.150
LASSO-I	8.66	100.0	13.40	0.832	2.489
Correlation structure: Compound symmetry ($\rho = 0.3$)					
SMA-Id	9.63	100.0	3.72	0.105	1.153
LASSO-U	253.00	75.5	0	0.107	1.161
LASSO-I	8.69	100.0	13.10	0.827	2.439
Correlation structure: Compound symmetry ($\rho = 0.5$)					
SMA-Id	9.56	100.0	4.40	0.152	1.173
LASSO-U	280.65	72.7	0	0.153	1.185
LASSO-I	8.48	100.0	15.30	0.972	2.603

NOTE: $p = 1000$ and the sample sizes range from 400–600.

Table 3.9: Comparisons under the homogeneous structure with small effect sizes

Method	$\sum_{k=1}^{5} \lvert\widehat{\mathcal{M}^{(k)}}\rvert/5$	Correct_0(%)	Incorre_0(%)	$\lvert\lvert\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\rvert\rvert^2/5$	Pred_err
Correlation structure: Auto-regressive ($\rho = 0.3$)					
Gold	9.96	100.0	0.84	0.005	1.011
SMA	9.99	100.0	0.66	0.005	1.011
SMA-Diag	10.23	99.9	0.54	0.006	1.013
aLASSO-U	106.60	49.2	0	0.018	1.029
aLASSO-I	8.78	100.0	12.20	0.147	1.217
Correlation structure: Auto-regressive ($\rho = 0.6$)					
Gold	9.98	100.0	0.60	0.005	1.011
SMA	10.15	99.9	0.52	0.005	1.011
SMA-Diag	12.62	98.6	0.22	0.014	1.023
aLASSO-U	101.17	52.0	0	0.020	1.027
aLASSO-I	8.64	100.0	13.60	0.162	1.240
Correlation structure: Compound symmetry ($\rho = 0.3$)					
Gold	9.97	100.0	0.74	0.005	1.011
SMA	10.07	99.9	0.54	0.005	1.011
SMA-Diag	10.28	99.8	0.54	0.007	1.014
aLASSO-U	106.48	49.2	0	0.019	1.029
aLASSO-I	8.78	100.0	12.30	0.142	1.210
Correlation structure: Compound symmetry ($\rho = 0.5$)					
Gold	9.95	100.0	0.92	0.006	1.011
SMA	10.07	99.9	0.66	0.006	1.011
SMA-Diag	11.27	99.3	0.26	0.011	1.019
aLASSO-U	105.50	49.7	0	0.023	1.029
aLASSO-I	8.52	100.0	14.90	0.176	1.258

NOTE: $p = 200$ and the sample sizes range from 1200–2000.

Table 3.10: Comparisons under the heterogeneous structure with small effect sizes

Method	$\sum_{k=1}^{5} \lvert\widehat{\mathcal{M}}^{(k)}\rvert$	Correct_0(%)	Incorrect_0(%)	$\lvert\lvert\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\rvert\rvert^2/5$	Pred_err
Correlation structure: Auto-regressive ($\rho = 0.3$)					
Gold	40.40	99.9	3.31	0.005	1.010
SMA	40.05	99.9	5.63	0.006	1.012
SMA-Diag	40.88	99.8	6.10	0.009	1.016
aLASSO-U	530.30	49.0	0	0.017	1.026
aLASSO-I	31.15	100.0	24.22	0.289	1.433
Correlation structure: Auto-regressive ($\rho = 0.6$)					
Gold	40.61	99.9	3.66	0.005	1.010
SMA	42.94	99.5	5.90	0.011	1.017
SMA-Diag	46.41	99.1	7.59	0.022	1.032
aLASSO-U	504.1	51.7	0.02	0.019	1.025
aLASSO-I	30.1	100.0	26.66	0.305	1.447
Correlation structure: Compound symmetry ($\rho = 0.3$)					
Gold	40.62	99.9	3.32	0.005	1.011
SMA	39.73	99.8	7.00	0.011	1.018
SMA-Diag	40.66	99.7	7.39	0.014	1.022
aLASSO-U	526.70	49.4	0	0.018	1.027
aLASSO-I	31.35	100.0	23.63	0.277	1.413
Correlation structure: Compound symmetry ($\rho = 0.5$)					
Gold	40.49	99.9	3.61	0.006	1.010
SMA	42.98	99.5	5.83	0.009	1.013
SMA-Diag	45.31	99.3	6.61	0.015	1.022
aLASSO-U	518.05	50.3	0	0.020	1.026
aLASSO-I	30.35	100.0	26.44	0.304	1.447

NOTE: $p = 200$ and the sample sizes range from 1200–2000.

3.7.3 Performance of the SMA under the sub-group structure

We first provide a motivating example for the sub-group structure. Suppose that we are analyzing 5 studies for a certain disease and two of them belong to a sub-category of the disease while the other three studies belong to another sub-category of the disease. This is particularly common in psychiatric diseases, where one major disease may include many sub-categories. Thus, the 5 studies can be treated as two subgroups based on their clinical information. For this situation, it is reasonable to assume that, in addition to the active covariates shared by all subgroups, each subgroup has its own set of active covariates.

Next, we describe simulation studies for the sub-group structure. Assume that there are H subgroups in the K studies, and let $\mathcal{S}_h$ denote the set of studies in the hth subgroup. In our simulation studies, we let $K = 5$, $H = 2$, $\mathcal{S}_1 = \{1, 2\}$ and $\mathcal{S}_2 = \{3, 4, 5\}$. That is, we let the 5 studies contain 2 subgroups, where subgroup 1 consists of studies 1 and 2, and subgroup 2 consists of studies 3, 4 and 5. The set of covariates with nonzero coefficients in subgroup 1 was set to $\{1, 4, 6, 9, 11, 14, 16, 21, 26, 31\}$, and the set of covariates with nonzero coefficients in subgroup 2 was set to $\{1, 4, 6, 9, 11, 14, 16, 21\}$. In other words, covariates 1, 4, 6, 9, 11, 14, 16 and 21 are active in both subgroups, while covariates 26 and 31 are active only in subgroup 1. We consider $p = 50, 200$, and 1000. The objective function for the subgroup structure was the same as equation (3.1) in the main text, except that the penalty weights were chosen as following:

$$w_{jk} = \left\{ \sum_{h=1}^{H} \left[I(k \in \mathcal{S}_h) |\mathcal{S}_h|^{-1} \sum_{l \in \mathcal{S}_h} |\widetilde{\beta}_{jl}|^{\alpha} \right] \right\}^{-1}, j = 1, ..., p, \text{ and } k = 1, ..., K,$$

where $|\mathcal{S}_h|$ is the cardinality of $\mathcal{S}_h$. The α is chosen to be 3 for $p = 200$, and 1 otherwise. The results are shown in Tables 3.11-3.13. It can be seen that our methods have a good chance of identifying the true model and tend to outperform the aLASSO-U and

Table 3.11: Comparisons under the subgroup structure for $p = 50$

Method	$\sum_{k=1}^{5} \lvert\widehat{\mathcal{M}}^{(k)}\rvert$	Correct_0(%)	Incorrect_0(%)	$\lvert\lvert\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\rvert\rvert^2/5$	Pred_err
Correlation structure: Auto-regressive ($\rho = 0.3$)					
Gold	43.83	100.0	0.50	0.004	1.002
SMA	44.03	99.9	0.45	0.004	1.002
SMA-Diag	44.13	99.9	0.39	0.005	1.003
aLASSO-U	119.95	63.1	0	0.007	1.006
aLASSO-I	36.85	100.0	16.25	0.944	2.401
Correlation structure: Auto-regressive ($\rho = 0.6$)					
Gold	43.91	100.0	0.32	0.004	1.002
SMA	44.62	99.6	0.41	0.004	1.002
SMA-Diag	46.43	98.8	0.30	0.008	1.007
aLASSO-U	127.35	59.5	0	0.008	1.006
aLASSO-I	37.35	100.0	15.11	0.906	2.345
Correlation structure: Compound symmetry ($\rho = 0.3$)					
Gold	43.87	100.0	0.50	0.004	1.002
SMA	44.22	99.8	0.41	0.004	1.003
SMA-Diag	44.14	99.8	0.52	0.005	1.004
aLASSO-U	127.40	59.5	0	0.008	1.006
aLASSO-I	36.90	100.0	16.20	0.932	2.372
Correlation structure: Compound symmetry ($\rho = 0.5$)					
Gold	43.83	100.0	0.48	0.005	1.002
SMA	44.70	99.6	0.36	0.005	1.002
SMA-Diag	45.26	99.3	0.50	0.007	1.006
aLASSO-U	136.75	55.0	0	0.010	1.006
aLASSO-I	36.85	100.0	16.32	0.935	2.375

NOTE: The sample sizes range from 1200–2000.

aLASSO-I (or LASSO-U and LASSO-I).

3.7.4 Performance of the SMA under smaller sample sizes

We assessed the performance of our methods when the sample sizes are relatively small. We considered two situations, '$p < n$' and '$p > n$'.

We let $n_k = 100$ for $k = 1, ..., 5$, and $p = 50$. The true model followed the heterogeneous structure II, as shown in Figure 3.2 (upper panel). The nonzero coefficients were simulated under a Uniform(0, 3.0). The results are shown in Table 3.14.

Table 3.12: Comparisons under the subgroup structure for $p = 200$

Method	$\sum_{k=1}^{5} \lvert\widehat{\mathcal{M}}^{(k)}\rvert$	Correct_0(%)	Incorrect_0(%)	$\lvert\lvert\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\rvert\rvert^2/5$	Pred_err
Correlation structure: Auto-regressive ($\rho = 0.3$)					
Gold	44.02	100.0	0.43	0.004	1.010
SMA	44.47	99.9	0.36	0.004	1.010
SMA-Diag	45.97	99.8	0.27	0.006	1.012
aLASSO-U	457.15	56.8	0	0.015	1.025
aLASSO-I	36.95	100.0	16.14	0.934	2.414
Correlation structure: Auto-regressive ($\rho = 0.6$)					
Gold	43.90	100.0	0.52	0.004	1.010
SMA	45.67	99.8	0.43	0.005	1.010
SMA-Diag	65.25	97.8	0.05	0.017	1.026
aLASSO-U	453.05	57.2	0	0.017	1.024
aLASSO-I	36.85	100.0	16.25	0.952	2.439
Correlation structure: Compound symmetry ($\rho = 0.3$)					
Gold	44.04	100.0	0.52	0.004	1.009
SMA	44.37	99.9	0.36	0.004	1.009
SMA-Diag	46.74	99.7	0.39	0.006	1.013
aLASSO-U	501.95	52.1	0	0.017	1.026
aLASSO-I	36.80	100.0	16.43	0.960	2.444
Correlation structure: Compound symmetry ($\rho = 0.5$)					
Gold	43.92	100.0	0.59	0.005	1.009
SMA	45.29	99.9	0.25	0.005	1.010
SMA-Diag	53.36	99.0	0.14	0.011	1.018
aLASSO-U	494.95	52.8	0	0.019	1.026
aLASSO-I	36.90	100.0	16.14	0.937	2.407

NOTE: The sample sizes range from 1200–2000.

Table 3.13: Comparisons under the subgroup structure for $p > n$

Method	$\sum_{k=1}^{5} \lvert\widehat{\mathcal{M}}^{(k)}\rvert$	Correct_0(%)	Incorrect_0(%)	$\lvert\lvert\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\rvert\rvert^2/5$	Pred_err
Correlation structure: Auto-regressive ($\rho = 0.3$)					
SMA-Id	42.57	100.0	3.25	0.070	1.133
LASSO-U	1037.10	80.0	0	0.071	1.132
LASSO-I	35.6	100.0	19.1	1.121	3.236
Correlation structure: Auto-regressive ($\rho = 0.6$)					
SMA-Id	42.38	100.0	3.68	0.087	1.142
LASSO-U	1092.90	78.8	0	0.086	1.142
LASSO-I	35.2	100.0	20.00	1.219	3.383
Correlation structure: Compound symmetry ($\rho = 0.3$)					
SMA-Id	42.44	100.0	3.55	0.097	1.147
LASSO-U	1174.50	77.2	0	0.096	1.151
LASSO-I	35.15	100.0	20.2	1.227	3.364
Correlation structure: Compound symmetry ($\rho = 0.5$)					
SMA-Id	42.12	100.0	4.27	0.141	1.162
LASSO-U	1320.80	74.2	0	0.141	1.174
LASSO-I	33.65	100.0	23.6	1.418	3.575

NOTE: $p = 1000$ and the sample sizes range from 400–600.

We also let $n_k = 100$ for $k = 1, ..., 5$, and $p = 1000$. The true model followed the heterogeneous structure III, as shown in Figure 3.2 (lower panel) in Supplemental Materials. The nonzero coefficients were simulated under a Uniform(1.0, 3.0). The results are shown in Table 3.15.

3.7.5 Supplementary Table for real data analysis

In Table 3.16, we show the models selected by the Gold method for the CARDIA and the MESA.

3.7.6 Supplementary Proofs

Proof for the equivalence of equations (3.1) **and** (3.5)

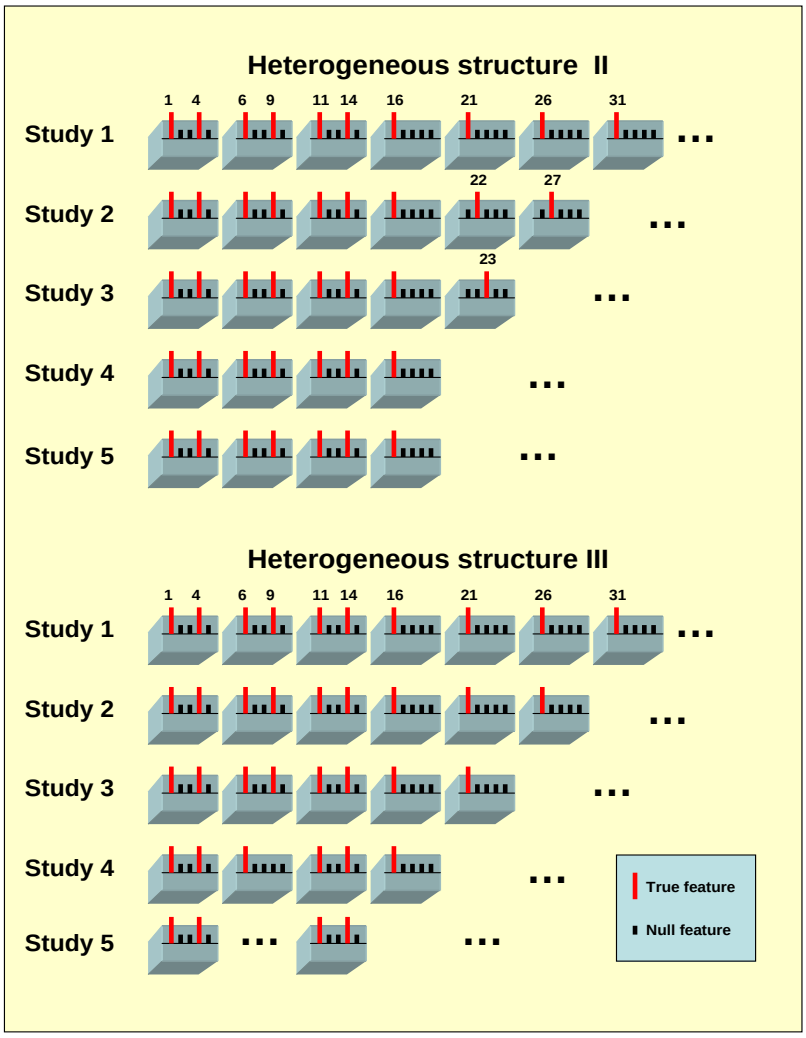

Figure 3.2: True models under heterogeneous structures II and III. Only blocks that harbor important covariates are shown.

Table 3.14: Comparisons of the SMA and other methods under small sample sizes and the heterogeneous structure II

Method	$\sum_{k=1}^{5} \lvert\widehat{\mathcal{M}}^{(k)}\rvert$	Correct_0(%)	Incorrect_0(%)	$\lvert\lvert\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\rvert\rvert^2/5$	Pred_err
Correlation structure: Auto-regressive ($\rho = 0.6$)					
Gold	41.63	98.6	5.59	0.099	1.132
SMA	60.22	89.5	6.61	0.229	1.238
aLASSO-U	183.90	31.6	0.37	0.255	1.266
aLASSO-I	27.25	100.0	33.63	7.281	11.983
Correlation structure: Compound symmetry ($\rho = 0.5$)					
Gold	40.84	98.8	6.32	0.112	1.132
SMA	57.13	90.8	7.49	0.224	1.234
aLASSO-U	187.05	30.1	0.20	0.269	1.284
aLASSO-I	26.00	100.0	36.83	7.855	13.156

NOTE: $p = 50$ and sample sizes are equal to 100.

