Penalized variable selection for gene-environment interactions

by

Yinhao Du

B.S., Emporia State University, 2013

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Abstract

Gene-environment (G×E) interaction is critical for understanding the genetic basis of complex disease beyond genetic and environment main effects. In addition to existing tools for interaction studies, penalized variable selection emerges as a promising alternative for dissecting G×E interactions. Despite the success, variable selection is limited in the following aspects. First, multidimensional measurements have not been taken into fully account in interaction studies. Published variable selection methods cannot accommodate structured sparsity in the framework of integrating multiomics data for disease outcomes. Second, in the big data context, no variable selection method has been developed so far to conduct tailored interaction analysis. Third, the solution to case control association G×E studies with high dimensional genomics variants in the big data context has not been made available so far. In this dissertation, we tackle these challenges rising from G×E interaction studies in the modern era through the following projects.

In the first project, we have developed a novel variable selection method to integrate multi-omics measurements in G×E interaction studies. Extensive studies have already revealed that analyzing omics data across multi-platforms is not only sensible biologically but also resulting in improved identification and prediction performance. Our integrative model can efficiently pinpoint important regulators of gene expressions through sparse dimensionality reduction and link the disease outcomes to multiple effects in the integrative G×E studies via accommodating a sparse bi-level structure. Simulation studies show the integrative model leads to better identification of G×E interactions and regulators than that of the alternative methods. In two G×E lung cancer studies with high dimensional multi-omics data, the integrative model leads to improved prediction and findings with important biological implications.

In the second project, we propose to conduct interaction studies in the big data context by adopting the divide-and-conquer strategy. In particular, the sparse group variable selection for important G×E effects has been developed within the framework of alternating direction method of multiplier (ADMM). To accommodate the large-scale data in terms of either samples or features, we have developed two novel parallel ADMM based variable selection methods across samples and features, respectively. The corresponding parallel algorithms can be efficiently implemented in distributed computing platforms. Simulation studies demonstrate that the parallel ADMM based penalization methods significantly improve the computational speed for analyzing large scale data from G×E interaction studies with satisfactory identification and prediction performance.

In the third project, we extend the proposed parallel ADMM based variable selection for G×E interactions in the case-control association study of type 2 diabetes. Within the parallel computation framework, we have developed a penalized logistic regression model accommodating the bi-level selection tailored for the case control G×E interaction study. The advantage of the proposed parallel penalization method has been fully illustrated in the distributed learning scenario. Simulation studies show the proposed method dramatically reduces the computational time while maintaining a competitive performance compared to the non-parallel counterparts. In the case study of type 2 diabetes with environmental factors and high dimensional SNP measurements, the proposed parallel penalization method leads to the identification of biologically important interaction effects.

Penalized variable selection for gene-environment interactions

by

Yinhao Du

B.S., Emporia State University, 2013

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
Cen Wu

Copyright

Abstract

Gene-environment (G×E) interaction is critical for understanding the genetic basis of complex disease beyond genetic and environment main effects. In addition to existing tools for interaction studies, penalized variable selection emerges as a promising alternative for dissecting G×E interactions. Despite the success, variable selection is limited in the following aspects. First, multidimensional measurements have not been taken into fully account in interaction studies. Published variable selection methods cannot accommodate structured sparsity in the framework of integrating multiomics data for disease outcomes. Second, in the big data context, no variable selection method has been developed so far to conduct tailored interaction analysis. Third, the solution to case control association G×E studies with high dimensional genomics variants in the big data context has not been made available so far. In this dissertation, we tackle these challenges rising from G×E interaction studies in the modern era through the following projects.

In the first project, we have developed a novel variable selection method to integrate multi-omics measurements in G×E interaction studies. Extensive studies have already revealed that analyzing omics data across multi-platforms is not only sensible biologically but also resulting in improved identification and prediction performance. Our integrative model can efficiently pinpoint important regulators of gene expressions through sparse dimensionality reduction and link the disease outcomes to multiple effects in the integrative G×E studies via accommodating a sparse bi-level structure. Simulation studies show the integrative model leads to better identification of G×E interactions and regulators than that of the alternative methods. In two G×E lung cancer studies with high dimensional multi-omics data, the integrative model leads to improved prediction and findings with important biological implications.

In the second project, we propose to conduct interaction studies in the big data context by adopting the divide-and-conquer strategy. In particular, the sparse group variable selection for important G×E effects has been developed within the framework of alternating direction method of multiplier (ADMM). To accommodate the large-scale data in terms of either samples or features, we have developed two novel parallel ADMM based variable selection methods across samples and features, respectively. The corresponding parallel algorithms can be efficiently implemented in distributed computing platforms. Simulation studies demonstrate that the parallel ADMM based penalization methods significantly improve the computational speed for analyzing large scale data from G×E interaction studies with satisfactory identification and prediction performance.

In the third project, we extend the proposed parallel ADMM based variable selection for G×E interactions in the case-control association study of type 2 diabetes. Within the parallel computation framework, we have developed a penalized logistic regression model accommodating the bi-level selection tailored for the case control G×E interaction study. The advantage of the proposed parallel penalization method has been fully illustrated in the distributed learning scenario. Simulation studies show the proposed method dramatically reduces the computational time while maintaining a competitive performance compared to the non-parallel counterparts. In the case study of type 2 diabetes with environmental factors and high dimensional SNP measurements, the proposed parallel penalization method leads to the identification of biologically important interaction effects.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost I would like to thank my major advisor Dr. Cen Wu. He has taught me, both consciously and unconsciously, how good research is done. I appreciate all his contributions of time and ideas to make my Ph.D. experience productive and stimulating. I am also thankful for the excellent example he has provided as a successful statistician and professor. Besides my advisor, I would like to thank my committee members: Dr. Weixing Song, Dr. Michael Higgins and Dr. Weiqun Wang for their time, interest, and valuable guidance through my study process. My appreciation also goes to Dr. David Renter for his willingness to serve as the chairperson of the examining committee for my doctoral degree.

I would like to thank the department of statistics for offering me graduate assistantships so that I could come to the states and complete my graduate studies at Kansas State University. I would like to thank everyone in the department for their kindness, thank all the professors in the department for the their excellent courses and for their help. There are no words coming to close to describing the gratitude to my friends. It has been a memorable time to study and discuss statistical questions with them. Lastly, I would like to thank my family for all their love and encouragement. For my parents who raised me with a full of love and supported me in all my pursuits. And most of all for my loving, supportive, encouraging, and patient husband Xiongya Li whose faithful support during the final stages of this Ph.D. is so appreciated

Chapter 1

Introduction

Recent human disease studies have shown that gene-environment interaction effects are associated with cancer outcomes beyond genetic and environment main effects. However, most of existing cancer research studies have only focused on genetic main effects, and fewer studies have considered the interaction effects between genetics and environmental factors. In this dissertation, we develop novel and powerful statistical models for identifying important genetics main and gene-environment (G×E) interaction effects. As interactions are described using the product between variables, the identification of important genetics main and G×E interaction effects is a high-dimensional problem, in which sample size (n) is much smaller than number of variables (p). Penalized variable selection is one of the most popular approaches for analyzing high-dimensional data, see Wu et al. (2019). This section provides a brief overview for penalized variable selection methods in high-dimensional data in general. We also summarize two major computational frameworks for penalized variable selection. We mainly review the two most popular frameworks, coordinate descent (CD) (Wu et al. (2008)) and alternative direction method of multipliers (ADMM) (Boyd et al. (2011)). In the last section, we discuss our motivations on G×E studies and the contributions in this dissertation.

1.1 Penalized Variable Selection

Variable selection through penalization has become popular for biomedical and bioinformatics studies in last two decades. As the number of genes is usually larger than the sample size, this type of study is of "large data dimensionality, small sample size" nature. For cancer (and complex disease) studies, one of the most important questions is how to select a subset of important genes that are associated with the disease. The question can be recasted as a variable selection problem. There's no doubt that high dimensional variable selection is one of the most important research topics in statistics (Fan and Lv (2010), Wu et al. (2019)).

Let X be the design matrix, where it has p columns of variables, such as gene expressions or single nucleotide polymorphisms (SNPs), and n rows of samples. Let Y be the disease outcome, such as continuous disease phenotype, categorical disease status or survival time. Penalization is one of the most important frameworks for variable selection. The model can be expressed as

$$L(\beta; Y, X) + P_\lambda(\beta),$$

where $L(\cdot)$ is the loss function measuring the goodness-of-fit, and $P_\lambda(\cdot)$ is the penalty function with tuning parameter λ to control shrinkage and sparsity on the coefficient β. By minimizing the above penalized loss function, parameter estimation and variable selection can be achieved simultaneously. LASSO (Tibshirani (1996)) is a well-known variable selection approach in high-dimensional data analysis. It has the following form

$$\frac{1}{2n}\|Y - X\beta\|^2 + \lambda|\beta|,$$

where $L(\cdot)$ is least square loss function and $P_\lambda(\cdot)$ is ℓ_1 penalty. It can select important variables through shrinking the coefficients of unimportant variables to zeros. Nowadays, LASSO is the baseline among the family of penalization methods. The development of penalties becomes more advanced for desired complicated data structure and estimator properties such as sparsity, continuity, unbiasedness and so on. For example, widely adopted penalties with unbiasedness property of estimators include smoothly clipped absolute deviation (SCAD)

(Fan and Li (2001)), the minimax concave penalty (MCP) (Zhang et al. (2010)), adaptive LASSO (Zou (2006)).

- SCAD

$$P_{\lambda,\gamma}(\beta) = \begin{cases} \lambda|\beta|, & |\beta| \leq \lambda \\ -\frac{\beta^2 - 2\gamma\lambda|\beta| + \lambda^2}{2(\gamma-1)}, & \lambda < |\beta| \leq \gamma\lambda \\ \frac{1}{2}(\gamma+1)\lambda^2, & |\beta| > \gamma\lambda \end{cases}$$

- MCP

$$P_{\lambda,\gamma}(\beta) = \begin{cases} \lambda|\beta| - (2\gamma)^{-1}\beta^2, & |\beta| \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & |\beta| > \gamma\lambda \end{cases}$$

- Adaptive LASSO

$$P_{\lambda,\gamma}(\beta) = \lambda(|\beta^{(0)}|^{-\gamma})|\beta|,$$

where $\gamma > 2$ for SCAD, $\gamma > 1$ for MCP and $\gamma > 0$ for adaptive LASSO. In adaptive LASSO, $\beta^{(0)}$ is the initial estimate of β. For a detailed review on variable selection and its applications in biology studies, please refer to Ma and Huang (2008), Fan and Lv (2010), and Wu and Ma (2014).

For complex data structures, more advanced penalty functions need to be considered. For example, group LASSO proposed by Yuan and Lin (2006) can be used to identify grouped variables. This penalty is

$$P_\lambda(\beta) = \lambda \sum_{g=1}^{G} \sqrt{d_g}\|\beta_g\|_2,$$

where β_g is the group coefficient vector with length d_g and G defines total number of groups. The term $\sqrt{d_g}$ adjusts the penalty function for the group size and $\|\cdot\|_2$ is the euclidean norm. Some other research studies also have been developed for grouping data structure (Huang et al. (2012)).

Another example is to account for high correlation among genetic features. Elastic net (Zou and Hastie (2005)) and fused-lasso (Tibshirani et al. (2005)) are among the most

widely used variable selection methods to analyze correlated genomic features. The elastic net penalty function is formulated as a combination of LASSO and ridge penalties,

$$P_{\lambda_1,\lambda_2}(\beta) = \lambda_1 \|\beta\|_2^2 + \lambda_2 |\beta|,$$

where ridge penalty accommodates correlations and LASSO penalty impose sparsity. While the fused-lasso has the following penalty function,

$$P_{\lambda_1,\lambda_2}(\beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|,$$

where LASSO penalty imposes sparsity and fusion penalty induces smoothness among the coefficients of neighboring features. In past few years, more advanced penalization methods have been developed to accommodate data structure in more efficient and flexible ways. For example, network-constrained regularization, including Li and Li (2008) and Huang et al. (2011), are among these methods. For example, Huang et al. (2011) has developed a sparse Laplacian shrinkage (SLS) penalty

$$P_{\lambda_1,\lambda_2,\gamma}(\beta) = \sum_{j=1}^{p} P_{\lambda_1,\gamma}(\beta_j) + \lambda_2 \sum_{1 \le j \le k \le p} |a_{jk}| (\beta_j - \text{sign}(a_{jk})\beta_k)^2$$

where $P_{\lambda_1,\gamma}(\beta_j)$ is the MCP (Zhang et al. (2010)) penalty to impose sparsity and Laplacian quadratic accounts for the correlation structure. In their paper, they also show that SLS has selection consistent property and its estimator is equivalent to oracle Laplacian shrinkage estimator with high probability. Other network-based penalization discussions can be found in Ren et al. (2017), Ren et al. (2019), Huang et al. (2018a), Kim and Sun (2019).

1.2 Computational Algorithms

With the development of the penalized variable selection approaches, efficient computation algorithms play a critical role. Many efficient computational algorithms have been initially

developed for LASSO, which is the baseline of the penalization approaches.Tibshirani (1996) introduced the inequality constrains to obtain feasible solution based on karush-Kuhn-Tucker (KKT) conditions. Efron et al. (2004) proposed a least angle regression (LARS) to obtain LASSO solution. Later, a couple of improved LARS algorithms have been developed, see Keerthi and Shevade (2007) and Zhou et al. (2013). MCMC has been the most widely adopted to achieve fast computation with the Gibbs samplers for Bayesian studies. Green (1990) derived a penalized EM algorithm, Beck and Teboulle (2009) derived an iterative shrinkage-thresholding algorithms (ISTA) for solving linear inverse problems and many other similar studies. Among all different computation frameworks, coordinate descent (CD) and alternative direction method of multipliers (ADMM) are the most widely adopted in penalized variable selection. In this dissertation, we'll focus on the frameworks of CD and ADMM.

Coordinate descent (CD) is one of the primary frameworks in high dimensional data analysis. The process of CD is to optimize the penalized loss function with respect to a single parameter at a time, iteratively cycling through all parameters until convergence is reached. For both robust and non-robust loss functions with convex and more complicated penalty function, CD can solve a large family of optimization problems based on first order methods including gradient-,sub gradient- and proximal- gradient based methods. To be specific, we derive the following CD for LASSO as an example. Consider a least square loss function with ℓ_1 penalty

$$L(\beta;\lambda) = \frac{1}{2n}\|Y - X\beta\|^2 + \lambda|\beta| = \frac{1}{2n}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

$$L(\beta;\lambda) = \frac{1}{2n}\sum_{i=1}^{N}(y_i - \sum_{k\neq j}^{p} x_{ik}\beta_k - x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

Consider the update of jth covariate coefficient, $j = 1, \cdots, p$,

$$L_j(\beta_j; \lambda) = \frac{1}{2n} \sum_{i=1}^{N} (r_i - x_{ik}\beta_j)^2 + \lambda|\beta_j|$$

where $r_i = y_i - \sum_{k \neq j}^{p} x_{ij}\beta_k$ is called partial residuals with respect to the jth covariate. Then the update of β_j is the minimizer of $L_j(\beta_j; \lambda)$. The process of LASSO CD is summarized in **Algorithm 1**.

Algorithm 1 LASSO Coordinate Descent Algorithm

Initialize $\beta_j^{(0)} = 0$, $m = 1$.
repeat
  for $j = 1, \cdots, p$ **do**
    Calculate $r_i = y_i - \sum_{k \neq j} x_{ik}\beta_k$ and $z_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij} r_i + \beta_j^{(m)}$.
    Update $\beta_j^{(m+1)} \leftarrow sgn(z_j)(|z_j| - \lambda)_+$.
    Update $r_i \leftarrow r_i - (\beta_j^{(m+1)} - \beta_j^{(m)})x_{ij}$, for $i = 1, \cdots, N$.
    Update $m = m + 1$
  end for
until convergence

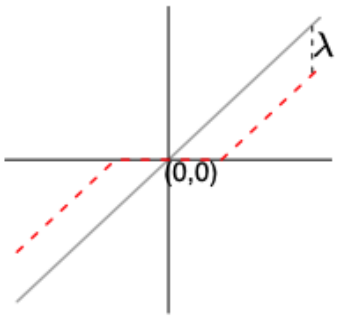


Figure 1.1: Soft threshold function (Hastie (2008))

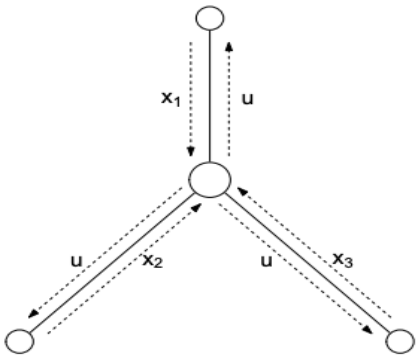


Figure 1.2: Dual decomposition (Boyd et al. (2011))

Besides, ADMM (Boyd et al. (2011)) can be considered as another major algorithm for penalized variable selection. It solves convex optimization based on dual ascent and augmented Lagrangian method. ADMM algorithm can handle multiple constraints in opti-

mization. To be specific, we show the following LASSO ADMM as an example. First, we need to introduce an extra variable z such that

$$\frac{1}{2n}\|Y - X\beta\|^2 + \lambda|z|, \text{ subject to } \beta - z = 0.$$

The augmented Lagrangian with dual variable u and predefined tuning parameter $1/\rho > 0$ is

$$\frac{1}{2n}\|Y - X\beta\|^2 + \lambda|z| + \frac{1}{\rho}u^T(\beta - z) + \frac{1}{2\rho}\|\beta - z\|^2.$$

Clearly, those two extra terms do not change the problem. To update β, we simplify the function with β related terms

$$\min_{\beta} \frac{1}{2n}\|Y - X\beta\|^2 + \frac{1}{\rho}u^T(\beta - z) + \frac{1}{2\rho}\|\beta - z\|^2,$$

which gives a closed form solution

$$\beta = \left(X^TX + \frac{1}{\rho}I\right)^{-1}\left(X^TY + \frac{1}{\rho}(z - u)\right).$$

To update z, we simplify the function with z related terms

$$\min_{z} \lambda|z| + \frac{1}{\rho}u^T(\beta - z) + \frac{1}{2\rho}\|\beta - z\|^2,$$

which also leads to a closed form solution (soft thresholding solution)

$$z = \text{sign}(\beta + u)(|\beta + u| - \lambda\rho)_+.$$

The dual update rule is

$$u^{(k)} = u^{(k-1)} + \frac{1}{\rho}\left(\beta^{(k)} - z^{(k)}\right)$$

where k is the iteration step. All the steps above can be done very efficiently. We summarize LASSO ADMM in **Algorithm 2**.

Regarding to the computation speed, ADMM is promising due to its nature as a distributed optimization method. In chapter 3, we demonstrate that by expressing traditional ADMM framework into a parallel form for large number of samples and/or features. The proposed parallel ADMM algorithm can be implemented on distributed computing platform such as Hadoop (Dean and Ghemawat (2008)) and Spark (Zaharia et al. (2010)). It's particularly useful for p, n, or both to be very large. In particular, most of existing penalized variable selection approaches for G×E interaction studies have been developed based on CD, but few studies have considered parallel ADMM. With similar variable selection accuracy, parallel methods are significantly faster (Yu et al. (2017)).

Algorithm 2 LASSO ADMM

Initialize $z^{(0)}, u^{(0)}, \rho, \lambda$, $k = 1$

repeat

$\beta^{(k)} = \left(X^T X + \frac{1}{\rho} I\right)^{-1} \left(X^T Y + \frac{1}{\rho}(z^{(k-1)} - u^{(k-1)})\right)$

$z^{(k)} = \text{sign}(\beta^{(k)} + u^{(k-1)})(|\beta^{(k)} + u^{(k-1)}| - \lambda\rho)_+$

$u^{(k)} = u^{(k-1)} + \frac{1}{\rho}\left(\beta^{(k)} - z^{(k)}\right)$

$k = k + 1$

until convergence

1.3 Gene-Environment (G×E) Interactions

It has been shown that beyond the genetic and environmental main effects, gene–environment (G×E) interactions play a critical role in understanding the genetic basis of complex disease. For instance, in type 2 diabetes, the interactions between gene TCF7L2 and environmental variables, such as physical activity and lifestyle changes has been reported to be associated with the risk of developing type 2 diabetes (Wu and Cui (2013), Wu et al. (2014) and Ren et al. (2020), Ren et al. (2021)). In lung cancer study, the interaction between susceptible genes and smoking status have been revealed to affect the prognosis (Wu et al. (2015) and Wu et al. (2018a)). The high dimensionality of genetic measurements makes identification and characterization of important G×E interactions especially challenging within the traditional

statistical framework.

Despite success, existing studies on G×E interactions still have limitations on identification accuracy, prediction ability and computation efficiency. We develop a bi-level selection to account for the hierarchical structure of main and interaction effects in G×E interactions. The proposed penalized variable selection can achieve better identification accuracy for G×E interactions under high-dimensional settings.

Furthermore, we improve prediction performance of cancer studies by integrating multiple types of omics data, such as gene expression, copy number variation, DNA methylation, mRNA and so on. In Chapter 2, we develop an integrative analysis to predict cancer outcome with multidimensional omics measurements and G×E interactions. The proposed analysis includes two steps. In the first step, a linear regulatory model (LRM) is constructed to analyze the relationship between different types of omics measurements. In the second step, cancer outcome is predicted by LRMs, residuals of LRMs, environment factors and G×E interactions. With penalization, we can select a subset of important genes, regulators, and G×E effects that can improve the prediction of cancer outcome. Simulation studies show that the proposed integrative approach outperforms alternative approaches. In real data analysis, we applied the proposed method to a survival cancer study under the accelerated failure time (AFT) model. The identification of important genetic main effects and G×E interaction effects has sensible biological meanings.

However, CD algorithm is not efficient enough in computation for conducting the proposed penalized variable selection. For a large-scale data analysis, we propose a parallel ADMM framework for G×E in Chapter 3. The parallel ADMM can be performed in a distributed computation platform more efficiently than other algorithms. In Chapter 3, we also show the parallel ADMM framework can improve computation without loosing identification accuracy or prediction ability. The detail derivation for both parallel in samples and features through ADMM is provided. Simulation studies show parallel ADMM can achieve similar prediction accuracy with much less computation time.

In chapter 4, we discuss the 3rd project of this dissertation. We'll develop a novel penalized bi-level selection method for binary outcomes in G×E interaction studies within the

parallel ADMM framework. The proposed method can be adopted to analyze the type 2 diabetes (T2D) case control data with high dimensional genomic variants from the Nurses's Health Studies (NHS). We expect that the distributted ADMM based sparse group identification can lead to important genetic main effects and corresponding G×E interactions that are related to T2D with much less amount of time.

Chapter 2

Integrating Multi–omics Data for Gene-Environment Interactions

2.1 Introduction

Gene-environment interactions reveal how the changes in environmental exposures mediate the contribution of genetic factors to influence the variations in disease traits, which makes it critical for understanding the comprehensive genetic architecture of complex diseases (Simonds et al. (2016), Dempfle et al. (2008)). Traditionally, G×E interaction studies have been mainly conducted within the framework of genetic association studies, in order to hunt down the important main and interaction effects associated with the disease phenotypes (Hirschhorn et al. (2002), Wu et al. (2012)).

Most of the existing G×E studies are one-dimensional in that the interactions between environmental factors and one type of genetic factor (such as gene expression or SNPs) have been considered. In the multi-omics era, there is a pressing need to account for multi-platform measurements in G×E studies. Consider a G×E analysis with environmental factors and gene expression (GE) as the G factors. In addition, DNA methylation (DM) and copy number alterations (CNA), which are the regulators of the genetic factors, are also available. A typical G×E analysis only focus on the interaction effects involving the G factor (GE)

and ignores its regulators, losing the extra power of elucidating the genetic basis of complex disease using multi-level omics data.

Integrating multi-omics data for prognostic outcomes has been mainly conducted using parallel and horizontal integration strategies (Wu et al. (2019)). With the parallel integration, different types of omics measurements are treated equally, and important associations between these measurements and the prognostic outcome are identified in a joint model. The hierarchical integration, on the other hand, fully accounts for the regulatory information by accommodating the indirect effects of regulators, such as DM and CNA, on the prognostic outcomes mediated through GEs. Meanwhile, the direct effects of regulators on the cancer outcomes, which have not been captured by GEs through other mechanisms such as post-transcriptional regulations, should also been taken into consideration.

Given the availability of multi-omics features, the major limitation of existing G×E interaction studies lies in the incapability of integrating regulators in the interaction model under prognostic outcomes, which has motivated us to develop a two stage integrative model for G×E interaction analysis using multi-level cancer omics data. At the first stage, the sparse regulatory relationship has been determined through penalization, where the linear regulatory modelling (Zhu et al. (2016)), or LRM, has been adopted to identify the sets of regulators that influence the sets of GEs, as well as the residuals of gene expression and residuals of regulators that cannot be captured by the LRMs. At the second stage, the LRMs and both types of residuals are treated as direct effects on cancer outcomes in the G×E model, and penalization has been conducted to identify important main and interaction effects.

In the past decade, the effectiveness of regularization for G×E interaction studies has been increasingly witnessed (Zhou et al. (2021)). Extension of the technique for an integrated interaction study is not trivial. Our method significantly advances from existing integration studies not tailored for interaction structures and interaction analysis ignoring the multidimensional omics measurements. Extensive simulation studies have been performed to demonstrate the advantage of the proposed method over multiple alternatives. In two case studies of the lung cancer data (LUSC and LUAD) from TCGA, our method leads

to main and interaction effects with sensible biological implications and improved prediction performance.

2.2 Method

Let $Y_{n\times 1}$ denote cancer outcome, $E_{n\times q} = (E_1, \cdots, E_q)$ denote the q environmental factors, $G_{n\times p_g} = (G_1, \cdots, G_{p_g})$ denote the p_g gene expressions, and $R_{n\times p_r} = (R_1, \cdots, R_{p_r})$ denote the p_r regulators. Suppose we have two measurements for the regulators, p_{r_1} DM and p_{r_2} CNA, then we can obtain $R_{n\times p_r}$ by stacking the measurements together with $p_r = p_{r_1} + p_{r_2}$. Next, we describe overall analysis framework and the integrative model.

2.2.1 Analysis Framework

First, consider a G×E model in the multi-omics scenario where the regulators of the G factors are also included, in addition to the main and interaction effects.

$$Y = \sum_{k=1}^{q} \alpha_k E_k + \sum_{j=1}^{p_g} \left(\beta_j G_j + \sum_{k=1}^{q} \eta_{jk} G_j E_k \right) + \sum_{t=1}^{p_r} \gamma_t R_t + \epsilon, \tag{2.1}$$

where α_k, β_j and η_{jk} are the regression coefficients for the kth environmental factor, jth gene expression and their interactions, respectively. Besides, γ_t is the regression coefficient for the tth regulator, and ϵ is the random error.

Model (2.1) shares the spirit of parallel integration by treating the genetic factor and its regulators equally. Although such a strategy has shown to be effective in several studies, a more attractive alternative is to conduct vertical integration via accounting for the regulatory information among different levels of omics measurements (Wu et al. (2019)). Typically, integrating multi–omics data in a main effect model with prognostic outcomes consists of two steps. At the first step, the sparse regulatory relationship can be identified, leading to gene expressions that are modulated and not modulated by regulators, which can then

be linked to clinical outcomes at the second step (Zhu et al. (2016), Wang et al. (2013)). Specifically, Zhu et al. (2016) proposed the linear regulatory model (LRM) to pinpoint the set of regulators that affect the corresponding set of GEs. Then clinical model incorporates the GEs, residual GEs and residual regulators. In this study, we extend the LRM to investigate the G×E interactions in the presence of multi–level omics measurements. In particular, the prognostic model at the second stage consists of : (1) a low dimensional environmental factors; (2) regulated GEs in the form of LRMs from the first stage and their interactions with those environmental factors; (3) Residual GEs and their interactions with environmental factors, and (4) residual effects of regulators.

