High-dimensional variable selection in longitudinal and nonlinear gene-environment interaction studies

by

Fei Zhou

M.S., Kansas State University, 2015

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

Variable selection from both the frequentist and Bayesian frameworks has gained increasing popularity in the analysis of high-dimensional genomic data. Despite the success of existing studies, challenges still remain as tailored methods for sparse interaction structures are not available when the response variables are repeatedly measured and/or have heavy-tailed distributions. These challenges have motivated the development of novel variable selection methods proposed in the following projects. Meanwhile, powerful software packages from these projects are publically available to facilitate fast and reliable computation, as well as reproducible research.

In the first project, we have developed a novel penalized variable selection method to identify important lipid–environment interactions in a longitudinal lipidomics study, where the environment factors refer to a group of dummy variables corresponding to a four-level treatment factor. An efficient Newton–Raphson based algorithm was proposed within the generalized estimating equation (GEE) framework. Simulation studies have demonstrated the superior performance of our method over alternatives, in terms of both identification accuracy and prediction performance. Analysis of the high-dimensional lipid datasets collected using mice from the skin cancer prevention study identified meaningful markers that provide fresh insight into the underlying mechanism of cancer preventive effects.

In the second project, we have proposed a sparse group penalization method for the bi-level G×E interaction study under the repeatedly measured phenotype to accommodate more general environment factors. Within the quadratic inference function (QIF) framework, the proposed method can achieve simultaneous identification of main and interaction effects on both the group and individual level. We conducted simulation studies to establish the advantage of the proposed regularization methods. In the case study, the environment factors include age, gender and treatment, which are either continuous or categorical. Our method

leads to improved prediction and identification of main and interaction effects with important implications.

In the third project, a sparse Bayesian quantile varying coefficient model has been developed for non-linear G×E studies. The proposed model can accommodate heavy-tailed errors and outliers from the disease phenotypes while pinpointing important non-linear interactions through Bayesian variable selection based on spike-and-slab priors. Fast computation has been facilitated by the efficient Gibbs sampler. Simulation studies and real data analysis with age as the univariate environment factor have been performed to show the superiority of the proposed method over multiple competing alternatives.

The open source R packages with C++ implementations of all the methods under comparison have been provided along this dissertation. The R packages interep and springer, for the first two projects respectively, are available on CRAN. The R package for the last project on Bayesian regularized quantile varying coefficient model will be released soon to the public.

High-dimensional variable selection in longitudinal and nonlinear gene-environment interaction studies

by

Fei Zhou

M.S., Kansas State University, 2015

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

Variable selection from both the frequentist and Bayesian frameworks has gained increasing popularity in the analysis of high-dimensional genomic data. Despite the success of existing studies, challenges still remain as tailored methods for sparse interaction structures are not available when the response variables are repeatedly measured and/or have heavy-tailed distributions. These challenges have motivated the development of novel variable selection methods proposed in the following projects. Meanwhile, powerful software packages from these projects are publically available to facilitate fast and reliable computation, as well as reproducible research.

In the first project, we have developed a novel penalized variable selection method to identify important lipid–environment interactions in a longitudinal lipidomics study, where the environment factors refer to a group of dummy variables corresponding to a four-level treatment factor. An efficient Newton–Raphson based algorithm was proposed within the generalized estimating equation (GEE) framework. Simulation studies have demonstrated the superior performance of our method over alternatives, in terms of both identification accuracy and prediction performance. Analysis of the high-dimensional lipid datasets collected using mice from the skin cancer prevention study identified meaningful markers that provide fresh insight into the underlying mechanism of cancer preventive effects.

In the second project, we have proposed a sparse group penalization method for the bi-level G×E interaction study under the repeatedly measured phenotype to accommodate more general environment factors. Within the quadratic inference function (QIF) framework, the proposed method can achieve simultaneous identification of main and interaction effects on both the group and individual level. We conducted simulation studies to establish the advantage of the proposed regularization methods. In the case study, the environment factors include age, gender and treatment, which are either continuous or categorical. Our method

leads to improved prediction and identification of main and interaction effects with important implications.

In the third project, a sparse Bayesian quantile varying coefficient model has been developed for non-linear G×E studies. The proposed model can accommodate heavy-tailed errors and outliers from the disease phenotypes while pinpointing important non-linear interactions through Bayesian variable selection based on spike-and-slab priors. Fast computation has been facilitated by the efficient Gibbs sampler. Simulation studies and real data analysis with age as the univariate environment factor have been performed to show the superiority of the proposed method over multiple competing alternatives.

The open source R packages with C++ implementations of all the methods under comparison have been provided along this dissertation. The R packages interep and springer, for the first two projects respectively, are available on CRAN. The R package for the last project on Bayesian regularized quantile varying coefficient model will be released soon to the public.

Table of Contents

List of Figures xi

List of Tables xiii

Acknowledgements xix

1 Introduction 1

1.1 Regularized Variable Selection 2

1.2 Bayesian Variable Selection 6

1.3 Works in this dissertation 7

2 Penalized variable selection for Lipid–environment interactions in a longitudinal lipidomics study 9

2.1 Introduction 9

2.2 Materials and Methods 12

2.2.1 Data and Model Settings 12

2.2.2 Generalized Estimating Equations 12

2.2.3 Penalized Identification 13

2.2.4 Computational Algorithms 16

2.3 Results 18

2.3.1 Simulation 18

2.3.2 Real Data Analysis 22

2.4 Discussion 24

3 Sparse group variable selection for gene–environment interactions in the longitudinal study . . . 27
3.1 Introduction . . . 27
3.2 Statistical Method . . . 30
3.2.1 Data and Model Settings for Longitudinal G×E Studies . . . 30
3.2.2 Quadratic Inference Function for Longitudinal G×E Interactions . . . 31
3.2.3 Penalized identification of G×E interactions in longitudinal studies . . . 33
3.2.4 Computational Algorithms for Sparse Group QIF . . . 34
3.2.5 Unbalanced Date Implementation . . . 37
3.3 Simulation . . . 38
3.4 Real Data Analysis . . . 42
3.5 Discussion . . . 45

4 The Regularized Bayesian Quantile Varying Coefficient Model . . . 47
4.1 Introduction . . . 47
4.2 Statistical Methods . . . 49
4.2.1 The Quantile Varying Coefficient Model . . . 49
4.2.2 The Regularized Bayesian Quantile Varying Coefficient Model . . . 52
4.3 The Gibbs Sampler . . . 53
4.4 Simulation . . . 57
4.5 Real Data Analysis . . . 65
4.6 Discussion . . . 69

5 Summary . . . 72

Bibliography . . . 74

A Appendices for Chapter 2 . . . 91

B Appendices for Chapter 3 . . . 104

B.1 Derivations of Alternative Methods 104
B.1.1 Penalized Group QIF 104
B.1.2 Penalized QIF 106
B.2 Other Simulation Results 107
B.3 Real Data Analysis 108

C Appendices for Chapter 4 116
C.1 Other simulation results 117
C.2 Hyper-parameters sensitivity analysis 123
C.3 Sensitivity analysis on smoothness specification 124
C.4 Posterior inference 130
C.4.1 Posterior inference for BQRVCSS 130
C.4.2 Posterior inference for BQRVC 136
C.4.3 Posterior inference for BVCSS 139
C.4.4 Posterior inference for BVC 144

List of Figures

3.1 Identification results under 25 important genetic main effects and G×E interactions (corresponding to 25 nonzero regression coefficients) in the 4 scenarios. TP/FP: true/false positives. mean(sd) of TP and FP based on 100 replicates. 41