Table 3.15: Comparisons of the SMA-Id and other methods under small sample sizes and the heterogeneous structure III

Method	$\sum_{k=1}^{5} \lvert\widehat{\mathcal{M}}^{(k)}\rvert$	Correct_0(%)	Incorrect_0(%)	$\lvert\lvert\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\rvert\rvert^2/5$	Pred_err
Correlation structure: Auto-regressive ($\rho = 0.6$)					
SMA-Id	37.79	100.0	0	0.396	1.774
LASSO-U	980.45	81.0	0	0.498	1.870
LASSO-I	20.25	100.0	45.51	14.550	32.234
Correlation structure: Compound symmetry ($\rho = 0.5$)					
SMA-Id	37.50	100.0	0	0.333	1.675
LASSO-U	898.70	82.6	0	0.419	1.756
LASSO-I	20.15	100.0	45.62	14.555	35.111

NOTE: $p = 1000$ and sample sizes are equal to 100.

Table 3.16: Variable selection for the CARDIA and MESA by the Gold method

Study	Model size	Selected SNPs
CARDIA-Black	11	rs4420638, rs445925, rs6511720, rs1713222 rs9302635, rs9534262, rs5752792, rs9348432 rs12401642, rs2479409, rs2760537
MESA-Black	5	rs660240, rs6511720, rs1713222 rs920184, rs9534262
CARDIA-White	13	rs4420638, rs660240, rs445925, rs6511720 rs1713222, rs2954021, rs9302635, rs920184 rs1203576, rs5752792, rs12401642, rs2479409 rs10445281
MESA-White	5	rs4420638, rs660240, rs445925, rs6511720 rs1713222

By the Cauchy-Schwartz inequality,

$$\begin{aligned} & \sum_{k=1}^{K}(\widetilde{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)^{\mathrm{T}}\widetilde{\boldsymbol{V}}_k^{-1}(\widetilde{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)+\lambda_1\sum_{j=1}^{p}\gamma_j+\sum_{j=1}^{p}\gamma_j^{-1}\left(\sum_{k=1}^{K}w_{jk}|\beta_{jk}|\right) \\ \geq & \sum_{k=1}^{K}(\widetilde{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)^{\mathrm{T}}\widetilde{\boldsymbol{V}}_k^{-1}(\widetilde{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k)+\sum_{j=1}^{p}\left(2\sqrt{\lambda_1\sum_{k=1}^{K}w_{jk}|\beta_{jk}|}\right), \end{aligned}$$

where the equality holds if and only if $\gamma_j=(1/\lambda_1)^{\frac{1}{2}}\left(\sum_{k=1}^{K}w_{jk}|\beta_{jk}|\right)^{\frac{1}{2}}$ for all j. Now, let $\lambda_1=(\lambda/2)^2$. Then the proof is completed.

Proof for the consistency of the proposed BIC criterion

Let $\mathcal{M}$ denote an arbitrary model, $\mathcal{M}_\lambda$ denote the model under λ, and $\mathcal{M}_T$ denote the true model. We say that $\mathcal{M}$ is an under-fitted model if $\mathcal{M}\not\supset\mathcal{M}_T$, and over-fitted if $\mathcal{M}\supset\mathcal{M}_T$ and $\mathcal{M}\neq\mathcal{M}_T$. Further define

$$\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}}\equiv\mathrm{argmin}_{\{\boldsymbol{\beta}_k\in R^p:\beta_{k,j}=0,\forall j\notin\mathcal{M}\}}(\boldsymbol{\beta}_k-\widetilde{\boldsymbol{\beta}}_k)^T(\boldsymbol{\beta}_k-\widetilde{\boldsymbol{\beta}}_k). \tag{3.9}$$

Note, in general, $\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}}\neq\widetilde{\boldsymbol{\beta}}_k$ because of the constraint that $\beta_{k,j}=0,\forall j\notin\mathcal{M}$.

Suppose that λ_0 yields the true model, λ_L yields an under-fitted model, and λ_H yields an over-fitted model. We wish to prove that with high probability, $BIC_{\lambda_L} > BIC_{\lambda_0}$, and $BIC_{\lambda_H} > BIC_{\lambda_0}$. We first prove the former.

Because (I) both $\widehat{\boldsymbol{\beta}}_{k,\lambda_0}$ and $\widetilde{\boldsymbol{\beta}}_k$ are consistent for $\boldsymbol{\beta}_k^0$ and (II) $q_{\lambda_0,k}\log(n_k)/n_k \to 0$, it is easy to verify that $BIC_{\lambda_0} = o_p(1)$. Next, for λ_L,

$$\begin{aligned}
BIC_{\lambda_L} &= \sum_{k=1}^{K}(\widehat{\boldsymbol{\beta}}_{k,\lambda_L} - \widetilde{\boldsymbol{\beta}}_k)^T(\widehat{\boldsymbol{\beta}}_{k,\lambda_L} - \widetilde{\boldsymbol{\beta}}_k) + \sum_{k=1}^{K} q_{\lambda_L,k} \times \log(n_k)/n_k \\
&\geq \sum_{k=1}^{K}(\widehat{\boldsymbol{\beta}}_{k,\lambda_L} - \widetilde{\boldsymbol{\beta}}_k)^T(\widehat{\boldsymbol{\beta}}_{k,\lambda_L} - \widetilde{\boldsymbol{\beta}}_k) \\
&\geq \sum_{k=1}^{K}(\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}_{\lambda_L}} - \widetilde{\boldsymbol{\beta}}_k)^T(\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}_{\lambda_L}} - \widetilde{\boldsymbol{\beta}}_k) \\
&\geq \min_{\mathcal{M}\not\supset\mathcal{M}_T} \sum_{k=1}^{K}(\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}} - \widetilde{\boldsymbol{\beta}}_k)^T(\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}} - \widetilde{\boldsymbol{\beta}}_k) \\
&\to \sum_{k=1}^{K}(\boldsymbol{\beta}_{k,\mathcal{M}}^0 - \boldsymbol{\beta}_k^0)^T(\boldsymbol{\beta}_{k,\mathcal{M}}^0 - \boldsymbol{\beta}_k^0) > 0.
\end{aligned}$$

The first inequality holds trivially. The second inequality holds because of the definition of $\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}_{\lambda_L}}$(see (3.9) for detail). The third inequality also holds trivially. The remaining part holds because I) $\hat{\boldsymbol{\beta}}_{k,\mathcal{M}} \to \boldsymbol{\beta}_{k,\mathcal{M}}^0$ II) $\tilde{\boldsymbol{\beta}}_k \to \boldsymbol{\beta}_k^0$ and III) $\mathcal{M}$ represents an underfitted model.

Next, we prove that $BIC_{\lambda_H} > BIC_{\lambda_0}$. Note that

$$\begin{aligned}
n(BIC_{\lambda_H} - BIC_{\lambda_0}) \geq\ & v_{\max}^{-1}\sum_{k=1}^{K} n_k(\widehat{\boldsymbol{\beta}}_{k,\lambda_H} - \widetilde{\boldsymbol{\beta}}_k)^T(\widehat{\boldsymbol{\beta}}_{k,\lambda_H} - \widetilde{\boldsymbol{\beta}}_k) - \\
& v_{\min}^{-1}\sum_{k=1}^{K} n_k(\widehat{\boldsymbol{\beta}}_{k,\lambda_0} - \widetilde{\boldsymbol{\beta}}_k)^T(\widehat{\boldsymbol{\beta}}_{k,\lambda_0} - \widetilde{\boldsymbol{\beta}}_k) + \\
& v_{\max}^{-1}\sum_{k=1}^{K}(q_{\lambda_H,k} - q_{\lambda_0,k})\log(n_k)
\end{aligned}$$

$$
\begin{aligned}
&\geq v_{\max}^{-1} \sum_{k=1}^{K} n_k (\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}_{\lambda_H}} - \widetilde{\boldsymbol{\beta}}_k)^T (\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}_{\lambda_H}} - \widetilde{\boldsymbol{\beta}}_k) - \\
&\quad v_{\min}^{-1} \sum_{k=1}^{K} n_k (\widehat{\boldsymbol{\beta}}_{k,\lambda_0} - \widetilde{\boldsymbol{\beta}}_k)^T (\widehat{\boldsymbol{\beta}}_{k,\lambda_0} - \widetilde{\boldsymbol{\beta}}_k) + \\
&\quad v_{\max}^{-1} \sum_{k=1}^{K} (q_{\lambda_H,k} - q_{\lambda_0,k}) \log(n_k) \\
&\geq v_{\max}^{-1} \sum_{k=1}^{K} n_k (\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}_{\lambda_H}} - \widetilde{\boldsymbol{\beta}}_k)^T (\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}_{\lambda_H}} - \widetilde{\boldsymbol{\beta}}_k) - \\
&\quad v_{\min}^{-1} \sum_{k=1}^{K} n_k (\widehat{\boldsymbol{\beta}}_{k,\lambda_0} - \widetilde{\boldsymbol{\beta}}_k)^T (\widehat{\boldsymbol{\beta}}_{k,\lambda_0} - \widetilde{\boldsymbol{\beta}}_k) + v_{\max}^{-1} \sum_{k=1}^{K} \log(n_k) \\
&\geq v_{\max}^{-1} \inf_{\mathcal{M} \supset \mathcal{M}_T} \sum_{k=1}^{K} n_k (\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}} - \widetilde{\boldsymbol{\beta}}_k)^T (\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}} - \widetilde{\boldsymbol{\beta}}_k) - \\
&\quad v_{\min}^{-1} \sum_{k=1}^{K} n_k (\widehat{\boldsymbol{\beta}}_{k,\lambda_0} - \widetilde{\boldsymbol{\beta}}_k)^T (\widehat{\boldsymbol{\beta}}_{k,\lambda_0} - \widetilde{\boldsymbol{\beta}}_k) + \\
&\quad v_{\max}^{-1} \sum_{k=1}^{K} \log(n_k) \qquad (3.10)
\end{aligned}
$$

The second inequality holds because of the definition of $\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}_{\lambda_H}}$. The third inequality holds because the considered model is an overfitted model. In (3.10), the first term is $O_p(1)$ because for any $\mathcal{M} \supset \mathcal{M}_T$, $\widehat{\boldsymbol{\beta}}_{k,\mathcal{M}}$ is $\sqrt{n_k}$ consistent; the second term is also $O_p(1)$; the third term goes to infinity. Hence, $BIC_{\lambda_H} > BIC_{\lambda_0}$ with probability tending to 1.

Proof of Theorem 1

By Fan and Li (2001), the existence of a $\sqrt{n}$-consistent local minimizer can be verified if, for an arbitrarily small $\epsilon > 0$, there exists a sufficiently large constant C such that

$$
\liminf_n P \left\{ \inf_{\|\boldsymbol{u}\|=C} Q_n(\boldsymbol{\beta}^0 + n^{-\frac{1}{2}} \boldsymbol{u}) > Q_n(\boldsymbol{\beta}^0) \right\} > 1 - \epsilon, \qquad (3.11)
$$

where $\boldsymbol{u} = (\boldsymbol{u}_1^{\mathrm{T}}, \ldots, \boldsymbol{u}_K^{\mathrm{T}})^{\mathrm{T}}$.

By the definition of Q_n,

$$
\begin{aligned}
& \left\{ Q_n(\boldsymbol{\beta}^0 + n^{-\frac{1}{2}}\boldsymbol{u}) - Q_n(\boldsymbol{\beta}^0) \right\} \\
= & \sum_{k=1}^{K} \left\{ \boldsymbol{u}_k{}^{\mathrm{T}} (n\widetilde{\boldsymbol{V}}_k)^{-1} \boldsymbol{u}_k + 2\boldsymbol{u}_k{}^{\mathrm{T}} (n\widetilde{\boldsymbol{V}}_k)^{-1} [\sqrt{n}(\boldsymbol{\beta}_k^0 - \widetilde{\boldsymbol{\beta}}_k)] \right\} \\
& + \lambda \sum_{j=1}^{p} \left[\sum_{k=1}^{K} w_{jk} |\beta_{jk}^0 + n^{-\frac{1}{2}} u_{jk}| \right]^{\frac{1}{2}} - \lambda \sum_{j=1}^{p_0} \left[\sum_{k \in \mathcal{M}_j} w_{jk} |\beta_{jk}^0| \right]^{\frac{1}{2}} \\
\geq & \sum_{k=1}^{K} \left\{ \boldsymbol{u}_k{}^{\mathrm{T}} (n\widetilde{\boldsymbol{V}}_k)^{-1} \boldsymbol{u}_k + 2\boldsymbol{u}_k{}^{\mathrm{T}} (n\widetilde{\boldsymbol{V}}_k)^{-1} [\sqrt{n}(\boldsymbol{\beta}_k^0 - \widetilde{\boldsymbol{\beta}}_k)] \right\} \\
& + \lambda \sum_{j=1}^{p_0} \left[\sum_{k \in \mathcal{M}_j} w_{jk} |\beta_{jk}^0 + n^{-\frac{1}{2}} u_{jk}| \right]^{\frac{1}{2}} - \lambda \sum_{j=1}^{p_0} \left[\sum_{k \in \mathcal{M}_j} w_{jk} |\beta_{jk}^0| \right]^{\frac{1}{2}} \\
= & \sum_{k=1}^{K} \left\{ \boldsymbol{u}_k{}^{\mathrm{T}} (n\widetilde{\boldsymbol{V}}_k)^{-1} \boldsymbol{u}_k + 2\boldsymbol{u}_k{}^{\mathrm{T}} (n\widetilde{\boldsymbol{V}}_k)^{-1} [\sqrt{n}(\boldsymbol{\beta}_k^0 - \widetilde{\boldsymbol{\beta}}_k)] \right\} \\
& + \lambda \sum_{j=1}^{p_0} \left\{ \left[\sum_{k \in \mathcal{M}_j} w_{jk} |\beta_{jk}^0 + n^{-\frac{1}{2}} u_{jk}| \right]^{\frac{1}{2}} - \left[\sum_{k \in \mathcal{M}_j} w_{jk} |\beta_{jk}^0| \right]^{\frac{1}{2}} \right\} \\
\equiv & A + B,
\end{aligned}
$$

where the first equality holds because $\beta_{jk}^0 = 0$ if (j,k) belongs to the set $\{j > p_0, k = 1, \ldots, K\}$ or the set $\{j = 1, \ldots, p_0,\ \ k \in \mathcal{M}_j^c\}$. We decompose A and B, respectively,

as $A = A_1 + A_2$ and $B = B_1 + B_2 + B_3$, where

$$\begin{aligned}
A_1 &= \sum_{k=1}^{K} \boldsymbol{u}_k{}^{\mathrm{T}} (n\widetilde{\boldsymbol{V}}_k)^{-1} \boldsymbol{u}_k, \quad A_2 = 2\boldsymbol{u}_k{}^{\mathrm{T}} (n\widetilde{\boldsymbol{V}}_k)^{-1} [\sqrt{n}(\boldsymbol{\beta}_k^0 - \widetilde{\boldsymbol{\beta}}_k)], \\
B_1 &= \lambda \sum_{j=1}^{p_0} \sum_{k \in \mathcal{M}_j} \frac{1}{2} \left\{ \sum_{l \in \mathcal{M}_j} w_{jl} |\beta_{jl}^0| \right\}^{-\frac{1}{2}} w_{jk} \mathrm{sgn}(\beta_{jk}^0) n^{-\frac{1}{2}} u_{jk}, \\
B_2 &= \lambda \sum_{j=1}^{p_0} \sum_{k \in \mathcal{M}_j} \sum_{k' \in \mathcal{M}_j} \frac{1}{2}(-\frac{1}{4}) \left\{ \sum_{l \in \mathcal{M}_j} w_{jl} |\beta_{jl}^0| \right\}^{-\frac{3}{2}} w_{jk} w_{jk'} \mathrm{sgn}(\beta_{jk}^0) \mathrm{sgn}(\beta_{jk'}^0) n^{-1} u_{jk} u_{jk'}, \\
B_3 &= \lambda \sum_{j=1}^{p_0} o_p \left(n^{-1} \sum_{k \in \mathcal{M}_j} u_{jk}^2 \right).
\end{aligned}$$

When p is fixed, the OLS estimators are root-n consistent and asymptotically normal, i.e., $||\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0|| = O_p(n_k^{-\frac{1}{2}})$ and $\sqrt{n_k}(\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0) \to_d N\big(\mathbf{0}, \sigma_k^2 \boldsymbol{\Sigma}^{(k)}\big)$ as $n \to \infty$. For each k, we have $n_k \widetilde{\boldsymbol{V}}_k \to_p \sigma_k^2 \boldsymbol{\Sigma}^{(k)}$ and $n_k/n \to \nu_k$ as $n \to \infty$, so $n\widetilde{\boldsymbol{V}}_k \to_p \sigma_k^2 \boldsymbol{\Sigma}^{(k)}/\nu_k$ and $(n\widetilde{\boldsymbol{V}}_k)^{-1} \to_p \nu_k \sigma_k^{-2} \{\boldsymbol{\Sigma}^{(k)}\}^{-1}$. Hence,

$$A_1 \geq 0.5 \sum_{k=1}^{K} \nu_k \sigma_k^{-2} \boldsymbol{u}_k^{\mathrm{T}} \{\boldsymbol{\Sigma}^{(k)}\}^{-1} \boldsymbol{u}_k \geq 0.5 \tau_{\min} \sigma_{\max}^{-2} \sum_{k=1}^{K} \nu_k \|\boldsymbol{u}_k\|^2 \geq 0.5 \tau_{\min} \nu_{\min} \sigma_{\max}^{-2} C^2,$$

where $\sigma_{\max} = \max_{k=1,\ldots,K}\{\sigma_k\}$, and

$$\begin{aligned}
|A_2| &\leq 2 \sum_{k=1}^{K} |\boldsymbol{u}_k^{\mathrm{T}} (n\widetilde{\boldsymbol{V}}_k)^{-1} [\sqrt{n}(\boldsymbol{\beta}_k^0 - \widetilde{\boldsymbol{\beta}}_k)]| \leq 2 \sum_{k=1}^{K} \|\boldsymbol{u}_k\| \cdot \|(n\widetilde{\boldsymbol{V}}_k)^{-1} [\sqrt{n}(\boldsymbol{\beta}_k^0 - \widetilde{\boldsymbol{\beta}}_k)]\| \\
&\leq 2C \sum_{k=1}^{K} \|(n\widetilde{\boldsymbol{V}}_k)^{-1} [\sqrt{n}(\boldsymbol{\beta}_k^0 - \widetilde{\boldsymbol{\beta}}_k)]\| = O_p(1) C.
\end{aligned}$$

Therefore, A_1 is bounded below by a term quadratic in C, and A_2 is uniformly bounded above by a term linear in C.