2.2.2 Stage 1: the Linear Regulatory Model (LRM)

Denote $g = (g_1, \cdots, g_{p_g})$ as the p_g gene expressions and denote $r = (r_1, \cdots, r_{p_r})$ as the p_r regulators. The LRM can be expressed as

$$E(gV_{p_g \times L}|r) = a_{1\times L} + rU_{p_r \times L}, \tag{2.2}$$

where a is the intercept, $V = (v_1, \cdots, v_L)$ and $U = (u_1, \cdots, u_L)$ both contain L columns of loading vectors (v_l and u_l for $l \in \{1, \cdots, L\}$). Denote L as the total number of LRMs. Here, we assume U and V have orthogonal columns, such that $u_l \perp u_{l^\top}$, $v_l \perp v_{l^\top}$, for $l \neq l^\top$. With this assumption, no overlap between gene expressions and regulators exists in LRM. We expect different LRMs represent different regulated relationship between gene expressions and regulators (Ciriello et al. (2012)). In addition, v_l and u_l are assumed as sparse loading vectors, as only a small number of gene expressions is regulated by at most a small number of regulators (Kristensen et al. (2014)).

For the jth gene expression, $j = 1, \cdots, p_g$, we right multiply $V^\top$ to both sides to simplify equation (2.2). Then the LRM can be formulated as a regression model with response

variable g_j and predictors r:

$$E(g_j) = a_j^\top + r\theta_j, \text{ for } j = 1, \cdots, p_g, \tag{2.3}$$

where $a_j^\top$ is an intercept and θ_j is the regression coefficient vector. Equation (2.3) indicates that one gene expression is regulated by a number of regulators. We impose sparsity on θ_j through penalization to identify sparse regulatory relationship. Then the penalized regression model can be written as

$$\frac{1}{2n}\left\|g_j - a_j^\top - r\theta_j\right\|_2^2 + \lambda|\theta_j|, \text{ for } j = 1, \cdots, p_g, \tag{2.4}$$

where λ is the tuning parameter. The LASSO is adopted for its computational simplicity and satisfactory performance (Tibshirani (1996)). Equation (2.4) leads to a regularized estimate of θ_j, indicating that one gene expression is regulated by a limited amount of regulators.

Next we further investigate the relationship between sets of gene expressions and regulators through singular value decomposition (SVD). The regression model (2.3) can be collectively written as

$$E(g) = \mathbf{a}^\top + r\Theta_{p_r \times p_g} \tag{2.5}$$

where $\mathbf{a}^\top$ is the vector of the intercept, $g_{1\times p_g} = (g_1, \cdots, g_{p_g})$, $r_{1\times p_r} = (r_1, \cdots, r_{p_r})$, and $\Theta_{p_r \times p_g} = (\theta_1, \cdots, \theta_{p_g})$ is the transition matrix. The SVD is performed on the transition matrix to separate regression coefficients representing gene expression and regulators:

$$\Theta = UDV^\top = (u_1, \cdots, u_L)D(v_1, \cdots, v_L)^\top \tag{2.6}$$

where $D = \text{diag}(d_1, \cdots, d_L)$ is a diagonal matrix with L diagonal elements. The diagonal matrix D can account for the dissimilarity among loading vectors in terms of different scaling factors. Subsequently, we can obtain estimated coefficients for gene expression and regulators by decomposing the estimated transition matrix $\hat{\Theta}$. Under the sparse condition, one gene

expression is regulated only by a few of regulators, and one regulator affects a few of gene expressions (Kristensen et al. (2014)). To impose sparsity, we adopt the sparse SVD method developed by Lee et al. (2010) where sparse singular vectors corresponding to the largest singular values are obtained recursively. Consider the first largest singular value (d_1, u_1, v_1), then the regularized sparse SVD can be expressed as

$$\frac{1}{2n}\left\|\hat{\Theta} - d_1 u_1 v_1\right\|_F^2 + \lambda|d_1 u_1| + \lambda|d_1 v_1| \tag{2.7}$$

where $\|\cdot\|_F$ is the Frobenius norm. Tuning parameter λ is same for u_1 and v_1 for computation efficiency. Here d_1 is treated as the scaling factor. After estimating (d_1, u_1, v_1), we update $\hat{\Theta} = \hat{\Theta} - \hat{d}_1\hat{u}_1\hat{v}_1^\top$ and recursively update (d_l, u_l, v_l), for $l = 2, \cdots, L$ in a similar manner. With sparse SVD, we can decompose coefficient and impose sparsity on p_z and p_x for every LRM. The standard LASSO is not applicable within the current LRM formulation since the shrinkage has been imposed on scaled singular vectors.

2.2.3 Stage 2: the Penalized G×E Interaction Model

Now we integrate multiomics measurements for G×E interactions. The regulated GEs, residual GEs, as well as residual regulators can be obtained through LRMs. The G factors are represented by regulated GEs and residual GEs, which are involved in the interaction with dimensional environmental factors. The partition of gene expressions into regulated and non-regulated components proceeds as follows. The L sets of regulated gene expressions (GV) are equivalent to the corresponding sets of regulators (RU). We include the L sets of regulated GEs (GV) in the G×E model since gene expressions are more directly related to cancer outcomes. The residual GEs, i.e. the non-regulated GEs which cannot be captured by LRMs, is denoted as $\tilde{G}_{n\times p_g}$. The G factors, consists of both GV and $\tilde{G}$, interact with q environmental factors. Denote $W_j = (G_jV_j, G_jV_jE_1, \cdots, G_jV_jE_q, \tilde{G}_j, \tilde{G}_jE_1, \cdots, \tilde{G}_jE_q)$, $(j = 1, \cdots, p_g)$. Then W_j corresponds to the interaction with respect to the jth GE. We only consider the main effect of residual regulators, because the influences of regulators on

cancer outcomes are mostly mediated by gene expressions, and investigating its interactions with environmental factors are not of interest.

The quantifications of the residuals $\tilde{G}$ and $\tilde{R}$ are conducted through perpendicular projection operation. As both can be calculated in the same manner, we take $\tilde{G}$ as an example. For the jth gene expression, define S_j as the set of all LRMs that contains the j th gene expression. If S_j is empty, then the jth gene expression is not regulated, which results in $\tilde{G}_j = G_j$. If S_j is not empty, we denote V_{S_j} as the sub-matrix of V that only contains columns (LRMs) of the jth gene expression. Following the perpendicular projection operation, we calculate the residual as $\tilde{G}_j = (I - GV_{S_j}((GV_{S_j})^\top (GV_{S_j}))^{-1}(GV_{S_j})^\top)G_j$, which is the projection of G_j onto the orthogonal space of GV_{S_j}.

Consider n subjects, p_g gene expressions and L LRMs. Then all the main and interaction effects can be collectively written as

$$W = (GV, GVE_1, \cdots, GVE_q, \tilde{G}, \tilde{G}E_1, \cdots, \tilde{G}E_q) = (X_1, X_2),$$

where $X_1 = (GV, GVE_1, \cdots, GVE_q)$ denotes the main effects of regulated GEs and their interactions with the environmental factors. Similarly, the effects corresponding to residual GEs are defined as $X_2 = (\tilde{G}, \tilde{G}E_1, \cdots, \tilde{G}E_q)$. Subsequently, we consider the following penalized regression models for G×E interactions:

$$\frac{1}{2n}\left\| Y - \sum_{k=1}^{q}\alpha_k E_k - \sum_{l=1}^{L} X_{1l}b_{1l} - \sum_{j=1}^{p_g} X_{2j}b_{2j} - \sum_{t=1}^{p_r}\gamma_t \tilde{R}_t \right\|_2^2 + \sum_{l=1}^{L} P_1\Big(b_{1l};\lambda_1\Big) + \sum_{j=1}^{p_g} P_2\Big(b_{2j};\lambda_2\Big) + \sum_{t=1}^{p_r} P_3\Big(\gamma_t;\lambda_3\Big) \quad (2.8)$$

where $X_{1l} = (GV_l, GV_lE_1, \cdots, GV_lE_q)$, $(l = 1, \cdots, L)$, represents the lth LRM and its interaction with q environmental factors, and $X_{2j} = (\tilde{G}_j, \tilde{G}_jE_1, \cdots, \tilde{G}_jE_q)$, $(j = 1, \cdots, p_g)$ denotes the main and interaction effects with respect to the jth residual GEs. Here, b_{1l} and b_{2j} are corresponding regression coefficients for X_{1l} and X_{2j}. γ_t is the coefficients for $\tilde{R}_t$ $(t = 1, \cdots, p_r)$, the residual of regulators. $P_i(\cdot;\lambda_i)$, $(i = 1, 2, 3)$, is the penalty function with

λ_i as the tuning parameter to impose sparsity. Since regression coefficients from the three components are on a similar scale, and different tunings dramatically increase the computational cost, the three tuning parameters are set as the same. Regularized identification in G×E interaction studies demands tailored penalty functions (Zhou et al. (2021)). For instance, b_{1l} stands for all the main and interaction effects with respect to the lth LRM. Selection of b_{1l} on the group levels determines if the lth LRM has any effect at all. If so, then selection of the individual effects within the group further determines the main and/or interactions that are associated with the cancer outcome. Therefore, penalized selection should accommodate the bi–level (or sparse group) structure. To be consistent with the analysis in stage 1, we still adopt LASSO as the baseline penalty function. Specifically, we have

$$P_1(b_{1l};\lambda_1) = \lambda_1\left\|b_{1l}\right\|_2 + \lambda_1\sum_{k=1}^{q+1}\left|b_{1lk}\right|,\ P_2(b_{2j};\lambda_2) = \lambda_2\left\|b_{2j}\right\|_2 + \lambda_2\sum_{k=1}^{q+1}\left|b_{2jk}\right|,$$

where $P_1(b_{1l};\lambda_1)$ and $P_2(b_{2j};\lambda_2)$ are sparse group LASSO. The L1 norm and L2 norm ($\|\cdot\|_2$) result in penalized identification on the individual and group level, respectively. The sparse group regularization has been adopted for bi-level selection of main and interaction effects on the individual and group level simultaneously. Its advantage over LASSO in G×E studies has been demonstrated in multiple studies (Zhou et al. (2021)). A corresponding price paid is computational cost as different bi-level regularization usually demands different tunings. As we only consider the main effect of residuals of regulators, the L1 norm penalty is adopted for γ_t ($t = 1, \cdots, p_r$). Since the number of environmental factors is usually low, the selection of them is not of interest. They are pre-determined with evidences of being associated with cancer from previous studies. The proposed regularization respects a weak hierarchy between main and interaction effects as the penalty has not been imposed on the environmental main effects. So once an interaction effect is selected, at least one of the two corresponding main effects will be in the model.

2.2.4 Computation

The equation (2.8) can be expressed as:

$$\frac{1}{2n}\left\|Y - E\alpha - X_1 b_1 - X_2 b_2 - \tilde{R}\gamma\right\|_2^2 + P_1(b_1;\lambda_1) + P_2(b_2;\lambda_2) + P_3(\gamma;\lambda_3) \tag{2.9}$$

where $\alpha_{q\times 1} = (\alpha_1, \cdots, \alpha_q)^\top$ is the coefficient vector for q environmental factors, $b_{1_{L(q+1)\times 1}} = (b_{1_1}, \cdots, b_{1_L})^\top$ and $b_{2_{p_g(q+1)\times 1}} = (b_{2_1}, \cdots, b_{2_{p_g}})^\top$ are the coefficient vectors for the main and interaction effects of the regulated and residual GEs, respectively. In addition, $\gamma_{p_r\times 1} = (\gamma_1, \cdots, \gamma_{p_r})\top$ is the coefficient vector for residual regulators.

The integrative analysis consists of two steps. In the first step, the loading matrices U and V are estimated through the construction of LRMs. The jth column of $\hat{\Theta}$, denoted as $\hat{\theta}_j$, $(j = 1, \cdots, p_g)$, is estimated by minimizing equation (2.4). For $l = 1, \cdots, L$, the singular vectors corresponding to the largest singular values, $(\hat{u}_l, \hat{v}_l, \hat{d}_l)$, are conducted through the rank-1 sparse SVD on $\hat{\Theta}$. The rank-1 sparse SVD is performed recursively for $l = 1, \cdots, L$, by updating $\hat{\Theta}^{(l+1)} = \hat{\Theta}^{(l)} - \hat{u}_l\hat{d}_l\hat{v}_l^\top$ at each l. In the second step, the shrinkage estimate of the regression coefficients can be obtained in the G×E model, where GV, RU, residuals of gene expressions $(\tilde{G})$, and residuals of regulators $(\tilde{R})$ are calculated accordingly. At the kth iteration, the vector of estimated regression coefficients for all environmental factors is computed by $\hat{\alpha}^{(k+1)} = (E^{(k)\top}E^{(k)})^{-1}E^{(k)\top}(Y - X_1\hat{b}_1^{(k)} - X_2\hat{b}_2^{(k)} - \tilde{R}\hat{\gamma}^{(k)})$. Given $\hat{\alpha}^{(k+1)}$ fixed at the current estimate, we obtain $(\hat{b}_1^{(k+1)}, \hat{b}_2^{(k+1)}, \hat{\gamma}^{(k+1)})$ by minimizing equation (2.9). The iteration stops until convergence. The outline of algorithm is shown in Table **Algorithm** 3:

Algorithm 3 The Integrative analysis for G×E Interaction

Step 1: Estimate the loading matrices of LRMs U and V: construct LRMs.

(a) For $j = 1, \cdots, p_g$, obtain $\hat{\theta}_j$ by minimizing equation (2.4). Then the estimate $\hat{\Theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_{p_g})$.

Initialize $l = 1$.

for $l = 1, \cdots, L$ **do**

(b) Apply rank-1 sparse SVD on $\hat{\Theta}$ to obtain the singular vectors corresponding to largest singular values (u_l, v_l, d_l).

(c) Update $\hat{\Theta}^{(l+1)} = \hat{\Theta}^{(l)} - u_l d_l v_l^\top$.

(e) $l = l + 1$.

end for

Step 2: Estimate regression coefficients α, b_1, b_2, γ: construct the penalized G×E interaction model.

(a) Calculate GV, RU, $\tilde{G}$ and $\tilde{R}$.

Initialize $\hat{b}_1^{(0)} = \hat{b}_2^{(0)} = \hat{\gamma}^{(0)} = 0$.

At the $(k+1)$th iteration.

repeat

(b) Compute $\hat{\alpha}^{(k+1)} = (E^{(k)\top} E^{(k)})^{-1} E^{(k)\top} (Y - X_1 \hat{b}_1^{(k)} - X_2 \hat{b}_2^{(k)} - \tilde{R}\hat{\gamma}^{(k)})$.

(c) Obtain $(\hat{b}_1^{(k+1)}, \hat{b}_2^{(k+1)}, \hat{\gamma}^{(k+1)})$ by minimizing equation (2.9) through bi-level selection.

until convergence

LASSO is adopted to conduct selection of important LRMs from the first stage. At the second stage, a sparse group LASSO has been formulated to accommodate the identification of main and interaction effects on both the group and individual level. We conjecture that other penalization methods, such as adaptive LASSO (Zou (2006)), SCAD (Fan and Li (2001)) and MCP (Zhang et al. (2010)) are also applicable in our framework. For example, MCP can be adopted to identify sparse regulatory relationship from the first stage, and a sparse group MCP is also tailored for the identification of important G×E interactions in the clinical model. We do not compare the performances of different baseline penalization methods within our framework as it is not the main interest here.

At the first step, we only use one tuning parameter λ for conducting sparse SVD due

to the similarity in scales between GE and its regulators. The three tuning parameters, $\lambda_1, \lambda_2, \lambda_3$, have been used in the second step, where λ_1 and λ_2 determine the sparsity of main and interaction effects with respective to the regulated and unregulated GEs correspondingly, and λ_3 controls the sparsity of the residuals from regulators. We choose the optimal tuning parameters using five-fold cross-validation in both the simulation study and real data analysis. The analysis has been implemented with statistical software R (version 3.6.3). In simulation, the average CPU time of running one replicated simulated data ($n = 500,\ p_g = p_r = 200,\ q = 4$) is 23.1 minutes on a regular desktop PC. The R codes are available from the corresponding author.

2.3 Simulation

We perform simulation to evaluate the utility of the proposed method integrative G×E model, termed as IGE. In addition, we consider three alternative methods: (1) The S-LASSO selects gene expressions and regulators separately using LASSO. (2) The J-LASSO selects gene expressions and regulators based on LASSO simultaneously. (3) ColReg, the collaborative regression (Gross and Tibshirani (2014)), identifies important GEs and regulators jointly in terms of explaining similar variation under the cancer outcome.

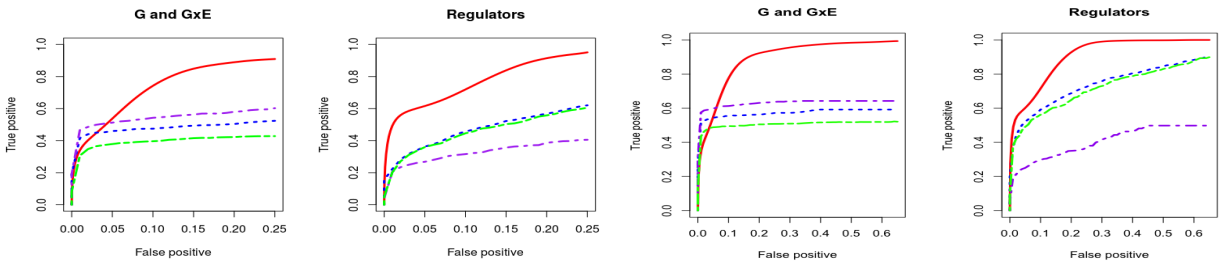
We generate the data as follows. First, each row of R is independently generated from a multivariate normal distribution with mean zero and one of the four covariance structures: (i) AR–1 structure with correlation coefficient $0.25^{|i-j|}$ for the ith and jth regulators; (ii) Banded correlation structure where the ith and jth regulators have $\rho = 0.33$ if $|i-j| = 1$ and $\rho = 0$ otherwise; (iii) the covariance extracted from TCGA lung squamous cell carcinoma (LUSC) data in Section 2.4, and (iv) the covariance structure of the lung adenocarcinoma (LUAD) from Section 2.4.

Choose $L = 20$ for the number of LRMs between gene expression and regulators. For $l = 1, \cdots, 20$, u_l or v_l is randomly assigned 5 non-zero entries, with values generated from unif$[2, 4]$. Then Θ is computed as $\sum_{l=1}^{20} u_l v_l^\top$ and G is generated as $G = R\Theta + \varepsilon$, where each row of matrix ε is independently generated from a multivariate normal distribution with

mean zero and the same covariance structure as R. To generate the cancer outcome, each row of E is generated independently from a multivariate normal distribution with marginal mean zero and AR-1 structure where the ith and jth components have correlation coefficient $0.5^{|i-j|}$. Subsequently, we generate the response from model (2.1) under standard normal errors.

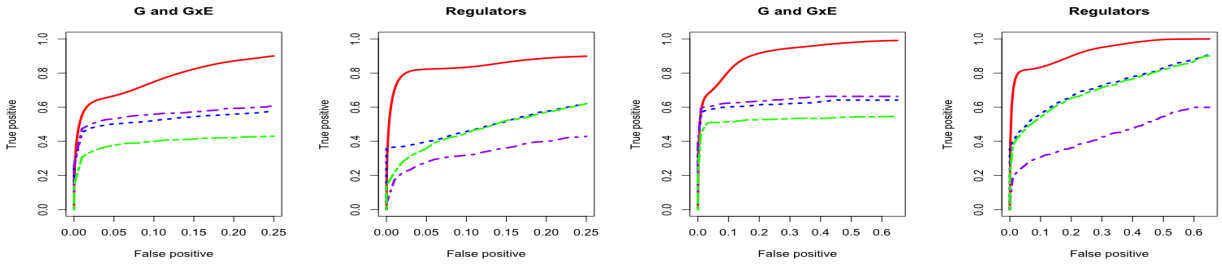
200 gene expression, 200 regulators and 4 environmental factors are simulated with two different sample sizes, 500 and 1000. To assign non-zeros effects in model (2.1), we randomly select 30 gene expressions. For every selected gene expression, 4 non-zero entries are randomly assigned to the coefficients of G factor or its corresponding G×E interactions. Those values are generated from unif$[0.25, 0.5]$ and unif$[0.5, 1]$ for weak and strong coefficient signals, respectively. The coefficients of regulators are randomly assigned with 30 non-zero coefficients generated from unif$[1, 2]$. The coefficients of environmental factors are generated from unif$[2, 3]$.

For a comprehensive evaluation, we consider a sequence of tuning parameter values (from 0 to 3, total 100 lambda values) and use the receiver operating characteristic (ROC) curve and partial area under the ROC curve (PAUC) to compare different methods. Total simulation replication is 100. All PAUCs are tabulated in Table 2.1 and Table 2.2. The ROC curves for AR-1 structure and estimated covariance from LUSC are shown in Figure 2.1 and Figure 2.2. Other scenarios of ROC curves are provided in Appendix A.1, respectively.



(a) n = 500, weak signal

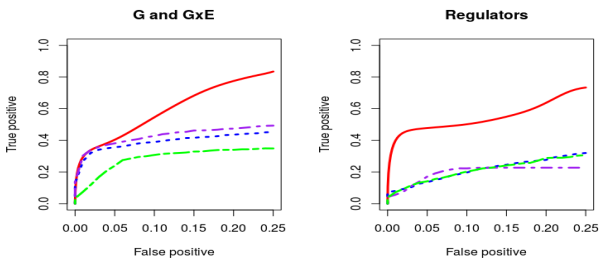
(b) n=1000, weak signal



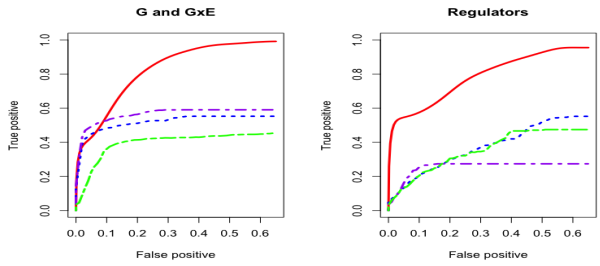
(c) n=500, strong signal

(d) n=1000, strong signal

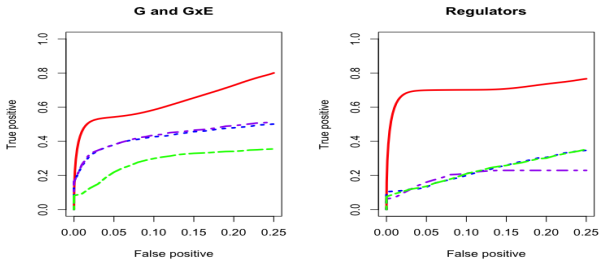
Figure 2.1: Four cases of ROC curves under AR-1 structure. The left panel corresponds to comparison under both weak and strong signals for 500 subjects. The right panel corresponds to comparison under both weak and strong signals for 1000 subjects. IGE, solid red; S-LASSO, dashed blue; J-LASSO, long dashed purple; ColReg, long dashed green.



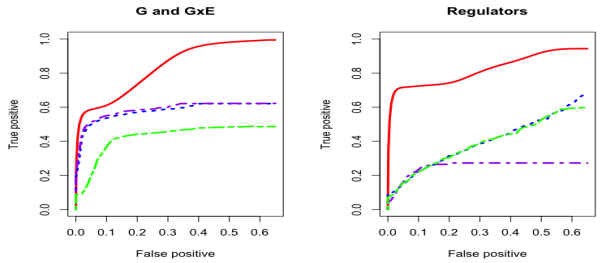
(a) n = 500, weak signal



(b) n=1000, weak signal



(c) n=500, strong signal



(d) n=1000, strong signal

Figure 2.2: Four cases of ROC curves under estimated covariance from LUSC. The left panel corresponds to comparison under both weak and strong signals for 500 subjects. The right panel corresponds to comparison under both weak and strong signals for 1000 subjects. IGE, solid red; S-LASSO, dashed blue; J-LASSO, long dashed purple; ColReg, long dashed green.

We consider using the receiver operating characteristic (ROC) curve and partial area under the ROC curve (PAUC) to compare different methods. Total simulation replicates is 100. All PAUCs are tabulated in Table 2.1 and Table 2.2. The ROC curves for AR-1 structure and estimated covariance from LUSC are shown in Figure 2.1 and Figure 2.2. The ROC curves in other scenarios are provided in Appendix A.1. For all simulation scenarios, the proposed method has higher PAUCs than the alternative methods. For example, in Table 2.1 with AR-1 correlation and weak signal, the proposed method has PAUC 0.73 (sd 0.07) for the identification of G and G×E effects, while J-LASSO, S-LASSO, ColReg have PAUCs 0.54 (sd 0.04), 0.47 (sd 0.04), and 0.39 (sd 0.03), respectively. For the identification of regulators, the proposed method has PAUC 0.76 (sd 0.10), while J-LASSO, S-LASSO and ColReg have PAUCs 0.32 (sd 0.05), 0.46 (sd 0.13), and 0.45 (sd 0.15), respectively. The similar pattern can be observed under settings with strong signals. When sample size increases, the identification

results of all methods become better. The proposed IGE outperforms alternative approaches across different scenarios. For instance, in Table 2.2 with AR-1 correlation and strong signal, the proposed method has PAUC 0.89 (sd 0.02) in the identification of G and G×E, while J-LASSO, S-LASSO, ColReg have PAUCs 0.62 (sd 0.04), 0.57 (sd 0.04), and 0.50 (sd 0.03), correspondingly. For the identification of regulators, the proposed method also outperforms the alternatives.

In addition, the proposed method outperforms the alternatives when the correlation is extracted from real data. For example, in Table 2.1 with estimated covariance from LUSC and weak signals, the proposed method has close PAUCs in both G and G×E and regulators, 0.59 (sd 0.09) and 0.55 (sd 0.15). Other methods have low accuracy in identifying main and interaction effects. In particular, J-LASSO, S-LASSO, ColReg have PAUCs 0.42 (sd 0.05) and 0.19 (sd 0.06), 0.39 (sd 0.04) and 0.21 (sd 0.06), and 0.28 (sd 0.04) and 0.21 (sd 0.07), respectively. When magnitude of the signals and sample size increase (e.g. with LUSC and strong signals), the proposed method still have the best performance in identification. Overall, the IGE model has much higher identification accuracy than other methods across different simulation settings by borrowing strength from accounting for regulatory relationship and bi-level selection in G×E interaction studies.