4.1 Simulation study for Error using the proposed method (BQRVCSS). Red line: true parameter values. Black line: median estimates of varying coefficients from BQRVCSS. Blue lines: 95% credible intervals for the estimated varying coefficients. 64

4.2 Potential scale reduction factor (PSRF) versus iteractions for the varying functions in Figure 4.1. Black line: PSRF. Red line: the threshold of 1.1. $\hat{\alpha}_{j1}$ to $\hat{\alpha}_{j5}(j = 0, ..., 3)$ represent the five estimated spline coefficients for the varying coefficient function γ_j, respectively. 65

4.3 Real data analysis using the proposed method (BQRVCSS). Black line: median estimates of varying coefficients for BQRVCSS. Blue dashed lines: 95% credible intervals for the estimated varying coefficients. 68

4.4 Real data analysis using the alternative method (BVCSS). Black line: median estimates of varying coefficients for BVCSS. Blue dashed lines: 95% credible intervals for the estimated varying coefficients. 69

A.1 Plot of the identification results for $n = 250$. $p = 75$ with an actual dimension of 304. $p = 150$ with an actual dimension of 604. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 95

A.2 Plot of the identification results for $n = 500$. $p = 150$ with an actual dimension 604. $p = 300$ with an actual dimension of 1204. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 96

B1 Prediction (MSE) results of the 4 scenarios. mean(sd) of prediction error based on 100 replicates. 108

List of Tables

3.1 Identification results for Scenario 1. TP/FP: true/false positives. mean(sd) of TP and FP based on 100 replicates. 40

3.2 Identification results for Scenario 2. TP/FP: true/false positives. mean(sd) of TP and FP based on 100 replicates. 41

4.1 Identification results for i.i.d. errors based on 100 replicates. C: correct-fitting proportion; O: overfitting proportion; U: underfitting proportion. 62

4.2 Estimation and prediction results for i.i.d. errors based on 100 replicates. TMSE: total mean squared equared error. pred: prediction error (check loss for quantile methods and squared loss for non-quantile methods). pred.mad: mean absolute prediction error. 63

A.1 Identification results for $n = 250$, $p = 75$ with an actual dimension of 304. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 91

A.2 Identification results for $n = 250$, $p = 150$ with an actual dimension of 604. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 92

A.3 Identification results for $n = 500$, $p = 150$ with an actual dimension of 604. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 93

A.4 Identification results for $n = 500$, $p = 300$ with an actual dimension of 1204. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 94

A.5 Estimation accuracy results for $n = 250$. $p = 75$ with an actual dimension of 304. $p = 150$ with an actual dimension of 604. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 97

A.6 Estimation accuracy results for $n = 500$. $p = 150$ with an actual dimension of 604. $p = 300$ with an actual dimension of 1204. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 98

A.7 Real data analysis result from Method A1 (method accommodating the lipid–environment interactions with exchangeable working correlation). 99

A.8 Real data analysis result from Method A4 (method not accommodating the lipid–environment interactions with exchangeable working correlation). . . . 99

A.9 Identification results for $n = 60$, $p = 30$ with an actual dimension of 124. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 100

A.10 Estimation accuracy results for $n = 60$, $p = 30$ with an actual dimension of 124. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 100

A.11 Data simulated based upon the underlying main effect only model. Identification results for $n = 250, p = 75, \rho = 0.8$ with an actual dimension of 304. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 101

A.12 Null models. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 101

A.13 Stability selection percentages for all the 17 true effects in the simulated data when $n = 250$, $p = 75$, $\rho = 0.8$ with an actual dimension of 304. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 102

A.14 Validation methods. Identification results for $n = 250$, $p = 75$ with an actual dimension of 304. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 103

A.15 Validation methods. Estimation accuracy results for $n = 250$, $p = 75$ with an actual dimension of 304. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. 103

B1 Identification results for Scenario 3. TP/FP: true/false positives. mean(sd) of TP and FP based on 100 replicates. 107

B2 Identification results for Scenario 4. TP/FP: true/false positives. mean(sd) of TP and FP based on 100 replicates. 107

C1 Identification results on CAMP data using the bi-level selection method under the exchangeable working correlation (sgQIF.exch). The identified SNPs and the corresponding genes are listed in the first two columns. The third column contains the coefficients of the main effects for each SNP. The last three columns correspond to the interactions between the SNPs and environmental factors. 108

C2 Identification results on CAMP data using the individual-level selection method under the exchangeable working correlation (iQIF.exch). The identified SNPs and the corresponding genes are listed in the first two columns. The third column contains the coefficients of the main effects for each SNP. The last three columns correspond to the interactions between the SNPs and environmental factors. 112

C1 Identification results for heterogeneous errors based on 100 replicates. C: correct-fitting proportion; O: overfitting proportion; U: underfitting proportion. 117

C2 Estimation and prediction results for heterogeneous errors based on 100 replicates. TMSE: total mean squared equared error. pred: prediction error (check loss or squared loss). pred.mad: mean absolute prediction error. 118

C3 Identification results for simulated SNPs with i.i.d. errors based on 100 replicates. C: correct-fitting proportion; O: overfitting proportion; U: underfitting proportion. 119

C4 Estimation and prediction results for simulated SNPs with i.i.d. errors based on 100 replicates. TMSE: total mean squared equared error. pred: prediction error (check loss or squared loss). pred.mad: mean absolute prediction error. 120

C5 Identification results for simulated SNPs with heterogeneous errors based on 100 replicates. C: correct-fitting proportion; O: overfitting proportion; U: underfitting proportion. 121

C6 Estimation and prediction results for simulated SNPs with heterogeneous errors based on 100 replicates. TMSE: total mean squared equared error. pred: prediction error (check loss or squared loss). pred.mad: mean absolute prediction error. 122

C7 Sensitivity analysis on the choice of the hyperparameter for π_0 using different Beta priors for the Laplace error dirstribution for the 30% quantile. 123

C8 Sensitivity analysis on the choice of the hyperparameter for η using different Gamma priors for the Laplace error dirstribution for the 30% quantile. . . . 123

C9 Sensitivity analysis on the choice of the hyperparameter for π_0 using different Beta priors for the Laplace error dirstribution for the 50% quantile. 124

C10 Sensitivity analysis on the choice of the hyperparameter for η using different Gamma priors for the Laplace error dirstribution for the 50% quantile. . . . 124

C11 Sensitivity analysis on smoothness specification for the Laplace error dirstribution for the 30% quantile. 125

C12 Sensitivity analysis on smoothness specification for the Normal error dirstribution for the 30% quantile. 126

C13 Sensitivity analysis on smoothness specification for the Laplace error dirstribution for the 50% quantile. 127

C14 Sensitivity analysis on smoothness specification for the Normal error dirstribution for the 50% quantile. 128

C15 Sensitivity analysis on smoothness specification for BVCSS with the Normal error dirstribution for the 30% quantile. 129

C16 Sensitivity analysis on smoothness specification for BVCSS with the Normal error dirstribution for the 50% quantile. 130

Acknowledgments

First of all, I would like to express my sincere gratitude and thanks to my advisor, Dr. Cen Wu, for his guidance and support in my Ph.D. projects. Dr. Wu is an encouraging mentor with brilliant ideas. The communications between us gave me inspirations in my study. His knowledge, expertise and enthusiasm toward research have set a good example that I will learn from in my future career.

I would like to extend my thanks to my Ph.D. committee members, Dr. Christopher Vahl, Dr. Haiyan Wang and Dr. Weiqun Wang for their support, advice and helpful insights in my study. I also want to thank Dr. Sonny TM Lee, for his willingness to serve as the chairperson of my examining committee.

I would like to thank the Department of Statistics and the families of Fryer and Siepman for offering me assistantships and scholarships. These financial support helped me a lot through my study. Also I would like to thank the professors in the department for offering the excellent courses. I thank all my friends for their help and encouragements.