For each $j = 1, \ldots, p_0$, we have $\sum_{l \in \mathcal{M}_j} w_{jl} |\beta_{jl}^0| \geq t_{2n} \sum_{l \in \mathcal{M}_j} |\beta_{jl}^0| \geq t_{2n} r_1$. This

implies that $\{\sum_{l\in\mathcal{M}_j} w_{jl}|\beta_{jl}^0|\}^{-1} \le (t_{2n}r_1)^{-1}$ for any $1 \le j \le p_0$. Then

$$\begin{aligned}
|B_1| &\le \lambda \sum_{j=1}^{p_0} \sum_{k\in\mathcal{M}_j} \frac{1}{2} \left\{ \sum_{l\in\mathcal{M}_j} w_{jl}|\beta_{jl}^0| \right\}^{-\frac{1}{2}} w_{jk} n^{-\frac{1}{2}} |u_{jk}| \le \frac{1}{2} n^{-\frac{1}{2}} \lambda (t_{2n}r_1)^{-\frac{1}{2}} \sum_{j=1}^{p_0} \sum_{k\in\mathcal{M}_j} w_{jk}|u_{jk}| \\
&\le \frac{1}{2} \lambda n^{-\frac{1}{2}} (t_{2n}r_1)^{-\frac{1}{2}} K^{\frac{1}{2}} t_{1n} \sum_{j=1}^{p_0} \|\boldsymbol{u}_j\| \le \frac{1}{2} p_0 K^{\frac{1}{2}} r_1^{-\frac{1}{2}} \lambda n^{-\frac{1}{2}} (t_{1n} t_{2n}^{-\frac{1}{2}}) C,
\end{aligned}$$

and

$$\begin{aligned}
|B_2| &= \lambda \sum_{j=1}^{p_0} \sum_{k\in\mathcal{M}_j} \sum_{k'\in\mathcal{M}_j} \frac{1}{8} \left\{ \sum_{l\in\mathcal{M}_j} w_{jl}|\beta_{jl}^0| \right\}^{-\frac{3}{2}} w_{jk} w_{jk'} n^{-1} |u_{jk}||u_{jk'}| \\
&\le \frac{1}{8} \lambda n^{-1} (t_{2n}r_1)^{-\frac{3}{2}} \sum_{j=1}^{p_0} \sum_{k\in\mathcal{M}_j} \sum_{k'\in\mathcal{M}_j} w_{jk} w_{jk'} |u_{jk}||u_{jk'}| \le \frac{1}{8} \lambda (t_{2n}r_1)^{-\frac{3}{2}} K t_{1n}^2 \left(\sum_{j=1}^{p_0} \|\boldsymbol{u}_j\| \right)^2 \\
&\le \frac{1}{8} p_0^2 K r_1^{-\frac{3}{2}} \lambda n^{-1} t_{1n}^2 t_{2n}^{-\frac{3}{2}} C^2.
\end{aligned}$$

Note that $B_3 = C^2 \lambda n^{-1} o_p(1)$. Note also that B_1 is bounded above by a term linear in C. Define $a_n = t_{1n} t_{2n}^{-\frac{1}{2}}$ and $b_n = t_{1n}^2 t_{2n}^{-\frac{3}{2}}$. Then A_1 is asymptotically positive and dominates A_2, A_3, B_1, B_2 and B_3, as long as $\lambda n^{-\frac{1}{2}} a_n = O_p(1)$, $\lambda n^{-1} b_n = o_p(1)$ and $\lambda n^{-1} = o_p(1)$. Therefore, as long as the constant C is sufficiently large, A_1 will always dominate the others with an arbitrarily large probability. This implies inequality (3.11), and the proof is completed.

Proof of Corollary 1

For the heterogeneous structure, we only need to show that the conditions in Theorem 1 are satisfied. Note that the OLS estimator is consistent with $\widetilde{\beta}_{jk} - \beta_{jk}^0 = O_p(n^{-1/2})$ for any $j = 1, \ldots, p$ and $k = 1, \ldots, K$. This implies that the weight $w_{jk} = |\beta_{jk}^0|^{-1} + O_p(n^{-1/2})$ for any (j, k) satisfying $1 \le j \le p_0$ and $k \in \mathcal{M}_j$. Then, by definition, $a_n = O_p(1)$ and $b_n = O_p(1)$. Therefore, as long as $\lambda/\sqrt{n} = O_p(1)$, the conditions in Theorem 1 are satisfied.

For the homogeneous structure, we have that $\mathcal{M}_j = \{1, \ldots, K\}$ for any $j = 1, \ldots, p_0$. Let $s_{1n} = \max\{w_j : j = 1, ..., p_0\}$ and $s_{2n} = \min\{w_j : j = 1, ..., p_0\}$. Following the proof of Theorem 1, we can show that, if $\lambda n^{-\frac{1}{2}} s_{1n} s_{2n}^{-\frac{1}{2}} = O_p(1)$, $\lambda n^{-1} s_{1n}^2 s_{2n}^{-\frac{3}{2}} = o_p(1)$ and $\lambda n^{-1} = o_p(1)$, then $\widehat{\boldsymbol{\beta}}^*$ is consistent. Next, we can show that $s_{1n} = O_p(1)$ and $s_{2n} = O_p(1)$. Then Corollary 1 follows.

Proof of Theorem 2

Let $\widehat{\boldsymbol{\beta}}_{k,\lambda} = (\widehat{\beta}_{1k,\lambda}, \ldots, \widehat{\beta}_{pk,\lambda})^{\mathrm{T}}$ for $k = 1, \ldots, K$ and $\widehat{\boldsymbol{\beta}}_{\lambda} = (\widehat{\boldsymbol{\beta}}_{1,\lambda}^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\beta}}_{K,\lambda}^{\mathrm{T}})^{\mathrm{T}}$. For any $(j,k) \in \mathcal{N}$, if $\widehat{\beta}_{jk,\lambda} \neq 0$, then by the KKT conditions, it must be true that

$$0 = n^{-\frac{1}{2}} \frac{\partial Q_n}{\partial \beta_{jk}}\Big|_{\widehat{\boldsymbol{\beta}}_{\lambda}} = 2(n^{-1}\widetilde{\boldsymbol{V}}_k^{-1})_{j\cdot}\sqrt{n}(\widehat{\boldsymbol{\beta}}_{k,\lambda} - \widetilde{\boldsymbol{\beta}}_k) + n^{-\frac{1}{2}}\lambda \frac{1}{2}\left\{\sum_{k'=1}^{K} w_{jk'}|\widehat{\beta}_{jk',\lambda}|\right\}^{-\frac{1}{2}} w_{jk}\mathrm{sgn}(\widehat{\beta}_{jk,\lambda}),$$

where $(\widetilde{\boldsymbol{V}}_k^{-1})_{j\cdot}$ represents the jth row of $\widetilde{\boldsymbol{V}}_k^{-1}$. Because $(n\widetilde{\boldsymbol{V}}_k)^{-1} \to_p \nu_k(\sigma_k^2\Sigma^{(k)})^{-1}$ and, under the conditions in Theorem 1, $\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{\lambda}) = O_p(1)$, the first term on the right side of the above equation is $O_p(1)$. Since $\widehat{\boldsymbol{\beta}}_{\lambda}$ is a root-n consistent estimator of $\boldsymbol{\beta}^0$, we have $|\widehat{\beta}_{jk,\lambda}| \leq r_2 + 1$ for all (j,k) with probability tending to 1. Thus,

$$\begin{aligned} & n^{-\frac{1}{2}}\lambda \frac{1}{2}\left\{\sum_{k'=1}^{K} w_{jk'}|\widehat{\beta}_{jk',\lambda}|\right\}^{-\frac{1}{2}} w_{jk} \geq \lambda n^{-\frac{1}{2}} \frac{1}{2}[K g_{1n}(1+r_2)]^{-\frac{1}{2}} g_{2n} \\ = \; & \frac{1}{2}[K(1+r_2)]^{-\frac{1}{2}} \lambda n^{-\frac{1}{2}} g_{2n} g_{1n}^{-\frac{1}{2}}. \end{aligned} \tag{3.12}$$

Given that $\lambda n^{-\frac{1}{2}} g_{2n} g_{1n}^{-\frac{1}{2}} \to \infty$, the right side of (3.12) $\to \infty$. Therefore, with probability tending to one, either $\widehat{\beta}_{jk} = 0$ for $(j,k) \in \mathcal{N}$ or the sign of (3.12) is equal to the sign of $\widehat{\beta}_{jk}$. The latter contradicts with the fact that $\widehat{\boldsymbol{\beta}}_{\lambda}$ is a minimizer of Q_n. Thus, we must have $P(\widehat{\beta}_{jk} = 0) \to 1$ for any $(j,k) \in \mathcal{N}$. This completes the proof.

Proof of Corollary 2

We first consider the heterogeneous structure. Note that $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) = O_p(1)$. This

implies that, for any $(j,k) \in \{j = 1, \ldots, p_0; k \in \mathcal{M}_j\}$,

$$w_{jk} = |\widetilde{\beta}_{jk}|^{-1} = \frac{1}{|\beta^0_{jk}|} + O_p(n^{-\frac{1}{2}}).$$

For any $(j,k) \in \mathcal{N}$, we have $\sqrt{n}(\widetilde{\beta}_{jk} - 0) = O_p(1)$, which implies that $w_{jk} = |\widetilde{\beta}_{jk}|^{-1} = O_p(n^{\frac{1}{2}})$. Thus, $g_{1n} = O_p(n^{\frac{1}{2}})$ and $g_{2n} = O_p(n^{\frac{1}{2}})$. It follows that $g_{2n} g_{1n}^{-\frac{1}{2}} = O_p(n^{\frac{1}{4}})$. Therefore, the last condition in Theorem 2 can be simplified to that $\lambda n^{-\frac{1}{4}} \to \infty$.

Next, we consider the homogeneous structure. Let $h_{1n} = \max\{w_j : j = p_0 + 1, ..., p\}$ and $h_{2n} = \min\{w_j : j = p_0 + 1, \ldots, p\}$. Following the proof of Theorem 2, we can show that, if $\lambda n^{-\frac{1}{2}} h_{2n} h_{1n}^{\frac{1}{2}} \to \infty$, then $P(\widehat{\beta}^*_{jk,\lambda} = 0) \to 1$ for any $(j,k) \in \mathcal{N}$. In addition, we can show that $h_{1n} = O_p(n^{\frac{1}{2}})$ and $h_{2n} = O_p(n^{\frac{1}{2}})$. Then Corollary 2 follows.

Proof of Theorem 3

Consider the heterogeneous structure. Let m_k be the cardinality of $\mathcal{A}_k$. Since we have shown that, with an arbitrarily large probability, the estimator of $\{\beta^0_{jk} : k = 1, \ldots, K; j \in \mathcal{A}_k^c\}$ must be 0, we can decompose $\widehat{\boldsymbol{\beta}}$ into $\{\widehat{\boldsymbol{\beta}}_{\mathcal{A}}, \mathbf{0}\}$. By the KKT conditions, it must be true that

$$\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{A}}}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = \mathbf{0}. \tag{3.13}$$

This implies that, for $k = 1, ..., K$, we have

$$\mathbf{0} = \{(n_k \widetilde{\boldsymbol{V}}_k)^{-1}\}_{\mathcal{A}_k \mathcal{A}_k}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k} - \widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k}) - \{(n_k \widetilde{\boldsymbol{V}}_k)^{-1}\}_{\mathcal{A}_k \mathcal{A}_k^c} \widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k^c} + n_k^{-1} \mathbf{e}_k, \tag{3.14}$$

where $\mathbf{e}_k$ is a vector of length m_k, with its sth component being $\frac{1}{2}\lambda \left\{\sum_{j \in \mathcal{A}_k} w_{sk} |\widehat{\beta}_{sj}|\right\}^{-\frac{1}{2}}$ $w_{sk} \text{sgn}(\widehat{\beta}_{sk})$. Since $\lambda n^{-\frac{1}{2}} a_n = o_p(1)$, each element of $\sqrt{n_k}(n_k^{-1} \mathbf{e}_k)$ is bounded by $o_p(1)$.

Define $\widetilde{\boldsymbol{F}}^{(k)} = (n_k \widetilde{\boldsymbol{V}}_k)^{-1}$, $\widetilde{\boldsymbol{F}}^{(k)}_{\mathcal{A}_k \mathcal{A}_k} = \{(n_k \widetilde{\boldsymbol{V}}_k)^{-1}\}_{\mathcal{A}_k \mathcal{A}_k}$ and $\widetilde{\boldsymbol{F}}^{(k)}_{\mathcal{A}_k A_k^c} = \{(n_k \widetilde{\boldsymbol{V}}_k)^{-1}\}_{\mathcal{A}_k A_k^c}$.

Then, (3.14) implies that

$$\sqrt{n_k}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k} - \boldsymbol{\beta}^0_{\mathcal{A}_k}) = \sqrt{n}_k(\widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k} - \boldsymbol{\beta}^0_{\mathcal{A}_k}) + \{\widetilde{\boldsymbol{F}}^{(k)}_{\mathcal{A}_k\mathcal{A}_k}\}^{-1}\widetilde{\boldsymbol{F}}^{(k)}_{\mathcal{A}_k\mathcal{A}_k^c}(\sqrt{n_k}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k^c}) + o_p(1).$$

Note that $\sqrt{n_k}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k^c} = O_p(1)$, $\mathrm{Cov}(\sqrt{n_k}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k}) \to_p \sigma_k^2\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k\mathcal{A}_k}$, $\mathrm{Cov}(\sqrt{n}_k\widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k^c}) \to_p \sigma_k^2\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k^c\mathcal{A}_k^c}$ and $\mathrm{Cov}(\sqrt{n_k}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k}, \sqrt{n}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k^c}) \to_p \sigma_k^2\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k\mathcal{A}_k^c}$. Therefore, we have the following results:

(1) In general, if $\widetilde{\boldsymbol{V}}_k$ is an arbitrary symmetric matrix specified by the user, then it is straightforward to show that

$$\sqrt{n_k}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k} - \boldsymbol{\beta}^0_{\mathcal{A}_k}) \to N(0, \sigma_k^2\boldsymbol{S}_k),$$

where $\boldsymbol{S}_k = \boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k} + 2\{\widetilde{\boldsymbol{F}}^{(k)}_{\mathcal{A}_k}\}^{-1}\widetilde{\boldsymbol{F}}^{(k)}_{\mathcal{A}_k\mathcal{A}_k^c}\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k\mathcal{A}_k^c} + \{\widetilde{\boldsymbol{F}}^{(k)}_{\mathcal{A}_k}\}^{-1}\widetilde{\boldsymbol{F}}^{(k)}_{\mathcal{A}_k\mathcal{A}_k^c}\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k^c}\widetilde{\boldsymbol{F}}^{(k)}_{\mathcal{A}_k^c\mathcal{A}_k}\{\widetilde{\boldsymbol{F}}^{(k)}_{\mathcal{A}_k}\}^{-1}$.

(2) If $n_k\widetilde{\boldsymbol{V}}_k \to \sigma_k^2\boldsymbol{\Sigma}^{(k)}$, then $\{\widetilde{\boldsymbol{F}}^{(k)}_{\mathcal{A}_k}\}^{-1}\widetilde{\boldsymbol{F}}^{(k)}_{\mathcal{A}_k\mathcal{A}_k^c} \to_p -\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k\mathcal{A}_k^c}\{\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k^c}\}^{-1}$. Consequently, the asymptotic covariance matrix can be simplified into $\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k} - \boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k\mathcal{A}_k^c}\{\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k^c}\}^{-1}\boldsymbol{\Sigma}^{(k)}_{\mathcal{A}_k^c\mathcal{A}_k}$, which is the oracle asymptotic covariance matrix for those nonzero coefficients.

The proof for the homogeneous structure is similar and thus is omitted.