Table 2.1: PAUC: mean (sd) based on 100 replicates. $p_g = p_r = 200, n = 500$

Covariance	Signal	Approaches	G and G×E	Regulators
AR-1	weak	IGE	0.73(0.07)	0.76(0.10)
		S-LASSO	0.47(0.04)	0.46(0.13)
		J-LASSO	0.54(0.04)	0.32(0.05)
		ColReg	0.39(0.03)	0.45(0.15)
	strong	IGE	0.77(0.07)	0.85(0.06)
		S-LASSO	0.52(0.05)	0.48(0.14)
		J-LASSO	0.55(0.04)	0.33(0.05)
		ColReg	0.39(0.03)	0.46(0.15)
Banded	weak	IGE	0.74(0.06)	0.74(0.10)
		S-LASSO	0.48(0.03)	0.44(0.11)
		J-LASSO	0.54(0.05)	0.32(0.04)
		ColReg	0.39(0.03)	0.43(0.12)
	strong	IGE	0.77(0.08)	0.84(0.06)
		S-LASSO	0.52(0.04)	0.46(0.11)
		J-LASSO	0.55(0.05)	0.32(0.04)
		ColReg	0.39(0.03)	0.43(0.12)
LUSC	weak	IGE	0.59(0.09)	0.55(0.15)
		S-LASSO	0.39(0.04)	0.21(0.06)
		J-LASSO	0.42(0.05)	0.19(0.06)
		ColReg	0.28(0.04)	0.21(0.07)
	strong	IGE	0.63(0.10)	0.71(0.13)
		S-LASSO	0.42(0.05)	0.22(0.07)
		J-LASSO	0.43(0.05)	0.19(0.06)
		ColReg	0.28(0.05)	0.22(0.07)
LUAD	weak	IGE	0.64(0.09)	0.62(0.15)
		S-LASSO	0.45(0.04)	0.21(0.06)
		J-LASSO	0.47(0.05)	0.19(0.05)
		ColReg	0.32(0.03)	0.22(0.07)
	strong	IGE	0.70(0.08)	0.77(0.11)
		S-LASSO	0.47(0.05)	0.23(0.08)
		J-LASSO	0.48(0.05)	0.18(0.05)
		ColReg	0.31(0.04)	0.23(0.08)

Table 2.2: PAUC: mean (sd) based on 100 replicates. $p_g = p_r = 200, n = 1000$

Covariance	Signal	Approaches	G and G×E	Regulators
AR-1	weak	IGE	0.89(0.02)	0.91(0.02)
		S-LASSO	0.57(0.04)	0.73(0.09)
		J-LASSO	0.62(0.04)	0.40(0.04)
		ColReg	0.50(0.03)	0.71(0.09)
	strong	IGE	0.91(0.02)	0.93(0.02)
		S-LASSO	0.61(0.04)	0.71(0.08)
		J-LASSO	0.64(0.05)	0.43(0.04)
		ColReg	0.52(0.03)	0.70(0.09)
Banded	weak	IGE	0.89(0.03)	0.91(0.03)
		S-LASSO	0.55(0.04)	0.73(0.07)
		J-LASSO	0.62(0.04)	0.40(0.05)
		ColReg	0.50(0.03)	0.71(0.08)
	strong	IGE	0.90(0.04)	0.92(0.02)
		S-LASSO	0.61(0.04)	0.72(0.08)
		J-LASSO	0.64(0.04)	0.44(0.06)
		ColReg	0.53(0.04)	0.70(0.08)
LUSC	weak	IGE	0.82(0.04)	0.78(0.06)
		S-LASSO	0.51(0.05)	0.36(0.07)
		J-LASSO	0.56(0.05)	0.25(0.07)
		ColReg	0.39(0.04)	0.35(0.08)
	strong	IGE	0.83(0.04)	0.82(0.06)
		S-LASSO	0.57(0.05)	0.39(0.07)
		J-LASSO	0.58(0.05)	0.25(0.08)
		ColReg	0.42(0.04)	0.38(0.07)
LUAD	weak	IGE	0.83(0.04)	0.80(0.06)
		S-LASSO	0.57(0.04)	0.43(0.06)
		J-LASSO	0.59(0.04)	0.25(0.06)
		ColReg	0.47(0.03)	0.43(0.06)
	strong	IGE	0.85(0.03)	0.84(0.04)
		S-LASSO	0.61(0.04)	0.46(0.07)
		J-LASSO	0.61(0.04)	0.26(0.06)
		ColReg	0.49(0.03)	0.46(0.07)

2.4 Analysis of TCGA Data

Lung cancer is a top rank cancer for both men and women. In this section, we apply the proposed method as well as the alternatives on lung adenocarcinoma (LUAD) data and lung squamous cell carcinoma (LUSC) data from the Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/).

LUAD is at present the most common lung cancer subtype among non-smokers and women, although it has been shown that smoking may increase the risk of LUAD (Subramanian and Govindan (2007), Couraud et al. (2012)). On the other hand, LUSC is closely associated with smoking, and is more common in men than in women (Kenfield et al. (2008)). LUAD grows more slowly with smaller masses than LUSC of the same stage, but LUAD tends to initiate metastasis at the early stages (Kumar et al. (2017)).

The processed level 3 data has been downloaded from TCGA data portal using package *cgdsr*. We match the multi-omics measurements with the clinical/environmental variables and survival outcome. LUSC and LUAD has 344 and 426 subjects, correspondingly. We first conduct screenings to reduce dimensionality so the regularization methods can be appropriately applied. Here, we select the top 200 mRNA with the largest marginal variances. As we matched the CNA and Methylation profiles with same mRNA, the corresponding 200 measurements on CNA and Methylation are selected at the same time. We select age, gender, smoking pack years, and pathologic tumor stage as environmental variables. The accelerated failure time (AFT) model (Appendix A.2) has been adopted to link the omics and clinical measurements to survival outcomes.

2.4.1 Lung Adenocarcinoma (LUAD) Data

The proposed method identifies 8 LRMs with 1 residual effect of gene expression (mRNA) and 14 residual effects of regulators (DM and CNA). Additionally, the proposed method results in identification of 7 LRM×E interactions and 11 G×E interactions from mRNA residual effects.

Table 2.3 provides the identified main effects of LRMs, residual GEs and regulators. We

Table 2.3: Analysis of the TCGA LUAD data: LRMs and residual effects for gene expression and regulators with the estimated coefficient or loadings in the parentheses.

	LRMs			
	#1(0.07)	#2(-0.01)	#3(-0.02)	#4(-0.03)
mRNA	PIK3R2(0.35)	PIK3R2(0.98)	ECT2(-0.98)	INTS7(-0.77)
	STK3(-0.74)	STK3(0.11)	PSMD2(-0.17)	PIK3R2(-0.62)
	NCKAP5L(0.74)	NCKAP5L(-0.08)		
	CUL9(0.14)			
CNA	NEK2(-0.22)	CECR1(0.65)	KPNA4(-0.44)	INTS7(-0.70)
	LPGAT1(0.22)	C1QTNF6(-0.75)	B3GALNT1(0.43)	DTL(0.70)
	INTS7(0.65)		PSMD2(-0.55)	
	DTL(-0.65)		LIPH(0.55)	
	CECR1(-0.19)			
	#5(-0.05)	#6(0.08)	#7(-0.06)	#8 (0.06)
mRNA	PIK3R2(0.12)	INTS7(0.73)	PIK3R2(-0.10)	PSMD2(0.31)
	STK3(-0.78)	PIK3R2(0.63)	STK3(-0.24)	TMOD3(0.61)
	NCKAP5L(0.57)	STK3(0.18)	CUL9(-0.96)	DIAPH3(0.72)
	CUL9(0.16)	NCKAP5L(-0.14)		
CNA	INTS7(-0.16)	NEK2(-0.69)	INTS7(-0.34)	MAPRE3(0.70)
	DTL(0.16)	LPGAT1(0.71)	DTL(0.36)	IFT172(-0.67)
	CECR1(-0.78)		CECR1(0.61)	PSMD2(0.09)
	C1QTNF6(-0.57)		C1QTNF6(-0.61)	ITGB1(0.09)
				ADAM10(0.14)
	Residual effects			
mRNA	MAST3(0.01)			
DM	ADSS(0.01)	SLC2A1(0.01)	PTCH2(0.01)	ECT2(0.09)
	TNS4(0.02)	MUSTN1(0.05)	DKK1(0.02)	FSCN1(0.05)
	GNPNAT1(0.04)	HPS1(-0.04)	MAPRE3(-0.02)	
CNA	LAMC2(-0.01)	CD5(-0.03)	E2F7(-0.01)	

can observe that LRMs does not contain effects from methylation, while most residual effects in regulators are from methylation.The identification results have important biological implications. As a representative example, gene PIK3R2 is identified by 6 different LRMs. From a recent study (Chen et al. (2020)), PIK3R2 is significantly associated with lung adenocarcinoma and its pathway plays a critical role in the progress of LUAD. Besides, gene STK3 is identified by 5 different LRMs. STK3 belongs to a large family of serine/threonine kinases, which are implicated in the regulation of signaling pathways involved in cell growth, cell differentiation cell death and cell volume (Huang et al. (2018b), Pombo et al. (2007)). The identified LRMs are also meaningful. For example, we observe the regulatory relationship between PIK3R2 and NEK2 from both LRM #1 and #6. One of the recent studies shows that this natural downstream regulation is significantly related to cancer outcome (Hameed and Ejaz (2020)). Among all the residual effects, we observe that most of them are from methylation. For example, SLC2A1, ECT2, TNS4, DKK1, GNPNAT1 are found to be associated with survival of lung cancer patients (Guo et al. (2020), Silva et al. (2019), Misono et al. (2019), Yang et al. (2019), Zhang et al. (2020b)).

Table 2.4: Analysis of the TCGA LUAD data: G×E interaction identifications from LRMs and gene expression with the estimated regression coefficients in the parentheses.

LRMs	AGE	GENDER	SMOKING
#1	0.08		-0.25
#2		0.02	
#3		0.01	
#4		0.01	0.01
#5			0.01
mRNA Residual	AGE	GENDER	SMOKING
MAST3			0.27
HPS1	0.01		
BBS5	-0.04		-0.03
TLE1	-0.01		
ADAM10		0.02	0.03
SLC16A3		0.07	
BTN2A2		-0.02	-0.06
FAM71E1			0.02

Table 2.4 provides the identification results for interaction effects. The proposed method selects variables with a sparse group nature. There are five LRMs interacting with environments. The first and fourth LRMs interact with two environment factors, and second, third

and fifth interact with one environment factor. Additionally, the proposed method can identify a total of 11 interactions involving mRNA residual effects. Note that here, the G factor is no longer in the usual sense from existing G×E studies. The G factor are represented by the LRMs and residual mRNAs which correspond to the regulated and un-regulated G factors, respectively.

In terms of prediction, we adopt a random sampling approach. More specifically, we randomly select 30% data as test set and the remaining as training set. Estimates are generated using the training set only and the predictions are made based on the testing set. We dichotomize the predicted response at the median, create two risk groups, and compute log-rank statistics, which measure the difference in survival between the two groups. Larger log-rank test statistic indicates better predictive performance. To avoid extreme splits, the procedure is repeated 100 times. The average log-rank test statistics are 5.97(IGE, sd 0.35), 4.76(S-LASSO, sd 0.25), 4.60(J-LASSO, sd 0.08), 3.74(ColReg, sd 0.26), respectively. The proposed method has the largest log-rank statistic, hence the best prediction performance.

2.4.2 Lung Squamous Cell Carcinoma (LUSC) Data

The proposed method identifies 8 LRMs with 2 residual effects from GEs and 17 residual effects from regulators (DM and CNA). The interactions involve 7 LRMs and 26 mRNAs.

Table 2.5 provides identified main effects by using the proposed method. As aforementioned, we aim to find a sparse relationship between gene expressions and regulators. Therefore, a small subset of regulators are related to genes and vice versa. Table 2.6 provides the identifications of G×E interaction effects. There's one LRM not interacting with any other environmental factors. The findings have important implications. For instance, gene RNF24 is identified by 2 different LRMs (#1, #2). RNF24 is a membrane protein, which interacts with TRPC protein (Lussier et al. (2008)). A recent study shows that RNF24 acts as one of the important factors for the prognosis of carcinoma (Lin et al. (2018)). RNF24 is also shown to be correlated with the occurrence of esophageal adenocarcinoma (Wang et al. (2014)). For DM, RGP1 is identified by 3 different LRMs (#4, #6, #7). According

Table 2.5: Analysis of the TCGA LUSC data: LRMs and residual effects for gene expression and regulators with the estimated coefficient or loadings in the parentheses.

	LRMs			
	#1(-0.01)	#2(0.01)	#3(0.01)	#4(-0.02)
mRNA	RNF24(-0.17)	SEC23B(0.23)	REEP3(-0.76)	AP2A2(-0.59)
	ESM1(-0.53)	RNF24(-0.97)	FUT11(-0.64)	PNPLA6(-0.37)
	RASAL2(-0.39)			RFX1(-0.55)
	LAMC1(-0.34)			XRN2(0.45)
	DLGAP4(-0.63)			
DM	DCBLD1(0.09)	TCF7L2(0.22)		RGP1(-0.52)
	CHI3L1(0.18)			NCOR2(0.27)
CNA	CD163L1(-0.16)	ENTPD6(0.68)	RERE(-0.89)	CD163L1(0.70)
	DLGAP4(-0.96)	ABHD12(-0.69)	DLGAP4(-0.43)	PARD6G(-0.39)
	#5(0.16)	#6(0.05)	#7(-0.05)	#8(0.01)
mRNA	COL5A3(0.45)	MGST3(0.33)	TPM4(0.68)	TCTN2(-0.45)
	DCBLD1(0.57)	OSBPL5(0.31)	UBB(0.59)	ANGPT2(-0.40)
	PDGFA(0.31)	SNX9(0.56)	NCOR2(-0.42)	UBE4B(-0.37)
	CHST15(0.45)	MYO1C(0.46)		MBTPS1(-0.47)
	LGALS1(0.39)	CCDC68(0.49)		FAM178B(-0.50)
DM	DCBLD1(-0.86)	CHST15(-0.97)	RGP1(-0.55)	NCOR2(0.16)
	FAM178B(-0.37)	RGP1(0.13)		
	CHST15(-0.17)	NCOR2(-0.10)		
		LGALS1(-0.15)		
CNA	DLGAP4(0.27)		STK40(-0.26)	CD163L1(-0.35)
			TCTN2(-0.78)	DLGAP4(-0.92)
	Residual effects			
mRNA	LRAT(-0.02)	PLEKHA6(-0.02)		
DM	BAMBI(0.01)	PYGB(0.02)	FUT11(-0.18)	ZNF394(0.03)
	CCIN(-0.01)	DEAF1(-0.10)	ACOT7(0.04)	KLK6(-0.12)
	LHX8(-0.01)	PLEKHB1(0.09)		
CNA	FGFRL1(-0.05)	DCBLD1(-0.04)	NEFL(-0.04)	CHST1(0.02)
	ULK1(-0.03)	FPR2(0.02)	PYGB(-0.10)	

to Anand et al. (2020), RGP1 belongs to the regulation of guanosine diphosphate (GDP) reaction exchange, and it acts as a prognostic factor in cancer. For CNA, CD163L1 is identified by 3 different LRMs (#1, #4, #8), and it can be used as a significant biomarker of cancer (Zahra et al. (2018)). The identified LRMs are also meaningful. For example, the regulatory relationship between NCOR2 and TCTN2 can be identified in LRM #7. This result has also been observed in a regulatory network analysis (Zeng et al. (2012)). Among all the residual effects, LRAT, PLEKHA6, ACOT7, KLK6, PLEKHB1, FGFRL1, FPR2 are associated with prognosis of LUSC patients from existing studies (Ke et al. (2020), Relli et al. (2019), Zhang et al. (2020a), Wang et al. (2016), Bae et al. (2020), Hu et al. (2019)).

To evaluate prediction, we adopt a random sampling approach and apply log-rank test for assessment. We adopt the similar procedure as previous real data analysis section. After repeating 100 times, the average log-rank test statistics are 33.20(IGE, sd 2.32), 25.06(S-LASSO, sd 1.84), 24.41(J-LASSO, sd 2.13), 27.88(ColReg, sd 2.45), respectively. The proposed method has superior prediction performance over alternatives.

2.5 Discussion

We have conducted an integrative gene-environment interaction analysis for multi-dimensional omics data based on the proposed two-step variable selection model. Specifically, at the first step, sparse regulatory relationship between the G factor and its regulators have been pinpointed via penalization, which leads to effects that can be directly linked to the prognostic outcomes. At the second step, a G$\times$ E prognostic model has been considered, where the G factor involved in the interaction consists of regulated (corresponding to the LRM) and unregulated (i.e., the residual GE) components. Besides, the residuals of the regulator are also included. The integrative G$\times$ E analysis fully takes the advantage of the multi-omics measurements, which distinguishes itself from most of the published studies.

Traditionally, statistical testing based marginal analysis has dominated the G$\times$ E studies. The paradigm shift to the joint analysis has been mainly motivated by the gene set and pathway based association analysis (Wang et al. (2011), Wu and Cui (2014), Jin et al.

Table 2.6: Analysis of the TCGA LUSC data: G×E interaction identifications from LRMs and gene expression with the estimated regression coefficients in the parentheses.

LRMs	AGE	GENDER	SMOKING
#1		0.02	0.03
#2		0.03	
#4	-0.02		
#5	0.01	0.05	-0.02
#6	0.01	-0.01	
#7		-0.36	
#8		0.02	
mRNA Residual	**AGE**	**GENDER**	**SMOKING**
LRAT		-0.17	
PLEKHA6		-0.30	
AP2A2	0.02		
SLC12A7	-0.10	0.07	
TCTN2	-0.15	-0.09	
CLEC5A	0.01		
RNF24	-0.06	0.04	
PRRX2	0.04		-0.04
CCDC74A	0.14	-0.13	
FGF9	0.03		-0.06
IGF2R	0.05	-0.02	
CHMP4C	0.24	0.13	-0.01
SLC45A4	-0.11		
SULF2	-0.05	-0.03	
UBB		-0.11	
DVL1		-0.07	
NID1		0.08	0.20
KLK8		0.01	
DOCK6		0.26	-0.10
FHDC1		0.01	-0.16
OPLAH		-0.12	
VSTM1			-0.02
SLC28A1			-0.07
TCF7L2			0.12
DLGAP4			-0.04
CRNKL1			-0.25

(2014), Jiang et al. (2017)). Recently, the effectiveness of regularized variable selection has been recognized not only in joint G× E studies when a large number of genetic factors are involved (Zhou et al. (2021)), but also in multi-level omics integrations (Wu et al. (2019), Du et al. (2021)). Therefore, it has been adopted for here.

This study can be improved from the following aspects. As strong correlations have been widely observed in among omics measurements, network based penalization can be imposed to accommodate the correlations among regulators at the first stage (Li and Li (2008), Sun and Wang (2012), Ren et al. (2017)). Besides, robustness can be incorporated at the first stage to model the regulatory relationship between GE and its regulators (Wu et al. (2018b)), and in the second stage for a robust prognostic model (Ren et al. (2019), Wu et al. (2018a)). Accounting for the form of environmental factors has received much attention in G× E studies, which results in the development of a wide range of nonparametric (Li et al. (2015), Wu and Cui (2013), Wu et al. (2018c)) and semiparametric (Wu et al. (2015), Ma and Xu (2015), Ren et al. (2020)) methods. However, in integrative G× E studies, capturing the nonlinear form of interaction is challenging. In this study, we focus on prognostic outcomes. With other types of outcomes, such as the longitudinal phenotypes (Zhou et al. (2019)), the G×E model in the second stage can be modified accordingly. We will postpone these further investigations to the future.

Chapter 3

Parallel Penalized Variable Selection for Large-Scale G×E Studies

3.1 Introduction

Many studies have shown that the interactions (G×E) between genetic and environmental risk factors play a critical role in predicting complex disease outcome beyond genetic and environmental main effects. In high-dimensional genetic studies, the identification of G×E interactions attracts more attentions. Penalized variable selection is one of the most popular approach. Despite success, the computation speed is still a challenge in G×E studies. For example, we evaluate different penalization methods by applying Lung Adenocaricinoma (LUAD) data from the Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/). The data has been collected from hundreds of subjects with thousands and thousands of genetic markers. If G×E interactions are taken into account, the total number of variables is huge, which can cost extreme computation time for many penalization methods.

Coordinate descent (CD) is one of the primary computation frameworks widely adopted, because it's simple, stable and efficient for a variety of penalized regression models. The process of CD is updating single parameter at a time and is iteratively cycling through all parameters until convergence. However, CD can't improve computation speed in either large

sample size, or large number of variables. The framework of CD lacks parallel computation capability.

In this chapter, we propose an alternative computation framework by adopting the alternating direction method of multipliers (ADMM). ADMM was first introduced by Gabay and Mercier (1976) and Glowinski and Marroco (1975). This algorithm framework solves optimization problems by splitting them into batches of samples or variables, then it can be solved in parallel. Hence, The parallel ADMM can conduct the optimization by using divide-and-conquer strategy without losing identification accuracy. Some existing research studies have developed ADMM framework for LASSO, sparse logistic regression, support vector machines, see examples in Bien et al. (2013), Ye et al. (2011), Peng et al. (2013), Zhu (2017) and Yu et al. (2017), Boyd et al. (2011). Few research studies discuss ADMM or parallel ADMM for penalized variable selection in G×E interactions. We will demonstrate ADMM for the identification of G×E interactions through penalization method and propose a parallel ADMM framework to improve computation speed in large-scale G×E studies.

3.2 A brief review of ADMM

Denote the response vector as $Y \in \mathbb{R}^n$ and the design matrix as $X \in \mathbb{R}^{n\times p}$, where n is the sample size and p is the number of variables. The penalized least square function can be expressed as

$$\frac{1}{2n}\|Y - X\beta\|^2 + P(\beta;\lambda), \tag{3.1}$$

where $\beta \in \mathbb{R}^p$ is the regression coefficients, λ is the tuning parameter, and $P(\beta;\lambda)$ is the penalty function. To formulate ADMM framework, we introduce an auxiliary variable $z \in \mathbb{R}^p$ and rewrite equation (3.1) as a constrained optimization

$$\frac{1}{2n}\|Y - X\beta\|^2 + P(z;\lambda), \text{ subject to } \beta - z = 0. \tag{3.2}$$

Equation (3.2) is also equivalent to the following augmented form with $\rho(>0)$ as the augmentation parameter

$$R(\beta, z, \tau) = \frac{1}{2n}\|Y - X\beta\|^2 + P(z;\lambda) + \tau^T(\beta - z) + \frac{\rho}{2}\|\beta - z\|^2. \tag{3.3}$$

To further simplify the equation, a scaled form, $u = \frac{1}{\rho}\tau$, is widely adopted in ADMM. Then equation (3.3) can be rewritten as

$$R(\beta, z, u) = \frac{1}{2n}\|Y - X\beta\|^2 + P(z;\lambda) + \frac{\rho}{2}\|\beta - z + u\|^2 - \frac{\rho}{2}\|u\|^2, \tag{3.4}$$

where equation (3.4) is called the Lagrangian formula with scaled form. The method of multipliers can be used to solve equation (3.2) by iteratively minimizing equation (3.4) over β, z, and maximizing it over u. At the $(k+1)$th iteration, the solutions are

$$\beta^{k+1} = \arg\min_{\beta} R(\beta, z^k, u^k) \tag{3.5}$$

$$z^{k+1} = \arg\min_{z} R(\beta^{k+1}, z, u^k) \tag{3.6}$$

$$u^{k+1} = u^k + (\beta^{k+1} - z^{k+1}) \tag{3.7}$$

We summarize this ADMM algorithm framework in Figure 3.1.

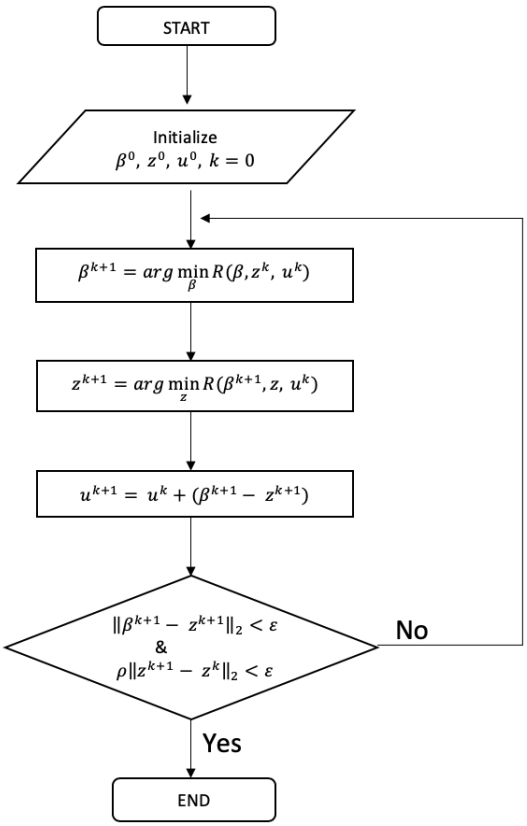


Figure 3.1: Flowchart of the ADMM algorithm framework for a penalized regression model.

3.3 Identification of G×E Interactions via ADMM

Penalized variable selection is one of the most popular approaches in high-dimensional data studies. Nowadays, many penalization methods have been developed to account for complicated data structures. Despite success, existing methods on G×E interactions still have limitations on identification. We develop a penalized bi-level selection method to better identify important G×E interactions and conduct it through ADMM. The bi-level selection penalization can determine the sparsity on both group and individual levels.

Denote $Y_{n\times 1}$ as the cancer outcome, and genetic effects and environment effects matrices as $G_{n\times p} = (G_1, \cdots, G_p)$ and $E_{n\times q} = (E_1, \cdots, E_q)$. The G×E interactions are denoted as $X_{n\times q} = (X_1, \cdots, X_q)$, where $X_j = (G_1E_j, \cdots, G_pE_j)$, $j = 1, \cdots, q$. The penalized least

square function for bi-level selection can be expressed as

$$R(\beta) = \frac{1}{2n}\|Y - X\beta\|^2 + \sum_{j=1}^{q} \lambda_1 \|\beta_j\| + \sum_{j=1}^{q}\sum_{h=1}^{p} \lambda_2 |\beta_{jh}|, \tag{3.8}$$

where $\beta_j = (\beta_{j1}, \cdots, \beta_{jp})^T \in \mathbb{R}^p$, for $j = 1, \cdots, q$ and $\beta = (\beta_1, \cdots, \beta_q)^T \in \mathbb{R}^{qp}$. There are two tuning parameters, where λ_1 controls the sparsity of genetic factors, and λ_2 controls the sparsity among G×E interactions.