Finally, I must thank my family, including my parents and grandparents, for their unconditional love and support.

Chapter 1

Introduction

Gene× Environment (G × E) interactions, in addition to the genetic and environmental main effects, have important implications for elucidating the etiology of complex diseases, such as cancer, type 2 diabetes and cardiovascular diseases (Cornelis and Hu (2012); Dempfle et al. (2008); Flowers et al. (2012); Hunter (2005); Simonds et al. (2016)). Multiple G × E studies have shown that the genetic contribution to the variation in disease phenotype or increase in disease risks are often mediated by environmental effects. Historically, G × E interactions have been examined from the perspective of assessing specific genetic effect under dichotomous environmental exposures (Ottman (1996)). With the availability of high-density genetic polymorphisms such as single nucleotide polymorphisms (SNPs), it has become possible to establish the statistical associations between millions of genetic variants and disease status or phenotype in genetic association studies (Hirschhorn et al. (2002); Huang and Liang (2019); Huang et al. (2018); Huang and Liang (2018); Lunetta (2008); Wu et al. (2012)), which has also made investigation of G × E interactions possible at the more comprehensive human genome scales (Cornelis et al. (2012); Du et al. (2021); Murcray et al. (2009); Winham and Biernacka (2013)).

The dissection of G × E interactions in genetic association studies, such as genome wide association study (GWAS), has been mainly conducted based on the assessment of statistical significance. For example, in the genome wide case-control association studies of

type 2 diabetes, with body mass index (BMI) as the environmental factor, the significance of the interaction between BMI and each one of the SNPs can be evaluated using p-values from the marginal test accounting for the interaction (Cornelis et al. (2012)). After multiple test adjustment, important interaction effects can be identified when the signals are beyond the genome-wide significance level.

Furthermore, the genetic association studies can be understood from a related but distinct perspective. Consider the data matrix where the columns are corresponding to features (or variables), such as all the main and interaction effects in a G $\times$ E study, and rows are corresponding to samples (or observations). As the number of columns is usually much larger than the sample size in a typical G $\times$ E interaction study, the data matrix is of "large dimensionality, small sample size" nature. Thus, the central statistical task is to hunt down the subset of important main and interaction effects that is associated with the disease outcome, which can be reformulated as a high dimensional variable selection problem in the regression framework. Specifically, the regression coefficients of variables (representing main and interaction effects) are continuously shrunk towards zero. A zero coefficient after shrinkage denotes that the corresponding effect is not included in the final model, and has no association with the response, such as the disease phenotype. Therefore, variable selection can be performed with parameter estimation simultaneously. Such a variable selection method is referred as penalization or regularization (Fan and Lv (2010); Wu and Ma (2015)).

1.1 Regularized Variable Selection

Generically, penalized regression coefficients can be defined as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} L(D;\beta) + P(\lambda;\beta),$$

where $L(D;\beta)$ is a loss function based on the observed data D and regression coefficients β to quantify the lack-of-fit. It can be a least square loss function or a negative log-likelihood function. The penalty function, $P(\lambda;\beta)$, measures the model complexity with tuning pa-

rameter λ. As $\lambda \to +\infty$, larger amount of penalty is imposed on $\hat{\beta}$, and more components of $\hat{\beta}$ become zeros. Accordingly, fewer features will be included in the final model. The phenomena of zeros in $\hat{\beta}$ is termed as sparsity in the literature of penalized variable selection. On the other hand, when $\lambda \to 0$, the model becomes more complex since more features are included in the final model, Tuning parameter λ balances the tendency towards two extremes. A properly tuned λ will lead to a reasonable number of variables with satisfactory interpretability and superior prediction performance.

LASSO (Tibshirani (1996))is one of the most popularly used penalized regression methods and it is a penlized least squares regression with $l1$ penalty, which is given in the following form:

$$||Y - X\beta||_2^2 + \lambda|\beta|,$$

where $||Y - X\beta||_2^2$ is the unpenalized loss function, $\beta = (\beta_1, ..., \beta_j)^\top$ and $\lambda|\beta| = \lambda \sum_{j=1}^{p} |\beta_j|$. Y is the response variable and X denotes the design matrix that contains p-dimensional genomic features, which can be gene expression, single nucleotide polymorphysm (SNP), copy number variation (CNV) and DNA mutation, etc. The solution to LASSO regression will yield a penalized estimator that is continuous (continuity) with small estimated coefficients shrunk to zero (sparsity). However, for large regression coefficients, the shrinkage will result in great bias toward 0. To overcome the problem of bias, alternative penalties have been proposed by other researchers. Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty and Zhang (2010) proposed the minimax concave penalty (MCP). The SCAD penalty is defines as

$$P_{SCAD}(\beta_j; \lambda, \gamma) = \begin{cases} \lambda|\beta_j| & \text{if } |\beta_j| \leq 0 \\ -\frac{\beta_j^2 - 2\gamma\lambda|\beta_j| + \lambda^2}{2(\gamma - 1)} & \text{if } \lambda < |\beta_j| \leq \gamma\lambda \\ \frac{1}{2}(\gamma + 1)\lambda^2 & \text{if } |\beta_j| > \gamma\lambda \end{cases},$$

where $\gamma > 2$ and $\lambda > 0$ are regularization parameters. The MCP panalty is defined as

$$P_{MCP}(\beta_j; \lambda, \gamma) = \begin{cases} \lambda|\beta_j| - \frac{\beta_j^2}{2\gamma} & \text{if } |\beta_j| \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2 & \text{if } |\beta_j| > \gamma\lambda \end{cases},$$

where the regularization parameter $\gamma > 1$. It has been proved that both penalties result in an estimator with the three properties: continuity, sparsity and unbiasedness.

Besides the high dimensionality issue, complex data structures bring more chanllenges to variable selection in G $\times$ E studies. For instance, in the selection of a group of factor level indicators for a categorical variable, the grouping structure is an important factor that needs to be taken into consideration. Yuan and Lin (2006) proposed the group LASSO regression method which enables a group-level variable selection. The group LASSO penalty is given as

$$||Y - X\beta||_2^2 + \lambda \sum_{k=1}^{m} \sqrt{L_k}||\beta_k||_2,$$

where $\beta_k = (\beta_{k1}, ..., \beta_{kL_k})^\top$ is a coefficient vector of length L_k and $\beta = (\beta_1^\top, ..., \beta_m^\top)^\top$. Besides group LASSO, other nonconvex group penalization methods, such as group SCAD and group MCP, have been developed to accommodate the group structure in variable selection (Huang et al. (2012)). The group SCAD penalty is defined as

$$P_{gSCAD}(||\beta_k||_2; \sqrt{L_k}\lambda, \gamma) = \begin{cases} \sqrt{L_k}\lambda||\beta_k||_2 & \text{if } ||\beta_k||_2 \leq 0 \\ -\frac{\beta_k^\top\beta_k - 2\sqrt{L_k}\gamma\lambda||\beta_k||_2 + L_k\lambda^2}{2(\gamma-1)} & \text{if } \sqrt{L_k}\lambda < ||\beta_k||_2 \leq \sqrt{L_k}\gamma\lambda \\ \frac{\sqrt{L_k}}{2}(\gamma+1)\lambda^2 & \text{if } ||\beta_k||_2 > \sqrt{L_k}\gamma\lambda \end{cases},$$

where the tuning parameters $\gamma > 2$ and $\lambda > 0$. The group MCP penalty is defined as

$$P_{gMCP}(||\beta_k||_2; \sqrt{L_k}\lambda, \gamma) = \begin{cases} \sqrt{L_k}\lambda||\beta_k||_2 - \frac{\beta_k^\top\beta_k}{2\gamma} & \text{if } ||\beta_k||_2 \leq \sqrt{L_k}\gamma\lambda \\ \frac{L_k}{2}\gamma\lambda^2 & \text{if } ||\beta_k||_2 > \sqrt{L_k}\gamma\lambda \end{cases},$$

where the regularization parameter $\gamma > 1$.