Chapter 4

Genetic risk prediction by Iterative SCAD-SVM (ISS)

4.1 Introduction

Genome-wide association studies provide a powerful platform for genetic risk prediction of human diseases (Kraft and Hunter, 2009). By harnessing the prediction power of single nucleotide polymorphisms (SNPs), it is anticipated that genetics risk prediction will have a profound impact on disease prevention and clinical practice in the foreseeable future (Collins, 2010).

Genetic risk prediction is a highly challenging task, as many complex human diseases are contributed by a large number of genetic variants, many of which with relatively small effects (Barret et al., 2008; Barret et al., 2009). The task is further complicated by the high correlations among SNPs (commonly referred to as the linkage disequilibrium (LD)), low penetrance of the causal variants, and unknown genetic models of the underlying risk loci. Indeed, the genetic risk prediction in the current literature is far from satisfactory and has triggered a debate on where to look for the "missing heritability" of complex diseases (Manolio et al., 2009).

For genetic risk prediction, using only those SNPs that reach the stringent genome-wide significance level is neither sufficient nor powerful, because other less significant SNPs may still carry nonnegligible predictive power. To improve the prediction power of GWAS, it is natural to consider as many SNPs as possible in the prediction model. One popular strategy is to reap all the SNPs that pass a pre-specified threshold in the univariate screening stage, and then build a prediction model based on the estimated marginal effects of those selected SNPs (Wray et al., 2007). Another common strategy is to simply count how many risk alleles each person carries, and then construct the prediction model by treating the counts of the risk alleles as the single covariate (James et al., 2008; Kang et al., 2010). We call the former as the *Marginal* method, and the latter as the *Count* method.

While these methods incorporate potentially many SNPs in the prediction model, they are associated with several problems. First, it is not clear what threshold should be specified during the screening stage, which hinders their practical use in analyzing real data; second, the Count method relies on the assumption that all the risk alleles in the prediction model bear the same effect sizes, which rarely holds in genetic studies (Evangelou et al., 2007); third, the Marginal method is built upon the marginal regression coefficients, but it is well-known that the marginal effects of covariates can deviate substantially from their joint effects. A better strategy is to consider a relatively large number of SNPs, and then conduct variable selection on those SNPs so that a smaller set of SNPs can be prioritized and the joint effects can be estimated.

Many variable selection methods have been developed in the last two decades, such as the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and the adaptive-LASSO (Zou, 2006), mainly for analyzing data with continuous outcomes. For data with binary outcomes, Zhang et al. (2006) proposed the SCAD support vector machine (SCAD-SVM) that imposes the SCAD penalty on the SVM. It is demonstrated

that the SCAD-SVM outperforms the SVM in both variable selection and prediction.

The SCAD-SVM was designed for the microarray data analysis, whose dimension is much lower than GWAS. To accommodate the extremely large dimension of GWAS, we adopt the Iterative Sure Independence Screening (ISIS) strategy by Fan and Lv (2008). This strategy is composed of two principles: 1) it reduces dimensions by a pre-screening utility before conducting variable selection; 2) it conducts a conditional screening procedure in order to capture important covariates that are marginally uncorrelated with the outcome. In addition to adopting the ISIS strategy, we design a new algorithm for implementing the SCAD-SVM. The original algorithm for the SCAD-SVM (Zhang et al., 2006) requires inversion of a large matrix, which may encounter numerical difficulties when the matrix is ill-conditioned. We adopt the recently developed local-linear approximation algorithm (Zou and Li, 2008) to tackle this issue. Our iterative SCAD-SVM (ISS) provides a novel tool for the task of genetic risk prediction with high dimensions.

The rest of the Chapter is organized as follows. In Section 4.2, we describe our method in detail. In Section 4.3, we conduct a wide array of simulation studies to examine the performance of our method and compare it to other methods. In Section 4.4, we show the performance of our method by applying it to real GWAS studies.

4.2 Methods

The data contain n subjects and p SNPs. For $i = 1, \ldots, n$, let y_i denote the disease outcome (1 for cases, -1 for controls), and x_{ij} denote the genotype of the jth SNP for the ith subject. The genotype of each SNP is represented by the number of minor alleles. We standardize the genotypes of each SNP by its sample standard derivation. The Iterative SCAD-SVM method mainly consists of three steps: 1) marginal hinge loss screening, 2) the SCAD-SVM, 3) conditional hinge loss screening and the SCAD-SVM.

4.2.1 Marginal Hinge Loss Screening

We first conduct the Hinge Loss Screening over all the SNPs. This step aims to prioritize a small set of SNPs based on their marginal significance in predicting the outcome. The hinge loss for the jth SNP is defined as

$$L_j \equiv n^{-1} \sum_{i=1}^{n} [1 - y_i(b_j + w_j \times x_{ij})]_+,$$

where b_j and w_j are unknown parameters, and g_+ represents the nonnegative part of g. For each SNP, we minimize L_j with respect to b_j and w_j to obtain $\widehat{L}_j$. The minimization can be conducted by linear programming.

The ISIS theory suggests to pick a subset of covariates whose cardinality t_0 is at the order of $O(n/\log n)$. We choose t_0 to be the integer part of $n/(8 \log n)$. We select the t_0 SNPs with the smallest $\widehat{L}_j$ to form a set $\mathcal{S}_1$. Next, we apply the SCAD-SVM to $\mathcal{S}_1$ to select important variables.

4.2.2 SCAD-SVM

For the ith subject, let z_i denote the t_0-vector consisting of the genotypes of the t_0 SNPs in $\mathcal{S}_1$. Let β_0 and $\beta = (\beta_1, ..., \beta_{t_0})^{\mathrm{T}}$ be the parameters for the directional vector, and λ be a tuning parameter. We aim to solve

$$\min_{\beta_0, \beta} \left\{ n^{-1} \sum_{i=1}^{n} \left[1 - y_i \left(\beta_0 + \beta^{\mathrm{T}} z_i\right)\right]_+ + \sum_{j=1}^{t_0} p_\lambda(|\beta_j|) \right\},$$

where

$$p_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j| & \text{if } 0 \le |\beta_j| < \lambda \\ \frac{(a^2-1)\lambda^2 - (|\beta_j| - a\lambda)^2}{2(a-1)}, i.e., -\frac{(|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2)}{2(a-1)} & \text{if } \lambda \le |\beta_j| < a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } a\lambda \le |\beta_j| \ , \end{cases}$$

and $a = 3.7$ (Fan and Li, 2001).

Zhang et al. (2006) introduced the Successive quadratic algorithm (SQA) to implement SCAD-SVM. The SQA algorithm involves an intermediate step that requires inversion of a square matrix, which may not be stable when the dimension of the square matrix is high. Instead, we develop a new algorithm, Successive local algorithm (SLA), which works as follows.

As in Zhang et al. (2006), we first approximate the hinge loss by

$$-\sum_{i=1}^{n}\frac{y_i(\beta_0+\beta^{\mathrm{T}}z_i)}{2n}-\frac{1}{2n}\sum_{i=1}^{n}\frac{y_i(\beta_0+\beta^{\mathrm{T}}z_i)}{|y_i-(\tilde{\beta}_0+\tilde{\beta}^{\mathrm{T}}z_i)|}+\frac{1}{4n}\sum_{i=1}^{n}\frac{(\beta_0+\beta^{\mathrm{T}}z_i)^2}{|y_i-(\tilde{\beta}_0+\tilde{\beta}^{\mathrm{T}}z_i)|}$$

up to a constant, where $\tilde{\beta}_0$ and $\tilde{\beta}$ are some arbitrarily chosen initial values. Next, we approximate the SCAD penalty by the local linear approximation (Zou and Li, 2008)

$$p_\lambda(|\beta_j|)\approx p_\lambda(|\tilde{\beta}_j|)+p'_\lambda(|\tilde{\beta}_j|)(|\beta_j|-|\tilde{\beta}_j|).$$

Removing the constant terms, we notice that minimizing the above formation is equivalent to minimizing

$$-\sum_{i=1}^{n}\frac{y_i(\beta_0+\beta^{\mathrm{T}}z_i)}{2n}-\frac{1}{2n}\sum_{i=1}^{n}\frac{y_i(\beta_0+\beta^{\mathrm{T}}z_i)}{|y_i-(\tilde{\beta}_0+\tilde{\beta}^{\mathrm{T}}z_i)|}+\frac{1}{4n}\sum_{i=1}^{n}\frac{(\beta_0+\beta^{\mathrm{T}}z_i)^2}{|y_i-(\tilde{\beta}_0+\tilde{\beta}^{\mathrm{T}}z_i)|}+\sum_{j=1}^{t_0}p'_\lambda(|\tilde{\beta}_j|)|\beta_j|.$$

We then solve this approximated objective function by the cyclic coordinate descent (CCD) algorithm (Friedman et al., 2010). The process is iterated until convergence. The detail is shown in the Appendix 2.

The tuning parameter is determined by a 5-fold cross-validation using the area under the ROC curve (AUC) as the evaluation criterion. The final estimators are $\hat{\beta}_0$ and $\widehat{\beta}$. Excluding covariates with zero estimates, the resulting model is named as $\mathcal{M}_1$.

For each subject, we calculate the liability score pertaining to $\mathcal{M}_1$ (i.e., the inner

product of SNPs in $\mathcal{M}_1$ and their estimated effect sizes), and denote the liability score for the ith subject as $\hat{\xi}_i$. We then use these $\hat{\xi}_i$'s in the following conditional screening.

4.2.3 Conditional Hinge Loss Screening and SCAD-SVM

Assuming that $\mathcal{M}_1$ contains t_1 SNPs, we label the set of the remaining $(p - t_1)$ SNPs as $\overline{\mathcal{M}}_1$. The first step is to screen all the SNPs in $\overline{\mathcal{M}}_1$ to identify a small set of candidate SNPs that are correlated with the outcome Y conditional on $\mathcal{M}_1$. To achieve this, we develop a Conditional Hinge Loss Screening procedure as follows. For $j \in \overline{\mathcal{M}}_1$, the conditional hinge loss for the j^{th} SNP can be written as

$$L_j^c \equiv n^{-1} \sum_{i=1}^{n} [1 - y_i(\hat{\xi}_i + b_j^* + w_j^* \times x_{ij})]_+,$$

where b_j^* and w_j^* are unknown parameters.

For each SNP in $\overline{\mathcal{M}}_1$, we minimize L_j^c with respect to b_j^* and w_j^* by linear programming to obtain $\widehat{L_j^c}$. Let t_2 be the integer part of $n/(8 \log n)$. We select the t_2 SNPs that have the smallest $\widehat{L_j^c}$ to form a set $\mathcal{S}_2$. In other words, $\mathcal{S}_2$ harbors important SNPs that are marginally uncorrelated (but conditionally correlated) with the disease.

Let ℓ be an integer ≤ 1000 (we choose the upper bound of model sizes to be 1000 because most of the joint prediction models have covariates less than 1000). Denote the set of SNPs with the smallest ℓ HL_j as $\mathcal{T}$. That is, $\mathcal{T}$ consists of the top ℓ SNPs from the marginal screening. We lump $\mathcal{T}$ and $\mathcal{S}_2$ together as a set $\mathcal{S}_3$, and then run the SCAD-SVM on $\mathcal{S}_3$ to obtain a model $\mathcal{M}_2$. $\mathcal{M}_2$ is anticipated to capture both marginally important SNPs and conditionally important SNPs.

4.3 Simulations

4.3.1 A motivating example

We first show a hypothetical example that, even if we know the true model beforehand and all the SNPs in the true model are independent, marginal regression estimators can be highly biased. Assume that the true model includes 100 independent SNPs, with effect sizes $\beta_j = 0.2 \times (-1)^j$ for $j = 1, ..., 100$. Further assume that all SNPs have minor allele frequency (MAF) of 0.3. Under the logistic regression model, 4000 cases and 4000 controls were simulated. We first fitted a multivariate logistic regression model for all the 100 SNPs to obtain the joint estimates for the regression coefficients, and then applied the marginal logistic regression for each SNP to obtain the marginal estimates. We conducted 100 simulations, and the averages of the estimates are plotted in Figure 4.1. Clearly, the marginal estimates are highly biased toward the null, while the joint estimates are much closer to the true parameters. This motivating example emphasizes the importance of estimating the true parameters under the joint model. In Appendix 2, we examine the asymptotic property of the marginal estimators and show that they are inconsistent estimators.

4.3.2 Data simulation and competing methods

In real GWAS data, SNPs tend to be correlated with each other and the number of noise SNPs far exceeds the number of causal SNPs (or disease loci (DL)). Our simulation studies have taken these issues into account. We simulated many more noise SNPs than the causal SNPs (to be described later), and the linkage disequilibrium was introduced into the SNPs according to a procedure described by He and Lin (2011). The causal SNPs were simulated under the logistic regression model. Given that there are numerous methods for risk prediction in the literature, it is not possible for us to

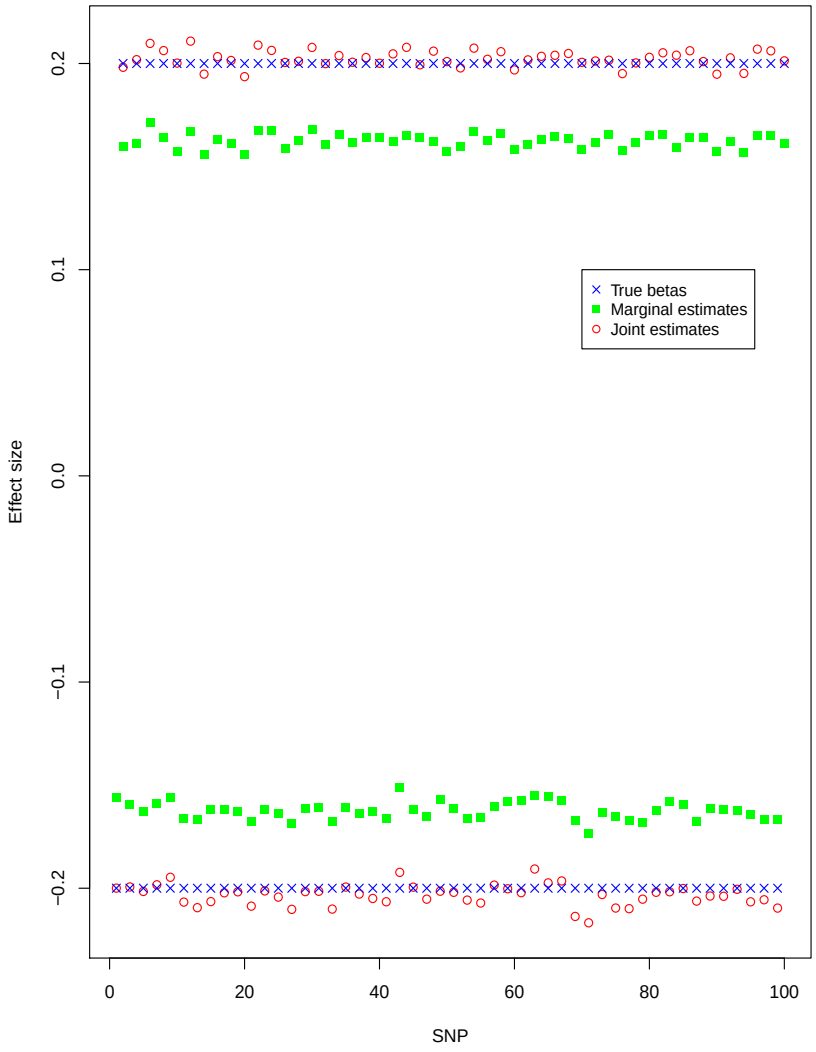

Figure 4.1: Comparison of the marginal regression and the joint regression estimates under the logistic model

survey all of them. Instead, we focused on the ones that are either currently being widely used or have the potential to be widely used for genetic risk prediction. First, the Armitage trend test (ATT) was conducted for all p SNPs and, unless otherwise stated, the top 1000 SNPs were collected as a set denoted by $\mathcal{G}$. We choose 1000 because most joint models contain no more than 1000 covariates (we also explored the top 500 in the latter part of this Chapter). Next, the following methods are applied to the SNPs in $\mathcal{G}$ to obtain the prediction model:

1) The Oracle method– Among the SNPs in $\mathcal{G}$, we identified the causal SNPs using our *a priori* information and then fit a joint logistic regression model to them. The joint coefficient estimates were used for the risk prediction. Note that this method is available only for simulation studies, but not for real data analysis. 2) The Marginal method– For each of the SNPs in $\mathcal{G}$, a logistic regression model was fitted. The obtained marginal estimates were used for prediction. 3) The Count method– Let g_j denote the jth SNP in $\mathcal{G}$, and $\hat{\gamma}_j$ denote the marginal estimate for g_j. The 'risk allele count' was calculated as $\sum_{j\in\mathcal{G}} g_j \text{sgn}(\hat{\gamma}_j)$. Then, a logistic regression model was fitted for the 'risk allele count'. 4) The Logistic-SCAD method– The hinge loss within the SCAD-SVM was replaced by the logistic regression likelihood to yield the Logistic-SCAD. The Logistic-SCAD was implemented via the local linear approximation algorithm, similar to the implementation of the SCAD-SVM as described in the Appendix 2.