To formulate ADMM framework, we start from expressing equation (3.8) as a constrained optimization problem with auxiliary variable $z \in \mathbb{R}^{qp}$

$$R(\beta, z) = \frac{1}{2n}\|Y - X\beta\|^2 + \sum_{j=1}^{q} \lambda_1 \|z_j\| + \sum_{j=1}^{q}\sum_{h=1}^{p} \lambda_2 |z_{jh}|, \text{ subject to } \beta - z = 0, \tag{3.9}$$

where z_j and z_{jh} are auxiliary variables for β_j and β_{jh}, for $j = 1, \cdots, q$ and $h = 1, \cdots, p$. Equation (3.9) is equivalent to the following augmented form with $\rho_1, \rho_2 (> 0)$ being the augmentation parameters

$$\begin{aligned} R(\beta, z, \tau) = \frac{1}{2n}\|Y - X\beta\|^2 &+ \sum_{j=1}^{q} \left[\lambda_1 \|z_j\| + \tau_j^T(\beta_j - z_j) + \frac{\rho_1}{2}(\beta_j - z_j)^2\right] \\ &+ \sum_{j=1}^{q}\sum_{h=1}^{p} \left[\lambda_2 |z_{jh}| + \tau_{jh}(\beta_{jh} - z_{jh}) + \frac{\rho_2}{2}(\beta_{jh} - z_{jh})^2\right], \end{aligned} \tag{3.10}$$

where β, z are the primal variables and τ is the dual variable. For simplicity, we impose a scaled form ($u = \frac{1}{\rho}\tau$) to simplify equation (3.10)

$$\begin{aligned} R(\beta, z, u) = \frac{1}{2n}\|Y - X\beta\|^2 &+ \sum_{j=1}^{q} \left[\lambda_1 \|z_j\| + \frac{\rho_1}{2}(\beta_j - z_j + u_j)^2 - \frac{\rho_1}{2}u_j^2\right] \\ &+ \sum_{j=1}^{q}\sum_{h=1}^{p} \left[\lambda_2 |z_{jh}| + \frac{\rho_2}{2}(\beta_{jh} - z_{jh} + u_{jh})^2 - \frac{\rho_2}{2}u_{jh}^2\right] \end{aligned} \tag{3.11}$$

At the $(k+1)$th iteration, the solutions of equation (3.11) can be derived in a similar manner

as equation (3.4)

$$\beta^{k+1} = \arg\min_{\beta} R(\beta, z^k, u^k), \tag{3.12}$$

$$z^{k+1} = \arg\min_{z} R(\beta^{k+1}, z, u^k), \tag{3.13}$$

$$u^{k+1} = u^k + (\beta^{k+1} - z^{k+1}). \tag{3.14}$$

Let's start to consider to update β. Equation (3.11) can be simplified as

$$\begin{aligned} R(\beta, z^k, u^k) = \frac{1}{2n}\|Y - X\beta\|^2 + \sum_{j=1}^{q} \frac{\rho_1}{2}\Big(\beta_j - z_j^k + u_j^k\Big)^2 \\ + \sum_{j=1}^{q}\sum_{h=1}^{p} \frac{\rho_2}{2}\Big(\beta_{jh} - z_{jh}^k + u_{jh}^k\Big)^2. \end{aligned} \tag{3.15}$$

The matrix form of equation (3.15) can be expressed as

$$R(\beta, z^k, u^k) = \frac{1}{2n}\|Y - X\beta\|^2 + \frac{\rho_1 + \rho_2}{2}(\beta - z^k + u^k)^2. \tag{3.16}$$

Hence, the minimizer of equation (3.16) with respect to β is

$$\begin{aligned} \frac{\partial R(\beta, z^k, u^k)}{\partial \beta} &= -\frac{1}{n}X^T Y + \frac{1}{n}X^T X\beta + (\rho_1 + \rho_2)\beta - (\rho_1 + \rho_2)(z^k - u^k) = 0, \\ \beta^{k+1} &= \Big[\frac{1}{n}X^T X + (\rho_1 + \rho_2)I\Big]^{-1}\Big[\frac{1}{n}X^T Y + (\rho_1 + \rho_2)(u^k - z^k)\Big]. \end{aligned} \tag{3.17}$$

Next let's consider to update z. Equation (3.11) can be simplified as

$$\begin{aligned} R(\beta^{k+1}, z, u^k) = \sum_{j=1}^{q}\Big[\lambda_1\|z_j\| + \frac{\rho_1}{2}(\beta_j^{k+1} - z_j + u_j^k)^2\Big] \\ + \sum_{j=1}^{q}\sum_{h=1}^{p}\Big[\lambda_2|z_{jh}| + \frac{\rho_2}{2}(\beta_{jh}^{k+1} - z_{jh} + u_{jh}^k)^2\Big]. \end{aligned} \tag{3.18}$$

For $j = 1, \cdots, q$, the minimizer of equation (3.18) with respect to z_j is

$$R(\beta_j^{k+1}, z_j, u_j^k) = \lambda_1 \|z_j\| + \frac{\rho_1}{2}(\beta_j^{k+1} - z_j + u_j^k)^2 + \sum_{h=1}^{p}\Big[\lambda_2|z_{jh}| + \frac{\rho_2}{2}(\beta_{jh}^{k+1} - z_{jh} + u_{jh}^k)^2\Big],$$

$$\begin{aligned}\frac{\partial R(\beta_j^{k+1}, z_j, u_j^k)}{\partial z_j} &= \frac{\lambda_1 z_j}{\|z_j\|} - \rho_1(\beta_j^{k+1} - z_j + u_j^k) + S \\ &= \Big(\frac{\lambda_1}{\|z_j\|} + \rho_1\Big) z_j - \rho_1(\beta_j^{k+1} + u_j^k) + S = 0\end{aligned}$$

where $S = \Big(\lambda_2 \text{sign}(z_{j1}) + \rho_2 z_{j1} - \rho_2(\beta_{j1}^{k+1} + u_{j1}^k), \cdots, \lambda_2 \text{sign}(z_{jp}) + \rho_2 z_{jp} - \rho_2(\beta_{jp}^{k+1} + u_{jp}^k)\Big)^T$. Define $g(z_j) = \Big(\frac{\lambda_1}{\|z_j\|} + \rho_1\Big)$. Fix z_j at current estimate, use g to denote $g(z_j)$. For $h = 1, \cdots, p$,

$$\lambda_2 \text{sign}(z_{jh}) + (g + \rho_2) z_{jh} - (\rho_1 + \rho_2)(\beta_{jh} + u_{jh}) = 0$$

Hence, the individual update of z_{jh} is

$$z_{jh}^{k+1} = S^1_{\frac{\lambda_2}{g+\rho_2}}\Big(\frac{\rho_1 + \rho_2}{g + \rho_2}(\beta_{jh}^{k+1} + u_{jh}^k)\Big) \tag{3.19}$$

where $S^1_\lambda(t) = \text{sign}(t)(|t| - \lambda)_+$. Define $t_{jh} = g z_{jh}^{k+1}, t_j = (t_{j1}, \cdots, t_{jp})$. The group wise update of z_j is

$$\begin{aligned}\Big(\frac{\lambda_1}{\|z_j\|} + \rho_1\Big) z_j &= t_j, \\ z_j^{k+1} &= S^2_{\frac{\lambda_1}{\rho}}\Big(\frac{t_j}{\rho 1}\Big),\end{aligned} \tag{3.20}$$

where $S^2_\lambda(t) = \Big(1 - \frac{\lambda}{\|t\|}\Big)_+ t$. We determine individual level sparsity in equation (3.19), indicating that only a subset of G×E interactions is related to cancer outcome for each gene. Group level sparsity is determined in equation (3.20), suggesting that only a subset of genetic main effects is related to cancer outcome. The tuning parameters are ρ_1, ρ_2, λ_1, and λ_2.

For large-scale genetics data, the computation speed becomes a big challenge for penalized variable selection. With the proposed penalization method, we adopt a parallel ADMM

framework to improve computation speed without satisfactory identification accuracy. The parallel ADMM can conduct the optimization by using divide-and-conquer strategy.

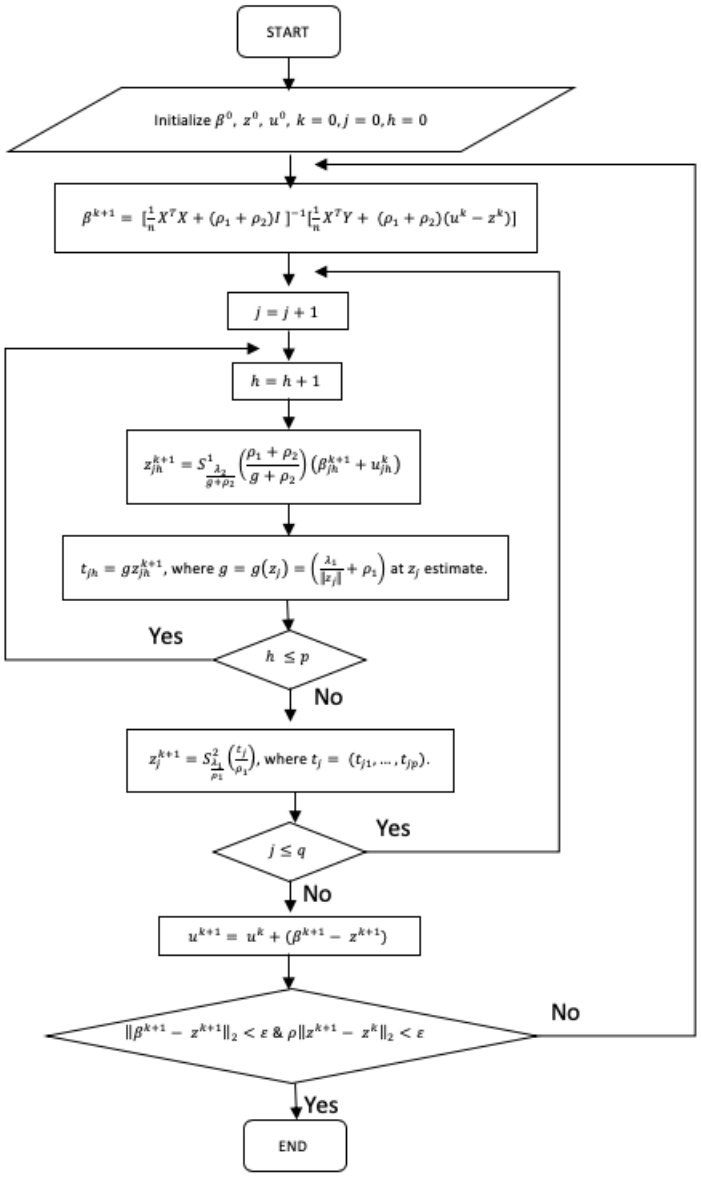


Figure 3.2: Flowchart of ADMM framework for bi-level selection from Section 3.3.

3.4 The Parallel ADMM

The ADMM can be well suited to distributed convex optimization (Boyd et al. (2011)). We can distribute different batches of data to different machines to improve computation speed. In this section, we demonstrate that by formulating equation (3.1) into specific parallel ADMM framework across samples or features. First, we review the discussion of parallel ADMM framework in Boyd et al. (2011).

3.4.1 Split across Samples

For classical statistical problems, we consider large amount of samples with small number of variables ($n \gg p$). In such case, we can divide the dataset into different batches of small numbers of samples with same number of variables for each batch. Consider the following distributed model fitting problem

$$\sum_{i=1}^{N} \frac{1}{2n_i} \|Y_i - X_i\beta_i\|^2 + P_\lambda(z), \text{ subject to } \beta_i - z = 0, \tag{3.21}$$

where $\beta_i \in \mathbb{R}^p$ and $z \in \mathbb{R}^p$. Hence, $Y_i \in \mathbb{R}^{n_i}$ and $X_i \in \mathbb{R}^{n_i \times p}$ represent the ith block of data, where $\sum_{i=1}^{N} = n$ and $i = 1, \cdots, N$. Each block of data can be processed in parallel by different machines. At the $(k+1)$th iteration, the solutions of the parallel ADMM across samples are

$$\beta_i^{k+1} = \arg\min_{\theta_i} \Big(\frac{1}{2n_i} \|Y_i - X_i\beta_i\|^2 + \frac{\rho}{2} \|\beta_i - z^k + u_i^k\|^2 \Big), \tag{3.22}$$

$$z^{k+1} = \arg\min_{z} \Big(P_\lambda(z) + \frac{N\rho}{2} \|z - \bar{\beta}^{k+1} - \bar{u}^k\|^2 \Big), \tag{3.23}$$

$$u_i^{k+1} = u_i^k + (\beta_i^{k+1} - z^{k+1}). \tag{3.24}$$

The update of β_i can be carried out in parallel for each block of data. Equation (3.23) shows that the update of z requires to collect variables to form the average. As examples, we conduct LASSO and group LASSO through the parallel ADMM across samples in Section

B.3.1 and Section B.3.2.

3.4.2 Split across Features

On the other side, there're many studies with a modest number of samples and a large number of features ($n \ll p$). For example, there're usually relatively few subjects with a very large number of SNPs in the cancer studies. Compared to Section 3.4.1, for all the samples, we can distribute the large number of variables into different batches. Consider the following distributed model fitting problem

$$\frac{1}{2n}\|Y - \sum_{i=1}^{N} X_i\beta_i\|^2 + \sum_{i=1}^{N} P_\lambda(z_i), \text{ subject to } X_i\beta_i - z_i = 0. \tag{3.25}$$

The design matrix X is divided as $X = [X_1, \cdots, X_N]$ with $X_i \in \mathbb{R}^{n\times p_i}$, and partition coefficient vector β as $\beta = (\beta_1, \cdots, \beta_N)^T$ with $\beta_i \in \mathbb{R}^{p_i}$, where $\sum_{i=1}^{N} p_i = p$. The auxiliary variable is $z_i \in \mathbb{R}^{p_i}$. The corresponding penalty function can be partitioned as $\sum_{i=1}^{N} P_\lambda(z_i)$. In addition, we use the same tuning parameter for different partitioned penalties because each block of variables should be in similar scale.

The parallel approach can be thought as partial prediction of Y using only the features referenced in β_i. At the $(k+1)$th iteration, the solutions of the parallel ADMM across features are

$$\begin{aligned}
\beta_i^{k+1} &= \arg\min_{\beta_i} \left(P_\lambda(\beta_i) + \frac{\rho}{2}\|X_i\beta_i - z_i^k + u_i^k\|^2\right) \\
z^{k+1} &= \arg\min_{z} \left(\frac{1}{2n}\|Y - \sum_{i=1}^{N} z_i\|^2 + \sum_{i=1}^{N}\frac{\rho}{2}\|X_i\beta_i^{k+1} - z_i^k + u_i^k\|^2\right) \\
u_i^{k+1} &= u_i^k + (X_i\theta_i^{k+1} - z_i^{k+1}).
\end{aligned}$$

Let $\overline{X\beta}^{k+1} = 1/N\sum_{i=1}^{N} X_i\beta_i^{k+1}$. To further simplify the solutions, the update of z can be

used by the average information such that

$$\bar{z}^{k+1} = \arg\min_{\bar{z}} \left(\frac{1}{2n} \|Y - qz\|^2 + \frac{N\rho}{2} \|\bar{z} - \overline{X\beta}^{k+1} - \bar{u}^k\|^2 \right)$$
$$z_i^{k+1} = \bar{z}^{k+1} + X_i\beta_i^{k+1} + u_i^k - \overline{X\beta}^{k+1} - \bar{u}^k.$$

Applying the above results into the update of u_i we'll have

$$u_i^{k+1} = \overline{X\beta}^{k+1} + \bar{u}^k - \bar{z}^{k+1},$$

where it indicats that all the dual variables are equal. Replacing z_i by the single dual variable u^k, the solutions of the parallel ADMM across features can be rewritten as

$$\beta_i^{k+1} = \arg\min_{\beta_i} \left(P_\lambda(\beta_i) + \frac{\rho}{2} \|X_i\beta_i - X_i\beta_i^k - \bar{z}^k + \overline{X\beta}^k + u_i^k\|^2 \right) \tag{3.26}$$
$$\bar{z}^{k+1} = \arg\min_{\bar{z}} \left(\frac{1}{2n} \|Y - Nz\|^2 + \frac{N\rho}{2} \|\bar{z} - \overline{X\beta}^{k+1} - \bar{u}^k\|^2 \right) \tag{3.27}$$
$$u^{k+1} = u^k + (\overline{X\beta}^{k+1} - \bar{z}^{k+1}) \tag{3.28}$$

Hence, the parallel ADMM across features can solve N parallel regularized least squares in p_i variables, respectively. From (3.26) to (3.27), we collect all the partial predictions $X_i\beta_i^{k+1}$ to form average prediction $\overline{X\beta}^{k+1}$. Then we update the average $\bar{z}^{k+1}$ by minimizing the quadratic function. The dual variable update is based on (3.28). Each update of β_i is a regularized problem with p_i variables, which can be solved in parallel. As examples, we conduct LASSO and group LASSO through parallel ADMM across features in Section B.2.1 and Section B.2.2.

3.5 Parallel ADMM for Bi-level Selection

In this section, we develop a parallel ADMM across samples and across features to conduct a penalized bi-level selection, so we can distribute different batches of data onto different

machines to imputation speed with satisfactory identification accuracy.

3.5.1 Parallel ADMM for Bi-level Selection across Samples

Consider the following distributed model fitting problem

$$\sum_{i=1}^{N} \frac{1}{2n_i} \|Y_i - X_i\beta_i\|^2 + \sum_{j=1}^{q} \lambda_1 \|z_j\| + \sum_{j=1}^{q} \sum_{h=1}^{p} \lambda_2 |z_{jh}|, \text{ subject to } \beta_i - z = 0 \tag{3.29}$$

where $\beta_i \in \mathrm{R}^{\mathrm{qp}}$, $z \in \mathrm{R}^{\mathrm{qp}}$, and i stands for the index of the ith block of data, for $i = 1, \cdots, N$. The data is divided into N blocks with the same number of variables. $Y_i \in \mathrm{R}^{\mathrm{n_i}}$ and $X_i \in \mathrm{R}^{\mathrm{n_i} \times \mathrm{qp}}$, where $\sum_{i=1}^{N} = n$. Then each block of data can be processed in parallel by different machines to improve computation speed. Following equation (3.11), we write equation (3.29) in Lagrangian form

$$\begin{aligned} R(\beta, z, u) = &\sum_{i=1}^{N} \frac{1}{2n_i} \|Y_i - X_i\beta_i\|^2 + \sum_{j=1}^{q} \left[\lambda_1 \|z_j\| + \frac{\rho_1}{2} (\beta_j - z_j + u_j)^2 - \frac{\rho_1}{2} u_j^2 \right] \\ &+ \sum_{j=1}^{q} \sum_{h=1}^{p} \left[\lambda_2 |z_{jh}| + \frac{\rho_2}{2} (\beta_{jh} - z_{jh} + u_{jh})^2 - \frac{\rho_2}{2} u_{jh}^2 \right] \end{aligned} \tag{3.30}$$

Then the updates of (β, z, u) follow equations (3.22), (3.23) and (3.24). For $i = 1, \cdots, N$,

$$\begin{aligned} \frac{\partial R(\beta_i, z^k, u_i^k)}{\partial \beta_i} &= -\frac{1}{n_i} X_i^T \Big(Y_i - X_i \beta_i \Big) + (\rho_1 + \rho_2)(\beta_i - z^k + u_i^k) = 0 \\ \beta_i^{k+1} &= \left[\frac{1}{n_i} X_i^T X_i + (\rho_1 + \rho_2) I \right]^{-1} \left[\frac{1}{n_i} X_i^T Y_i + (\rho_1 + \rho_2)(z^k - u_i^k) \right] \end{aligned} \tag{3.31}$$

Next, we conduct the solution of z update through penalized bi-level selection. For $j = 1, \cdots, q$,

$$\begin{aligned}\frac{\partial R(\beta_j^{k+1}, z_j, u_j^k)}{\partial z_j} &= \frac{\lambda_1 z_j}{\|z_j\|} - N\rho_1\Big(\bar{\beta}_j^{k+1} - z_j + \bar{u}_j^k\Big) + S \\ &= \Big(\frac{\lambda_1}{\|z_j\|} - N\rho_1\Big)z_j - N\rho_1\Big(\bar{\beta}_j^{k+1} + \bar{u}_j^k\Big) + S \\ &= g(z_j)z_j - N\rho_1\Big(\bar{\beta}_j^{k+1} + \bar{u}_j^k\Big) + S = 0\end{aligned}$$

where $S = \Big(\lambda_2 \text{sign}(z_{j1}) + N\rho_2 z_{j1} - N\rho_2(\bar{\beta}_{j1}^{k+1} + \bar{u}_{j1}^k), \cdots, \lambda_2 \text{sign}(z_{jp}) + N\rho_2 z_{jp} - N\rho_2(\bar{\beta}_{jp}^{k+1} + \bar{u}_{jp}^k)\Big)^T$. Fix z_j at current estimate and use g to denote $g(z_j)$. For $h = 1, \cdots, p$,

$$\begin{aligned}&\lambda_2 \text{sign}(z_{jh}) + (g + N\rho_2)z_{jh} - (N\rho_1 + N\rho_2)(\bar{\beta}_{jh}^{k+1} + \bar{u}_{jh}^k) = 0 \\ &z_{jh}^{k+1} = S^1_{\frac{\lambda_2}{g+N\rho_2}}\Big(\frac{N\rho_1 + N\rho_2}{g + N\rho_2}(\bar{\beta}_{jh}^{k+1} + \bar{u}_{jh}^k)\Big)\end{aligned}$$

Let $t_{jh} = g z_{jh}^{k+1}$ and $t_j = (t_{j1}, \cdots, t_{jp})$. With $t_j = g(z_j)z_j$, we have

$$\Big(\frac{\lambda_1}{\|z_j\|} + N\rho_1\Big)z_j = t_j, \;\; z_j^{k+1} = S^2_{\frac{\lambda_1}{N\rho_1}}\Big(\frac{t_j}{N\rho_1}\Big) \tag{3.32}$$

The parallel ADMM across samples iteratively updates β_i on different machines. For each machine, group level and individual level sparsity are controlled by bi-level selection penalization.

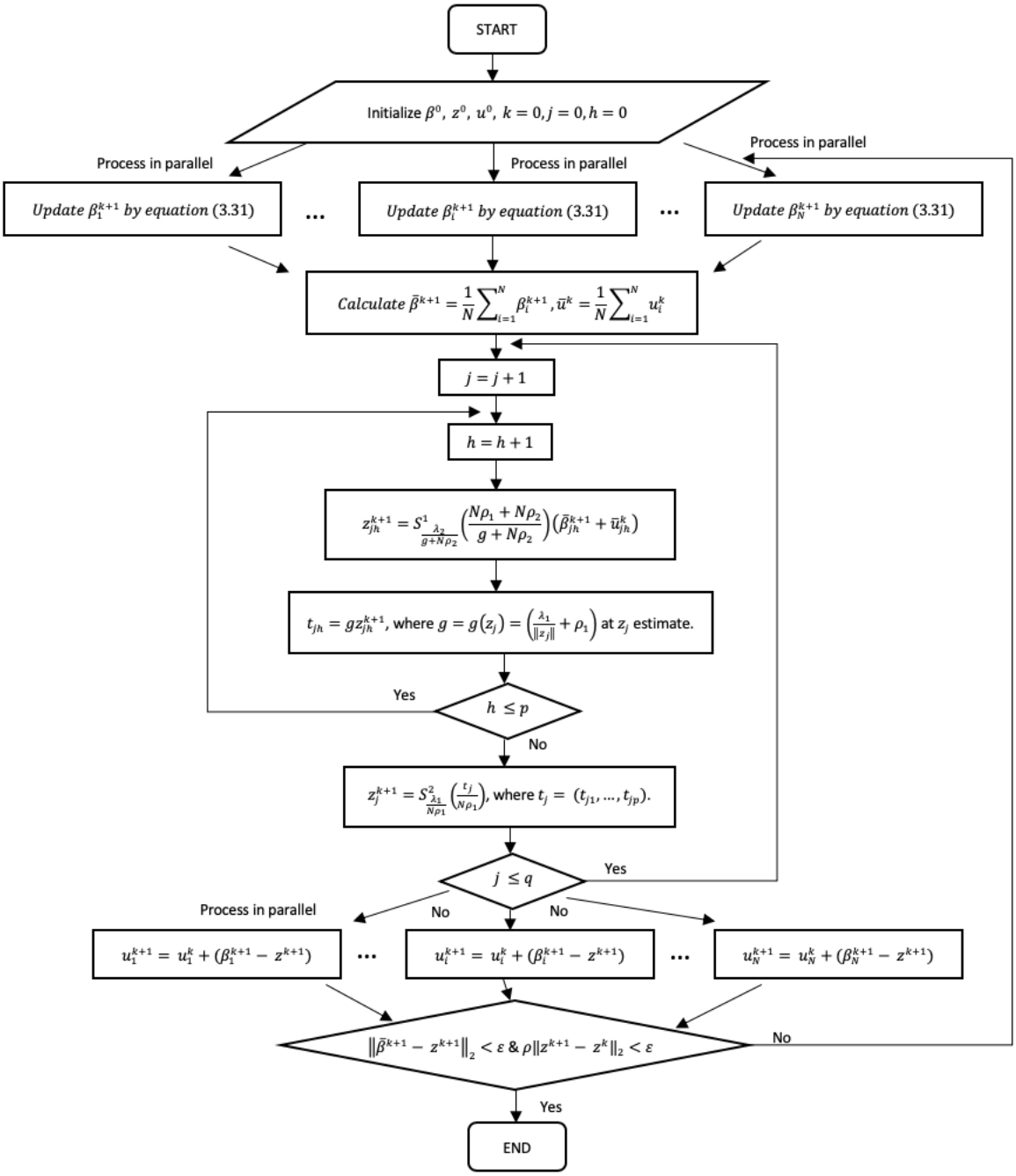


Figure 3.3: Flowchart of parallel ADMM framework for bi-level selection across samples from Section 3.5.1.

3.5.2 Parallel ADMM for Bi-level Selection across Features

Consider the following distributed model fitting problem

$$\frac{1}{2n}\|Y-\sum_{i=1}^{q}X_i\beta_i\|^2+\sum_{i=1}^{q}\left[\lambda_1\|z_i\|+\sum_{h=1}^{p}\lambda_2|z_{ih}|\right],\ \text{subject to } X_i\beta_i-z_i=0, \tag{3.33}$$

where the columns of the data are divided into q groups with $X_i\in\mathbb{R}^{n\times p}$ and $\beta_i\in\mathbb{R}^p$, for $i=1,\cdots,q$. Then the update of (β,z,u) follow corresponding equations (3.26), (3.27) and (3.28). Equation (3.27) shows that the update of $\bar{z}$ is not involved any penalization. Then the update of $\bar{z}$ can be derived

$$\begin{aligned}\frac{\partial R(\beta^{k+1},\bar{z},\bar{u}^k)}{\partial\bar{z}}&=-\frac{N}{n}\Big(Y-N\bar{z}\Big)-N\Big(\rho_1+\rho_2\Big)\Big(\overline{X\beta}^{k+1}-\bar{z}+\bar{u}^k\Big)=0,\\ \bar{z}^{k+1}&=\Big[N+(\rho_1+\rho_2)n\Big]^{-1}\Big[Y+(\rho_1+\rho_2)n\overline{X\beta}^{k+1}+(\rho_1+\rho_2)n\bar{u}^k\Big].\end{aligned} \tag{3.34}$$

Next, let's consider the update of β_i. Let $M=X_i\beta_i^k-\overline{X\beta}^k+\bar{z}^k-u^k$. For $i=1,\cdots,N$,

$$\frac{\partial R(\beta_i,z^k,u^k)}{\partial\beta_i}=\frac{\lambda_1\beta_i}{\|\beta_i\|}+\rho_1X_i^T(X_i\beta_i-M)+S=g(\beta_i)\beta_i-\rho_1X_i^TM+S=0,$$

where $S=(\lambda_2\text{sign}(\beta_{i1})+\rho_2x_{i1}(x_{i1}-M),\cdots,\lambda_2\text{sign}(\beta_{ip})+\rho_2x_{ip}(x_{ip}-M))^T$. Fix β_i at the current estimate and use g to denote $g(\beta_i)$. For $h=1,\cdots,p$,

$$\begin{aligned}&\lambda_2\text{sign}(\beta_{ih})+(g+\rho_2x_{ih}^2)\beta_{ih}-(\rho_1+\rho_2)x_{ih}M=0\\ &\beta_{ih}^{k+1}=S^1_{\frac{\lambda_2}{g+\rho_2x_{ih}^2}}\Big(\frac{\rho_1+\rho_2}{g+\rho_2x_{ih}^2}x_{ih}M\Big)\end{aligned}$$

Let $t_{ih}=g\beta_{ih}^{k+1}$ and $t_i=(t_{i1},\cdots,t_{ip})$. With $g(\beta_i)\beta_i=t_i$, we have

$$\Big(\frac{\lambda_1}{\|\beta_i\|}+\rho_1n\Big)\beta_i=t_i,\ \beta_i^{k+1}=S^2_{\frac{\lambda_1}{\rho_1n}}\Big(\frac{t_i}{\rho_1n}\Big) \tag{3.35}$$

The parallelization of the ADMM across features can be implemented on a distributed computing framework. Each β_i update can be solved on different machines.