While the group LASSO method gives a sparse set of groups, maintaining the "group-in, group-out" characteristic, sometimes it is still necessary to achive sparsity within group. For example, in a G $\times$ E model that involves p genetic factors and q environment factors and the main effect and the interactions with the q environment factors of each genetic factor forms a group of $(1+q)$ terms. In order to determine whether a genetic factor is associated with the response variable, first of all, a group level selection should be performed. Moreover, if a genetic factor has been found to be associated with the response, then it's also necessary to carry out an individual level selection within the group. Therefore, the sparse-group LASSO (Simon et al. (2013)) has been proposed based on a combination of the LASSO and group LASSO penalties:

$$||Y - X\beta||_2^2 + \lambda_1 \sum_{k=1}^{m} \sqrt{L_k}||\beta_k||_2 + \lambda_2|\beta|,$$

where λ_1 and λ_2 are the tuning parameters for the group LASSO and LASSO penalties, respectively. The sparse group LASSO type penalties sparse group SCAD and sparse group MCP have also been established. Similar to sparse group LASSO, the sparse group SCAD and sparse group MCP penalties perform a bi-level selection on the group level and individual level simultaneously. The sparse group SCAD penalty is defined as

$$P_{sgSCAD}(\beta; \sqrt{L_k}\lambda_1, \lambda_2, \gamma) = \sum_{k=1}^{m} P_{gSCAD}(||\beta_k||_2; \sqrt{L_k}\lambda_1, \gamma) + \sum_{k=1}^{m} P_{SCAD}(\beta_k; \lambda_2, \gamma)$$

and the sparse group MCP penalty is defined as

$$P_{sgMCP}(\beta; \sqrt{L_k}\lambda_1, \lambda_2, \gamma) = \sum_{k=1}^{m} P_{gMCP}(||\beta_k||_2; \sqrt{L_k}\lambda_1, \gamma) + \sum_{k=1}^{m} P_{MCP}(\beta_k; \lambda_2, \gamma).$$

When analyzing omics data, the problem of model-misspecification and heterogeneity exists, such as data contamination in the predictors, heavy-tailed errors and outliers in the response variables, which motivate the development of roubust methods that are robust to these problems. In penalized regression, robustness can be achieved via the "unpenalized

robust loss function + penalty" form. The robust loss function includes the least absolute deviation (LAD) loss function, the check loss function, the rank-based loss function and their variants (Wu and Ma (2015)).

1.2 Bayesian Variable Selection

Bayesian variable selection has been another classical statistical strategy for analyzing high dimensional data. O'Hara and Sillanpaa (2009) categorized Bayesian variable selection approaches into four categories: (1) adaptive shrinkage, (2) indicator model selection, (3) stochastic search variable selection (SSVS) and (4) model space approach. Bayesian methods have been applied to cancer genomics data and adaptive shrinkage is closely connected with the variable selection methods in the frequentist perspective.

According to Tibshirani (1996), the LASSO estimate is equivalent to the posterior estimate when the regression coefficients adopt the independent and identical Laplace prior from the Bayesian perspective. The Laplace prior is given as

$$\pi(\beta_j|\tau) = \frac{1}{2\tau}e^{-|\beta_j|/\tau}, j = 1, ..., p,$$

where $\tau = 1/\lambda$. Park and Casella (2008) proposed Bayesian LASSO by imposing a conditional Laplace prior on the regression coefficients:

$$\pi(\beta_j|\sigma^2) = \frac{\lambda}{2\sqrt{\sigma^2}}e^{-\lambda|\beta_j|/\sqrt{\sigma^2}},$$

with σ^2 having an independent priori $\pi(\sigma^2)$, which guarantees the unimodality of the posterior distribution. Kyung et al. (2010) extends this rationale of specifying the prior other LASSO type of penalization methods, such as group LASSO, fused LASSO and the elastic net. In particular, the Bayesian group LASSO can be achieved by inroducing a multivariate Laplace prior:

$$\pi(\beta_k|\sigma^2) \propto \exp(-\frac{\sqrt{L_k\lambda}}{\sqrt{\sigma^2}}||\beta_k||_2),$$

where β_k is a coefficient vector of length L_k and $\beta = (\beta_1^\top, ..., \beta_m^\top)^\top$. $\frac{\sqrt{L_k\lambda}}{\sqrt{\sigma^2}}$ is the scale parameter in multivariate Laplace distribution. These aforementioned methods have a drawback that they cannot shrink a posterior estimate to exactly 0. Therefore, the spike-and-slab priors have been adopted to overcome this problem (Mitchell and Beauchamp (1988)). The spike-and-slab priors have been defined in the follwoing form:

$$\beta_j|\phi_j \sim \phi_j\pi_0(\beta_j) + (1-\phi_j)\pi_1(\beta_j), j = 1, ..., p,$$

where $\phi_j \in \{0, 1\}$ is an auxiliary indicator variable. $\pi_0(\cdot)$ denotes a spike distribution for zero coefficients corresponding to neligibly small effects and $\pi_1(\cdot)$ denotes a flat slab distribution for nonzero effects. In practice, $\pi_1(\cdot)$ adopts a normal distribution with large variance. Kuo and Mallick (1998) sets $\pi_0(\cdot)$ to a point mass prior, which is defined as $\delta_0(\beta_j)$, and the coefficients of unimportant effects are set to zero in the spike part. When $\phi_j = 0$, $\beta_j \sim \pi_1(\beta_j)$, which implies the jth genetic factor has nonzero coefficient in the model. Then $\phi_j = 1$ implies the absence of the jth genetic factor.

Besides, George and McCulloch (1993) proposed the SSVS method which adopts a combination of two normal distributions as the spike-and-slab prior: $\phi_j \mathrm{N}(0, c_j\tau_j^2) + (1-\phi_j)\mathrm{N}(0, \tau_j^2)$, where the spike part corresponds to the second density which is centered about zero with a small variance. Ročková and George (2018) adopted a mixture of two Laplace distributions as prior in the SSVS method. Many other methods use the Laplace and point mass mixture prior in variable selection Xu and Ghosh (2015); Yuan and Lin (2005); Zhang et al. (2016).

1.3 Works in this dissertation

In Chapter 2, we developed a novel penalized variable selection method for lipid-environment interactions in a longitudinal lipidomics study. Lipid species are critical components of eukaryotic membranes. They play key roles in many biological processes such as signal transduction, cell homeostasis, and energy storage. Investigations of lipid–environment interactions, in addition to the lipid and environment main effects, have important implica-

tions in understanding the lipid metabolism and related changes in phenotype. An efficient Newton–Raphson based algorithm was proposed within the generalized estimating equation (GEE) framework. We conducted extensive simulation studies to demonstrate the superior performance of our method over alternatives, in terms of both identification accuracy and prediction performance. As weight control via dietary calorie restriction and exercise has been demonstrated to prevent cancer in a variety of studies, analysis of the high-dimensional lipid datasets collected using mice from the skin cancer prevention study identified meaningful markers that provide fresh insight into the underlying mechanism of cancer preventive effects.