We calculated the True Positive Cluster (TPC) (He and Lin, 2011) to gauge how many causal SNPs were captured. To evaluate the prediction accuracy of the compared methods, we further simulated an independent testing data set under the same logistic model as the data set used for the model building. This independent testing data set also contains 4000 subjects (2000 cases and 2000 controls). We calculated the AUC, the 10%-extreme-err (He and Lin, 2011), the difference between the predicted and the true liability (excluding the intercept), the correlation between the predicted and the

Table 4.1: Prediction accuracy under a moderate number of noise features

	Oracle	ISS	Logistic-SCAD	Marginal	Count
Model size	10.0	12.2	11.8	50	50
TPC	-	10.0	10.0	10.0	10.0
10%-extreme-err	0.131	0.133	0.132	0.192	0.222
AUC	0.760	0.759	0.759	0.710	0.687
liab-dif	0.175	0.258	0.283	2.247	1.021
liab-corr	0.996	0.990	0.991	0.809	0.721

true liability (excluding the intercept). The latter two quantities are named as liab-diff and liab-corr. For each of the following experiments, 100 simulation were conducted.

4.3.3 Models with a moderate number of noise SNPs

We first test these methods under a moderate number of noise SNPs to gain some insight into their performance. We let the total number of SNPs be 600, among which there are 10 causal SNPs. The causal SNPs are independent, and their effects are set to be $(0.5, -0.5, 0.5, -0.5, 0.5, -0.5, 0.5, -0.5, 0.5, -0.5)$. The MAFs for the causal SNPs follow a Uniform(0.25, 0.5) distribution. To ensure a good coverage of the causal SNPs, we let $\mathcal{G}$ be the set of the top 50 SNPs in the ATT. The results are shown in Table 4.1. It can be seen that the Oracle method performs the best in prediction, followed by Logistic-SCAD and ISS. The Marginal and Count methods are not based on joint estimates and, as expected, have the lowest prediction accuracy. We also carried an experiment where the testing samples were simulated under the prospective sampling, and a similar trend was observed (Table 4.10). Thus, in the following experiments, we only use retrospective samples for the testing data.

4.3.4 Models with a large number of noise SNPs

Next, we incorporated nearly 60,000 noise SNPs into the model to make it more challenging for risk prediction. We first studied the situation where the number of causal SNPs is 100, and their effects are moderate (with the effect sizes equal to 0.5 or -0.5). The results are shown in Table 4.2. It can be seen that, the model sizes of the ISS and the Logistic-SCAD are much smaller than those of the Marginal and the Count methods, indicating that the former two were able to remove many noise SNPs from their models. At the same time, it appears that the former two methods captured a large proportion of the causal SNPs as indicated by their TPCs. This explains why they have better prediction performance than the Marginal and the Count methods. The Logistic-SCAD has a high AUC but an extremely high core-liab-dif, suggesting that although this method preserves the rank of the risk liabilities quite well, it tends to generate biased estimates by shifting the regression coefficients in the same direction. Thus, our results suggest that when the number of noise features is high, the Logistic-SCAD is less ideal than the ISS. The Marginal and the Count methods have the lowest prediction accuracy among all the compared methods, verifying that marginal estimates are inferior to joint estimates for prediction. To allow for heterogeneous effect sizes of the 100 causal SNPs, we further simulated those nonzero regression coefficients under the Uniform(0.1, 0.5) with alternate signs (Table 4.3). Under this situation, the performance of the Logistic-SCAD further deteriorated, while the ISS is second only to the Oracle method in prediction accuracy.

4.3.5 Models with marginally uncorrelated SNPs

One of the main advantages of the ISIS strategy is to capture causal SNPs that are marginally uncorrelated (but conditionally correlated) with the outcome. To test the effectiveness of our ISS in capturing conditionally correlated SNPs, we let three of

Table 4.2: Prediction accuracy of moderate predictors under a large number of noise features (p=60000, DL=100)

	Oracle	ISS	Logistic-SCAD	Marginal	Count
Model size	99.5	241.1	290.2	1000	1000
TPC	-	99.6	99.4	99.7	99.7
10%-extreme-err	0.005	0.010	0.025	0.058	0.073
AUC	0.944	0.924	0.891	0.834	0.816
liab-dif	1.147	2.136	9.536	7.102	2.671
liab-corr	0.988	0.946	0.878	0.759	0.721

Table 4.3: Prediction accuracy of weak to moderate predictors under a large number of noise features (p=60000, DL=100)

	Oracle	ISS	Logistic-SCAD	Marginal	Count
Model size	77.5	374.6	479.6	1000	1000
TPC	-	80.8	80.1	80.8	80.8
10%-extreme-err	0.032	0.064	0.137	0.117	0.143
AUC	0.874	0.828	0.754	0.772	0.748
liab-dif	1.063	1.614	10.003	6.891	2.631
liab-corr	0.959	0.848	0.665	0.710	0.649

Table 4.4: Prediction accuracy when some causal SNPs are uncorrelated with the outcome (p=60000, DL=100)

	Oracle	ISS	Logistic-SCAD	Marginal	Count
Model size	77.4	332.6	464.8	1000	1000
TPC	-	81.8	79.6	80.5	80.5
Conditional	-	78%	6%	8%	8%
10%-extreme-err	0.029	0.048	0.128	0.098	0.125
AUC	0.882	0.852	0.764	0.791	0.765
liab-dif	1.081	1.542	9.735	7.374	2.250
liab-corr	0.949	0.877	0.670	0.733	0.670

the 100 causal SNPs to be marginally uncorrelated with the outcome. The details on simulating such a correlation structure have been described by He and Lin (2011). The total number of SNPs was kept at 60000, and the effects sizes of the causal SNPs still followed Uniform(0.1, 0.5) with alternate signs. Table 4.4 shows that the ISS has a probability of 78% to capture the conditionally correlated causal SNPs, while the other methods have little chance to capture those SNPs. Furthermore, the difference in AUC appears to be widened between the ISS method and the other three methods, i.e., the Logistic-SCAD, the Marginal and the Count methods.

4.3.6 Models that deviate from the logistic model

In reality, the underlying genetic model may not be the logistic model. To test the robustness of our method, we perturbed the logistic model with a random intercept for each subject. The results are shown in Table 4.5. The introduced random intercept apparently has a negative impact on all the methods, but the ISS remains to achieve the highest AUC (excluding the Oracle method). We also simulated the data under the probit model and observed a similar pattern of performance for the compared methods (Table 4.11 in Supplemental Materials).

Table 4.5: Prediction accuracy in the presence of random effects (p=60000)

	Oracle	ISS	Logistic-SCAD	Marginal	Count
Model size	74.0	423.5	512.6	1000	1000
TPC	-	78.9	77.8	78.8	78.8
10%-extreme-err	0.047	0.102	0.182	0.148	0.176
AUC	0.849	0.785	0.716	0.743	0.721
liab-dif	1.110	1.718	10.984	6.468	2.387
liab-corr	0.948	0.785	0.606	0.677	0.618

4.3.7 Other considerations

In most of the above experiments, we chose the top 1000 candidate SNPs as the starting set of SNPs. We also explored using the top 500 candidates as the starting set of SNPs, and the results are consistent with what we have observed above (Tables 4.12-4.13 in Supplemental Materials).

4.4 Real Data Analysis

We applied our method to the Wellcome Trust Case-Control Consortium (WTCCC) (2007) data. The WTCCC data include 2000 cases for each of seven diseases, the type 1 diabetes (T1D), the type 2 diabetes (T2D), coronary heart disease, hypertension, bipolar disorder, rheumatoid arthritis (RA) and Crohn's diseases. The WTCCC data also include 3000 controls shared by all the 7 diseases. We focus on the T1D and the RA, which are known to have strong genetic components. The other 5 diseases have weaker genetic effects and their genetic risk prediction in general is poor; current literature suggests that other factors, such as environmental factors, need to be taken into account for better risk prediction (Janssens and van Duijn, 2008; Collins et al., 2011).

For the T1D data, we split the data into three parts, and used two parts as the

Table 4.6: Prediction accuracy for the WTCCC-T1D data

	ISS	Logistic-SCAD	Marginal	Count
	stringent SNP exclusion criteria			
10%-extreme-err	0.081	0.116	0.268	0.258
AUC	0.852	0.813	0.705	0.675
	less-stringent SNP exclusion criteria			
10%-extreme-err	0.056	0.082	0.254	0.239
AUC	0.867	0.839	0.721	0.681

training data and one part as the testing data. We excluded SNPs with low MAF (< 0.05) as well as SNPs in departure from the Hardy-Weinberg equilibrium ($p < 10^{-3}$). We first applied the ISS, the Logistic-SCAD, the Marginal method and the Count method to the training data to obtain the prediction models, and then calculated the the prediction error for the testing data. Because the true model is unknown for real data, it is not possible for us to calculate the liab-diff and liab-corr. Instead, we report the AUC and 10%-extreme-err in Table 4.6 (upper panel). It can be seen that the ISS method outperforms all other methods in both AUC and 10%-extreme-err. To test the robustness of our method, we divided the data into 5 parts (4 parts as training data and 1 part as testing data), and excluded SNPs with less stringent criteria (MAF< 0.01, or $p < 10^{-5}$ for HWE). Again, the ISS method appears to be more accurate than the other methods (Table 4.6, lower panel).

It is well known that the HLA region on Chromosome (Chr) 6 contains strong genetic variants contributing to T1D, and many SNPs in this region are in high LD. It is also well known that excluding highly correlated predictors from the prediction models can lead to enhanced stability of the models. Hence, we decided to prune the SNPs on Chr6 and examine how the pruning affects the prediction accuracy. We implemented

Table 4.7: Prediction accuracy for the WTCCC-T1D data with the HLA pruned

	ISS	Logistic-SCAD	Marginal	Count
		pruning at $r^2 = 0.64$		
10%-extreme-err	0.081	0.144	0.203	0.250
AUC	0.854	0.769	0.728	0.689
		pruning at $r^2 = 0.5$		
10%-extreme-err	0.079	0.147	0.195	0.245
AUC	0.854	0.766	0.735	0.687

the pruning mechanism in the PLINK (i.e., the sliding window mechanism), and conducted the pruning at different threshold levels: $r^2 = 0.64, r^2 = 0.5$, and $r^2 = 0.05$. The value 0.5 is the default pruning-threshold used by PLINK, while the other two values have been used in the literature (Hoggart et al., 2008; He and Lin, 2011). The results in Table 4.7 show that under mild and moderate pruning ($r^2 = 0.64$ and $r^2 = 0.5$), the performance of the ISS, the Marginal and the Count methods is slightly improved (compared to the upper panel of Table 4.6). Under heavy pruning ($r^2 = 0.05$), all methods yielded AUC under 0.7 (data not shown), indicating that inappropriate pruning can lead to severe loss of information for prediction.

Next, we analyzed the RA data in a similar manner, and the results are shown in Table 4.8. Again, regardless of the stringency of the SNP exclusion criteria, the ISS is leading all other compared methods in prediction accuracy. Since RA is also believed to be influenced by the HLA region (though to a less extent compared to the T1D), we also tested the effect of the pruning on the prediction performance of the compared methods (Table 4.9). Again, it appears that slight to moderate pruning can sometimes improve the prediction accuracy, but heavy pruning results in a considerable loss of prediction power. It is to be noted that our results for the Marginal method and the Count method agree well with those by Evans et al. (2009), reassuring that these two

Table 4.8: Prediction accuracy for the WTCCC-RA data

	ISS	Logistic-SCAD	Marginal	Count
	stringent SNP exclusion criteria			
10%-extreme-err	0.222	0.316	0.330	0.267
AUC	0.701	0.615	0.678	0.660
	less-stringent SNP exclusion criteria			
10%-extreme-err	0.245	0.310	0.327	0.265
AUC	0.690	0.625	0.681	0.662

Table 4.9: Prediction accuracy for the WTCCC-RA data with the HLA pruned

	ISS	Logistic-SCAD	Marginal	Count
	pruning at $r^2 = 0.64$			
10%-extreme-err	0.238	0.335	0.259	0.281
AUC	0.702	0.625	0.681	0.655
	pruning at $r^2 = 0.5$			
10%-extreme-err	0.244	0.322	0.251	0.276
AUC	0.694	0.627	0.679	0.649
	pruning at $r^2 = 0.05$			
10%-extreme-err	0.303	0.336	0.365	0.394
AUC	0.633	0.608	0.586	0.567

methods may not be ideal in real practice.

4.5 Discussion

We previously developed a method, GWASelect, for variable selection in GWAS (He and Lin, 2011). GWASelect was mainly designed for disease gene hunting, where false discovery rate is a major concern. The ISS developed herein is for the task of prediction, where prediction power is our primary interest. Our simulation studies clearly show that some noise SNPs can be tolerated in the prediction model as long as 1) a large

number of causal SNPs are captured and 2) the number of noise SNPs is not too high. Our experiments suggest that the task of disease gene hunting may be quite different from the task of disease prediction, and different strategies may need to be considered for the two tasks.

We have shown both analytically and numerically that marginal estimates are sub-optimal for risk prediction. Marginal estimation provides a quick way to reduce the high dimensions of GWAS, and hence can still be helpful in certain circumstances to get some preliminary idea of the prediction power of SNPs. We highly recommend that joint estimation be performed to potentially harvest more prediction power.

GWAS offer a useful platform for genetic risk prediction, but there are many other types of data that may need to be considered for improving the prediction accuracy. For example, family structures, environmental factors, rare genetic variants, copy number variation and epigenetic elements may play important roles in the development of diseases (Jirtle and Skinner, 2007; Ruderfer et al., 2010). While these data may potentially help to retrieve part of the "missing inheritability", how to integrate them together under high dimensions poses a tremendous challenge for model building. Furthermore, current prediction models may need to be revised as the underlying biology of many SNPs are unveiled. Therefore, the task of constructing statistically more powerful and biologically more informative prediction models will require a joint effort from multiple disciplines.

Genetic risk prediction is a highly complicated topic, and we have not attempted to provide a panacea. We have mainly focused on the issue of joint effects estimation. Other issues, such as genetic heterogeneity, secondary outcomes and the time of disease-onset, also need to be taken into account for better prediction (Webb et al., 2011). In addition, biological pathway analysis may also help to predict genetic risks (Hu et al., 2011). These issues are beyond the range of this work, but merit further investigations.

Table 4.10: Prediction accuracy under prospective sampling with a moderate number of noise features (p=600, DL=10)

	Oracle	ISS	Logistic-SCAD	Marginal	Count
Model size	10.0	12.2	11.8	50	50
TPC	10.0	10.0	10.0	10.0	10.0
10%-extreme-err	0.291	0.293	0.293	0.337	0.357
AUC	0.759	0.758	0.758	0.708	0.684
liab-dif	0.173	0.254	0.280	2.437	0.895
liab-corr	0.995	0.989	0.990	0.797	0.707

4.6 Supplemental Materials

4.6.1 Prediction under the prospective sampling

In parallel with the experiment whose results are shown in Table 4.1, we conducted a similar experiment in which the testing samples were simulated under the prospective sampling instead of the retrospective sampling. The results are shown in Table 4.10. It can be seen that the performance of the compared methods has a similar trend as that when the testing samples were simulated under the retrospective sampling.

4.6.2 Prediction when the true model is the probit model

We simulated the data under the probit model and then tested the performance of the compared methods. The effect sizes of the causal SNPs follow the Uniform(0.1, 0.5) with alternate signs. The results are shown in Table 4.11. The Oracle method still achieves the highest AUC, but its liab-dif is elevated, indicating some estimation bias generated by the Oracle method. This is not surprising because the underlying true model is the probit model, while the Oracle method adopted the joint logistic regression model. The ISS method appears to be quite robust to the altered underlying true model, as its liab-dif is quite low. In terms of prediction accuracy, the ISS method

Table 4.11: Prediction accuracy under the probit model (p=60000, DL=100)

	Oracle	ISS	Logistic-SCAD	Marginal	Count
Model size	84.7	138.4	216.8	1000	1000
TPC	-	84.7	84.1	87.1	87.1
10%-extreme-err	0.003	0.005	0.011	0.041	0.063
AUC	0.956	0.947	0.922	0.857	0.825
liab-dif	2.835	1.289	10.325	8.003	2.775
liab-corr	0.980	0.962	0.912	0.783	0.718

Table 4.12: Prediction by starting with the top 500 SNPs under a large number of noise features (p=60000, DL=100)

	Oracle	ISS	Logistic-SCAD	Marginal	Count
Model size	70.9	193.7	255.7	500	500
TPC	-	75.4	73.7	73.8	73.8
10%-extreme-err	0.036	0.061	0.081	0.105	0.116
AUC	0.869	0.833	0.808	0.783	0.773
liab-dif	1.166	1.663	3.961	5.704	2.203
liab-corr	0.946	0.860	0.801	0.741	0.715

still achieves the second highest AUC among all the methods.