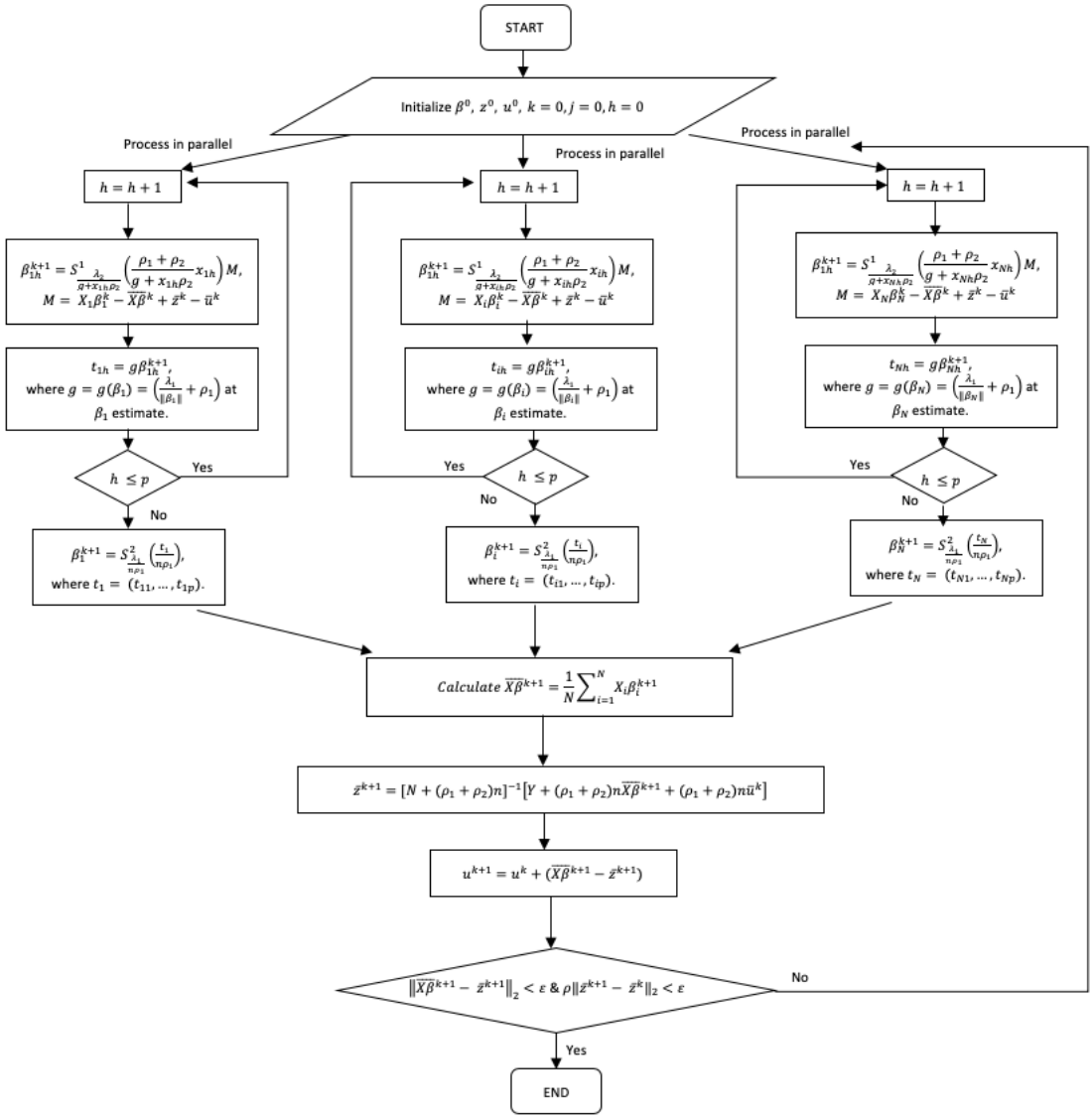


Figure 3.4: Flowchart of parallel ADMM framework for bi-level selection across features from Section 3.5.2.

3.6 Simulation

We compare the performance of the bi-level selection (SGLASSO) to two alternatives. LASSO is an individual level penalty without considering the grouping structure in G×E.

Group LASSO (GLASSO) is a group level penalty without achieving sparsity within groups, that is, once a gene is selected, its main effect and all interactions are selected. To evaluate the computation speed, we evaluate the performance of all three methods through parallel and non-parallel ADMM frameworks. Denote PLASSO, PGLASSO and PSGLASSO as parallel approaches for LASSO, GLASSO and SGLASSO, respectively.

To set up different simulation scenarios, we generate datasets with different correlation structures and correlation levels, each with n subjects, p genes and q environmental factors, respectively. For each subject, we simulate genetic matrix through $G_{n\times p} \sim N(0, \Sigma)$, where we consider the following four covariance structures for Σ:

1. (AR1-3) Autoregressive correlation structure with correlation coefficients $0.3^{|i-j|}$ for the ith and jth variables.

2. (AR1-5) Autoregressive correlation structure with correlation coefficients $0.5^{|i-j|}$ for the ith and jth variables.

3. (banded 1) Banded correlation structure, in which variable i and j have correlation coefficients $\rho = 0.11$ if $|i-j| = 1$ and $\rho = 0$ otherwise.

4. (banded 2) Banded correlation structure, in which variable i and j have correlation coefficients $\rho = 0.2$ if $|i-j| = 1$, 0.11 if $|i-j| = 2$ and 0 otherwise.

The environmental factors are from $E_{n\times q} \sim N(0, \Sigma_E)$, where we choose Σ_E to be autoregressive with correlation coefficient 0.9. To assign nonzero coefficients, we randomly select 4 groups. For each group, we randomly select 5 entries with coefficients generated from Unif[0, 0.5]. Denote coefficient vector $\beta = (\beta_1, \cdots, \beta_p) \in \mathbb{R}^{pq}$, where $\beta_j \in \mathbb{R}^q$, for $j = 1, \cdots, p$. Then the response variable Y is generated from regression model $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, 1)$.

Simulations are evaluated under 6 different settings: we choose $(n, p, q) = (800, 100, 10)$, $(800, 50, 10)$, $(1500, 100, 10)$, $(1500, 50, 10)$, $(5000, 100, 10)$, $(10000, 100, 10)$, respectively. First three settings are used to evaluate the performance of parallel ADMM across features and last three settings are evaluated for performance across samples. Evaluation of feature selection accuracy is based on true positive (TP) and false positive (FP). Evaluation of prediction

is based on mean square error between fitted values and true values. Computation time is reported in seconds. The simulation is repeated 100 times and conducted on a PC with Core i7 4-core processor and 8GB RAM.

The simulation results of the six approaches are tabulated in table 3.1, B.1, B.2, 3.2, B.3. B.4. In general the bi-level selection has better identification than alternatives. For example, in table 3.1, given $n = 800$, $p = 100$, $q = 10$, under covariance structure 1 (AR1-3). LASSO identifies 17.80 (SD 1.90) TPs with 10.50 (SD 1.50) FPs. GLASSO identifies a larger number of TPs, 18.23 (SD 2.35), with a larger number of FPs, 49.50 (SD 27.94). However, the SGLASSO can identify larger number of TPs, 18.53 (SD 1.96), with a small number of FPs, 2.63 (SD 1.24). Among three approaches, the difference in performance comes from identifying G×E interactions. SGLASSO can determine the sparsity on both group and individual levels. For prediction accuracy, SGLASSO outperforms LASSO with lower FPs. GLASSO can obtain highest prediction accuracy because it identifies a larger number of FPs. Overall, we can observe the similar patterns for the other covariance structures and simulation settings.

In terms of computation speed, we compare parallel ADMM and traditional ADMM frameworks under either large p or large n setting. In a word, the parallel ADMM can significantly improve the computation speed and the proposed PSGLASSO has better identifications than alternatives. For example, in table 3.1, under $p > n$ setting. For computation time, all parallel approaches are almost four times faster than non-parallel approaches. For identification under covariance structure 1 (AR1-3), PLASSO identifies 18.43 (SD 1.97) TPs, which is better than the 17.80 (SD 1.90) TPs identified by LASSO. However, similar FPs are identified by both approaches. PGLASSO identifies 18.66 (SD 5.07) TPs and 49.03 (SD 27.90) FPs, which are similar as the identification of GLASSO. The proposed PSGLASSO identifies 18.36 (SD 2.79) TPs and 2.96 (SD 1.52) FPs, which are also similar as the identification of SGLASSO. Hence, the parallel ADMM across features can improve computation speed without loosing identification accuracy.

Another example, in table 3.2, under $n > p$ setting. For computation time, all parallel approaches are almost two times faster than non-parallel approaches. For identification

under covariance structure 1 (AR1-3), PLASSO identifies a slightly larger FPs, 13.50 (SD 2.75), but it identifies similar TPs, 18.03 (SD 1.12). PGLASSO identifies smaller FPs, 54.20 (SD 20.16), and similar TPs, 18.86 (SD 1.13). PSGLASSO identifies 3.23 (SD 2.45) FPs, which is slightly larger than SGLASSO. However, PSGLASSO identifies similar TPs as SGLASSO, which is 18.63 (SD 1.15). Overall, the parallel ADMM across samples can improve computation speed with satisfactory identification accuracy.

The comparison of parallel ADMM in different numbers (M) of chunks is shown in Figure 3.5 from a large sample case. In particular, we evaluate the performance of differences for LASSO penalty with $(n, p, q) = (10000, 100, 10)$. Other methods have similar trends. The number (M) of chunks is set to 1, 10 and 100, respectively. As shown in figure 3.5, the parallel algorithm maintains prediction accuracy and it significantly reduces computation time.

Figure 3.5: Comparison of parallel ADMM for splitting across samples with LASSO penalty for different numbers (M) of subset of data. (n, p, q) = (10000, 100, 10).

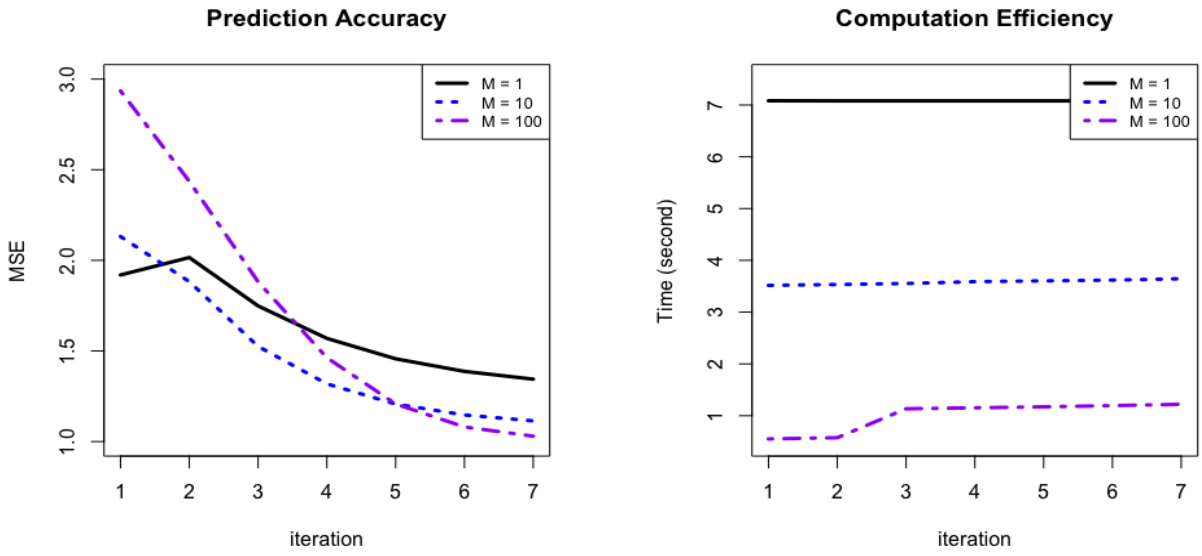


Table 3.1: Comparison between ADMM and parallel ADMM in splitting features for $(n, p, q) = (800, 100, 10)$ with LASSO, GLASSO and SGLASSO penalties. Numbers are mean (standard deviation) of true positives (TP), false positives (FP), prediction and time.

Approach	Correlation	TP	FP	Prediction	Time
LASSO	AR1-3	17.80(1.90)	10.50(1.50)	0.82(0.04)	20.36(0.60)
	AR1-5	18.10(1.88)	12.00(1.91)	0.81(0.05)	20.79(0.91)
	banded1	18.10(2.09)	10.06(1.79)	0.82(0.05)	20.38(0.46)
	banded2	17.70(1.93)	9.46(1.27)	0.85(0.05)	20.49(0.56)
GLASSO	AR1-3	18.23(2.35)	49.50(27.94)	0.95(0.07)	20.34(1.20)
	AR1-5	18.03(2.20)	43.13(21.70)	0.91(0.07)	20.60(0.94)
	banded1	17.63(2.71)	50.13(24.58)	0.98(0.09)	20.20(0.77)
	banded2	18.26(3.26)	52.73(24.54)	0.96(0.07)	20.42(0.92)
SGLASSO	AR1-3	18.53(1.96)	2.63(1.24)	0.84(0.04)	22.85(1.03)
	AR1-5	18.56(1.71)	3.60(1.99)	0.82(0.03)	23.30(1.07)
	banded1	18.73(1.68)	2.30(1.20)	0.83(0.05)	22.81(0.30)
	banded2	18.56(1.61)	1.90(1.02)	0.81(0.05)	23.02(0.75)
Approach	Correlation	TP	FP	Prediction	Time
PLASSO	AR1-3	18.43(1.97)	10.90(2.04)	0.84(0.05)	5.09(0.92)
	AR1-5	18.26(2.58)	8.70(3.15)	0.83(0.05)	5.20(1.08)
	banded1	18.46(2.37)	9.20(1.24)	0.84(0.05)	5.09(0.53)
	banded2	18.20(2.15)	8.46(1.30)	0.84(0.04)	5.12(0.54)
PGLASSO	AR1-3	18.66(5.07)	49.03(27.90)	0.91(0.47)	5.08(1.06)
	AR1-5	19.06(3.94)	54.30(29.56)	0.91(0.30)	5.15(1.39)
	banded1	17.63(5.54)	46.33(23.90)	0.86(0.49)	5.05(0.58)
	banded2	17.50(5.11)	48.26(25.81)	0.85(0.52)	5.10(0.46)
PSGLASSO	AR1-3	18.36(2.79)	2.96(1.52)	0.86(0.04)	5.71(2.81)
	AR1-5	18.80(3.42)	2.86(2.14)	0.87(0.08)	5.82(2.75)
	banded1	18.23(2.95)	2.53(0.90)	0.87(0.06)	5.70(1.01)
	banded2	18.83(2.92)	2.96(1.18)	0.87(0.05)	5.75(0.86)

Table 3.2: Comparison between ADMM and parallel ADMM in splitting samples for $(n, p, q) = (10000, 50, 10)$ with LASSO, GLASSO and SGLASSO penalties. Numbers are mean (standard deviation) of true positives (TP), false positives (FP), prediction and time.

Approach	Correlation	TP	FP	Prediction	Time
LASSO	AR1-3	18.93(1.55)	10.60(2.65)	0.81(0.01)	55.30(1.08)
	AR1-5	18.70(1.68)	9.70(2.12)	0.81(0.01)	55.39(1.25)
	banded1	18.13(2.25)	10.00(3.08)	0.81(0.01)	55.91(0.61)
	banded2	18.13(1.71)	11.26(3.06)	0.81(0.01)	55.59(1.44)
GLASSO	AR1-3	18.90(2.21)	57.46(17.32)	0.81(0.02)	52.89(1.37)
	AR1-5	18.86(2.06)	55.90(15.08)	0.83(0.02)	50.61(0.73)
	banded1	18.13(2.31)	56.93(16.06)	0.82(0.02)	50.65(0.91)
	banded2	18.23(1.75)	52.10(16.76)	0.82(0.01)	53.41(1.90)
SGLASSO	AR1-3	18.76(1.10)	2.36(1.60)	0.80(0.01)	74.68(2.06)
	AR1-5	18.60(1.35)	2.30(1.97)	0.80(0.01)	73.38(0.72)
	banded1	18.06(1.33)	2.67(1.25)	0.80(0.01)	73.21(1.53)
	banded2	18.23(1.22)	2.20(1.90)	0.80(0.01)	73.29(2.04)
Approach	Correlation	TP	FP	Prediction	Time
PLASSO	AR1-3	18.03(1.12)	13.50(2.75)	0.80(0.01)	25.65(1.55)
	AR1-5	18.23(1.38)	10.16(2.81)	0.80(0.01)	25.69(1.01)
	banded1	18.43(1.61)	12.83(1.91)	0.80(0.01)	25.45(1.07)
	banded2	18.76(1.07)	11.56(1.54)	0.80(0.01)	25.79(1.74)
PGLASSO	AR1-3	18.86(1.13)	54.20(20.16)	0.81(0.01)	26.44(4.08)
	AR1-5	18.70(1.86)	55.06(29.37)	0.81(0.01)	25.30(0.94)
	banded1	18.93(2.25)	55.50(22.61)	0.81(0.01)	25.32(1.01)
	banded2	18.46(1.54)	56.86(27.47)	0.80(0.01)	26.70(1.74)
PSGLASSO	AR1-3	18.63(1.15)	3.23(2.45)	0.80(0.01)	34.04(2.47)
	AR1-5	18.56(1.35)	2.91(2.06)	0.81(0.01)	33.35(1.18)
	banded1	18.03(1.35)	2.53(1.73)	0.80(0.01)	33.26(0.87)
	banded2	18.23(1.22)	2.34(1.86)	0.80(0.01)	36.99(7.52)

Table 3.3: Summary of ADMM Frameworks

Framework	Penalized Least Square Loss	Solutions	Method
General ADMM	Equation (3.1)	Equation (3.5), (3.6), (3.7)	Existing
General Parallel ADMM across Samples	Equation (3.21)	Equation (3.22), (3.23), (3.24)	Existing
General Parallel ADMM across Features	Equation (3.25)	Equation (3.26), (3.27), (3.28)	Existing
ADMM for SGLASSO	Equation (3.8)	Equation (3.17), (3.8), (3.14)	New
Parallel ADMM for SGLASSO across Samples	Equation (3.29)	Equation (3.31), (3.32), (3.24)	New
Parallel ADMM for SGLASSO across Features	Equation (3.33)	Equation (3.35), (3.34), (3.28)	New

Chapter 4

Parallel Penalized Variable Selection for GxE Interactions in Case Control Study of Type 2 Diabetes

4.1 Introduction

Type 2 Diabetes (T2D) is a common human disease, which is related to both environmental and genetic factors. Many studies have shown that gene-environment (G×E) interaction effects are associated with disease traits beyond genetic and environment main effects. For example, the interaction between gene TCF7L2 and environmental variables, such as physical activity and lifestyle changes has been reported to be associated with the risk of developing T2D (Wu and Cui (2013), Wu et al. (2014) and Ren et al. (2020)). With high dimensionality of genetic data, it has been a challenge to identify important genetic main effects and G×E interactions.

Penalized variable selection is one of the most popular approaches in high-dimensional data studies. Nowadays, many penalization methods have been developed to account for complicated data structures. Despite success, existing methods on G×E interactions still have limitations on identification and computation speed. We develop a penalized bi-level

selection method to better identify important genetics main effects and G×E interactions for binary response in large-scale data.

With the proposed penalization method, we adopt a parallel ADMM framework to improve computation speed for a large-scale data analysis. The parallel ADMM can conduct the optimization by using divide-and-conquer strategy without losing identification accuracy. From the simulation studies, the proposed penalized bi-level selection method outperforms other alternatives with better identification of genetic main effects and G×E interactions for binary response. Additionally, the proposed parallel ADMM framework can significantly improve computation speed with satisfactory identification for binary response. In the case control study of T2D from the Nursers's Health Studies (NHS), the proposed parallel penalzied bi-level selection method has meaningful identifications of important genetic main effects and G×E interactions with a much faster speed.

The rest of chapter is organized as follows. Section 4.2 describes the model in detail. In Section 4.3, we evaluate the performance of proposed approach through simulation studies and compare with alternative approaches. A case control study of T2D is presented in Section 4.4.

4.2 Method

Denote $G_{n\times p} = (G_1, \cdots, G_p)$ as the p genes and $E_{n\times q} = (E_1, \cdots, E_q)$ as the q environment factors. Denote $Y_{n\times 1} = (y_1, \cdots, y_n)^\top$ as the binary response, where $y_i = 1$ indicate the case of disease, and 0 otherwise, $i = 1, \cdots, n$. Consider the following G×E model with the joint effects of all E and G and their interactions

$$\begin{aligned}\mu &= \alpha_0 + \sum_{h=1}^{q} \alpha_h E_h + \sum_{j=1}^{p} \beta_{j0} G_j + \sum_{j=1}^{p}\sum_{h=1}^{q} \beta_{jh} G_j E_h \\ &= \alpha_0 + \sum_{h=1}^{q} \alpha_h E_h + \sum_{j=1}^{p} \left(\beta_{j0} G_j + \sum_{h=1}^{q} \beta_{jh} G_j E_h \right) \\ &= \alpha_0 + \sum_{h=1}^{q} \alpha_h E_h + \sum_{j=1}^{p} X_j \beta_j, \end{aligned} \tag{4.1}$$

where $X_j = (G_j, G_j E_1, \cdots, G_j E_q)$, α_0 is the intercept, α_h is the regression coefficient for the hth environment factor, and β_{j0} is the regression coefficient for the jth gene. $\beta_j = (\beta_{j0}, \beta_{j1}, \cdots, \beta_{jq})^\top$ corresponds to the jth genetic main effect and its interactions with q environment factors.

The logistic regression model can be expressed as

$$P(y_i = 1|\mu_i) = \pi_i = \frac{e^{\mu_i}}{1 + e^{\mu_i}},\ i = 1, \cdots, n.$$

where μ_i is the ith component of μ. The corresponding loss function for logistic regression is the negative log-likelihood

$$\begin{aligned} L(\mu) = \frac{1}{n}\sum_{i=1}^{n} L_i(\mu_i) &= -\frac{1}{n}\sum_{i=1}^{n} log P(Y_i = y_i|\mu_i) \\ &= -\frac{1}{n}\sum_{i=1}^{n} \big\{ y_i log\pi_i + (1 - y_i) log(1 - \pi_i) \big\}. \end{aligned} \tag{4.2}$$

4.2.1 Penalized Logistic Regression for G×E Interactions

With the evidence of previous studies, a small number of environment factors are pre-selected. They may be always included in the model, but the selection is not our interest. Hence, we consider a weak hierarchical structure between main and interaction effects, under which if an interaction term is selected as important, then at least one of the two corresponding main effects is selected.

The number of genetic factors and G×E interactions is much larger than the sample size, and only a small subset of important genetic factors and G×E interactions that are associated with disease status. We proposed a penalization method to identify important effects in the G×E interaction study. Consider the following penalized likelihood

$$R(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\left\{y_i log\pi_i + (1-y_i)log(1-\pi_i)\right\} + \sum_{j=1}^{p}\lambda_1\|\beta_j\| + \sum_{j=1}^{p}\sum_{h=0}^{q}\lambda_2|\beta_{jh}|, \qquad (4.3)$$

where $\beta_j = (\beta_{j0}, \beta_{j1}, \cdots, \beta_{jq}) \in \mathbb{R}^{q+1}$ and $\beta = (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^{p(q+1)}$, for $j = 1, \cdots, p$. λ_1 and λ_2 are tuning parameters. On the group level, β_j is the coefficient vector for the jth genetic main effect and its interactions with q environment factors. The group LASSO penalty determines whether the jth genetic factor is associated with the binary trait. If β_j is nonzero, then either the corresponding main effect or G×E interactions or both can be associated with the outcome. On the individual level, the LASSO penalty further determines the specific individual effects that are associated with disease outcome. Hence, the bi-level selection can identify important main and interaction effects on group and individual levels simultaneously. Both types of effects are penalized on group level and individual levels. The environment main effects always exist in the model. So once at least one interaction effect is nonzero, the genetic main effect can be either zero or nonzero, which respects the weak hierarchical structure.

Penalized variable selection has been widely adopted for high dimensional genetic data analysis. For G×E studies, tailored penalization methods need to be considered for the interaction structure. Here, we develop the bi-level selection based on LASSO and group LASSO, but other baseline penalties can also be developed, such as MCP and group MCP.

Computation is a challenge for the development of penalized variable selection approaches in large-scale data. Coordinate descent algorithm (CD) is one of the most popular algorithms for penalized variable selection in G×E studies, but it cannot be developed in parallel form to improve computation speed. Here, we propose an alternative algorithm by applying alternating direction method of multiplier (ADMM) due to its nature as a distributed optimization method. Specificly, we develop a parallel ADMM framework for bi-level selection across fea-

tures. The proposed method conducts the optimization by using divide-and-conquer strategy while achieving good efficiency and accuracy. The parallel ADMM can be further implemented on distributed computing platform such as Hadoop (Dean and Ghemawat (2008)) and Spark (Zaharia et al. (2010)).

4.2.2 Parallel ADMM for Bi-level Selection across Features

The ADMM can be well suited to distributed convex optimization (Boyd et al. (2011)). We develop a parallel ADMM across features to conduct a penalized bi-level selection with binary outcome, so we can distribute different batches of data to different machines to improve computation speed. The ADMM is constructed through a iteratively reweighed least square algorithm for binary outcome, which yields a same form as the quadratic approximation to the penalized objective function based on Taylor expansion about current estimates. Denote $\alpha^{(k)}$, $\beta^{(k)}$, as the estimates of regression coefficients at kth iteration . Let $X = (X_1, \cdots X_p) \in \mathbb{R}^{n\times p(q+1)}$ represent all genes and their corresponding G×E interactions. Then the quadratic approximation to (4.3)

$$R(\alpha,\beta) \approx -\frac{1}{2n}(\tilde{Y} - E\alpha - X\beta)^\top W(\tilde{Y} - E\alpha - X\beta) + \sum_{j=1}^{p}\lambda_1\|\beta_j\| + \sum_{j=1}^{p}\sum_{h=0}^{q}\lambda_2|\beta_{jh}|, \quad (4.4)$$

where $\alpha = (\alpha_1, \cdots, \alpha_q)^\top \in \mathbb{R}^q$ is the regression coefficients for the environmental factors, and $\beta = (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^{p(q+1)}$ is the regression coefficients for all genetic main effects and their G×E interactions. $W_{n\times n}$ is a diagonal matrix of weights with elements $w_i = \pi_i(1-\pi_i)$, which is evaluated at $\beta^{(k)}$. Define $\tilde{Y}$ as the working response, where $\tilde{Y} = E\alpha^{(k)} + X\beta^{(k)} + W^{-1}(Y - \pi)$.

We develop an iterative algorithm to update parameters. Given the current estimate parameter $\beta^{(k)}$, we minimize the weighted least square function with respect to α. Then compute $\alpha^{(k+1)} = (E^\top E)^{-1}E^{-1}W(\tilde{Y} - X\beta^{(k)})$. Next, fix α at the current estimate $\alpha^{(k+1)}$ in $R(\alpha,\beta)$. To formulate the parallel ADMM framework, we start from expressing equation

(4.4) as a constrained optimization problem with auxiliary variable $z = (z_1, \cdots, z_B)$.

$$R(\alpha^{(k+1)}, \beta, z) \approx -\frac{1}{2n}\left(\tilde{Y} - E\alpha^{(k+1)} - \sum_{b=1}^{B} X_b\beta_b\right)^\top W \left(\tilde{Y} - E\alpha^{(k+1)} - \sum_{b=1}^{B} X_b\beta_b\right) + \sum_{b=1}^{B}\sum_{j=1}^{p/B}\left\{\lambda_1\|z_{bj}\| + \sum_{h=0}^{q}\lambda_2|z_{bjh}|\right\} \text{ subject to } X_b\beta_b - z_b = 0, \quad (4.5)$$

where we split the data columns into B groups with $X_b \in \mathbb{R}^{n\times p(q+1)/B}$ and $\beta_b \in \mathbb{R}^{p(q+1)/B}$, for $b = 1, \cdots, B$. Therefore, the bi-level penalty function is partitioned as B groups, correspondingly. As each partition of data should be on similar scale, we search two dimensional grid of (λ_1, λ_2) to find optimal pair of tuning parameters through V-fold cross validation.