In Chapter 3, we developed a sparse group penalization method to conduct the bi-level G×E interaction study under the repeatedly measured phenotype. Penalized variable selection for high dimensional longitudinal data has received much attention as accounting for the correlation among repeated measurements can provide additional and essential information for improved identification and prediction performance. Despite the success, in longitudinal studies, the potential of penalization methods is far from fully understood for accommodating structured sparsity. Within the quadratic inference function (QIF) framework, the proposed method can achieve simultaneous identification of main and interaction effects on both the group and individual level. Simulation studies have shown that the proposed method outperforms major competitors. In the case study of asthma data from the Childhood Asthma Management Program (CAMP), we conduct G×E study by using high dimensional SNP data as the Genetic factor and the longitudinal trait, forced expiratory volume in one second (FEV1), as phenotype. Our method leads to identification of improved prediction and main and interaction effects with important implications.

In Chapter 4, we propose a novel regularized Bayesian method to identify important nonlinear G×E interactions in quantile regression model. This is an on-going project and we have successfully proposed the statistical model and obtained extensive simulation results that demonstrate the superority of the proposed method over the alternative methods in terms of identification and estimation accuracy in the case there are heavy-tailed distributions in the response.

Chapter 2

Penalized variable selection for Lipid–environment interactions in a longitudinal lipidomics study

2.1 Introduction

Longitudinal data are frequently observed in a diversity of scientific research areas, including economics, biomedical studies, and clinical trials. A common characteristic of the longitudinal data is that the same subject is measured repeatedly over a certain period of time; thus, the repeated measurements are correlated. Many modeling techniques have been proposed to accommodate the multivariate correlated nature of the data (Bandyopadhyay et al. (2011); Verbeke et al. (2014)). The emergence of new types of data has brought constant challenges to the development of novel statistical methods for longitudinal studies. One representative example is the high-dimensional data where the number of variables is much larger than the sample size. As penalization has been demonstrated as an effective way for conducting variable selection in linear and generalized linear models with a univariate response (Fan and Lv (2010); Wu and Ma (2015)), substantial efforts have been devoted to developing penalized variable selection methods with longitudinal responses, such as Cho and Qu (2013); Ma et al.

(2013); Wang et al. (2012), among many others.

This study was partially motivated by overcoming the limitations of existing penalization methods in order to analyze the high-dimensional lipidomics data from longitudinal studies. Lipids are a broad group of biomolecules in eukaryotic membranes, involved in various critical biological roles such as energy storage, cellular membrane structure, or cell signaling and homeostasis (Barona et al. (2005); Berridge (1987); Goñi and Alonso (1999); Thiam et al. (2013)).Lipid metabolism has been found to be associated with several diseases, especially chronic diseases such as diabetes, cancer, inflammatory disease, and Alzheimer (Markgraf et al. (2016); Stephenson et al. (2017); Zhou et al. (2012)).

The lipid data were obtained from our previous work on the lipid changes in weight controlled CD-1 mice (King et al. (2015)). In the current study, the phenotype of interest is the body weight of experimental animals, which was measured every week for 10 weeks. The environmental factor was exercise and/or dietary restriction, which had four different levels, control (ad libitum feeding and sedentary), AE (exercise and ad libitum feeding), PE (exercise and pair feeding), and DCR (sedentary and 20% dietary calorie restriction). Both triacylglycerol (TG) and diacylglycerol (DG) profiles in the plasma were measured using electrospray ionization MS/MS (King et al. (2015)). Here, we focused on the DG profiles and treated them as lipid factors. Besides the lipid main effects, we were particularly interested in investigating the interactions between lipids and environment/treatment effects, which will shed novel insight in the understanding of weight changes in a longitudinal setting beyond studies solely focusing on the main lipidomics effects. With the control as the baseline, we created a group of three dummy variables to represent the four levels of the treatment factor that can be treated as environmental factors in general. The product between the dummy variable group and lipid denotes the lipid–environment interactions. The formulation of the interaction group in our study shared the spirit of group LASSO, which was primarily motivated by the selection of important dummy variable groups from ANOVA problems (Yuan and Lin (2006)). As the total number of main and interaction effects was much larger than the sample size, penalized variable selection was a natural choice to identify the important subset of effects. Such methods for G×E interactions, including Wu et al. (2014,

2018), however, cannot be adopted for the longitudinal studies.

On the other hand, existing penalization methods in longitudinal studies have been mostly developed for the identification of important main effects only. For instance, Wang et al. Wang et al. (2012) proposed the penalized generalized estimating equation (PGEE) to select predictors that are associated with the longitudinal response. Ma et al. Ma et al. (2013) considered the selection of important predictors and estimation of non-parametric effects through splines for repeated measures data. Cho and Qu (2013) developed a penalized quadratic inference function (PQIF) method to conduct variable selection on main effects. Fan et al. (2012) developed robust variable selection through a penalized robust estimating equation to incorporate the correlation structure for repeated measurements. These studies have ignored the interaction effects and cannot be adopted to analyze our data directly. In addition, our limited search also suggests that user-friendly software packages for variable selection methods in longitudinal studies have been relatively underdeveloped. For penalization methods, only two R packages (PGEE and pgee.mixed) are available, and both packages have focused on the selection of important main effects. The codes for most studies in this area have not even been made publicly available.

To accommodate simultaneously the selection of individual and group structure corresponding to the main lipid effect and interaction effect respectively, we propose a novel penalized variable selection method for longitudinal clustered data. Our method significantly advances the existing penalization methods by considering the interaction effects. Through incorporating the group structure, selection of both main and interaction effects can be efficiently conducted within the generalized estimating equation framework (LIANG and ZEGER (1986)). Furthermore, to facilitate fast computation and reproducible research, we implement the proposed and benchmark methods in the R package (Zhou et al. (2020)). The software package is open-source, and the core module has been developed in C++. The advantage of our method over alternatives has been demonstrated in extensive simulation studies. Analysis of the motivating dataset yields findings with important implications.

2.2 Materials and Methods

2.2.1 Data and Model Settings

Consider a longitudinal study with n subjects and k_i observations measured repeatedly across time for the i^{th} subject ($1 \leqslant i \leqslant n$). Without loss of generality, we set $k_i = k$. Y_{ij} denotes the response for the i^{th} subject at time j ($1 \leqslant j \leqslant k$). $X_{ij} = (X_{ij1}, ..., X_{ijp})^\top$ is the p-dimensional vector of lipid factors. In our study, $E_{ij} = (E_{ij1}, ..., E_{ijq})^\top$ denotes the q-dimensional treatment factor. Suppose that the lipid factors, treatment factors, and their interactions are associated with the longitudinal phenotype through the following model:

$$Y_{ij} = \beta_0 + E_{ij}^\top \beta_1 + X_{ij}^\top \beta_2 + (X_{ij} \otimes E_{ij})^\top \beta_3 + \epsilon_{ij} = Z_{ij}^\top \beta + \epsilon_{ij} \tag{2.1}$$

where $\beta = (\beta_0, \beta_1^\top, \beta_2^\top, \beta_3^\top)^\top$ and $Z_{ij} = (1, E_{ij}^\top, X_{ij}^\top, (X_{ij} \otimes E_{ij})^\top)^\top$ are $(1 + q + p + pq)$-dimensional vectors that represent all the main and interaction effects. The interactions between lipids and treatment factors are modeled through $X_{ij} \otimes E_{ij}$, the Kronecker product of the p-dimensional vector X_{ij}, and the q-dimensional vector E_{ij} within the following form:

$$X_{ij} \otimes E_{ij} = [X_{ij1}E_{ij1}, X_{ij1}E_{ij2}, ..., X_{ij1}E_{ijq}, X_{ij2}E_{ij1}, ..., X_{ijp}E_{ijq}]^\top$$

which is a pq-dimensional vector. β_0 is the intercept. β_1, β_2, and β_3 are unknown coefficient vectors of dimensions q, p, and pq, respectively. We assume that the observations are dependent within the same subject, and independent if they are from different subjects. $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{ik_i})^T$ follows a multivariate normal distribution $N_k(0, \Sigma_i)$, with Σ_i as the covariance matrix for the repeated measure of the i^{th} subject across the k time points.