4.6.3 Prediction by starting with the top 500 candidate SNPs

In Table 4.3, the top 1000 candidate SNPs were considered as the starting set of SNPs for the compared methods. Here, we conducted an experiment in which the top 500 candidate SNPs were set as the starting set. Briefly, we simulated 60000 SNPs, among which 100 are causal SNPs. The effect sizes of the 100 SNPs follow the Uniform(0.1, 0.5) with alternate signs. The results in Table 4.12 demonstrate that the performance of the compared methods is largely consistent with that in Table 4.3. Hence, regardless whether the number of starting SNPs is 1000 or 500, our ISS method competes favorably with the other compared methods (except the Oracle method).

Table 4.13: Prediction by starting with the top 500 candidates in the presence of random effects

	Oracle	ISS	Logistic-SCAD	Marginal	Count
Model size	67.3	223.9	280.5	500	500
TPC	-	72.8	71.0	71.1	71.1
10%-extreme-err	0.051	0.097	0.117	0.132	0.145
AUC	0.843	0.791	0.769	0.756	0.746
liab-dif	1.170	1.598	4.121	5.223	2.090
liab-corr	0.932	0.801	0.744	0.711	0.686

In Table 4.13, we conducted an experiment in which the true underlying model is the logistic model but with the perturbation of a random intercept. This experiment is in parallel with the experiment whose results are shown in Table 4.5 except that we started with the top 500 candidate SNPs instead of the top 1000 candidate SNPs. Again, our ISS method is only second to the Oracle method in terms of the prediction accuracy.

Chapter 5

Future Research

5.1 Variable Selection for Multivariate-outcome Data

Multivariate-outcome data are often encountered in genome-wide association studies. The outcomes collected for each subject usually include both binary traits (such as disease status) and a series of quantitative traits. One can, of course, analyze each trait separately, but there are at least two advantages to analyze all the traits jointly. First, multiple traits tend to capture the etiological characteristics of a disease better than a single trait; second, by harnessing the correlations among the outcomes, one usually achieves higher statistical power to detect genetic associations. A number of methods have been proposed for the analysis of multivariate-outcome for GWAS, such as the one by Ferreira and Purcell (2009) and the one by Avery et al. (2011).

While these methods are useful, they were designed to analyze each SNP individually. To better estimate the joint effects of multiple SNPs, a variable selection method for multivariate regression is needed. However, existing variable selection methods for multivariate-outcome data are quite limited. Recently, Peng et al. (2010) proposed a variable selection method for the analysis of microarray data, and named their method as remMAP. This approach imposes two types of penalties on the objective function to achieve sparsity on the regression coefficient matrix, but its theoretical properties,

such as the parameter estimation consistency and the model selection consistency, are entirely unclear.

There are at least two reasons to explain why few variable selection methods have been developed for multivariate-outcome data: first, it is often difficult to specify the joint distribution of the multiple outcomes, and hence the joint likelihood is hard to be constructed; second, computation is much more challenging than univariate-outcome data due to the enlarged dimension of the outcomes. Motivated by Johnson et al. (2008), Wang et al. (2011) proposed to use the penalized generalized estimating equation plus the SCAD penalty for longitudinal data with high dimensions, where they allow the dimension p_n to grow with the sample size n at the order of $O(n)$ (referred to as GEE-SCAD hereafter). While GEE-SCAD can be potentially borrowed to analyze multivariate-outcome data, it has some limitations. First, it is not straightforward to apply GEE-SCAD to data with both binary and continuous outcomes, because GEE-SCAD implicitly assumes that the regression parameters for all outcomes are at the same scale. Second, it remains challenging to establish model selection criteria, such as BIC, in the high-dimensional GEE setting. Third, it can be awkward to impose more complicated penalty terms, such as the group penalty, to GEE. Fourth, when n is large, say at thousands, GEE-SCAD may become infeasible in computation.

Another strategy to deal with multivariate-outcome data is to model each trait by its marginal distribution, and then estimate the covariances of the regression coefficients from the data (Wei et al., 1989). We will propose a variable selection method based on this strategy. We will 1) allow the outcomes to be a mixture of both binary and continuous outcomes, and the regression parameters for different outcomes to be at different scales; 2) impose a group penalty on the regression coefficients to borrow strengths in the shared information of multiple outcomes. In addition, we will allow p_n to be at the order of $O(n)$.

Appendix 1: Chapter 3 Proofs

We provide the proofs of Theorems 4∼6 and Corollaries 3 and 4 in this appendix. The proofs for Theorems 1∼3 and Corollaries 1 and 2 are relegated to Section 3.8.6 in the Supplemental Materials of Chapter 3.

Proof of Theorem 4

By the definition of $\widehat{\boldsymbol{\beta}}_k$,

$$\begin{aligned} & \sum_{k=1}^{K}(\widehat{\boldsymbol{\beta}}_k-\widetilde{\boldsymbol{\beta}}_k)^{\mathrm{T}}\widetilde{\boldsymbol{V}}_k^{-1}(\widehat{\boldsymbol{\beta}}_k-\widetilde{\boldsymbol{\beta}}_k)+\lambda\sum_{j=1}^{p_n}\left(\sum_{k=1}^{K}w_{jk}|\widehat{\beta}_{jk}|\right)^{\frac{1}{2}} \\ \leq & \sum_{k=1}^{K}(\boldsymbol{\beta}_k^0-\widetilde{\boldsymbol{\beta}}_k)^{\mathrm{T}}\widetilde{\boldsymbol{V}}_k^{-1}(\boldsymbol{\beta}_k^0-\widetilde{\boldsymbol{\beta}}_k)+\lambda\sum_{j=1}^{p_n}\left(\sum_{k=1}^{K}w_{jk}|\beta_{jk}^0|\right)^{\frac{1}{2}}. \end{aligned}$$

Write $(\widehat{\boldsymbol{\beta}}_k-\widetilde{\boldsymbol{\beta}}_k)=(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^0)-(\widetilde{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^0)$. Then simple calculations yield that

$$\sum_{k=1}^{K}\left\{(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^0)^{\mathrm{T}}\widetilde{\boldsymbol{V}}_k^{-1}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^0)+2(\boldsymbol{\beta}_k^0-\widehat{\boldsymbol{\beta}}_k)^{\mathrm{T}}\widetilde{\boldsymbol{V}}_k^{-1}(\widetilde{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^0)\right\}\leq\lambda\sum_{j=1}^{p_n}\left(\sum_{k=1}^{K}w_{jk}|\beta_{jk}^0|\right)^{\frac{1}{2}}.$$

Note that $\widetilde{\boldsymbol{V}}_k^{-1}=\hat{\sigma}_k^{-2}\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k$, where $\hat{\sigma}_k$ is the OLS estimator of σ_k. For each k, define

$$\begin{aligned} h_{1k} &\equiv (\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^0)^{\mathrm{T}}(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k)(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^0)=\sum_{i=1}^{n_k}[\boldsymbol{x}_{ik}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^0)]^2, \\ h_{2k} &\equiv 2(\boldsymbol{\beta}_k^0-\widehat{\boldsymbol{\beta}}_k)^{\mathrm{T}}(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k)(\widetilde{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^0)=2(\boldsymbol{\beta}_k^0-\widehat{\boldsymbol{\beta}}_k)^{\mathrm{T}}\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{\varepsilon}_k=2\sum_{i=1}^{n_k}\epsilon_{ik}\boldsymbol{x}_{ik}^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_k-\boldsymbol{\beta}_k^0), \end{aligned}$$

where $\boldsymbol{\varepsilon}_k = (\epsilon_{1k}, \ldots, \epsilon_{n_k k})^{\mathrm{T}}$. Also, define $\psi_n = \lambda \sum_{j=1}^{p_n} \left(\sum_{k=1}^{K} w_{jk} |\beta_{jk}^0| \right)^{\frac{1}{2}}$. Then

$$\sum_{k=1}^{K} \hat{\sigma}_k^{-2} (h_{1k} + h_{2k}) \leq \psi_n. \tag{5.1}$$

We now generalize the proof of Theorem 1 in Huang et al. (2009) to multiple studies. For each k, define $\boldsymbol{\delta}_k = (\boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{X}_k)^{1/2} (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0)$ and $\boldsymbol{D}_k = (\boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{X}_k)^{-1/2} \boldsymbol{X}_k^{\mathrm{T}}$. It is easy to verify that

$$h_{1k} + h_{2k} = \|\boldsymbol{\delta}_k\|^2 - 2(\boldsymbol{D}_k \boldsymbol{\varepsilon}_k)^{\mathrm{T}} \boldsymbol{\delta}_k = \|\boldsymbol{\delta}_k - \boldsymbol{D}_k \boldsymbol{\varepsilon}_k\|^2 - \|\boldsymbol{D}_k \boldsymbol{\varepsilon}_k\|^2.$$

Combining the above equation with (5.1), we have that

$$\sum_{k=1}^{K} \hat{\sigma}_k^{-2} \|\boldsymbol{\delta}_k - \boldsymbol{D}_k \boldsymbol{\varepsilon}_k\|^2 \leq \sum_{k=1}^{K} \hat{\sigma}_k^{-2} \|\boldsymbol{D}_k \boldsymbol{\varepsilon}_k\|^2 + \psi_n. \tag{5.2}$$

Using the inequality $\|\boldsymbol{\delta}_k\|^2 \leq 2\|\boldsymbol{\delta}_k - \boldsymbol{D}_k \boldsymbol{\varepsilon}_k\|^2 + 2\|\boldsymbol{D}_k \boldsymbol{\varepsilon}_k\|^2$ and (5.2), we have

$$\begin{aligned} \sum_{k=1}^{K} \hat{\sigma}_k^{-2} \|\boldsymbol{\delta}_k\|^2 &\leq 2 \sum_{k=1}^{K} \hat{\sigma}_k^{-2} \|\boldsymbol{\delta}_k - \boldsymbol{D}_k \boldsymbol{\varepsilon}_k\|^2 + 2 \sum_{k=1}^{K} \hat{\sigma}_k^{-2} \|\boldsymbol{D}_k \boldsymbol{\varepsilon}_k\|^2 \\ &= 4 \sum_{k=1}^{K} \frac{\hat{\sigma}_k^{-2}}{\sigma_k^{-2}} \sigma_k^{-2} \|\boldsymbol{D}_k \boldsymbol{\varepsilon}_k\|^2 + 2\psi_n \leq 4 \sum_{k=1}^{K} (1+1) \sigma_k^{-2} \|\boldsymbol{D}_k \boldsymbol{\varepsilon}_k\|^2 + 2\psi_n. \end{aligned}$$

Let d_i be the ith column of $\boldsymbol{D}_k$. Since $E\|\boldsymbol{D}_k \boldsymbol{\varepsilon}_k\|^2 = \sum_{i=1}^{n} \|d_i\|^2 E\epsilon_{ik}^2 = \sigma_k^2 \mathrm{tr}(\boldsymbol{D}_k \boldsymbol{D}_k^{\mathrm{T}}) =$

$\sigma_k^2 p_n$, we have $E(\sum_{k=1}^K \hat{\sigma}_k^{-2}\|\boldsymbol{\delta}_k\|^2) \le 8Kp_n + 2\psi_n$. On the other hand,

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|^2 &= \sum_{k=1}^K (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0)^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0) \le \sum_{k=1}^K (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0)^{\mathrm{T}}(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k/(bn_k))(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0) \\
&= n^{-1}b^{-1}\sum_{k=1}^K \frac{n}{n_k}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0)^{\mathrm{T}}(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k)(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0) \\
&\le n^{-1}b^{-1}\nu_{\min}^{-1}\sum_{k=1}^K (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0)^{\mathrm{T}}(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k)(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0) \\
&\le \frac{\sigma_{\max}^2}{nb\nu_{\min}}\sum_{k=1}^K \hat{\sigma}_k^{-2}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0)^{\mathrm{T}}(\boldsymbol{X}_k^{\mathrm{T}}\boldsymbol{X}_k)(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0) = \frac{\sigma_{\max}^2}{nb\nu_{\min}}\sum_{k=1}^K \hat{\sigma}_k^{-2}\|\boldsymbol{\delta}_k\|^2.
\end{aligned}$$

Therefore,

$$E\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|^2 \le n^{-1}b^{-1}\nu_{\min}^{-1}\sigma_{\max}^2(8Kp_n + 2\psi_n).$$

Note that $0 < \psi_n \le \lambda p_0\sqrt{Kt_{1n}r_2}$, where $t_{1n} = \max\{w_{jk} : 1 \le j \le p_0, k \in \mathcal{M}_j\}$. Thus,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p\left(\left\{\frac{p_n + \lambda\sqrt{t_{1n}}}{n}\right\}^{1/2}\right).$$

Consequently, if $\lambda\sqrt{t_{1n}} = O_p(p_n)$, then $\widehat{\boldsymbol{\beta}}$ is root-(n/p_n) consistent. The proof is completed.

Proof of Corollary 3

For the heterogeneous structure, we only need to show that the conditions in Theorem 4 are satisfied. Note that the OLS estimator is consistent with $\widetilde{\beta}_{jk} - \beta_{jk}^0 = O_p(\sqrt{p_n/n})$ for any $j = 1, \ldots, p$ and $k = 1, \ldots, K$. This implies that the weight $w_{jk} \equiv (\widetilde{\beta}_{jk})^{-(2+\nu)} = |\beta_{jk}^0|^{-(2+\nu)} + O_p(\sqrt{p_n/n})$ for any (j,k) satisfying $1 \le j \le p_0$ and $k \in \mathcal{M}_j$. Then, by definition, $t_{1n} = O_p(1)$. Therefore, as long as $\lambda = O_p(p_n)$, the conditions in Theorem 4 are satisfied.

For the homogeneous structure, let $s_{1n} = \max\{w_j : j = 1, ..., p_0\}$. Following the proof of Theorem 4, we can show that, if $\lambda\sqrt{s_{1n}} = O_p(p_n)$, then $\widehat{\boldsymbol{\beta}}^*$ is consistent.

Likewise, we can show that $s_{1n} = O_p(1)$, which implies $\lambda = O_p(p_n)$.

Proof of Theorem 5

For any $(j,k) \in \mathcal{N}$, if $\widehat{\beta}_{jk,\lambda} \neq 0$, then by the Karush-Kuhn-Tucker (KKT) conditions, it must be true that

$$\begin{aligned}
0 &= \frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_{jk}}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_\lambda} \\
&= 2n_k \left[(n_k \widetilde{\boldsymbol{V}}_k)^{-1}\right]_{j.} (\widehat{\boldsymbol{\beta}}_{k,\lambda} - \widetilde{\boldsymbol{\beta}}_k) + \lambda \frac{1}{2} \left\{ \sum_{k'=1}^{K} w_{jk'} |\widehat{\beta}_{jk',\lambda}| \right\}^{-\frac{1}{2}} w_{jk} \mathrm{sgn}(\widehat{\beta}_{jk,\lambda}) \\
&= 2n_k \left[\boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{X}_k / n_k\right]_{j.} (\widehat{\boldsymbol{\beta}}_{k,\lambda} - \widetilde{\boldsymbol{\beta}}_k) + \lambda \frac{1}{2} \left\{ \sum_{k'=1}^{K} w_{jk'} |\widehat{\beta}_{jk',\lambda}| \right\}^{-\frac{1}{2}} w_{jk} \mathrm{sgn}(\widehat{\beta}_{jk,\lambda}) \\
&\equiv E_1 + E_2,
\end{aligned}$$

where $[\boldsymbol{H}]_{j.}$ represents the jth row of $\boldsymbol{H}$. Since $||\widehat{\boldsymbol{\beta}}_{k,\lambda} - \boldsymbol{\beta}_k^0|| = O_p(\sqrt{p_n/n})$ and $||\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0|| = O_p(\sqrt{p_n/n})$, we have that $||\widehat{\boldsymbol{\beta}}_{k,\lambda} - \widetilde{\boldsymbol{\beta}}_k|| = O_p(\sqrt{p_n/n})$. Define $\boldsymbol{T} = \boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{X}_k / n_k$. Then

$$|E_1| \leq 2nO_p(\sqrt{p_n/n}) \left\{ \sum_{i=1}^{p_n} \boldsymbol{T}_{ij}^2 \right\}^{1/2} = O_p(\sqrt{np_n})$$

because the eigenvalues of T are bounded.

Since $\widehat{\boldsymbol{\beta}}_\lambda$ is a $\sqrt{n/p_n}$ consistent estimator of $\boldsymbol{\beta}^0$, we have that $|\widehat{\beta}_{jk,\lambda}| \leq r_2 + 1$ for all (j,k) with probability tending to 1. Then

$$|E_2| \geq \lambda \frac{1}{2} [K g_{1n}(1+r_2)]^{-\frac{1}{2}} g_{2n} = \frac{1}{2} [K(1+r_2)]^{-\frac{1}{2}} \lambda g_{2n} g_{1n}^{-\frac{1}{2}}. \tag{5.3}$$

By condition C3(ii), the right side of (5.3)$\to \infty$. Therefore, with probability tending to one, either $\widehat{\beta}_{jk} = 0$ for $(j,k) \in \mathcal{N}$ or the sign of (5.3) is equal to the sign of $\widehat{\beta}_{jk}$. The latter contradicts with the fact that $\widehat{\boldsymbol{\beta}}_\lambda$ is a minimizer of Q_n. Thus, $P(\widehat{\beta}_{jk} = 0) \to 1$

for any $(j,k) \in \mathcal{N}$.