Equation (4.5) is equivalent to the following augmented form with $\rho_1, \rho_2 (> 0)$ being the augmentation parameters

$$R(\alpha^{(k+1)}, \beta, z, \tau) \approx -\frac{1}{2n}\left(\tilde{Y} - E\alpha^{(k+1)} - \sum_{b=1}^{B} X_b\beta_b\right)^\top W \left(\tilde{Y} - E\alpha^{(k+1)} - \sum_{b=1}^{B} X_b\beta_b\right) + \sum_{b=1}^{B}\sum_{j=1}^{p/B}\left\{\lambda_1\|z_{bj}\| + \tau_{bj}^\top(\beta_{bj} - z_{bj}) + \frac{\rho_1}{2}(\beta_{bj} - z_{bj})^2 + \sum_{h=0}^{q}\left[\lambda_2|z_{bjh}| + \tau_{bjh}(\beta_{bjh} - z_{bjh}) + \frac{\rho_2}{2}(\beta_{bjh} - z_{bjh})^2\right]\right\}, \quad (4.6)$$

where β, z are the primal variables and τ is the dual variable. For simplicity, we impose a scaled form ($u = \frac{1}{\rho}\tau$) to simplify equation (4.6)

$$R(\alpha^{(k+1)}, \beta, z, u) \approx -\frac{1}{2n}\left(\tilde{Y} - E\alpha^{(k+1)} - \sum_{b=1}^{B} X_b\beta_b\right)^\top W \left(\tilde{Y} - E\alpha^{(k+1)} - \sum_{b=1}^{B} X_b\beta_b\right) + \sum_{b=1}^{B}\sum_{j=1}^{p/B}\left\{\lambda_1\|z_{bj}\| + \frac{\rho_1}{2}(\beta_{bj} - z_{bj} + u_{bj})^2 - \frac{\rho_1}{2}u_{bj}^2 + \sum_{h=0}^{q}\left[\lambda_2|z_{bjh}| + \frac{\rho_2}{2}(\beta_{bjh} - z_{bjh} + u_{bjh})^2 - \frac{\rho_2}{2}u_{bjh}^2\right]\right\}.$$

For each batch of data, the estimation can be treated as partial prediction of Y with the features of β_b. According to the discussion in Boyd et al. (2011) , the parallel ADMM solutions at $(k+1)$th iteration can be derived as

$$\beta_b^{(k+1)} = \arg\min_{\beta_b} \left\{ P(\beta_b; \lambda_1, \lambda_2) + \frac{\rho}{2} \left\| X_b\beta_b - X_i\beta_b^{(k)} - \bar{z}^{(k)} + \overline{X\beta}^{(k)} + u_b^{(k)} \right\|^2 \right\} \tag{4.7}$$

$$\bar{z}^{(k+1)} = \arg\min_{\bar{z}} \left\{ -\frac{1}{2n} \left(\tilde{Y} - E\alpha^{(k+1)} - Nz \right)^\top W \left(\tilde{Y} - E\alpha^{(k+1)} - Nz \right) + \frac{N\rho}{2} \|\bar{z} - \overline{X\beta}^{(k+1)} - \bar{u}^{(k)}\|^2 \right\} \tag{4.8}$$

$$u^{(k+1)} = u^{(k)} + \left(\overline{X\beta}^{(k+1)} - \bar{z}^{(k+1)} \right) \tag{4.9}$$

With the parallel ADMM framework, all the partial predictions $X_b\beta_b^{(k+1)}$ are collected to form average prediction $\overline{X\beta}^{(k+1)}$ from (C.15) to (C.16). Then we update the average $\bar{z}^{k+1}$ by minimizing the quadratic function. The dual variable update is based on (C.17). Each update of β_b is a regularized problem with $p(q+1)/B$ variables, which can be solved in a parallel fashion.

With fixed tuning parameters, the proposed parallel ADMM algorithm proceeds as follows (**Algorithm 4**) .

4.3 Simulation

We compare the performance of the bi-level selection (SGLASSO) to two alternatives. LASSO is an individual level penalty without considering the grouping structure in G×E. Group LASSO (GLASSO) is a group level penalty without achieving sparsity within groups, that is, once a gene is selected, its main effect and all interactions are selected. To evaluate the computation speed, we evaluate the performance of all three methods through parallel and non-parallel ADMM frameworks. Denote PLASSO, PGLASSO and PSGLASSO as parallel approaches for LASSO, GLASSO and SGLASSO, respectively.

To set up different simulation scenarios, we generate datasets with different correlation

Algorithm 4 Parallel ADMM for BI-level Selection across Features

Initialize $\beta^{(1)}, z^{(1)}, u^{(1)}$, $k = 1$.

repeat

Compute $\alpha^{(k+1)} = (E^\top E)^{-1} E^{-1} W(\tilde{Y} - X\beta^{(k)})$.

for $b = 1, \cdots, B$ **do**

$$\beta_b^{(k+1)} = \arg\min_{\beta_b} \left\{ P(\beta_b; \lambda_1, \lambda_2) + \frac{\rho}{2} \left\| X_b\beta_b - X_i\beta_b^{(k)} - \bar{z}^{(k)} + \overline{X\beta}^{(k)} + u_b^{(k)} \right\|^2 \right\}$$

end for

$$\bar{z}^{(k+1)} = \arg\min_{\bar{z}} \left\{ -\frac{1}{2n} \left(\tilde{Y} - E\alpha^{(k+1)} - Nz \right)^\top W \left(\tilde{Y} - E\alpha^{(k+1)} - Nz \right) + \frac{N\rho}{2} \| \bar{z} - \overline{X\beta}^{(k+1)} - \bar{u}^{(k)} \|^2 \right\}$$

$$u^{(k+1)} = u^{(k)} + \left(\overline{X\beta}^{(k+1)} - \bar{z}^{(k+1)} \right)$$

$k = k + 1$

until convergence

structures and correlation levels, each with n subjects, p genes and q environmental factors, respectively. For each subject, we simulate genetic matrix through $G_{n\times p} \sim N(0, \Sigma)$, where we consider the following four covariance structures for Σ:

1. (AR1-3) Autoregressive correlation structure with correlation coefficients $0.3^{|i-j|}$ for the ith and jth variables.

2. (AR1-5) Autoregressive correlation structure with correlation coefficients $0.5^{|i-j|}$ for the ith and jth variables.

3. (banded 1) Banded correlation structure, in which variable i and j have correlation coefficients $\rho = 0.11$ if $|i - j| = 1$ and $\rho = 0$ otherwise.

4. (banded 2) Banded correlation structure, in which variable i and j have correlation coefficients $\rho = 0.2$ if $|i - j| = 1$, 0.11 if $|i - j| = 2$ and 0 otherwise.

The environmental factors are from $E_{n\times q} \sim N(0, \Sigma_E)$, where we choose Σ_E to be autoregressive with correlation coefficient 0.9. To assign nonzero coefficients of all genetic main effects and interactions, we randomly select 10 groups. For each group, we randomly select

2 elements with coefficients generated from Unif[0.5, 1]. The nonzero coefficients of environment factors are generated from Unif[0,1]. Follow equation 4.1, the binary response can subsequently be simulated.

Simulations are evaluated under three different settings: we choose $(n, p, q) = (500, 100, 4)$, $(500, 200, 4)$, $(800, 100, 4)$, respectively. Evaluation of feature selection accuracy is based on true positive (TP) and false positive (FP). Computation time is reported in seconds. The simulation is repeated 100 times and conducted on a PC with Core i7 4-core processor and 8GB RAM.

The simulation results of the six approaches are tabulated in Table 4.1, Table C.1 and Table C.2. In general, the bi-level selection has better identification than alternatives. For example, in Table 4.1, given $n = 500$, $p = 100$, $q = 4$, under covariance structure (1). LASSO identifies 15.96 (SD 1.29) TPs with 10.18 (SD 3.34) FPs. GLASSO identifies a larger number of TPs, 18.52 (SD 1.26), with a larger number of FPs, 28.06 (SD 3.77). However, the SGLASSO can identify larger number of TPs, 18.20 (SD 1.40), with a smaller number of FPs, 6.30 (SD 2.92). Among three approaches, the difference in identification performance comes from G×E interactions. SGLASSO can accommodate the group level selection and individual level selection simultaneously. We can observe the similar patterns for the other covariance structures in Table 4.1. As the dimension decreases (Table C.1) or the sample size increases (Table C.2), the identification performance can be improved for all approaches. However, we can observe that overall SGLASSO outperforms alternatives with a higher TP and a lower FP under each setting.

In terms of computation efficiency, we compare parallel ADMM and traditional ADMM frameworks. In a word, the parallel ADMM can significantly improve the computation efficiency and the proposed PSGLASSO has better identification performance than alternatives. For example, in Table 4.1, under covariance structure (AR1-3). For computation, both PLASSO and PSGLASSO are almost five times faster than LASSO and SGLASSO. PGLASSO is almost three times faster than GLASSO. For identification, PLASSO identifies 14.02 (SD 1.50) TPs, which is less than the 15.96 (SD 1.29) TPs identified by LASSO, while similar FPs are identified by both approaches. PGLASSO identifies 31.22 (SD 4.64) FPs,

Table 4.1: Binary Response: Comparison between ADMM and parallel ADMM in splitting features for $(n, p, q) = (500, 200, 4)$ with LASSO, GLASSO and SGLASSO penalties. Numbers are mean (standard deviation).

Approach	Correlation	TP	FP	Time
LASSO	AR1-3	15.96(1.29)	10.18(3.34)	11.24(0.12)
	AR1-5	16.10(1.60)	6.64(2.31)	14.72(0.16)
	banded 1	14.84(1.49)	14.82(4.06)	10.44(0.12)
	banded 2	15.52(1.48)	10.74(3.23)	10.86(0.12)
GLASSO	AR1-3	18.52(1.26)	28.06(3.74)	12.06(0.13)
	AR1-5	18.02(1.85)	26.40(3.58)	15.82(0.24)
	banded 1	17.12(1.89)	27.98(5.17)	10.61(0.11)
	banded 2	18.24(1.55)	26.94(2.91)	11.09(0.10)
SGLASSO	AR1-3	18.20(1.40)	6.30(2.92)	23.99(0.24)
	AR1-5	18.00(1.60)	3.98(1.96)	33.01(0.57)
	banded 1	16.74(1.70)	7.70(3.50)	20.81(0.14)
	banded 2	17.88(1.63)	6.32(2.42)	22.44(0.20)
Approach	Correlation	TP	FP	Time
PLASSO	AR1-3	14.02(1.50)	10.16(3.09)	2.25(0.06)
	AR1-5	14.22(1.56)	9.00(3.01)	2.45(0.08)
	banded 1	13.26(1.54)	11.06(3.18)	2.31(0.07)
	banded 2	13.78(1.52)	9.34(3.00)	2.20(0.05)
PGLASSO	AR1-3	18.60(1.29)	31.22(4.64)	4.09(0.10)
	AR1-5	18.08(1.80)	28.78(4.04)	4.16(0.09)
	banded 1	17.36(1.87)	31.20(6.28)	4.19(0.09)
	banded 2	18.48(1.31)	30.22(4.51)	3.75(0.08)
PSGLASSO	AR1-3	18.06(1.49)	6.34(2.98)	4.66(0.10)
	AR1-5	17.84(1.68)	4.04(2.09)	4.79(0.10)
	banded 1	16.68(1.77)	7.66(3.54)	4.64(0.08)
	banded 2	17.86(1.65)	6.24(2.17)	4.49(0.09)

which is more than the FPs identified by GLASSO, 28.06 (SD 3.74), while similar TPs are identified by both approaches. The proposed PSGLASSO identifies 18.06 (SD 1.49) TPs with 6.34 (SD 2.98) FPs, which is similar to SGLASSO identification results. Hence, the PSGLASSO is more stable in the identification of G×E interactions.

4.4 Real Data Analysis

We applied our model to the Nurse Health Study (NHS), a nested case control cohort study of type 2 diabetes (T2D). This data set is from the Gene, Environment Association Studies

Consortium (GENVEA) (Cornelis et al. (2010)). As a Genome Wide Association Study (GWAS), the NHS aims at examining multiple hypotheses on the lifestyle and dietary factors leading to complex diseases including type 2 diabetes. More details of the study can sbe found from Colditz and Hankinson (2005) and Rimm et al. (1991). In this section, we focus on studying SNPs from chromosome 10 and compare identifications of G×E interactions between proposed methods and alternatives.

After matching phenotypes and genotypes, the dataset contains 3224 subjects and 17037 SNPs. Conducting screenings to reduce feature dimensionality and improve stability is very common in many studies. For example, Li et al. (2015) applied single SNP analysis before his downstream analysis to filter important SNPs in the genome-wide association study. In this study, we apply a marginal logistic regression with binary response (1 indicating the case of disease and 0 otherwise) to evaluate every single SNP significance by using genetic main and G×E interaction effects as predictors. Then we select top 200 most significant SNPs. The remaining environment factors are pre-selected. With evidences of being associated with T2D from previous studies (Hu et al. (2001)), we select four environment factors: the total physical activity (ACT), glycemic load (GL), age (AGE) and alcohol intake (ALCOHOL).

We analyze data using SGLASSO and LASSO through non-parallel ADMM. In addition, we apply PSGL and PLASSO to compare identifications and computation efficiency. As shown in Table 4.4 shows, the proposed SGL identifies 5 genetic main effects and 25 total G×E interactions. From Table 4.4, we also observe that not all genes have main effects. Some genes are identified because some of their corresponding G×E interactions are associated to T2D. Table C.4 shows the result from approach LASSO, which it identifies 4 genetic main effects and 28 total G×E interactions. The results of approaches PSGLASSO and PLASSO are provided in Table C.3 and Table C.5, respectively. The parallel ADMM framework can generate similar identification results as traditional ADMM framework, but it sufficiently improved computation time. The PSGL identifies 6 genetic main effects and total 25 G×E interactions. The PLASSO identifies 8 genetic main effects and total 28 G×E interactions. We use 5-fold cross validation to select tuning parameters. The computation time for 4 approaches are 28.41(SGLASSO), 16.30(PSGLASSO), 35.60(LASSO) and 12.19 (PLASSO)

seconds, respectively.

To further compare the identifications by different approaches, we provide Table 4.2 and Table 4.3 to show the numbers of identifications and overlaps as well as the RV coefficients. The RV coefficient measures the similarity of overlapping information, with a larger value indicating a higher similarity. As shown in Table 4.2 and Table 4.3 , the proposed approach identifies more similar number of G×E interactions between traditional ADMM and parallel ADMM frameworks. All approaches identify similar genetic main effects. In addition, SGLASSO and PSGLASSO can identify similar G×E interactions as LASSO. For parallel ADMM framework, we improve computation time by distributing different batches of data to different machines. Without loosing satisfactory identifications, the proposed PSGL is more stable than PLASSO to identify G×E interactions.

Table 4.2: Analysis of NHS T2D: numbers of main effects identified by different approaches and their overlaps. RV coefficients are in the parentheses.

Approach	LASSO	PLASSO	SGL	PSGL
LASSO	4	4(0.99)	3(0.99)	3(0.99)
PLASSO		8	4(0.99)	4(0.96)
SGL			5	4(0.98)
PSGL				6

Table 4.3: Analysis of NHS T2D: numbers of total G×E interactions identified by different approaches and their overlaps. RV coefficients are in the parentheses.

Approach	LASSO	PLASSO	SGL	PSGL
LASSO	28	18(0.65)	20(0.7)	19(0.52)
PLASSO		28	13(0.76)	14(0.67)
SGL			25	15(0.72)
PSGL				25

Table 4.4: Analysis of NHS T2D: G×E interaction identifications from SGL. Numbers are estimated regression coefficients for genetic main effect and G×E interactions.

Gene Name	SNP ID	Main	ACT	GL	AGE	ALCOHOL
CAMK1D	rs12763487			-0.0866		
PRPF38AP1	rs1538511			-0.1180		
RP11-195B3.1	rs7070200				0.0573	-0.0725
UPF2	rs11257429			0.0144		
VIM	rs359296			-0.1335		
PHYH	rs1556718		-0.0594		0.0753	
RP5-1119O21.2	rs1538246			-0.0876		
AKR1C2	rs10904384				0.0100	
RP11-556E13.1	rs1919738			-0.1115		
REEP3	rs10822184	-0.0898				
SVIL	rs1247093	0.0843		-0.1031		
RP11-135D11.2	rs12354667					-0.1079
8-Mar	rs7081687					-0.0796
LUZP4P1	rs12262659	0.0564				
RP11-543F8.2	rs2152963					-0.0651
SNRPEP8	rs2891427		0.0041			
PRKCQ	rs650652	-0.1085		-0.0707		
MIR4675	rs7922148			-0.0808		
AL512640.1	rs1545844		-0.0720			-0.0910
DNAJC1	rs2666775	0.0729		-0.0851		
RP11-445P17.5	rs10904468			-0.1079		
RP11-490O24.2	rs12252306				0.0948	
PRKG1	rs10995831			-0.0930		
ASAH2	rs7908522			-0.1234		

Chapter 5

Summary

Gene-environment (G×E) interaction is critical for understanding the genetic basis of complex disease beyond genetic and environment main effects. This dissertation focuses on developing penalized variable selection methods to conduct efficient variable selection of G×E interactions.

The multidimensional measurements are gaining significant popularity in cancer studies. In Chapter 2, we conduct an integrative G×E interaction analysis for multidimensional omics data based on a two-step variable selection model. Specifically, at the first step, the sparse regulatory relationship between gene expression (GE) and its regulators have been pinpointed via penalization. At the second step, the G×E prognostic model consists of: (1) environmental factors; (2) regulated GEs and their interactions with environmental factors; (3) residual effects of GEs and their interactions with environmental factors; (4) residual effects of regulators. Our method distinguishes from most of the published studies, which takes the advantage of the two-step model to integrate multi-omics measurements in G×E studies. The paper associated with this study is under review.

With the development of the penalized variable selection approaches, efficient computation algorithms play a critical role in G×E studies. In Chapter 3, we propose an alternative computation framework by adopting the alternative direction method of multipliers (ADMM). Compared to coordinate descent (CD) algorithm, the ADMM can conduct the

optimization in parallel. To accommodate the large scale data in terms of either samples or features, we develop two novel ADMM based variable selection methods across samples or features in G×E studies. Simulation studies demonstrate that the proposed methods significantly improve the computation speed with satisfactory identification and prediction performance.

In the last chapter of the dissertation, we utilize the proposed parallel ADMM based variable selection for G×E interactions in the case-control study of type 2 diabetes. Our method significantly distinguishes from others: (1) we adopt the logistic regression model for the binary response; (2) the bi-level selection is considered to be the tailored penalization method for the G×E interaction structure; (3) a parallel ADMM framework for bi-level selection across features is developed to conduct the optimization by using divide-and-conquer strategy while achieving good efficiency and accuracy. The proposed method can be further implemented on distributed computing platform such as Hadoop.

Bibliography

Shashi Anand, Mohammad Aslam Khan, Moh' Khushman, Santanu Dasgupta, Seema Singh, Ajay Pratap Singh, et al. Comprehensive analysis of expression, clinicopathological association and potential prognostic significance of rabs in pancreatic cancer. *International journal of molecular sciences*, 21(15):5580, 2020.

Jeong Mo Bae, Xianyu Wen, Tae-Shin Kim, Yoonjin Kwak, Nam-Yun Cho, Hye Seung Lee, and Gyeong Hoon Kang. Fibroblast growth factor receptor 1 (fgfr1) amplification detected by droplet digital polymerase chain reaction (ddpcr) is a prognostic factor in colorectal cancers. *Cancer research and treatment: official journal of Korean Cancer Association*, 52(1):74, 2020.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111, 2013.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Yichuan Chen, Jingqun Tang, Ting Lu, and Fang Liu. Capn1 promotes malignant behavior and erlotinib resistance mediated by phosphorylation of c-met and pik3r2 via degrading ptpn1 in lung adenocarcinoma. *Thoracic Cancer*, 2020.

Giovanni Ciriello, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, 22(2):398–406, 2012.

Graham A Colditz and Susan E Hankinson. The nurses' health study: lifestyle and health among women. *Nature Reviews Cancer*, 5(5):388–396, 2005.

Marilyn C Cornelis, Arpana Agrawal, John W Cole, Nadia N Hansel, Kathleen C Barnes, Terri H Beaty, Siiri N Bennett, Laura J Bierut, Eric Boerwinkle, Kimberly F Doheny, et al. The gene, environment association studies consortium (geneva): maximizing the knowledge obtained from gwas by collaboration across studies of multiple conditions. *Genetic epidemiology*, 34(4):364–372, 2010.

Sébastien Couraud, Gérard Zalcman, Bernard Milleron, Franck Morin, and Pierre-Jean Souquet. Lung cancer in never smokers–a review. *European journal of cancer*, 48(9):1299–1311, 2012.

Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

Astrid Dempfle, André Scherag, Rebecca Hein, Lars Beckmann, Jenny Chang-Claude, and Helmut Schäfer. Gene–environment interactions for complex traits: definitions, methodological requirements and challenges. *European Journal of Human Genetics*, 16(10):1164–1172, 2008.

Yinhao Du, Kun Fan, Xi Lu, and Cen Wu. Integrating multi–omics data for gene-environment interactions. *BioTech*, 10(1):3, 2021.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & mathematics with applications*, 2(1):17–40, 1976.

Roland Glowinski and A Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 9(R2):41–76, 1975.

Peter J Green. On use of the em algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(3):443–452, 1990.

Samuel M Gross and Robert Tibshirani. Collaborative regression. *Biostatistics*, 16(2):326–338, 2014.

Wei Guo, Sijin Sun, Lei Guo, Peng Song, Xuemin Xue, Hao Zhang, Guochao Zhang, Renda Li, Yibo Gao, Bin Qiu, et al. Elevated slc2a1 expression correlates with poor prognosis in patients with surgically resected lung adenocarcinoma: A study based on immunohistochemical analysis and bioinformatics. *DNA and Cell Biology*, 39(4):631–644, 2020.

Yasir Hameed and Samina Ejaz. Up-regulation of fn1, activation of maturation promoting factor and associated signaling pathway facilitates epithelial-mesenchymal transition, inhibits apoptosis and elevates proliferation rate of breast cancer cells. *Silico Analysis of Microarray Datasets*, 2020.

Trevor Hastie. Fast regularization paths via coordinate descent. In *The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Denver*, volume 2009, 2008.

Joel N Hirschhorn, Kirk Lohmueller, Edward Byrne, and Kurt Hirschhorn. A comprehensive review of genetic association studies. *Genetics in medicine*, 4(2):45–61, 2002.

Frank B Hu, JoAnn E Manson, Meir J Stampfer, Graham Colditz, Simin Liu, Caren G

Solomon, and Walter C Willett. Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *New England journal of medicine*, 345(11):790–797, 2001.

Jing Hu, Lutong Xu, Tao Shou, and Qiang Chen. Systematic analysis identifies three-lncrna signature as a potentially prognostic biomarker for lung squamous cell carcinoma using bioinformatics strategy. *Translational Lung Cancer Research*, 8(5):614, 2019.

Hai-Hui Huang, Jing-Guo Dai, and Yong Liang. Clinical drug response prediction by using a lq penalized network-constrained logistic regression method. *Cellular Physiology and Biochemistry*, 51(5):2073–2084, 2018a.

Jian Huang, Shuangge Ma, Hongzhe Li, and Cun-Hui Zhang. The sparse laplacian shrinkage estimator for high-dimensional regression. *Annals of statistics*, 39(4):2021, 2011.

Jian Huang, Patrick Breheny, and Shuangge Ma. A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4), 2012.

Ningyu Huang, Wenbo Lin, Xiuyu Shi, and Tao Tao. Stk24 expression is modulated by dna copy number/methylation in lung adenocarcinoma and predicts poor survival. *Future Oncology*, 14(22):2253–2263, 2018b.

Yu Jiang, Yuan Huang, Yinhao Du, Yinjun Zhao, Jie Ren, Shuangge Ma, and Cen Wu. Identification of prognostic genes and pathways in lung adenocarcinoma using a bayesian approach. *Cancer Informatics*, 1(7), 2017.

Lv Jin, Xiao-Yu Zuo, Wei-Yang Su, Xiao-Lei Zhao, Man-Qiong Yuan, Li-Zhen Han, Xiang Zhao, Ye-Da Chen, and Shao-Qi Rao. Pathway-based analysis tools for complex diseases: a review. *Genomics, proteomics & bioinformatics*, 12(5):210–220, 2014.

Di Ke, Qiang Guo, Teng-Yang Fan, and Xue Xiao. Analysis of the role and regulation mechanism of hsa-mir-147b in lung squamous cell carcinoma based on the cancer genome atlas database. *Cancer biotherapy & radiopharmaceuticals*, 2020.

S Sathiya Keerthi and Shirish Shevade. A fast tracking algorithm for generalized lars/lasso. *IEEE Transactions on Neural Networks*, 18(6):1826–1830, 2007.

Stacey A Kenfield, Esther K Wei, Meir J Stampfer, Bernard A Rosner, and Graham A Colditz. Comparison of aspects of smoking among the four histological types of lung cancer. *Tobacco control*, 17(3):198–204, 2008.

Kipoong Kim and Hokeun Sun. Incorporating genetic networks into case-control association studies with high-dimensional dna methylation data. *BMC bioinformatics*, 20(1):510, 2019.

Vessela N Kristensen, Ole Christian Lingjærde, Hege G Russnes, Hans Kristian M Vollan, Arnoldo Frigessi, and Anne-Lise Børresen-Dale. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299, 2014.

Vinay Kumar, Abul K Abbas, and Jon C Aster. *Robbins basic pathology e-book*. Elsevier Health Sciences, 2017.

Mihee Lee, Haipeng Shen, Jianhua Z Huang, and JS Marron. Biclustering via sparse singular value decomposition. *Biometrics*, 66(4):1087–1095, 2010.

Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.

Jiahan Li, Zhong Wang, Runze Li, and Rongling Wu. Bayesian group lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *The annals of applied statistics*, 9(2):640, 2015.

Ting Lin, Jingxian Gu, Kai Qu, Xing Zhang, Xiaohua Ma, Runchen Miao, Xiaohong Xiang, Yunong Fu, Wenquan Niu, Junjun She, et al. A new risk score based on twelve hepatocellular carcinoma-specific gene expression can predict the patients' prognosis. *Aging (Albany NY)*, 10(9):2480, 2018.

Marc P Lussier, Pascale K Lepage, Simon M Bousquet, and Guylain Boulay. Rnf24, a new

trpc interacting protein, causes the intracellular retention of trpc. *Cell calcium*, 43(5): 432–443, 2008.

Shuangge Ma and Jian Huang. Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics*, 9(5):392–403, 2008.

Shujie Ma and Shizhong Xu. Semiparametric nonlinear regression for detecting gene and environment interactions. *Journal of Statistical Planning and Inference*, 156:31–47, 2015.

Shunsuke Misono, Naohiko Seki, Keiko Mizuno, Yasutaka Yamada, Akifumi Uchida, Hiroki Sanada, Shogo Moriya, Naoko Kikkawa, Tomohiro Kumamoto, Takayuki Suetsugu, et al. Molecular pathogenesis of gene regulation by the mir-150 duplex: mir-150-3p regulates tns4 in lung adenocarcinoma. *Cancers*, 11(5):601, 2019.