2.2.2 Generalized Estimating Equations

The joint likelihood function for longitudinally clustered response Y_{ij} is generally difficult to specify. LIANG and ZEGER (1986) developed the generalized estimating equations (GEE) method to account for the intra-cluster correlation. It models the marginal instead of the conditional distribution given the previous observations and only requires a working corre-

lation structure for Y_{ij} to be specified.

The first two marginal moments of Y_{ij} are denoted by $\mathrm{E}(Y_{ij}) = \mu_{ij} = Z_{ij}^T\beta$ and $\mathrm{Var}(Y_{ij}) = \upsilon(\mu_{ij})$, respectively, where υ is a known variance function. Then, the estimating equation for β is defined as:

$$\sum_{i=1}^{n} \frac{\partial \mu_i(\beta)}{\partial \beta} V_i^{-1}(Y_i - \mu_i(\beta)) = 0, \tag{2.2}$$

where $\mu_i(\beta) = (\mu_{i1}(\beta), ..., \mu_{ik}(\beta))^\top$, $Y_i = (Y_{i1}, ..., Y_{ik})^\top$ and V_i is the covariance matrix of Y_i. The first term in (2.2), $\frac{\partial \mu_i(\beta)}{\partial \beta}$, reduces to $Z_i = (Z_{i1}, ..., Z_{ik})^\top$, which corresponds to the $k \times (1 + q + p + pq)$ matrix of the main and interaction effects.

V_i is often unknown in practice and difficult to estimate especially when the number of variance components is large. In GEE, the covariance matrix V_i is specified through a simplified working correlation matrix $R(\eta)$ as $V_i = A_i^{\frac{1}{2}} R(\eta) A_i^{\frac{1}{2}}$, with the diagonal marginal variance matrix $A_i = \mathrm{diag}\{\mathrm{Var}(Y_{i1}), ..., \mathrm{Var}(Y_{ik})\}$. $R(\eta)$ is characterized by a finite-dimensional nuisance parameter η. Commonly adopted correlation structures for $R(\eta)$ can be independent, AR(1), and exchangeable, among others. LIANG and ZEGER (1986) showed that if η can be consistently estimated, the GEE estimator of the regression coefficient is consistent. Furthermore, the consistency holds even when the working correlation structure is misspecified.

2.2.3 Penalized Identification

When the dimensionality of lipid factors is high, the total number of main and interaction effects is even higher. However, only a small subset of important effects is associated with the phenotype, which naturally leads to a variable selection problem. Penalized GEE based methods, including Wang et al. (2012) and Ma et al. (2013), have been proposed for conducting selection under correlated longitudinal responses. However, those published studies focus on the main effects and ignore the interactions. As shown in (2.1), the lipid–environment interactions are modeled on the group level, that is the interaction between all the q treatment factors and the h^{th} lipidomics measurement ($1 \leqslant h \leqslant p$). Such a group structure cannot be accommodated by variable selection methods from existing longitudinal studies. This fact motivates us to develop a method for the interaction analysis of repeated measures data,

termed as interep, with the following penalized generalized estimating equation:

$$Q(\beta) = U(\beta) - \sum_{g=1}^{p} \rho'(|\beta_{2g}|; \lambda_1, \gamma)\text{sign}(\beta_{2g}) - \sum_{h=1}^{p} \rho'(||\beta_{3h}||_{\Sigma_h}; \sqrt{q}\lambda_2, \gamma), \tag{2.3}$$

where $U(\beta)$ is the score equation in GEE and $\rho'(\cdot)$ is the first derivative of the minimax concave penalty (MCP) (Zhang (2010)). Since the environmental factors are usually of low dimension and are predetermined as important ones, they are not subject to penalized selection. $U(\beta)$ is defined as:

$$U(\beta) = \sum_{i=1}^{n} Z_i^T V_i^{-1}(Y_i - \mu_i(\beta)),$$

and the MCP can be expressed as:

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

where λ is the tuning parameter and γ is the regularization parameter. The first derivative function of the MCP penalty is:

$$\rho'(t; \lambda, \gamma) = (\lambda - \frac{t}{\gamma}) I(t \leq \gamma\lambda).$$

MCP can be adopted for the regularized selection on both individual and group level effects. It is fast, continuous, and nearly unbiased (Zhang (2010)).

In (2.3), the vector $\beta_2 = (\beta_{21}, ..., \beta_{2p})^\top$ denotes the regression parameters for all the p lipid factors. $\beta_3 = (\beta_{31}^\top, .., \beta_{3p}^\top)^\top$, which denotes the regression parameters for lipid–environment interactions, is a vector of length pq. β_{3h} is a vector of length q ($h = 1, 2, ..., p$), corresponding to the interactions between the h^{th} lipid feature and the environment factors. With the control as the baseline, the environment factors have been formulated as a group of dummy variables. With high-dimensional main and interaction effects, penalization is critical for the identification of important effects out of the large number of candidates. In the penalized generalized estimating equation (2.3), the first penalty term adopts MCP directly to conduct

the selection of main lipid effects on the individual level. The second penalty, in the forms of group MCP, imposes shrinkage on the product between the lipid factors and dummy variable group, which corresponds to the lipid–environment interactions. The group level selection of interaction effects is consistent with the mechanism of creating the dummy variable group of environmental factors. Note that such a rationale of formulating the penalized generalized estimating equation (2.3) is deeply rooted in group LASSO (Yuan and Lin (2006)).

In particular, λ_1 and λ_2 in (2.3) are tuning parameters. $\rho'(||\beta_{3h}||_{\Sigma_h}; \sqrt{q}\lambda_2, \gamma)$ is the group MCP penalty that corresponds to the interactions between the h^{th} $(h = 1, 2, ..., p)$ lipid factor and the q environment factors. The empirical norm $||\beta_{3h}||_{\Sigma_h}$ is defined as: $||\beta_{3h}||_{\Sigma_h} = (\beta_{3h}^\top \Sigma_h \beta_{3h})^{1/2}$ with $\Sigma_h = n^{-1} B_h^\top B_h$. $B_h = Z[, (2+q+p+(h-1)\times q) : (1+q+p+h\times q)]$, and it contains the q columns in Z that correspond to the interactions from the h^{th} lipid factor with the q environment factors.

A variety of penalized variable selection methods for high-dimensional longitudinal data have been developed in the past two decades for analyzing high-dimensional omics data, such as gene expressions, single nucleotide polymorphisms (SNPs), and copy number variations (CNVs) (Ma et al. (2013); Wang et al. (2012)). However, lipidomics data have been rarely investigated by using high-dimensional variable selection methods. We developed a package, (interep `https://cran.r-project.org/package=interep`) that incorporates our recently developed penalization procedures to conduct interaction analysis for high-dimensional lipidomics data with repeated measurements (Zhou et al. (2020)).

The uniqueness of the proposed study lies in accounting for the group structure of lipid–environment interactions through penalized identification. Therefore, the main lipid effects and lipid–environment interactions are penalized on individual and group levels, separately, which leads to a formulation of both MCP and group MCP penalties. Although our model has been motivated from a specific lipidomics profiling study in weight controlled mice (King et al. (2015)), it can be readily extended to accommodate more general cases in interaction studies where the environmental factors are not dummy variables formulated from the ANOVA setting. In such a case, for each lipid factor, the main lipid effects and lipid–environment interactions form a group, with the leading component of the group being a

vector of 1s. As not all the effects in the group are expected to be associated with the phenotype, a sparse group type of variable selection is demanded. Such a formulation has been investigated in survival analysis (Wu et al. (2018)), but not in longitudinal studies yet. With a simple modification of our model to penalize the main and interaction effects on the individual and group level simultaneously, the proposed one becomes a penalized sparse group GEE model and can be adopted to handle general environmental factors in high-dimensional cancer genomics studies.