Proof of Corollary 4

We first consider the heterogeneous structure. Under condition (C1), Yohai and Maronna (1979) showed that $\sqrt{n/p_n}||\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0|| = O_p(1)$ for $k = 1, \ldots, K$. Hence, for any $(j,k) \in \mathcal{N}$, we have $\sqrt{n/p_n}(\widetilde{\beta}_{jk} - 0) = O_p(1)$, implying that $w_{jk} = |\widetilde{\beta}_{jk}|^{-(2+\vartheta)} = O_p((n/p_n)^{\frac{2+\vartheta}{2}})$. Thus, $g_{1n} = O_p((n/p_n)^{\frac{2+\vartheta}{2}})$ and $g_{2n} = O_p((n/p_n)^{\frac{2+\vartheta}{2}})$. This implies that $g_{2n} g_{1n}^{-\frac{1}{2}} = O_p((n/p_n)^{\frac{2+\vartheta}{4}})$. If $\vartheta = 1$, the last condition in Theorem 5 can be simplified to that $\lambda n^{-\frac{1}{4}} p^{-\frac{5}{4}} \to \infty$.

Next, we consider the homogeneous structure. Let $h_{1n} = \max\{w_j : j = p_0+1, ..., p\}$ and $h_{2n} = \min\{w_j : j = p_0+1, \ldots, p\}$. Following the proof of Theorem 5, we can show that, if $\lambda(np_n)^{-\frac{1}{2}} h_{2n} h_{1n}^{\frac{1}{2}} \to \infty$, then $P(\widehat{\beta}^*_{jk,\lambda} = 0) \to 1$ for any $(j,k) \in \mathcal{N}$. In addition, we can show that $h_{1n} = O_p((n/p_n)^{\frac{1}{2}})$ and $h_{2n} = O_p((n/p_n)^{\frac{1}{2}})$. Then Corollary 4 follows.

Proof of Theorem 6

Consider the heterogeneous structure. Since we have shown that, with an arbitrarily large probability, the estimator of $\{\beta_{jk}^0 : (j,k) \in \mathcal{N}\}$ must be 0, we can decompose $\widehat{\boldsymbol{\beta}}$ into $\{\widehat{\boldsymbol{\beta}}_{\mathcal{A}}, \mathbf{0}\}$. By the KKT conditions, $\widehat{\boldsymbol{\beta}}$ should satisfy

$$\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{A}_k}}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = \mathbf{0}.$$

By the definition of Q_n and the fact that $\widetilde{\boldsymbol{V}}_k^{-1} = \hat{\sigma}_k^{-2} \boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{X}_k$, we have

$$\begin{aligned} Q_n(\boldsymbol{\beta}) \;=\; & \sum_{k=1}^{K} [(\boldsymbol{\beta}_{\mathcal{A}_k} - \widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k})^{\mathrm{T}} (\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k})(\boldsymbol{\beta}_{\mathcal{A}_k} - \widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k}) \\ & + 2(\boldsymbol{\beta}_{\mathcal{A}_k} - \widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k^c})^{\mathrm{T}} (\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k^c})(\boldsymbol{\beta}_{\mathcal{A}_k^c} - \widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k^c}) \\ & + (\boldsymbol{\beta}_{\mathcal{A}_k^c} - \widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k^c})^{\mathrm{T}} (\boldsymbol{X}_{\mathcal{A}_k^c}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k^c})(\boldsymbol{\beta}_{\mathcal{A}_k^c} - \widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k^c})]/\hat{\sigma}_k^2 + \lambda \sum_{j=1}^{p} \left(\sum_{k=1}^{K} w_{jk} |\beta_{jk}| \right)^{\frac{1}{2}}. \end{aligned}$$

Define $\boldsymbol{y}_k = (y_{1k}, \ldots, y_{n_k k})^{\mathrm{T}}$. Then for each $k = 1, \ldots, K$, we have

$$\begin{aligned}
\mathbf{0} = \frac{\hat{\sigma}_k^2}{2} \frac{\partial Q_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{\mathcal{A}_k}}\Big|_{(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k}, \mathbf{0})} &= (\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k})(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k} - \widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k}) + (\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k^c})(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k^c} - \widetilde{\boldsymbol{\beta}}_{\mathcal{A}_k^c}) + \widehat{\boldsymbol{e}}_k, \\
&= (\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k})\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k} + (\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k^c})\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k^c} - \boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_k \widetilde{\boldsymbol{\beta}}_k + \widehat{\boldsymbol{e}}_k, \\
&= (\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k})\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k} + (\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k^c})\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k^c} \\
&\quad - \boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_k (\boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{X}_k)^{-1} \boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{y}_k + \widehat{\boldsymbol{e}}_k, \\
&= (\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k})\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k} + (\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k^c})\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k^c} \\
&\quad - \boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{y}_k + \widehat{\boldsymbol{e}}_k, \qquad (5.4)
\end{aligned}$$

where $\widehat{\boldsymbol{e}}_k$ is a vector of length $|\mathcal{A}_k|$ with its sth component being $\lambda \hat{\sigma}_k^2 \{\sum_{l=1}^K w_{sl} |\widehat{\beta}_{sl}|\}^{-\frac{1}{2}}$ $w_{sk} \mathrm{sgn}(\widehat{\beta}_{sk})/4$. In the last equation, we use the fact that $\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_k (\boldsymbol{X}_k^{\mathrm{T}} \boldsymbol{X}_k)^{-1} \boldsymbol{X}_k^{\mathrm{T}} = \boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}}$. Since $\boldsymbol{\beta}_{\mathcal{A}_k^c} = \mathbf{0}$ for each $k = 1, \ldots, K$, we have $\boldsymbol{y}_k = \boldsymbol{\varepsilon}_k + \boldsymbol{X}_{\mathcal{A}_k} \boldsymbol{\beta}_{\mathcal{A}_k}^0$. Then (5.4) implies that

$$\mathbf{0} = (\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k})(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k} - \boldsymbol{\beta}_{\mathcal{A}_k}^0) + \boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k^c} \widehat{\boldsymbol{\beta}}_{\mathcal{A}_k^c} - \boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{\varepsilon}_k + \widehat{\boldsymbol{e}}_k. \qquad (5.5)$$

It follows from Theorem 5 that $P(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k^c} = \mathbf{0}) \to 1$. Therefore, the second term on the right side of (5.5) equals to zero with probability tending to one. Consequently,

$$(\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k})(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k} - \boldsymbol{\beta}_{\mathcal{A}_k}^0) = \sum_{i=1}^n \boldsymbol{x}_{i_{\mathcal{A}_k}} \epsilon_{ik} - \widehat{\boldsymbol{e}}_k, \qquad (5.6)$$

where $\boldsymbol{x}_{i_{\mathcal{A}_k}}$ represents the subvector corresponding to the $\mathcal{A}_k$ part of $\boldsymbol{x}_{ik}$. Thus,

$$\sqrt{n}_k \boldsymbol{\gamma}_n^{\mathrm{T}} (\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k} - \boldsymbol{\beta}_{\mathcal{A}_k}^0) = \frac{1}{\sqrt{n}_k} \sum_{i=1}^n \epsilon_{ik} \boldsymbol{\gamma}_n^{\mathrm{T}} \left(\frac{\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k}}{n_k} \right)^{-1} \boldsymbol{x}_{i_{\mathcal{A}_k}} - \frac{1}{\sqrt{n}_k} \boldsymbol{\gamma}_n^{\mathrm{T}} \left(\frac{\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k}}{n_k} \right)^{-1} \widehat{\boldsymbol{e}}_k.$$

Note that

$$\begin{aligned} & \left| \frac{1}{\sqrt{n}_k} \boldsymbol{\gamma}_n^{\mathrm{T}} \left(\frac{\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k}}{n_k} \right)^{-1} \widehat{\boldsymbol{e}}_k \right| \\ \leq & \frac{1}{\sqrt{n}_k} ||\boldsymbol{\gamma}_n|| \times b^{-1} \times ||\widehat{\boldsymbol{e}}_k|| \\ = & O_p(n_k^{-1/2} \lambda). \end{aligned}$$

The equality holds because $||\boldsymbol{\gamma}_n|| = 1$, $||\widehat{\boldsymbol{e}}_k|| = O_p(\lambda)$. Under condition (C3)(i), $n_k^{-1/2}\lambda = o_p(1)$, so it follows that

$$\frac{1}{\sqrt{n}_k} \boldsymbol{\gamma}_n^{\mathrm{T}} \left(\frac{\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k}}{n_k} \right)^{-1} \widehat{\boldsymbol{e}}_k = o_p(1).$$

Therefore,

$$\sqrt{n}_k s_{n,k}^{-1} \boldsymbol{\gamma}_n^{\mathrm{T}} (\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k} - \boldsymbol{\beta}_{\mathcal{A}_k}^0) = \frac{s_{n,k}^{-1}}{\sqrt{n}_k} \sum_{i=1}^{n} \epsilon_{ik} \boldsymbol{\gamma}_n^{\mathrm{T}} \left(\frac{\boldsymbol{X}_{\mathcal{A}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{A}_k}}{n_k} \right)^{-1} \boldsymbol{x}_{i_{\mathcal{A}_k}} + o_p(1).$$

This equation is equivalent to equation (14) in the proof of Theorem 2 in Huang et al. (2008). Following their arguments, we can show that, under conditions (C3) and (C4),

$$\sqrt{n}_k s_{n,k}^{-1} \boldsymbol{\gamma}^{\mathrm{T}} (\widehat{\boldsymbol{\beta}}_{\mathcal{A}_k} - \boldsymbol{\beta}_{\mathcal{A}_k}^0) \to N(0, 1).$$

Appendix 2: Chapter 4 Proofs

Implementation of the SCAD-SVM

Suppose that the data consist of n subjects, each with d covariates. Let $x_i \equiv (x_{i1}, \ldots, x_{id})^{\mathrm{T}}$ denote the covariate-vector for the ith person. Let b and $w \equiv (w_1, \ldots, w_d)^{\mathrm{T}}$, respectively, represent the intercept and the coefficient-vector. The objective function for the SCAD-SVM is

$$\min_{b,w} \frac{1}{n} \sum_{i=1}^{n} [1 - y_i(b + w^{\mathrm{T}} x_i)]_+ + \sum_{j=1}^{d} p_\lambda(|w_j|),$$

where $p_\lambda(|w_j|)$ is the SCAD penalty for w_j.

Outer loop:

At a given point $(\hat{b}, \hat{w})$, we approximate the hinge loss by a local quadratic term and approximate the SCAD penalty by a local linear term (as shown in Section 4.2.2 in the main text). After removing those constant terms, the objective function for the SCAD-SVM becomes

$$\widetilde{A}(b, w) = -\sum_{i=1}^{n} \frac{y_i(b + w^{\mathrm{T}} x_i)}{2n} - \frac{1}{2n} \sum_{i=1}^{n} \frac{y_i(b + w^{\mathrm{T}} x_i)}{|y_i - (\hat{b} + \hat{w}^{\mathrm{T}} x_i)|} + \frac{1}{4n} \sum_{i=1}^{n} \frac{(b + w^{\mathrm{T}} x_i)^2}{|y_i - (\hat{b} + \hat{w}^{\mathrm{T}} x_i)|} + \sum_{j=1}^{d} p'_\lambda(|\hat{w}_j|)|w_j|.$$

Inner loop:

Plugging in $\hat{w}$, we obtain $p'_\lambda(|\hat{w}_j|)$. Then, we fix $\hat{b}, \hat{w}$ and $p'_\lambda(|\hat{w}_j|)$ throughout the inner loop. The CCD algorithm is applied to solve the optimization problem. For

$j = 1, \ldots, d$, the partial derivative with respect to w_j is

$$\begin{aligned}\frac{\partial \widetilde{A}}{\partial w_j} &= -\sum_{i=1}^{n} \frac{y_i x_{ij}}{2n} - \frac{1}{2n} \sum_{i=1}^{n} \frac{y_i x_{ij}}{|y_i - (\hat{b} + \hat{w}^{\mathrm{T}} x_i)|} \\ &\quad + \frac{1}{2n} \sum_{i=1}^{n} \frac{(b + w^{\mathrm{T}} x_i) x_{ij}}{|y_i - (\hat{b} + \hat{w}^{\mathrm{T}} x_i)|} \\ &\quad + p'_{\lambda}(|\hat{w}_j|)\mathrm{sgn}(w_j).\end{aligned}$$

With a little algebra, it can be shown that w_j can be updated by

$$\frac{S\left\{\frac{1}{2n}\sum_{i=1}^{n} y_i x_{ij} + \frac{1}{2n}\sum_{i=1}^{n} \frac{(y_i - b - w^{(j)\mathrm{T}} x_i^{(j)}) x_{ij}}{|y_i - (\hat{b} + \hat{w}^{\mathrm{T}} x_i)|}, \quad p'_{\lambda}(|\hat{w}_j|)\right\}}{\frac{1}{2n}\sum_{i=1}^{n} \frac{x_{ij}^2}{|y_i - (\hat{b} + \hat{w}^{\mathrm{T}} x_i)|}},$$

where $S(\cdot, \cdot)$ is the soft-thresholding operator (Friedman et al., 2010), and $w^{(j)}$ is the vector w by excluding w_j. Similarly, the b can be updated by

$$\frac{\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \frac{y_i - w^{\mathrm{T}} x_i}{|y_i - (\hat{b} + \hat{w}^{\mathrm{T}} x_i)|}}{\sum_{i=1}^{n} \frac{1}{|y_i - (\hat{b} + \hat{w}^{\mathrm{T}} x_i)|}}.$$

We keep updating the w_j's and b until the inner loop reaches convergence.

Back to Outer loop: We re-approximate the objective function by the newly obtained $(\hat{b}, \hat{w})$. Then, the whole algorithm is iterated until the outer loop reaches convergence.

The Asymptotic Property of the Marginal Estimators

Suppose that we observe a random sample $(\mathbf{x_1}, y_1), ..., (\mathbf{x_n}, y_n)$ from the distribution of a vector $(\mathbf{X}, Y)$, which follows the logit model:

$$P(Y = 1|\mathbf{X}) = \frac{\exp(\alpha^0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)}{\exp(\alpha^0 + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0) + 1},$$

where $\mathbf{X} = (X_1, ..., X_p)^{\mathrm{T}}$ and $\boldsymbol{\beta}^0 = (\beta_1^0, ..., \beta_p^0)^{\mathrm{T}}$. Here, α^0 represents the true intercept and $\boldsymbol{\beta}^0$ represents the vector of the true regression-coefficients. It is well-known that if one fits a joint logistic regression model for all the p covariates (SNPs), then the estimator for $\boldsymbol{\beta}^0$ is consistent. Now, suppose that one fits the marginal logistic regression model, i.e., the model with only one SNP.(Not to lose generality, assume that the first SNP is included in the model). Under this mis-specified model, we show in the sequel that, even if all the p SNPs are independent, the marginal estimator of β_1^0 is not consistent. For simplicity, we first consider the situation where $\alpha^0 = 0$.

Situation I: $\alpha^0 = 0$

Assume that we include only X_1 in the model (without intercept). Under this situation, we maximize the 'mis-specified' log-likelihood

$$\sum_{i=1}^{n} [x_{i1}\beta_1 y_i - \log(1 + e^{x_{i1}\beta_1})]$$

with respect to β_1, and name the maximizer as $\hat{\beta}_1$.

To derive the limiting quantity of $\hat{\beta}_1$, define $m_{\beta_1} \equiv [X_1\beta_1 Y - \log(1 + e^{X_1\beta_1})]$. Then

$$E(m_{\beta_1}|\mathbf{X}) \;\; = \;\; E[(X_1\beta_1 Y - \log(1 + e^{X_1\beta_1})|\mathbf{X}],$$

which, evaluated at $Y = 1$ and $Y = 0$, equals to

$$\left[(X_1\beta_1 \times 1 - \log(1 + e^{X_1\beta_1})\right] \times P(Y = 1) + [(X_1\beta_1 \times 0 - \log(1 + e^{X_1\beta_1})] \times P(Y = 0).$$

Now, plugging $P(Y = 1) = \frac{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)}{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)+1}$ and $P(Y = 0) = \frac{1}{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)+1}$ into the above

equation, we have

$$E(m_{\beta_1}|\mathbf{X}) \;=\; X_1\beta_1\frac{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)}{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)+1} - \log(1+e^{X_1\beta_1}).$$

Then, $E(m_{\beta_1}) = E(E(m_{\beta_1}|\mathbf{X})) = E\left[X_1\beta_1\frac{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)}{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)+1} - \log(1+e^{X_1\beta_1})\right]$. Next, we need to find the maximizer of $E(m_{\beta_1})$.

$$0 = \frac{\partial E(m_{\beta_1})}{\partial\beta_1} = E[X_1\frac{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)}{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)+1} - \frac{e^{X_1\beta_1}}{1+e^{X_1\beta_1}}X_1]$$

Therefore,

$$E[X_1\frac{e^{X_1\beta_1}}{1+e^{X_1\beta_1}}] = E[X_1\frac{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)}{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)+1}],$$

which, after a little algebra, is equivalent to

$$E[X_1\frac{1}{1+e^{X_1\beta_1}}] = E[X_1\frac{1}{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)+1}]. \tag{5.7}$$

Name the solution to the above equation as β_1^*. By the Theorem 5.23 of Van Der Vaart (1998), β_1^* is the limiting quantity of $\hat{\beta}_1$.