Zhimin Peng, Ming Yan, and Wotao Yin. Parallel and distributed sparse optimization. In *2013 Asilomar conference on signals, systems and computers*, pages 659–646. IEEE, 2013.

Celia María Pombo, Thomas Force, John Kyriakis, Emilio Nogueira, Miguel Fidalgo, and Juan Zalvide. The gck ii and iii subfamilies of the ste20 group kinases. *Front Biosci*, 12 (3):850–859, 2007.

Valeria Relli, Marco Trerotola, Emanuela Guerra, and Saverio Alberti. Abandoning the notion of non-small cell lung cancer. *Trends in molecular medicine*, 25(7):585–594, 2019.

Jie Ren, Tao He, Ye Li, Sai Liu, Yinhao Du, Yu Jiang, and Cen Wu. Network-based regularization for high dimensional snp data in the case–control study of type 2 diabetes. *BMC genetics*, 18(1):44, 2017.

Jie Ren, Yinhao Du, Shaoyu Li, Shuangge Ma, Yu Jiang, and Cen Wu. Robust network-based regularization and variable selection for high-dimensional genomic data in cancer prognosis. *Genetic epidemiology*, 43(3):276–291, 2019.

Jie Ren, Fei Zhou, Xiaoxi Li, Qi Chen, Hongmei Zhang, Shuangge Ma, Yu Jiang, and

Cen Wu. Semiparametric bayesian variable selection for gene-environment interactions. *Statistics in Medicine*, 39(5):617–638, 2020.

Jie Ren, Fei Zhou, Xiaoxi Li, Shuangge Ma, Yu Jiang, and Cen Wu. Robust bayesian variable selection for gene-environment interactions. *Biometrics*, 2021.

Eric B Rimm, Edward L Giovannucci, Walter C Willett, Graham A Colditz, Alberto Ascherio, Bernard Rosner, and Meir J Stampfer. Prospective study of alcohol consumption and risk of coronary disease in men. *The Lancet*, 338(8765):464–468, 1991.

Victor Menezes Silva, Jessica Alves Gomes, Liliane Patrícia Gonçalves Tenório, Genilda Castro de Omena Neta, Karen da Costa Paixão, Ana Kelly Fernandes Duarte, Gabriel Cerqueira Braz da Silva, Ricardo Jansen Santos Ferreira, Bruna Del Vechio Koike, Carolinne de Sales Marques, et al. Schwann cell reprogramming and lung cancer progression: a meta-analysis of transcriptome data. *Oncotarget*, 10(68):7288, 2019.

Naoko I Simonds, Armen A Ghazarian, Camilla B Pimentel, Sheri D Schully, Gary L Ellison, Elizabeth M Gillanders, and Leah E Mechanic. Review of the gene-environment interaction literature in cancer: what do we know? *Genetic epidemiology*, 40(5):356–365, 2016.

Winfried Stute and J-L Wang. The strong law under random censorship. *The Annals of statistics*, pages 1591–1607, 1993.

Janakiraman Subramanian and Ramaswamy Govindan. Lung cancer in never smokers: a review. *Journal of clinical oncology*, 25(5):561–570, 2007.

Hokeun Sun and Shuang Wang. Penalized logistic regression for high-dimensional dna methylation data with case-control studies. *Bioinformatics*, 28(10):1368–1375, 2012.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity

and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Lily Wang, Peilin Jia, Russell D Wolfinger, Xi Chen, and Zhongming Zhao. Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics*, 98(1):1–8, 2011.

Wenting Wang, Veerabhadran Baladandayuthapani, Jeffrey S Morris, Bradley M Broom, Ganiraju Manyam, and Kim-Anh Do. ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159, 2013.

Xing Wei Wang, Wei Wei, Wei Qiang Wang, Xiao Yan Zhao, Hong Guo, and Dian Chun Fang. Ring finger proteins are involved in the progression of barrett esophagus to esophageal adenocarcinoma: a preliminary study. *Gut and liver*, 8(5):487, 2014.

Yihan Wang, Jingyu Zhang, Xingjun Xiao, Hongbo Liu, Fang Wang, Song Li, Yanhua Wen, Yanjun Wei, Jianzhong Su, Yunming Zhang, et al. The identification of age-associated cancer markers by an integrative analysis of dynamic dna methylation changes. *Scientific reports*, 6:22722, 2016.

Cen Wu and Yuehua Cui. A novel method for identifying nonlinear gene–environment interactions in case–control association studies. *Human genetics*, 132(12):1413–1425, 2013.

Cen Wu and Yuehua Cui. Boosting signals in gene-based association studies via efficient snp selection. *Briefings in bioinformatics*, 15(2):279–291, 2014.

Cen Wu and Shuangge Ma. A selective review of robust variable selection with applications in bioinformatics. *Briefings in bioinformatics*, 16(5):873–883, 2014.

Cen Wu, Shaoyu Li, and Yuehua Cui. Genetic association studies: an information content perspective. *Current genomics*, 13(7):566–573, 2012.

Cen Wu, Yuehua Cui, and Shuangge Ma. Integrative analysis of gene–environment interactions under a multi-response partially linear varying coefficient model. *Statistics in medicine*, 33(28):4988–4998, 2014.

Cen Wu, Xingjie Shi, Yuehua Cui, and Shuangge Ma. A penalized robust semiparametric approach for gene–environment interactions. *Statistics in medicine*, 34(30):4016–4030, 2015.

Cen Wu, Yu Jiang, Jie Ren, Yuehua Cui, and Shuangge Ma. Dissecting gene-environment interactions: A penalized robust approach accounting for hierarchical structures. *Statistics in medicine*, 37(3):437–456, 2018a.

Cen Wu, Qingzhao Zhang, Yu Jiang, and Shuangge Ma. Robust network-based analysis of the associations between (epi) genetic measurements. *Journal of multivariate analysis*, 168:119–130, 2018b.

Cen Wu, Ping-Shou Zhong, and Yuehua Cui. Additive varying-coefficient model for nonlinear gene-environment interactions. *Statistical applications in genetics and molecular biology*, 17(2), 2018c.

Cen Wu, Fei Zhou, Jie Ren, Xiaoxi Li, Yu Jiang, and Shuangge Ma. A selective review of multi-level omics data integration using variable selection. *High-throughput*, 8(1):4, 2019.

Tong Tong Wu, Kenneth Lange, et al. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008.

Jilin Yang, Yuedong Liu, Xiaoshi Mai, Shun Lu, Li Jin, and Xiang Tai. Stat1-induced up-regulation of linc00467 promotes the proliferation migration of lung adenocarcinoma cells by epigenetically silencing dkk1 to activate wnt/β-catenin signaling pathway. *Biochemical and biophysical research communications*, 514(1):118–126, 2019.

Gui-Bo Ye, Yifei Chen, and Xiaohui Xie. Efficient variable selection in support vector machines via the alternating direction method of multipliers. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 832–840, 2011.

Liqun Yu, Nan Lin, and Lan Wang. A parallel algorithm for large-scale nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics*, 26(4):935–939, 2017.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. *HotCloud*, 10(10-10):95, 2010.

Andleeb Zahra, Itrat Rubab, Sumaira Malik, Amina Khan, Muhammad Jawad Khan, and M Qaiser Fatmi. Meta-analysis of mirnas and their involvement as biomarkers in oral cancers. *BioMed research international*, 2018, 2018.

Lingyao Zeng, Jian Yu, Tao Huang, Huliang Jia, Qiongzhu Dong, Fei He, Weilan Yuan, Lunxiu Qin, Yixue Li, and Lu Xie. Differential combinatorial regulatory network analysis related to venous metastasis of hepatocellular carcinoma. *BMC genomics*, 13(S8):S14, 2012.

Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Qian Zhang, Rui Huang, Hanqing Hu, Lei Yu, Qingchao Tang, Yangbao Tao, Zheng Liu, Jiaying Li, and Guiyu Wang. Integrative analysis of hypoxia-associated signature in pan-cancer. *Iscience*, 23(9):101460, 2020a.

Shusen Zhang, Yuanyuan Lu, Zhongxin Liu, Xiaopeng Li, Zhihua Wang, and Zhigang Cai. Identification six metabolic genes as potential biomarkers for lung adenocarcinoma. *Journal of Computational Biology*, 2020b.

Fei Zhou, Jie Ren, Gengxin Li, Yu Jiang, Xiaoxi Li, Weiqun Wang, and Cen Wu. Penalized variable selection for lipid–environment interactions in a longitudinal lipidomics study. *Genes*, 10(12):1002, 2019.

Fei Zhou, Jie Ren, Xi Lu, Shuangge Ma, and Cen Wu. Gene-environment interaction: A variable selection perspective. *Methods in Molecular Biology. Humana Press*, 2212:191–223, 2021.

Quan Zhou, Shiji Song, Cheng Wu, and Gao Huang. Kernelized lars–lasso for constructing radial basis function neural networks. *Neural Computing and Applications*, 23(7-8):1969–1976, 2013.

Ruoqing Zhu, Qing Zhao, Hongyu Zhao, and Shuangge Ma. Integrating multidimensional omics data for cancer outcome. *Biostatistics*, 17(4):605–618, 2016.

Yunzhang Zhu. An augmented admm algorithm with application to the generalized lasso problem. *Journal of Computational and Graphical Statistics*, 26(1):195–204, 2017.

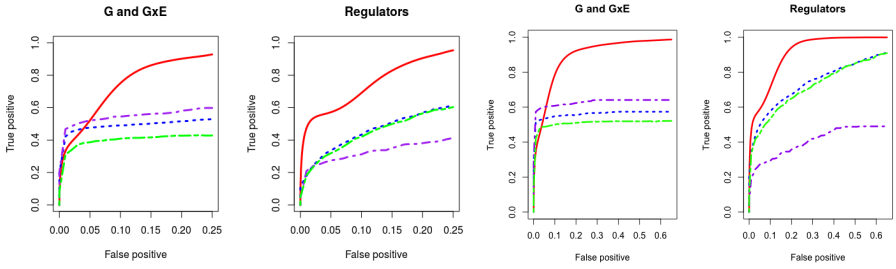
Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

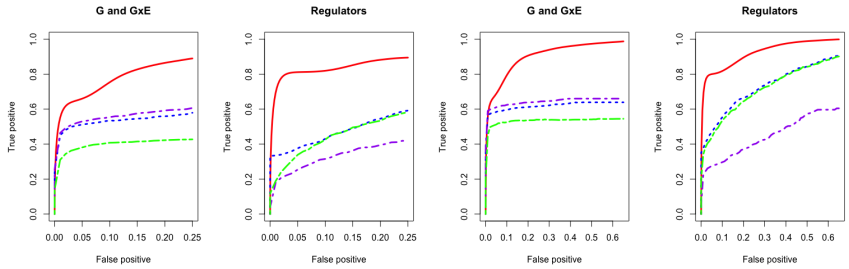
Appendix A

Appendix for Chapter 2

A.1 Other Simulation Results

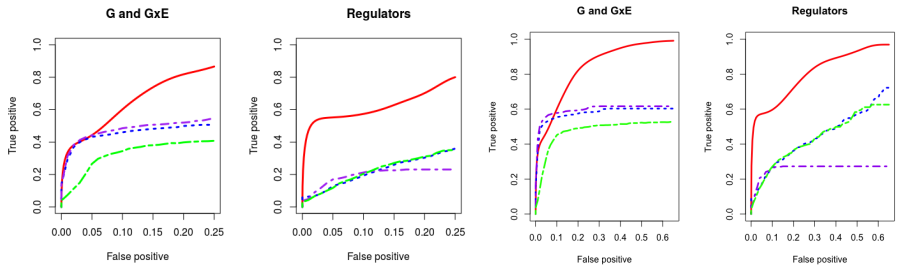


(a) n = 500, weak signal (b) n=1000, weak signal

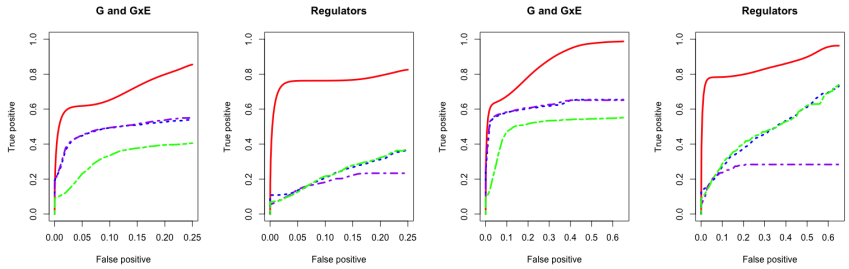


(c) n=500, strong signal (d) n=1000, strong signal

Figure A.1: Four scenarios ROC curves under banded covariance structure. Left two columns are 500 subjects to compare weak and strong signal performance. Right two columns are 1000 subjects to compare weak and strong signal performance. IGE, solid red; S-LASSO, dashed blue; J-LASSO, long dashed purple; ColReg, long dashed green.



(a) n = 500, weak signal (b) n=1000, weak signal



(c) n=500, strong signal (d) n=1000, strong signal

Figure A.2: Four scenarios ROC curves under LUAD covariance structure. Left two columns are 500 subjects to compare weak and strong signal performance. Right two columns are 1000 subjects to compare weak and strong signal performance. IGE, solid red; S-LASSO, dashed blue; J-LASSO, long dashed purple; ColReg, long dashed green.

A.2 AFT Model

Denote T as the logarithm of the failure time and denote C as the logarithm of the censoring time. Under right censoring, we observe $Y = \min(T, C)$, $\delta = I(T \leq C)$. We adopt the Kaplan-Meier weights for censoring. Let $\hat{F}$ be the Kaplan-Meier estimator of the distribution function F of T. According to Stute and Wang (1993), we have $\hat{F}(y) = \sum_{i=1}^{n} w_i I\{Y_{(i)} \leq y\}$, where w_i can be computed as

$$w_1 = \frac{\delta_{(1)}}{n}, w_i = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_j}, \ i = 2, ..., n,$$

where $Y_{(1)} \leq \cdots \leq Y_{(n)}$ are the order statistics of Y_i and $\delta_{(1)}, \cdots, \delta_{(n)}$ are the corresponding censoring indicators. Denote $(E_{(i)}, \ X_{1_{(i)}}, \ X_{2_{(i)}}, \tilde{R}_{(i)})$ as the measurements associated with

$(Y_{(i)}, \delta_{(i)})$, where the notations are from equation (2.9). We center $E_{(i)}$, $X_{1_{(i)}}$, $X_{2_{(i)}}$, $\tilde{R}_{(i)}$, $Y_{(i)}$ using w_i-weighted mean as follows:

$$\bar{E}_w = \sum_{i=1}^{n} w_i E_{(i)} / \sum_{i=1}^{n} w_i, \ \bar{X}_{1_w} = \sum_{i=1}^{n} w_i X_{1_{(i)}} / \sum_{i=1}^{n} w_i, \bar{X}_{2_w} = \sum_{i=1}^{n} w_i X_{2_{(i)}} / \sum_{i=1}^{n} w_i$$

$$\bar{\tilde{R}}_w = \sum_{i=1}^{n} w_i \tilde{R}_{(i)} / \sum_{i=1}^{n} w_i, \ , \bar{Y}_w = \sum_{i=1}^{n} w_i Y_{(i)} / \sum_{i=1}^{n} w_i.$$

Then the centered predictors and responses are $E_{w_{(i)}} = \sqrt{w_i}(E_{(i)} - \bar{E}_w)$, $X_{1_{w_{(i)}}} = \sqrt{w_i}(X_{1_{(i)}} - \bar{X}_{1_w})$, $X_{2_{w_{(i)}}} = \sqrt{w_i}(X_{2_{(i)}} - \bar{X}_{2_w})$, $\tilde{R}_{w_{(i)}} = \sqrt{w_i}(\tilde{R}_{(i)} - \bar{\tilde{R}}_w)$ and $Y_{w_{(i)}} = \sqrt{w_i}(Y_{(i)} - \bar{Y}_w)$. Hence, $Y = (Y_{w(1)}, \cdots, Y_{w(n)})^T$, $E = (E_{w(1)}, \cdots, E_{w(n)})^T$, $X_1 = (X_{1_{w(1)}}, \cdots, X_{1_{w(n)}})^T$, $X_2 = (X_{2_{w(1)}}, \cdots, X_{2_{w(n)}})^T$, and $\tilde{R} = (\tilde{R}_{w(1)}, \cdots, \tilde{R}_{w(n)})^T$.

Appendix B

Appendix for Chapter 3

B.1 ADMM

B.1.1 Lasso

Let's consider following target function with Lasso penalty function

$$R(\beta) = \frac{1}{2n}\|Y - X\beta\|^2 + \lambda|\beta|$$

Let's apply ADMM with scaled form ($u = \frac{1}{\rho}\tau$)

$$R(\beta, z) = \frac{1}{2n}\|Y - X\beta\|^2 + \lambda|z| + \frac{\rho}{2}(\beta - z + u)^2 - \frac{\rho}{2}u^2$$

Hence, the solutions can be achieved through gradient ascent algorithm

$$\frac{\partial R(\beta)}{\beta} = -\frac{1}{n}X^T(Y - X\beta) + \rho(\beta - z + u),\ \hat{\beta} = \Big(\frac{1}{n}X^TX + \rho I\Big)^{-1}\Big(\frac{1}{n}X^TY + \rho(z - u)\Big)$$

$$\frac{\partial R(z)}{z} = \lambda\text{sign}(z) - \rho(\beta - z + u),\ \hat{z} = S^1_{\frac{\lambda}{\rho}}\Big(\beta + u\Big)$$

$$\hat{u}^{k+1} = \hat{u}^k + (\hat{\beta}^{k+1} - \hat{z}^{k+1})$$

B.1.2 Group Lasso

Let's consider following target function with group Lasso penalty function

$$R(\beta) = \frac{1}{2n}\|Y - X\beta\|^2 + \lambda \sum_{j=1}^{q} \|\beta_j\|$$

Let's apply ADMM with scaled form ($u = \frac{1}{\rho}\tau$)

$$R(\beta, z) = \frac{1}{2n}\|Y - X\beta\|^2 + \sum_{j=1}^{q} \Big[\lambda\|z_j\| + \frac{\rho}{2}(\beta_j - z_j + u_j)^2 - \frac{\rho}{2}u_j^2\Big]$$

Note that we can write group wise β_j in a matrix form and the solution is same as Lasso β update. The residual update is same for any penalty function approach. Hence, we only show the derivation of z update. For $j = 1, \cdots, q$,

$$\frac{\partial R(z_j)}{\partial z_j} = \frac{\lambda z_j}{\|z_j\|} - \rho(\beta_j - z_j + u_j),\ \hat{z}_j = S^2_{\frac{\lambda}{\rho}}\Big(\beta_j + u_j\Big)$$

B.2 Parallel ADMM across Features

From the general derivation forms, we can see different penalty functions can only provide different β updates. Hence, we only show the derivation of β update.

B.2.1 Lasso

Let's consider following target function with Lasso penalty

$$R(\beta_i) = \lambda|\beta_i| + + \frac{\rho}{2}\Big(X_i\beta_i - X_i\beta_i^k - \bar{z}^k + \overline{X\beta}^k + u_i^k\Big)^2$$

Then solution can be achieved through first derivative of target function equals to zero

$$\frac{\partial R(\beta_i)}{\partial \beta_i} = \lambda \text{sign}(\beta_i) + \rho X_i^T X_i\beta_i - \rho X_i^T(X_i\beta_i^k + \bar{u}^k - \overline{X\beta} =^k) = 0$$

Let's assume the design matrix group wise columns are orthonormalized, that is, $X_i^T X_i/n = I$. Then the solution is $\hat{\beta}_i = S^1_{\lambda/n\rho}\Big(X_i^T M/n\Big)$, where $M = X_i\beta_i^k - \overline{X\beta}^k + \bar{z}^k - u^k$. For $i = 1, \cdots, N$.

B.2.2 Group Lasso

Let's consider following target function with group Lasso penalty and assume N is the number of groups

$$R(\beta_i) = \lambda\|\beta_i\| + +\frac{\rho}{2}\Big(X_i\beta_i - X_i\beta_i^k - \bar{z}^k + \overline{X\beta}^k + u_i^k\Big)^2$$

Then solution can be achieved through first derivative of target function equals to zero

$$\frac{\partial R(\beta_i)}{\partial \beta_i} = \frac{\lambda\beta_i}{\|\beta_i\|} + \rho X_i^T\Big(X_i\beta_i - M\Big) = 0$$

Let's assume the design matrix group wise columns are orthonormalized, that is, $X_i^T X_i/n = I$. Then the solution is $\hat{\beta}_i = S^2_{\lambda/n\rho}\Big(X_i^T M/n\Big)$, where $M = X_i\beta_i^k - \overline{X\beta}^k + \bar{z}^k - u^k$. For $i = 1, \cdots, N$.

B.3 Parallel ADMM across Samples

From the general derivation forms, we can see different penalty functions can only provide different z updates. Hence, we only show the derivation of z update.

B.3.1 Lasso

Let's consider following target function with Lasso penalty

$$R(z) = \lambda|z| + \sum_{i=1}^{N}\frac{\rho}{2}(\beta_i - z + u_i)^2$$

Then it's not hard to obtain the solution by taking the first derivatives

$$\begin{aligned}
\frac{\partial R(z)}{\partial z} &= \lambda \text{sign}(z) - \sum_{i=1}^{N} \rho(\beta_i - z + u_i) \\
&= \lambda \text{sign}(z) - \rho(\sum_{i=1}^{N} \beta_i - Nz + \sum_{i=1}^{N} u_i) \\
&= \lambda \text{sign}(z) - \rho(\bar{\beta} - Nz + \bar{u}) \\
\hat{z} &= S^1_{\lambda/N\rho}\Big(\bar{\beta} + \bar{u}\Big)
\end{aligned}$$

B.3.2 Group Lasso

Let's consider following target function with Group Lasso penalty

$$R(z) = \sum_{j=1}^{q} \Big[\lambda \|z_j\| + \sum_{i=1}^{N} \frac{\rho}{2}(\beta_{ij} - z_j + u_{ij})^2\Big]$$

For $j = 1, \cdots, q$,

$$\begin{aligned}
R(z_j) &= \lambda \|z_j\| + \sum_{i=1}^{N} \frac{\rho}{2}(\beta_{ij} - z_j + u_{ij})^2 \\
\frac{\partial R(z_j)}{\partial z_j} &= \frac{\lambda z_j}{\|z_j\|} - N\rho(\bar{\beta}_j - z_j + \bar{u}_j) \\
\hat{z}_j &= S^2_{\lambda/N\rho}\Big(\bar{\beta}_j + \bar{u}_j\Big)
\end{aligned}$$

B.4 Other Simulation Results

Table B.1: Comparison between ADMM and parallel ADMM in splitting features for $(n, p, q) = (800, 50, 10)$ with LASSO, GLASSO and SGLASSO penalties. Numbers are mean (standard deviation).

Approach	Correlation	TP	FP	Prediction	Time
LASSO	AR1-5	18.133(2.013)	11.667(1.322)	0.836(0.051)	5.796(0.610)
	AR1-8	18.100(1.807)	10.467(1.570)	0.842(0.061)	5.911(0.568)
	banded1	17.933(1.701)	11.300(1.179)	0.847(0.035)	5.755(0.722)
	banded2	17.833(2.245)	11.767(1.654)	0.847(0.053)	5.850(0.564)
GLASSO	AR1-3	18.167(2.627)	51.167(1.704)	0.879(0.034)	5.854(0.560)
	AR1-5	18.533(2.713)	57.100(10.067)	0.893(0.084)	5.773(0.602)
	banded1	19.133(4.066)	50.300(31.456)	0.854(0.057)	5.642(0.727)
	banded2	19.567(2.635)	54.000(21.158)	0.880(0.057)	5.628(0.609)
SGLASSO	AR1-3	18.300(2.336)	3.067(0.907)	0.841(0.064)	6.728(0.725)
	AR1-5	18.167(1.642)	5.200(2.091)	0.840(0.052)	6.735(0.484)
	banded1	17.733(1.999)	2.633(0.765)	0.825(0.061)	6.531(0.749)
	banded2	17.967(1.956)	2.933(1.081)	0.850(0.049)	6.613(0.533)
Approach	Correlation	TP	FP	Prediction	Time
PLASSO	AR1-3	18.600(2.430)	11.800(2.007)	0.841(0.056)	3.864(0.696)
	AR1-5	18.600(1.831)	12.767(2.515)	0.853(0.056)	3.941(0.221)
	banded1	18.433(1.942)	10.633(0.765)	0.854(0.047)	3.837(0.533)
	banded2	18.567(2.300)	11.300(1.393)	0.847(0.057)	3.900(0.672)
PGLASSO	AR1-3	18.333(4.901)	58.667(22.046)	1.020(0.359)	3.902(0.689)
	AR1-5	18.667(4.498)	58.667(22.200)	1.021(0.292)	3.848(0.448)
	banded1	18.00(6.017)	45.200(26.945)	0.926(0.434)	3.761(0.622)
	banded2	18.467(5.084)	48.600(39.718)	1.028(0.370)	3.788(0.835)
PSGLASSO	AR1-3	18.667(2.808)	2.233(1.612)	0.860(0.079)	4.520(0.659)
	AR1-5	18.533(2.488)	2.467(1.583)	0.881(0.067)	4.490(0.400)
	banded1	18.433(2.700)	2.300(0.466)	0.861(0.054)	4.354(0.671)
	banded2	18.667(3.209)	3.067(1.172)	0.865(0.046)	4.408(0.346)

Table B.2: Comparison between ADMM and parallel ADMM in splitting features for $(n, p, q) = (1500, 100, 10)$ with LASSO, GLASSO and SGLASSO penalties. Numbers are mean (standard deviation).

Approach	Correlation	TP	FP	Prediction	Time
LASSO	AR1-3	18.800(1.990)	9.467(0.776)	0.842(0.039)	24.878(1.707)
	AR1-5	18.833(1.821)	12.300(1.466)	0.831(0.036)	24.883(2.033)
	banded1	19.633(2.125)	10.067(1.254)	0.833(0.029)	24.352(2.156)
	banded2	18.900(1.647)	10.167(1.379)	0.830(0.028)	25.138(1.553)
GLASSO	AR1-3	18.100(5.732)	67.467(32.765)	0.912(0.133)	24.928(1.660)
	AR1-5	17.900(2.496)	56.933(37.115)	0.911(0.066)	24.858(2.143)
	banded1	17.600(4.264)	59.867(27.973)	0.892(0.110)	23.503(0.517)
	banded2	17.767(5.900)	53.733(32.192)	0.926(0.139)	25.178(2.762)
SGLASSO	AR1-3	18.433(1.870)	2.400(0.814)	0.829(0.038)	29.453(1.769)
	AR1-5	18.567(1.569)	3.567(1.695)	0.830(0.040)	29.258(2.110)
	banded1	18.833(1.840)	2.067(1.254)	0.825(0.038)	28.110(0.666)
	banded2	18.700(1.557)	2.300(1.794)	0.819(0.035)	28.096(4.884)
Approach	Correlation	TP	FP	Prediction	Time
PLASSO	AR1-3	18.933(1.946)	10.800(2.398)	0.827(0.040)	9.218(3.262)
	AR1-5	18.667(1.539)	11.867(3.884)	0.829(0.025)	8.627(2.703)
	banded1	18.033(1.903)	10.533(1.819)	0.824(0.041)	8.615(1.396)
	banded2	17.433(1.654)	9.267(1.799)	0.825(0.036)	8.074(1.014)
PGLASSO	AR1-3.	18.000(4.472)	63.033(39.670)	0.803(0.091)	8.852(3.328)
	AR1-5	18.133(3.848)	61.333(39.207)	0.823(0.084)	8.77(3.197)
	banded1	17.333(3.698)	59.667(36.230)	0.868(0.085)	8.727(1.297)
	banded2	18.167(4.292)	58.100(34.816)	0.841(0.102)	9.949(4.551)
PSGLASSO	AR1-3	18.467(2.556)	3.867(1.193)	0.844(0.036)	9.707(3.924)
	AR1-5	17.933(1.639)	3.100(1.458)	0.842(0.032)	9.737(2.154)
	banded1	18.600(2.044)	3.467(1.730)	0.835(0.034)	9.188(1.612)
	banded2	17.933(1.837)	2.700(1.466)	0.829(0.043)	9.014(1.862)

Table B.3: Comparison between ADMM and parallel ADMM in splitting samples for $(n, p, q) = (5000, 100, 10)$ with LASSO, GLASSO and SGLASSO penalties. Numbers are mean (standard deviation).