2.2.4 Computational Algorithms

We developed an efficient Newton–Raphson type of algorithm to obtain the penalized estimate $\hat{\beta}$. Starting with an initialized value, we can solve the penalized GEE iteratively. The estimated $\hat{\beta}^{(d+1)}$ in the $(d+1)^{\text{th}}$ iteration can be solved as:

$$\hat{\beta}^{(d+1)} = \hat{\beta}^{(d)} + [T^{(d)} + nW^{(d)}]^{-1}[U^{(d)} - nW^{(d)}\hat{\beta}^{(d)}], \tag{2.4}$$

where $U^{(d)}$ is the score function expressed in terms of $\hat{\beta}^{(d)}$ at the d^{th} iteration and $T^{(d)}$ is the corresponding first derivative function of $U^{(d)}$:

$$T^{(d)} = \sum_{i=1}^{n} Z_i^T V_i^{-1} Z_i,$$

which is also a function of $\hat{\beta}^{(d)}$. The MCP penalty was imposed on both the individual level (main lipid effects) and group level (lipid–environment interactions). Therefore, $W^{(d)}$ is a diagonal matrix that contains the first derivative of the MCP penalty for the lipid factors and the first derivative of the group MCP penalty for the lipid–environment interactions. We define $W^{(d)}$ as:

$$W^{(d)} = \text{diag}\{\underbrace{0, ..., 0}_{1+q}, \frac{\rho'(|\hat{\beta}_{21}^{(d)}|; \lambda_1, \gamma)}{\epsilon + |\hat{\beta}_{21}^{(d)}|}, ..., \frac{\rho'(|\hat{\beta}_{2p}^{(d)}|; \lambda_1, \gamma)}{\epsilon + |\hat{\beta}_{2p}^{(d)}|},$$
$$\frac{\rho'(||\hat{\beta}_{31}^{(d)}||_{\Sigma_1}; \sqrt{q}\lambda_2, \gamma)}{\epsilon + ||\hat{\beta}_{31}^{(d)}||_{\Sigma_1}}, ..., \frac{\rho'(||\hat{\beta}_{3p}^{(d)}||_{\Sigma_p}; \sqrt{q}\lambda_2, \gamma)}{\epsilon + ||\hat{\beta}_{3p}^{(d)}||_{\Sigma_p}}\},$$

where ϵ is a small positive number set to 10^{-6} to avoid the numerical instability when the denominator is zero. The first (1 + q) elements on the diagonal of W are zero, suggesting that there is no shrinkage imposed on the coefficients for the intercept and the environmental factors. We can use $nW\hat{\beta}$ and nW to approximate the first derivative function of MCP in the penalized score equation and the second derivative function of the MCP penalty, respectively. Given a fixed tuning parameter, the regression parameter $\hat{\beta}^{(d+1)}$ can be updated iteratively till convergence. The stopping criterion is that the L1 norm for the L1 difference between two consecutive iterations is less than 10^{-3}, and convergence can usually be achieved within 10 iterations.

There are two tuning parameters λ_1 and λ_2 and a regularization parameter γ. λ_1 controls the sparsity of lipid factors, and λ_2 determines sparsity among lipid–environment interactions. We chose the optimal tuning parameters λ_1 and λ_2 using five-fold cross-validation in both the simulation study and real data analysis. The regularization parameter γ was obtained via a data driven approach. In our numerical study, we examined a sequence of values, such as 1.8, 3, 4.5, 6, and 10, suggested by published studies, and found that the results were not sensitive to the choice of the value of γ, and then set the value at 3. We split the dataset into five equally sized subsets and took four of them as the training dataset, leaving the last subset as the testing dataset. The penalized estimates were obtained from the training data, and then, prediction performance was evaluated on the testing data. A joint search over a two-dimensional grid of (λ_1, λ_2) was conducted to find the optimal pair of tuning parameters.

Given fixed tuning parameters, we implemented the algorithm as follows:

(1) Set the initial coefficient vector $\beta^{(0)}$ using LASSO;

(2) Update $\beta^{(d+1)}$ using equation (2.4) at the $(d+1)$th interation;

(3) Repeat Step (2) until the convergence criterion is satisfied.

In our study, we considered the methods considering both lipid main effects and lipid–environment interactions with exchangeable working correlation (A1), AR(1) working correlation (A2), and independence working correlation (A3). For comparison with the methods that cannot accommodate the identification of lipid–environment interactions, we also in-

cluded A4–A6, which incorporate the exchangeable, AR(1), and independence working correlation, respectively. The alternative methods A4–A6 do not ignore the interaction effects. Instead, they treat the interaction effects individually, so the group structure considered in A1–A3 does not exist. We computed the CPU running time for 100 replicates of simulated lipidomics data with $n = 250, \rho = 0.8, p = 75$ (with a total dimension of 304) and fixed tuning parameters on a regular laptop for A1–A6, which can be implemented using our developed package: (interep `https://cran.r-project.org/package=interep`) Zhou et al. (2020). The CPU running time in seconds was 48.8 (A1), 40.2 (A2), 29.0 (A3), 49.3 (A4), 39.7 (A5), and 27.9 (A6), respectively.

2.3 Results

2.3.1 Simulation

We evaluated the performance of all six methods (A1–A6) through extensive simulation studies. Among them, A1–A3 were developed for accommodating the interaction structures with different working correlations, while A4–A6 were only focused on the identification of main effects so the structure of the group level interaction effects were not respected. Note that there are existing studies that can also achieve the selection of main effects in longitudinal studies. For example, Wang et al. Wang et al. (2012) adopted the smoothly clipped absolute deviation (SCAD) penalty for conducting the selection of main effects. Since the MCP is incorporated as the baseline penalty in A1–A3, A4–A6 have thus been developed based on MCP and used as benchmark methods for comparison.

The responses were generated from the model (2.2) with sample size $n = 250$ and 500. The number of time points k was set to five. The dimensions for lipid factors X_{ij} were $p = 75$, 150 and 300. With $q = 3$ for E_{ij}, we first simulated a vector of length n from the standard normal distribution. A group of three binary dummy variables for environmental factors could then be generated after dichotomizing the vector at the 30^{th} and 70^{th} percentiles. In addition, the lipids were simulated from a multivariate normal distribution with mean zero

and the AR1 covariance matrix with marginal variance one and auto-correlation coefficient 0.5. We simulated the random error ϵ from a multivariate normal distribution by assuming a zero mean vector and an AR1 covariance structure with $\rho = 0.5$ and 0.8. Note that when considering the interactions, the actual dimensionality was much larger than p. For instance, given $n = 250$, $p = 150$, and $q = 3$, the total dimension for all the main and interaction effects was 604.

The coefficients were simulated from $U[0.4, 0.8]$ for 17 nonzero effects, consisting of the intercept, 3 environmental dummy variables, 4 lipid main effects, and 3 groups of lipid–environment interactions (9 interaction effects). We generated 100 replicates for the four settings: (1) $n = 250$ and $p = 75$, (2) $n = 250$ and $p = 150$, (3) $n = 500$ and $p = 150$, and (4) $n = 500$ and $p = 300$. All the rest of the coefficients were set to zero. For each setting, we considered two correlation coefficients ($\rho = 0.5$ and 0.8) for the random error. The number of true positives (TP) and false positives (FP) was recorded.