In general, there is no closed form for β_1^*. In the context of SNP study, it is common to assume that X_j follows a multinomial distribution with the outcome being (0,1,2), and the associated probabilities being $(q_j^2, 2q_j(1-q_j), (1-q_j)^2)$. Then, the LHS of equation (5.7) is equal to

$$0\times q_1^2 + \frac{1}{1+e^{\beta_1}}\times 2q_1(1-q_1) + \frac{1}{1+e^{2\beta_1}}\times(1-q_1)^2. \tag{5.8}$$

Now, we evaluate the RHS of (5.7). For $j = 2,\ldots,p$, let $k_j \in \{0,1,2\}$. Let $P_{k_j} = q_j^2 I(k_j=0) + 2q_j(1-q_j)I(k_j=1) + (1-q_j)^2 I(k_j=2)$. Then, if all the p SNPs

are independent, the RHS of (5.7)

$$E[\frac{X_1}{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)+1}] = 2q_1(1-q_1)\sum_{k_2=0}^{2}\cdots\sum_{k_p=0}^{2}\left\{\frac{1}{1+e^{\beta_1^0+k_2\beta_2^0+\ldots+k_p\beta_p^0}}\prod_{j=2}^{p}P_{k_j}\right\}+ (1-q_1)^2\sum_{k_2=0}^{2}\cdots\sum_{k_p=0}^{2}\left\{\frac{2}{1+e^{2\beta_1^0+k_2\beta_2^0+\ldots+k_p\beta_p^0}}\prod_{j=2}^{p}P_{k_j}\right\}. \quad (5.9)$$

Relating (5.8) to (5.9), it is clear that β_1^* is not equal to β_1^0. This implies that β_1^* is not a consistent estimator for β_1^0. It is not difficult to show that, β_1^* is monotone with respect to β_1^0. While there is no closed form solution for β_1^*, we give an approximate solution here. Approximating both e^{β_1} and $e^{2\beta_1}$ by $e^{1.5\beta_1}$ in (5.8), we get the solution

$$\beta_1^* \approx \frac{2}{3}\log\left\{\frac{(1-q_1^2)}{E[\frac{X_1}{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0)+1}]}-1\right\}.$$

This suggests that the limiting quantity for the marginal estimator $\hat{\beta}_1$ is related not only to SNP X_1 and β_1^0, but also to all the other SNPs in the true model as well as their effect sizes.

Situation II: $\alpha^0 \neq 0$

Now, assume that the true intercept $\alpha^0 \neq 0$. Suppose that we fit the marginal model with the first SNP and an intercept α, and let $\hat{\beta}_1$ and $\hat{\alpha}$ denote the corresponding coefficients estimators. Under this situation, it can be shown that the limiting equations contain two parts,

$$E[\frac{1}{1+e^{X_1\beta_1+\alpha}}] = E[\frac{1}{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0+\alpha^0)+1}] \quad (5.10)$$

and

$$E[\frac{X_1}{1+e^{X_1\beta_1+\alpha}}] = E[\frac{X_1}{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0+\alpha^0)+1}]. \tag{5.11}$$

The joint solution to the above two limiting equations, $\alpha^\dagger$ and $\beta_1^\dagger$, would be the limiting quantity for $\hat{\alpha}$ and $\hat{\beta}_1$. Under the assumption that X_j follows a multinomial distribution, one can expand the above two equations in a similar manner as that for equation (5.7), which results in

$$\frac{1}{1+e^{\alpha}}q_1^2 + \frac{1}{1+e^{\beta_1+\alpha}}2q_1(1-q_1) + \frac{1}{1+e^{2\beta_1+\alpha}}(1-q_1)^2 = E[\frac{1}{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0+\alpha^0)+1}]$$

and

$$0 + \frac{1}{1+e^{\beta_1+\alpha}}2q_1(1-q_1) + \frac{2}{1+e^{2\beta_1+\alpha}}(1-q_1)^2 = E[\frac{X_1}{\exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^0+\alpha^0)+1}].$$

With a similar argument as that in Situation I, we can show that

(1) in general, $\hat{\alpha}$ is not a consistent estimator for α^0, nor is $\hat{\beta}_1$ consistent for β_1^0;

(2) because $\alpha^\dagger$ and $\beta_1^\dagger$ are intertwined, $\beta_1^\dagger$ is not necessarily monotone with β_1^0.

Bibliography

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. *Second International symposium on information theory*, pp. 267-281.

Altshuler,D. *et al.* (2008) Genetic mapping in human disease. *Science*, **322**, 881-888.

ARIC Investigators. (1989) The Atherosclerosis Risk in Communities (ARIC) Study: Design and Objectives. *Am. J. Epidemiol.*, **129**, 687-702.

Avery, C. L. *et al.* (2011) A Phenomics-Based Strategy Identifies Loci on APOC1, BRAP, and PLCG1 Associated with Metabolic Syndrome Phenotype Domains. *PLos Genet.*, **7**, e1002322.

Barrett, J.C. *et al.* (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.*, **40**, 955-962.

Barrett, J.C. *et al.* (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.*, **41**, 703-707.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289-300.

Bild, D.E. *et al.* (2002) Multi-ethnic Study of Atherosclerosis: Objectives and Design. *Am. J. of Epidemiol.*, **156**, 871-881.

Breiman, L. (1995) Better subset regressionusing the nonnegative garrote. *Technometrics*, **37**, 373-384.

Burton,P.R.*et al.* (2009) Size matters: just how big is BIG?: Quantifying realisitic sample size requirements for human genome epedimiology. *Int. J. Epidemiol.*, **38**, 263-273.

Candes E. and Tao T. (2007) The Dantzig selector: statistical estimatoion when p is much larger than n. *Ann. Statist.*, **35**, 2313-2351.

Chen,J. and Chen,Z. (2008) Extended Baysian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759-771.

Collins, F. (2010) Has the revolution arrived? *Nature*, **464**, 674-675.

Collins, G.S. et al. (2011) Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Medicine*, **9**, 103.

Dobra,A. (2007) Variable selection and dependency networks for genomewide data. *Biostatistics*, **8**, 1-18.

Efron, B.*et al.*(2004) Least angle regression. *Ann. Statist.*, **32**, 407-499.

Evangelou, E. *et al.* (2007) Meta-analysis in genome-wide association datasets: strategies and application in Parkinson disease. *PLos ONE*, **2**: e196.

Evans, D.M. *et al.* (2009) Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum. Mol. Genet.*, **2009**: 3525-3531.

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

Fan,J and Lv,J. (2008) Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B*, **70**, 849-911.

Fan, J. and Peng, H. (2004) Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.*, **32**, 928-961.

Fan J. and Song R. (2010) Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, **38**, 3567-3604.

Fan,J. *et al.* (2009) Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.*, **10**, 2013-2038.

Ferreira M. and Purcell S. (2009) A multivariate test of association. *Bioinformatics*, **25**, 132–133.

Frank,I.E. and Friedman,J.H. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109-135.

Franke,L. *et al.* (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011-1025.

Friedman, G.D.*et al.* (1988) CARDIA: Study Design, Recruitment, and Some Characteristics of the Examined Subjects. *J. Clin. Epidemiol.*, **41**, 1105-1116.

Friedman, R. *et al.* (2007) Pathwise coordinate optimization. *Ann. Appl. Stat.* **1**, 302–332.

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1-22.

Furey, T.S. *et al.* (2000) Support vector machine classfication and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906-914.

Guyon,I. *et al.* (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389-422.

Genkin,A. *et al.* (2007) Large-scale Bayesian logistic regression for text categorization. *Technometrics*, **49**, 291-304.

Han,B. *et al.* (2010) A Markov blanket-based method for detecting causal SNPs in GWAS. *BMC Bioinform.*, **11**(Suppl. 3), S5.

Hastie,T. *et al.* (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd edn. Springer-Verlag, New York, pp.75-76.

He, Q. and Lin, D.Y. (2011) A variable selection method for genome-wide association studies. *Bioinformatics*, **27**, 1–8.

Herrmann,J. *et al.* (2009) Isomer-specific effects of CLA on gene expression in human adipose tissue depending on PPARγ P12A polymorphism: a double blind, randomized, controlled cross-over study. *Lipids Health Dis.*, **8**:35.

Hoggart,C.J. *et al.* (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.*, **4**, e1000130.

Hofmann,S.M. *et al.* (2007) Adipocyte LDL receptorrelated protein1 expression modulates postprandial lipid transport and glucose homeostasis in mice. *J. Clin. Invest.*, **117**, 3271-3282.

Hu, P. *et al.* (2011) Pathway-based joint effects analysis of rare genetic variants using Genetic Analysis Workshop 17 exon seqence data. *BMC Proc.*, **5**(Suppl 9): S45.

Huang, J. *et al.* (2008) Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587-613.

Huang, J. *et al.* (2009) A Group Bridge Approach for Variable Selection. *Biometrika*, **96**, 339-355

Hudson,R.R. (2002) Generating samples under a Wright-Fisher neutral model. *Bioinformatics*, **18**, 337-378.

Ioannidis J.P. (2005) Why most publised research findings are false. *PLoS Med*, **2**: e124.

Ioannidis, J.P.*et al.* (2007) Heterogeneity in meta-analyses of genome-wide asscociation investigations. *PLos ONE*, **2**: e841.

James, B.M. *et al.* (2008) Genotype score in addition to common risk factors for prediction of Type 2 Diabetes. *N. Engl. J. Med.*, **359**, **359**, 2208-2219.

Janssens A.C. and van Duijn C.M. (2008) Genome-based prediction of common diseases: advances and prospects. *Hum. Mol. Genet.*, **17**, R166-R173.

Jirtle R.L. and Skinner M.K. (2007) Environmental epigenomics and disease susceptibility. *Nat. Rev. Genet.*, **8**, 253-262.

Johnson B. *et al.* (2008) Penalized estimating functions and variable selection in semi-parametric regression models. *J. Am. Stat. Ass.*, **103**, 672-680.

Josins, L. and Barret, J.C. (2011) Genetic risk prediction in complex disease. *Hum. Mol. Genet.*, Advanced publication.

Kang, J. *et al.* (2010) Practical issues in building risk-prediction models for complex diseases. *J. Biopharm. Stat.*, **20**, 415-440.

Knight K. and Fu W. (2000) Asymptotics for lasso-type estimators. *Ann. Statist.*, **28**, 1356-1378.

Kooperberg C. *et al.* (2010) Risk prediction using genome-wide association studies. *Genet. Epi.*, **34**, 643-652.

Kraft, P. and Hunter, D.J. (2009) Genetic risk prediction-Are we there yet? *N. Engl. J. MED.*, **360**, 1701-1703.

Ku C.S. *et al.* (2010) The pursuit of genome-wide association studies: where are we now? *J. Hum. Genet.*, **55**, 195-206.

Li,C. and Li,M. (2008) GWAsimulator: a rapid whole-genome simulation program. *Bioinformatics*, **24**, 140-142.

Li,K.-C. (2002) Genome-wide coexpression dynamics: theory and application. *Proc. Natl. Acad. Sci. USA*, **99**, 16875-16880.

Lin, D.Y. and Zeng, D. (2010) On the relative efficiency of using summary statistics versus individual level data in meta-analysis. *Biometrika* **97**, 321-332.

Lin, D.Y. and Zeng, D. (2010) Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genet. Epid.*, **34**, 60-66.

Lindgren, C.M. *et al.* (2009) Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLos Genet.*, **5**:e1000508.

Lu, Q. and Elston, R. (2008) Using the optimal receiver operating characteristic curve to design a predictive genetic test exemplified with Type 2 Diabetes. *Am. J. Hum. Genet.*, **82**, 641-651.

Lulianella,A. *et al.* (2008) Cux2 (Cutl2) integrates neural progenitor development with cell-cycle progression during spinal cord neurogenesis . *Development*, **135**, 729-741.

Lusted, L.B. (1971) Signal dectectabiligy and medical dicision-making. *Science*, **171**, 1217-1219.

Ma, S. *et al.* (2007) Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics*, **8**:60.

Ma, S. *et al.* (2011) Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics* Advance publication.

Machiela, M.J.*et al.* (2011) Evaluation of Polygenic risk scores for predicting breast and prostate cancer risk. *Genet. Epid.*, **35**, 506-514.

Manolio, T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747-753.

McCarthy M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356-369.

Meier, L. *et al.* (2008) The group lasso for logistic regression. *J. R. Stat. Soc. Ser. B*, **70**, 53-71.

Meinshausen,N. *et al.* (2009) P-values for high-dimensional regression. *J. Am. Stat. Ass.*, **104**, 1671-1681.

Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. R. Stat. Soc. Ser. B*, **72**, 417-448.

Noble J. (2006). Meta-analysis: methods, strengths, weaknesses, and political uses. *J. Lab. Clinic. Med.* **147**, 7–20.

Peng J.*et al.* (2010) Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.*, **4**, 53–77.

Ruderfer, D.M. *et al.* (2010) Family-based genetic risk prediction of multifactorial disease. *Genome Med.*, **2**, 1:7.

Scott, L.J. *et al.* (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, **316**: 1341-1345.

Seddon, J.M. *et al.* (2009) Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and enviomental variables. *Invest. Ophth. Vis. Sci.*, **50**: 2044-2053.

Su,H.P. *et al.* (2002) Interaction of CED-6/GULP, an adapter protein involved in engulfment of apoptotic cells with CED-1 and CD91/low density lipoprotein receptor-related protein (LRP). *J. Biol. Chem.*, **281**, 12081-12092.

Schwarz, D. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464.

The international schizophrenia consortium. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 748-752.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58**, 267-288.

van der Vaart, A.W. (1998) Asymptotic Statistics. Cambridge University Press.

Vapnik, V.N. (1999) An overiew of statistical learning theory. *IEEE transactions on neural networks*, **10**, 988-999.

Wang, H. and Leng, C. (2007). Unified LASSO estimation by least squares approximation. *J. Am. Stat. Ass.* **102**, 1039–1048.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Am. Stat. Ass.* **104**, 1512–1524

Wang L. (2011) GEE analysis of clustered binary data with diverging number of covariates. *Ann. Statist.*, in press.

Wang L. *et al.* (2011) Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, in press.

Wang, S. *et al.* (2009) Hierarchically penalized Cox regression for censored data with grouped variables and its oracle property. *Biometrika*, **96**, 307-322.

Webb, R. *et al.* (2011) Early disease onset is predicted by a higher genetic risk for lupus and is associated with a more severe phenotype in lupus patients. *Ann. Rheum. Dis.*, **70**, 151-156.

Wei L.J. *et al.* (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Am. Stat. Ass.*, **84**, 1065-1073.

Wei,R. *et al.* (2009) From disease association to risk assesement: an optimistic view from genome-wide association studies on type 1 diabetes. *PLos Genet.*, **5**, e1000678.

Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661-678.

Wray,N.R. *et al.* (2007) Prediction of individual risk to disease from genome-wide association studies. *Genome Res.*, **17**, 1520-1528.

Wu,T.T. and Lange,K. (2008) Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.*,**2**, 224-244.

Wu,T.T. *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714-721.

Yohai, V.J., and Maronna, R.A. (1979). Asymptotic Behavior of M-estimator for the Linear Model. *Ann. Statist.*, **7**, 258-268.

Yuan, M. and Lin, Y. (2006) Model Selection and Estimation in Regression with Grouped Variables. *J. R. Stat. Soc. Ser. B*, 68, 49-67.

Yuan,M. and Lin, Y. (2007) On the non-negative garrotte estimator. *J. R. Stat. Soc. Ser. B*, **69**, 143-161.

Zeggini,E. *et al.* (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, **316**, 1336-1341.

Zeggini,E. *et al.* (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.*, **40**, 638-645.

Zeggini E. and Ioannidis JP. (2009) Meta-analysis in genome-wide association studies. *Pharmacogenomics* **10**, 191–201.

Zeisel S. (2007) Nutrigenomics and metabolomics will change clinical nutrition and public health practice: insights from studies on dietary requirements for choline. *Am. J. Clinic. Nutri.* **86**, 542–548.

Zhang H. *et al.* (2006) Gene selection using support vector machines with non-convex penalty. *Bioinformatics*, **22**, 88-95.

Zhao, S. and Li, Y. (2010) Principled sure independence screening for Cox models with ultra-high-dimensional covariates. Tech. rep., Harvard School of Public Heath.

Zhou *et al.* (2010) Association screening of common and rare genetic variants by peanlized regression. *Bioinformatics*, **26**, 2357-2382.

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

Zou,H. (2006) The adaptive Lasso and its oracle properties. *J. Am. Stat. Ass.*, **101**, 1418-1429.

Zou,H.*et al.* (2007) On the degrees of freedom of the LASSO. *Ann. Statist.*, **35**, 2173-2192.

Zou,H. and Li,R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, **36**, 1509-1533.

Zou,H. and Zhang, H.H. (2011) On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.*, **37**, 1733-1751.