Approach	Correlation	TP	FP	Prediction	Time
LASSO	AR1-3	18.740(1.536)	10.492(2.624)	0.807(0.016)	50.273(1.070)
	AR1-5	18.509(1.668)	9.601(2.107)	0.810(0.013)	50.357(1.241)
	banded1	17.948(2.232)	9.997(3.054)	0.804(0.016)	50.549(0.608)
	banded2	18.146(1.699)	11.152(3.034)	0.806(0.011)	50.533(1.431)
GLASSO	AR1-3	18.707(2.195)	56.881(17.147)	0.809(0.020)	48.079(1.363)
	AR1-5	18.675(2.042)	55.330(14.935)	0.824(0.029)	46.006(0.726)
	banded1	17.948(2.291)	56.352(15.900)	0.819(0.021)	46.044(0.909)
	banded2	18.047(1.737)	51.569(16.593)	0.817(0.016)	48.557(1.881)
SGLASSO	AR1-3	18.576(1.093)	2.343(1.590)	0.796(0.016)	67.890(2.048)
	AR1-5	18.410(1.340)	2.277(1.953)	0.799(0.013)	66.704(0.718)
	banded1	17.883(1.323)	2.643(1.241)	0.796(0.012)	66.549(1.516)
	banded2	18.047(1.211)	2.178(1.882)	0.797(0.017)	66.628(2.021)
Approach	Correlation	TP	FP	Prediction	Time
PLASSO	AR1-3	17.849(1.117)	13.362(2.717)	0.801(0.016)	23.319(1.537)
	AR1-5	18.047(1.374)	10.063(2.786)	0.799(0.015)	23.360(1.003)
	banded1	18.245(1.596)	12.700(1.892)	0.801(0.016)	23.139(1.062)
	banded2	18.576(1.062)	11.449(1.531)	0.801(0.012)	23.449(1.731)
PGLASSO	AR1-3	18.675(1.125)	53.647(19.960)	0.807(0.015)	24.039(4.042)
	AR1-5	18.509(1.841)	54.505(29.073)	0.811(0.016)	23.003(0.938)
	banded1	18.740(2.235)	54.934(22.379)	0.806(0.016)	23.022(1.003)
	banded2	18.279(1.532)	56.287(27.197)	0.801(0.017)	24.278(1.728)
PSGLASSO	AR1-3	18.443(1.147)	3.200(2.434)	0.795(0.014)	30.948(2.454)
	AR1-5	18.378(1.343)	2.880(2.044)	0.803(0.012)	30.323(1.176)
	banded1	17.849(1.337)	2.507(1.716)	0.796(0.012)	30.242(0.870)
	banded2	18.047(1.211)	2.318(1.841)	0.800(0.015)	33.632(7.445)

Table B.4: Comparison between ADMM and parallel ADMM in splitting samples for $(n, p, q) = (10000, 100, 10)$ with LASSO, GLASSO and SGLASSO penalties. Numbers are mean (standard deviation).

Approach	Correlation	TP	FP	Prediction	Time
LASSO	AR1-3	19.122(1.568)	10.706(2.678)	0.823(0.016)	55.859(1.092)
	AR1-5	18.887(1.702)	9.797(2.150)	0.826(0.013)	55.952(1.267)
	banded1	18.314(2.278)	10.201(3.116)	0.820(0.016)	56.166(0.620)
	banded2	18.516(1.734)	11.380(3.096)	0.822(0.011)	56.148(1.460)
GLASSO	AR1-3	19.089(2.240)	58.042(17.497)	0.825(0.020)	53.421(1.391)
	AR1-5	19.056(2.084)	56.459(15.240)	0.840(0.029)	51.118(0.740)
	banded1	18.314(2.338)	57.502(16.225)	0.835(0.021)	51.161(0.927)
	banded2	18.415(1.773)	52.621(16.932)	0.833(0.016)	53.952(1.919)
SGLASSO	AR1-3	18.955(1.115)	2.391(1.622)	0.812(0.016)	75.434(2.090)
	AR1-5	18.786(1.368)	2.323(1.993)	0.815(0.013)	74.116(0.732)
	banded1	18.248(1.350)	2.697(1.267)	0.812(0.012)	73.943(1.547)
	banded2	18.415(1.235)	2.222(1.920)	0.813(0.017)	74.031(2.062)
Approach	Correlation	TP	FP	Prediction	Time
PLASSO	AR1-3	18.213(1.140)	13.635(2.772)	0.817(0.016)	25.910(1.569)
	AR1-5	18.415(1.402)	10.269(2.843)	0.815(0.015)	25.956(1.023)
	banded1	18.617(1.628)	12.959(1.931)	0.817(0.016)	25.710(1.084)
	banded2	18.955(1.084)	11.683(1.562)	0.817(0.012)	26.054(1.766)
PGLASSO	AR1-3	19.056(1.148)	54.742(20.368)	0.823(0.015)	26.710(4.125)
	AR1-5	18.887(1.879)	55.618(29.667)	0.827(0.016)	25.559(0.957)
	banded1	19.122(2.281)	56.055(22.836)	0.822(0.016)	25.580(1.023)
	banded2	18.652(1.563)	57.436(27.752)	0.817(0.017)	26.976(1.763)
PSGLASSO	AR1-3	18.819(1.171)	3.265(2.484)	0.811(0.014)	34.386(2.504)
	AR1-5	18.753(1.371)	2.931(2.086)	0.819(0.012)	33.693(1.200)
	banded1	18.213(1.365)	2.558(1.751)	0.812(0.012)	33.602(0.888)
	banded2	18.415(1.235)	2.365(1.879)	0.816(0.015)	37.369(7.597)

Appendix C

Appendix for Chapter 4

C.1 Penalized Logistic Regression

C.1.1 ADMM for Bi-level selection (SGLASSO)

Consider the following quadratic approximation to the penalized likelihood with sparse group LASSO penalty

$$R(\alpha,\beta) \approx -\frac{1}{2n}(\tilde{Y}-E\alpha-X\beta)^{\top}W(\tilde{Y}-E\alpha-X\beta)+\sum_{j=1}^{p}\lambda_1\|\beta_j\|+\sum_{j=1}^{p}\sum_{h=0}^{q}\lambda_2|\beta_{jh}|, \quad \text{(C.1)}$$

Given the current estimate parameter $\beta^{(k)}$, we minimize the weighted least square function with respect to α. Then compute $\alpha^{(k+1)} = (E^{\top}E)^{-1}E^{-1}W(\tilde{Y}-X\beta^{(k)})$. Next, fix α at the current estimate $\alpha^{(k+1)}$ in $R(\alpha,\beta)$. To formulate the parallel ADMM framework, we start from expressing equation (C.1) as a constrained optimization problem with auxiliary variable z.

$$\begin{aligned} R(\alpha^{(k+1)},\beta,z) \approx -\frac{1}{2n}\left(\tilde{Y}-E\alpha^{(k+1)}-X\beta\right)^{\top}W\left(\tilde{Y}-E\alpha^{(k+1)}-X\beta\right) \\ +\sum_{j=1}^{p}\lambda_1\|z_j\|+\sum_{j=1}^{p}\sum_{h=0}^{q}\lambda_2|z_{jh}|, \text{ subject to } \beta-z=0, \quad \text{(C.2)} \end{aligned}$$

where z_j and z_{jh} are auxiliary variables for β_j and β_{jh}, for $j = 1, \cdots, p$ and $h = 0, \cdots, q$. Equation (C.2) is equivalent to the following augmented form with $\rho_1, \rho_2 (> 0)$ being the augmentation parameters

$$\begin{aligned} R(\alpha^{(k+1)}, \beta, z, \tau) \approx & -\frac{1}{2n}\left(\tilde{Y} - E\alpha^{(k+1)} - X\beta\right)^\top W \left(\tilde{Y} - E\alpha^{(k+1)} - X\beta\right) \\ & + \sum_{j=1}^{p} \Bigg\{ \lambda_1 \|z_j\| + \tau_j^\top (\beta_j - z_j) + \frac{\rho_1}{2}(\beta_j - z_j)^2 \\ & \quad + \sum_{h=0}^{q} \left[\lambda_2 |z_{jh}| + \tau_{jh}(\beta_{jh} - z_{jh}) + \frac{\rho_2}{2}(\beta_{jh} - z_{jh})^2 \right] \Bigg\}, \end{aligned} \tag{C.3}$$

where β, z are the primal variables and τ is the dual variable. For simplicity, we impose a scaled form ($u = \frac{1}{\rho}\tau$) to simplify equation (C.3)

$$\begin{aligned} R(\alpha^{(k+1)}, \beta, z, u) \approx & -\frac{1}{2n}\left(\tilde{Y} - E\alpha^{(k+1)} - X\beta\right)^\top W \left(\tilde{Y} - E\alpha^{(k+1)} - X\beta\right) \\ & + \sum_{j=1}^{p} \Bigg\{ \lambda_1 \|z_j\| + \frac{\rho_1}{2}(\beta_j - z_j + u_j)^2 - \frac{\rho_1}{2} u_j^2 \\ & \quad + \sum_{h=0}^{q} \left[\lambda_2 |z_{jh}| + \frac{\rho_2}{2}(\beta_{jh} - z_{jh} + u_{jh})^2 - \frac{\rho_2}{2} u_{jh}^2 \right] \Bigg\}. \end{aligned}$$

According to the discussion in Boyd et al. (2011) , the ADMM solutions at $(k+1)$th iteration can be derived as

$$\beta^{(k+1)} = \arg\min_{\beta} R(\alpha^{(k+1)}, \beta, z^{(k)}, u^{(k)}), \tag{C.4}$$

$$z^{(k+1)} = \arg\min_{z} R(\alpha^{(k+1)}, \beta^{(k+1)}, z, u^{(k)}), \tag{C.5}$$

$$u^{(k+1)} = u^k + (\beta^{k+1} - z^{k+1}). \tag{C.6}$$

C.1.2 ADMM for LASSO

Consider the following quadratic approximation to the penalized likelihood with LASSO penalty

$$R(\alpha, \beta) \approx -\frac{1}{2n}(\tilde{Y} - E\alpha - X\beta)^\top W(\tilde{Y} - E\alpha - X\beta) + \lambda|\beta|, \tag{C.7}$$

Given the current estimate parameter $\beta^{(k)}$, we minimize the weighted least square function with respect to α. Then compute $\alpha^{(k+1)} = (E^\top E)^{-1}E^{-1}W(\tilde{Y} - X\beta^{(k)})$. Next, fix α at the current estimate $\alpha^{(k+1)}$ in $R(\alpha, \beta)$. To formulate the parallel ADMM framework, we start from expressing equation (C.7) as a constrained optimization problem with auxiliary variable z.

$$\begin{aligned} R(\alpha^{(k+1)}, \beta, z) \approx & -\frac{1}{2n}\left(\tilde{Y} - E\alpha^{(k+1)} - X\beta\right)^\top W\left(\tilde{Y} - E\alpha^{(k+1)} - X\beta\right) + \lambda|z| \\ & \text{subject to } \beta - z = 0, \end{aligned} \tag{C.8}$$

Equation (C.8) is equivalent to the following augmented form with $\rho(> 0)$ being the augmentation parameters

$$\begin{aligned} R(\alpha^{(k+1)}, \beta, z, \tau) \approx & -\frac{1}{2n}\left(\tilde{Y} - E\alpha^{(k+1)} - X\beta\right)^\top W\left(\tilde{Y} - E\alpha^{(k+1)} - X\beta\right) \\ & + \lambda\Big|z\Big| + \tau^\top\Big(\beta - z\Big) + \frac{\rho_1}{2}\Big(\beta - z\Big)^2 \end{aligned}$$

where β, z are the primal variables and τ is the dual variable. For simplicity, we impose a scaled form ($u = \frac{1}{\rho}\tau$) to simplify equation (C.9)

$$\begin{aligned} R(\alpha^{(k+1)}, \beta, z, u) \approx & -\frac{1}{2n}\left(\tilde{Y} - E\alpha^{(k+1)} - X\beta\right)^\top W\left(\tilde{Y} - E\alpha^{(k+1)} - X\beta\right) \\ & + \lambda\Big|z\Big| + \frac{\rho}{2}\Big(\beta - z + u\Big)^2 - \frac{\rho}{2}u^2 \end{aligned}$$

According to the discussion in Boyd et al. (2011) , the ADMM solutions at $(k+1)$th iteration can be derived as

$$\beta^{(k+1)} = \arg\min_{\beta} R(\alpha^{(k+1)}, \beta, z^{(k)}, u^{(k)}), \tag{C.9}$$

$$z^{(k+1)} = \arg\min_{z} R(\alpha^{(k+1)}, \beta^{(k+1)}, z, u^{(k)}), \tag{C.10}$$

$$u^{(k+1)} = u^{k} + (\beta^{k+1} - z^{k+1}). \tag{C.11}$$

C.1.3 Parallel ADMM for LASSO (PLASSO)

Consider the following quadratic approximation to the penalized likelihood with LASSO penalty

$$R(\alpha, \beta) \approx -\frac{1}{2n}(\tilde{Y} - E\alpha - X\beta)^{\top} W (\tilde{Y} - E\alpha - X\beta) + \lambda|\beta|, \tag{C.12}$$

Given the current estimate parameter $\beta^{(k)}$, we minimize the weighted least square function with respect to α. Then compute $\alpha^{(k+1)} = (E^{\top}E)^{-1}E^{-1}W(\tilde{Y} - X\beta^{(k)})$. Next, fix α at the current estimate $\alpha^{(k+1)}$ in $R(\alpha, \beta)$. To formulate the parallel ADMM framework, we start from expressing equation (C.12) as a constrained optimization problem with auxiliary variable $z = (z_1, \cdots, z_B)$.

$$\begin{aligned} R(\alpha^{(k+1)}, \beta, z) \approx -\frac{1}{2n}\left(\tilde{Y} - E\alpha^{(k+1)} - \sum_{b=1}^{B} X_b\beta_b\right)^{\top} W \left(\tilde{Y} - E\alpha^{(k+1)} - \sum_{b=1}^{B} X_b\beta_b\right) \\ + \sum_{b=1}^{B} \lambda\Big|z_b\Big|, \text{ subject to } X_b\beta_b - z_b = 0, \end{aligned} \tag{C.13}$$

where we split the data columns into B groups with $X_b \in \mathbb{R}^{n\times p(q+1)/B}$ and $\beta_b \in \mathbb{R}^{p(q+1)/B}$, for $b = 1, \cdots, B$. Equation (C.13) is equivalent to the following augmented form with $\rho(> 0)$

being the augmentation parameters

$$R(\alpha^{(k+1)},\beta,z,\tau) \approx -\frac{1}{2n}\left(\tilde{Y}-E\alpha^{(k+1)}-\sum_{b=1}^{B}X_b\beta_b\right)^{\top} W\left(\tilde{Y}-E\alpha^{(k+1)}-\sum_{b=1}^{B}X_b\beta_b\right) + \sum_{b=1}^{B}\left[\lambda|z_b|+\tau_b^{\top}(\beta_b-z_b)+\frac{\rho}{2}(\beta_b-z_b)^2\right], \tag{C.14}$$

where β, z are the primal variables and τ is the dual variable. For simplicity, we impose a scaled form ($u=\frac{1}{\rho}\tau$) to simplify equation (C.14)

$$R(\alpha^{(k+1)},\beta,z,u) \approx -\frac{1}{2n}\left(\tilde{Y}-E\alpha^{(k+1)}-\sum_{b=1}^{B}X_b\beta_b\right)^{\top} W\left(\tilde{Y}-E\alpha^{(k+1)}-\sum_{b=1}^{B}X_b\beta_b\right) + \sum_{b=1}^{B}\left[\lambda|z_b|+\frac{\rho}{2}(\beta_b-z_b+u_b)^2-\frac{\rho}{2}u_b^2\right]$$

For each batch of data, the estimation can be treated as partial prediction of Y with the features of β_b. According to the discussion in Boyd et al. (2011) , the parallel ADMM solutions at $(k+1)$th iteration can be derived as

$$\beta_b^{(k+1)} = \arg\min_{\beta_b}\left\{\lambda|\beta_b|+\frac{\rho}{2}\left\|X_b\beta_b-X_i\beta_b^{(k)}-\bar{z}^{(k)}+\overline{X\beta}^{(k)}+u_b^{(k)}\right\|^2\right\} \tag{C.15}$$

$$\bar{z}^{(k+1)} = \arg\min_{\bar{z}}\left\{-\frac{1}{2n}\left(\tilde{Y}-E\alpha^{(k+1)}-Nz\right)^{\top} W\left(\tilde{Y}-E\alpha^{(k+1)}-Nz\right) + \frac{N\rho}{2}\|\bar{z}-\overline{X\beta}^{(k+1)}-\bar{u}^{(k)}\|^2\right\} \tag{C.16}$$

$$u^{(k+1)} = u^{(k)}+\left(\overline{X\beta}^{(k+1)}-\bar{z}^{(k+1)}\right) \tag{C.17}$$

C.2 Other Simulation Results

Table C.1: Binary Response: Comparison between ADMM and parallel ADMM in splitting features for $(n, p, q) = (500, 100, 4)$ with LASSO, GLASSO and SGLASSO penalties. Numbers are mean (standard deviation).

Approach	Correlation	TP	FP	Time
LASSO	AR1-3	16.94(1.64)	8.88(2.66)	2.29(0.09)
	AR1-5	17.28(1.19)	6.86(2.68)	2.63(0.09)
	banded 1	16.32(1.28)	12.58(3.05)	2.14(0.08)
	banded 2	16.72(1.52)	11.26(2.96)	2.33(0.08)
GLASSO	AR1-3	15.40(2.33)	21.18(3.43)	3.96(0.10)
	AR1-5	16.68(2.01)	22.66(2.63)	5.56(1.06)
	banded 1	12.72(2.22)	17.24(2.75)	3.11(0.08)
	banded 2	15.22(1.96)	20.62(2.67)	3.82(0.03)
SGLASSO	AR1-3	18.18(1.52)	3.00(2.02)	3.07(0.08)
	AR1-5	18.52(1.24)	2.02(1.30)	3.48(0.08)
	banded 1	16.90(1.69)	4.66(2.64)	3.17(0.08)
	banded 2	17.74(1.54)	3.54(2.02)	3.09(0.09)
Approach	Correlation	TP	FP	Time
PLASSO	AR1-3	15.14(1.71)	7.76(2.45)	0.95(0.03)
	AR1-5	15.00(1.51)	7.80(2.66)	0.88(0.02)
	banded 1	14.90(1.65)	8.94(3.33)	1.06(0.05)
	banded 2	14.88(1.33)	8.44(2.74)	0.98(0.05)
PGLASSO	AR1-3	13.40(2.00)	19.94(3.11)	2.37(0.09)
	AR1-5	15.04(1.81)	22.32(2.72)	2.69(0.10)
	banded 1	11.95(2.27)	14.16(3.40)	2.12(0.08)
	banded 2	11.92(2.42)	17.70(3.56)	2.32(0.09)
PSGLASSO	AR1-3	17.86(1.81)	2.70(2.01)	2.77(0.09)
	AR1-5	18.42(1.37)	1.66(1.11)	2.88(0.08)
	banded 1	16.52(1.89)	4.08(2.32)	2.65(0.09)
	banded 2	17.58(1.51)	3.02(1.90)	2.68(0.09)

Table C.2: Binary Response: Comparison between ADMM and parallel ADMM in splitting features for $(n, p, q) = (800, 100, 4)$ with LASSO, GLASSO and SGLASSO penalties. Numbers are mean (standard deviation).

Approach	Correlation	TP	FP	Time
LASSO	AR1-3	18.74(0.98)	7.42(2.50)	5.53(0.10)
	AR1-5	18.84(0.91)	5.50(2.40)	7.06(0.09)
	banded 1	18.30(1.09)	10.88(2.79)	4.93(0.07)
	banded 2	18.64(1.10)	8.58(3.13)	5.22(0.09)
GLASSO	AR1-3	15.54(2.34)	20.10(3.36)	4.76(0.09)
	AR1-5	16.74(1.79)	22.64(2.89)	6.19(0.12)
	banded 1	11.96(2.36)	14.90(3.08)	3.84(0.10)
	banded 2	14.28(2.24)	18.62(3.16)	4.47(0.09)
SGLASSO	AR1-3	19.42(0.73)	2.34(1.67)	5.13(0.15)
	AR1-5	19.36(0.98)	1.26(1.24)	5.46(0.27)
	banded 1	18.78(0.97)	3.02(1.49)	5.03(0.14)
	banded 2	19.34(0.79)	1.92(1.15)	5.76(0.17)
Approach	Correlation	TP	FP	Time
PLASSO	AR1-3	17.90(1.26)	6.88(2.38)	1.50(0.07)
	AR1-5	17.66(1.04)	6.20(2.23)	1.30(0.04)
	banded 1	17.58(1.26)	7.86(2.35)	1.53(0.07)
	banded 2	17.78(1.32)	7.12(2.52)	1.53(0.07)
PGLASSO	AR1-3	16.60(1.71)	26.22(2.60)	3.03(0.11)
	AR1-5	16.76(1.64)	26.54(2.46)	3.75(0.11)
	banded 1	15.60(1.76)	23.26(2.60)	2.83(0.10)
	banded 2	16.00(1.94)	25.42(2.92)	2.87(0.09)
PSGLASSO	AR1-3	19.44(0.83)	2.54(2.07)	3.96(0.13)
	AR1-5	19.38(0.85)	1.58(1.14)	4.22(0.11)
	banded 1	19.80(1.14)	3.22(2.00)	4.11(0.13)
	banded 2	19.46(0.86)	2.24(1.46)	3.68(0.11)

C.3 Real Data Analysis: other approaches

Table C.3: Analysis of NHS T2D: G×E interaction identifications from PSGL. Numbers are estimated regression coefficients for genetic main effect and G×E interactions.

Gene Name	SNP ID	Main	ACT	GL	AGE	ALCOHOL
WDFY4	rs2663058				0.0190	
RP11-445P17.5	rs10904468			-0.0304		
SVIL	rs914279	0.0224			0.0312	-0.0496
DKK1	rs7093925	-0.0192				
RP11-490O24.2	rs12252306				0.0308	
REEP3	rs10822184	-0.0291				
CAMK1D	rs12763487			-0.0277		
UPF2	rs2062982			-0.0321		
RP11-523O18.7	rs2663058				0.0190	
PRPF38AP1	rs1538511			-0.0486		
DNAJC1	rs2666775	0.0251		-0.0329		
CHAT	rs1917810					-0.0153
RP11-543F8.2	rs2646425	-0.0125		-0.0247		-0.0112
RNU6-413P	rs875598			-0.0173		
CUBN	rs11254275			-0.0269		
RP11-174J11.1	rs10761833		-0.0469			
RP11-195B3.1	rs7070200					-0.0143
PRKCQ	rs650652	-0.0433				
VIM	rs359296			-0.0554		
LRRC18	rs2663058				0.0190	
RP5-1119O21.2	rs1538246			-0.0213		
AKR1C2	rs2518044				0.0291	
PRKG1	rs10995831			-0.0284		
SNRPEP8	rs2891427		-0.0199			
ASAH2	rs7908522			-0.0423		
PHYH	rs1556718				0.0192	

Table C.4: Analysis of NHS T2D: G×E interaction identifications from LASSO. Numbers are estimated regression coefficients for genetic main effect and G×E interactions.

Gene Name	SNP ID	Main	ACT	GL	AGE	ALCOHOL
REEP3	rs10822184	-0.0867				
RP11-478B11.2	rs9804334					-0.0864
RP5-1119O21.2	rs1538246			-0.0794		
MIR4675	rs7922148			-0.0893		
RP11-174J11.1	rs2578070		0.0229			
ASAH2	rs7908522			-0.1146		
PRKCQ	rs650652	-0.1007				
PHYH	rs1556718				0.0808	
LINC00838	rs1331690				0.0675	
AKR1C2	rs10904384				0.0249	
RP11-543F8.2	rs2152963					-0.0466
RP11-490O24.2	rs12252306				0.0871	
RP11-71J2.1	rs10825013			-0.0797		
CHAT	rs6537547					
DKK1	rs7093925	-0.0777				
DNAJC1	rs2066270			-0.0066		
RNU6-413P	rs875598			-0.0856		
8-Mar	rs7081687					-0.0948
CAMK1D	rs12763487			-0.0965		
SNRPEP8	rs2891427		-0.0345			
CUBN	rs11254275			-0.0822		
VIM	rs359294			0.0429		
AKR1C1	rs7915338					
PRKG1	rs10995831			-0.0893		
PRPF38AP1	rs1538511			-0.1115		
AL512640.1	rs1545844					-0.1002
RP11-195B3.1	rs7070200					-0.0677
RP11-445P17.5	rs10904464			-0.0567		
RP11-556E13.1	rs1919738			-0.0944		
LUZP4P1	rs12262659	0.0586				
UPF2	rs11257429			0.0101		
AMD1P1	rs7918915				0.0936	
RP11-135D11.2	rs12354667					-0.1069
SVIL	rs6481643					0.0136

Table C.5: Analysis of NHS T2D: G×E interaction identifications from PLASSO. Numbers are estimated regression coefficients for genetic main effect and G×E interactions.

Gene Name	SNP ID	Main	ACT	GL	AGE	ALCOHOL
RP11-174J11.1	rs2256778		0.1133			
RP11-71J2.1	rs10825013			-0.0937		
WDFY4	rs2943246	0.0562				
RP11-543F8.2	rs2152963			-0.0968		-0.0427
PHYH	rs1556718				0.0783	
AKR1C1	rs7915338				0.0059	
RP11-523O18.7	rs2943246	0.0562				
RP11-445P17.5	rs10904455				0.0722	
PRKCQ	rs650652	-0.1055				
SVIL	rs6481643	0.0387				
ANTXRL	rs7895458					-0.0935
AMD1P1	rs7918915				0.0879	
LUZP4P1	rs12262659	0.0725				
RPL7P37	rs2767051					0.1234
REEP3	rs10822184	-0.0963				
ASAH2	rs7908522			-0.1184		0.0565
ADARB2	rs4554799			-0.0657		
RP11-556E13.1	rs1919738			-0.0790		
CAMK1D	rs12763487			-0.1125		
8-Mar	rs7081687					-0.1202
LRRC18	rs2943246	0.0562				
ANTXRLP1	rs11259760				0.0671	
LINC00838	rs1331690				0.0667	
RP11-195B3.1	rs7070200					-0.0811
PRPF38AP1	rs1538510			-0.1198		
MIR4675	rs7922148			-0.0878		
VIM	rs359294			-0.1242		
RNU6-163P	rs7906409			-0.0906		
RNU6-413P	rs875598			-0.1230		
CHAT	rs6537547					-0.2351
RP11-490O24.2	rs12252306				0.0785	
PRKG1	rs10995831			-0.0940		
DKK1	rs7093925	-0.0769				
AKR1C2	rs10904384				0.0720	