In addition to identification results, we also calculated the estimation accuracy in terms of the difference between estimated and true coefficients. In particular, the mean squared error corresponding to the true nonzero coefficients and true zero coefficients (for noisy effects) were termed as MSE and NMSE, respectively. The total mean squared error for the coefficient vector, or TMSE, is computed as:

$$\text{TMSE} = \frac{1}{100}\sum_{r=1}^{100} ||\hat{\beta}^{(r)} - \beta||^2/p_\beta$$

where p_β is the dimension of β and $\hat{\beta}^{(r)}$ is the estimated value of β in the r^{th} simulated dataset. MSE and NMSE were calculated in a similar way as for TMSE.

Identification results of the six methods (A1–A6) are tabulated in Tables A.1–A.4. In general, A1–A3, which account for both the lipid main effects and lipid–environment interactions, had better performance than A4–A6, which only accommodated the main effects. For example, in Table A.1, given $n = 250$, $\rho = 0.5$, $p = 75$, the actual dimension is 304. A1 identified 14.5 (sd 1.9) nonzero effects out of all the 17 true positives, with a relatively small number of false positives of 4.8 (sd 3.1). On the other hand, A4 identified a smaller number

of true positives, 1.3 (sd 1.5), with a larger number of false positives, 6.6 (sd 4.2). Among the identified effects, A1 identified 7.4 (sd 1.5) interactions, with 3.1 (sd 2.6) false positives. A4 identified a smaller TP of 6.1 (sd 1.1) and a higher FP of 5.1 (sd 3.3) of the lipid–environment interactions. We could observe that the difference in identification performance between A1 and A4 came mainly from the interaction effects, which was due to the fact that A4 could not accommodate the group level selection corresponding to the lipid–environment interactions. As the dimension increased, A1 outperformed A4 more significantly. For instance, in Table A.4, the overall dimension for $n = 500$, $\rho = 0.8$, $p = 300$ is 1204. A1 had a TP of 15.9 (sd 1.2) and an FP of 3 (sd 2.6), while A4 had a smaller TP 14.5 (sd 1.2) and a higher FP 4.5 (sd 3.0). Figures A.1 and A.2 are plotted based on the identification results from Tables A.1–A.4. We can observe that overall, A1–A3 outperformed A4–A6 with a higher TP and a lower FP under each setting.

In terms of estimation accuracy, A1–A3 also had a better performance compared with A4–A4, as shown in Tables A.5 and A.6. For the panel corresponding to $n = 250$, $\rho = 0.5$, and $p = 75$ in Table A.5, the mean squared error for the nonzero coefficients of A1 was 0.1055, which was less than half of that of A4 (0.2321). Besides, A1 also had a smaller total mean squared error (TMSE). All the pieces of evidence suggested that A1 had higher estimation accuracy than A4. We can observe the pattern for the rest of the four methods. As the dimension increased to $n = 500$, $\rho = 0.8$, and $p = 300$ (so the total dimension was 1204) in Table A.6, the MSE of A1 (0.0688) was also smaller than that of A4 (0.1949). There were no obvious differences in NMSE among these settings.

Another important conclusion we make from the simulation study is that, for the methods that differ only in working correlation, i.e., A1 (exchangeable), A2 (AR1), and A3 (independence), there was no significant difference in terms of either identification or estimation accuracy, as shown by Tables A.1–A.6, as well as Figures A.1 and A.2. Such an observation suggests that the proposed methods under the GEE framework were robust to the misspecification of the working correlation, and this is consistent with the conclusions from main effects only models in longitudinal studies (Cho and Qu (2013)).

To mimic the sample size and number of lipid factors in the case study, we also conducted

a simulation in settings with $n = 60$, $p = 30$, and $q = 3$. Therefore, the overall dimension of main and interaction effects was 124. The coefficients were generated from $U[1.4,1.8]$ for 17 nonzero effects. The identification and prediction results are summarized in Tables A.7 and A.8 in the Appendix, respectively. Consistent patterns were observed. For example, in terms of identification, under $\rho = 0.5$, A1 had a higher TP of 13.6 (sd 2.5) compared to the 11.1 (sd 2.6) of A4, and a lower FP of 4.7 (sd 2.7), compared to the FP of 5.4 (sd 2.8) identified by A4.

Evaluation of all the methods, especially A1–A3, was also conducted when the true underlying model was misspecified. We generated the response (phenotype) from a main effect only model with eight true main effects when $n = 250$, $p = 75$, $\rho = 0.8$ with a total dimension of 304. Results are provided in Table A.11. When the interaction effects did not exist, A1 had only identified a very small number of false interaction effects, with 0.7 (sd 1.7) false positives. A2–A6 performed similarly in terms of identifying false interaction effects. All six methods identified a comparable number of true main effects. Overall, all methods had similar performance in identification, as well as prediction, when the data generating model had only main effects. Such a phenomenon is reasonable by further examining the results in Table A.1. We found that the major difference between A1–A3 and A4–A6 was due to the identification of interaction effects. Therefore, when only main effects were present, all the methods had comparable performances.

Penalized regression and hypothesis testing are two related, but distinct aspects in statistical analysis. The proposed study was not aimed at developing test statistics, computing the power functions, and assessing the control of type 1 error, so these statistical test related results are not available, just like most of the studies on penalized regression. Recently, efforts devoted to bridging the two areas have been mainly restricted to linear models under high-dimensional settings (Lee et al. (2016); Lockhart et al. (2014); Taylor and Tibshirani (2015)). Extensions to interaction models have not been reported so far. In particular, we are not aware of results reported for longitudinal models. Nevertheless, we conducted the simulation by assuming the null model and tabulate the identification results in Table A.12. The results should be interpreted as identification with misspecified models. As we observed,

under the null model, all six methods led to a very small number of false positives.

To assess the consistency of variable selection in longitudinal settings, we carried out the stability selection (Meinshausen and Bühlmann (2010)) under $n = 250$, $p = 75$, and $\rho = 0.8$. Each time, we selected 200 out of the total of 250 subjects without replacement and then conducted selection. The process was repeated 100 times, which yielded a proportion of selected effects. Larger proportions of being selected suggested stable results. Stability selection is well known for assessing the stability of penalized selection, and it alleviates the concern that the effects have only been identified by chance. We investigate the selection proportions of the 17 true main and interaction effects for all six methods in Table A.13. A1 identified 14 true effects with proportions above 70%, which is consistent with the results shown in the lower panel of Table A.1, where 13.7 TPs (sd 2.3) were identified. Such a consistent pattern can be observed across all six methods.

Although no consensus on the optimal criterion of selecting tuning parameters has been reached so far, cross-validation is perhaps the most well accepted criterion to select tuning parameters in the community of high-dimensional data analysis (Fan and Lv (2010); Wu and Ma (2015)). To further justify its appropriateness, under the setting of $n = 250$ and $p = 75$, we performed the analysis by selecting tuning parameters using an independently generated testing dataset with a sample size of 1000 and $p = 75$. The models were fitted on the training dataset, and prediction was assessed based on the independently generated testing dataset, so no data were used in training the model. The identification and prediction results are tabulated in Tables A.14 and A.15, respectively. A comparison to Tables A.1 and A.5 demonstrates that the results obtained by cross-validation and validation were very close.

2.3.2 Real Data Analysis

We applied the proposed and alternative methods on a dataset from one of our previous studies in animal models (King et al. (2015)). In the study, 60 female CD-1 mice were assigned to four different treatment groups, which were control (ad libitum feeding and sedentary), AE (exercise and ad libitum feeding), PE (exercise and pair feeding), and DCR

(sedentary and 20% dietary calorie restriction). The phenotype of interest was mice's body weight, which was measured every week for 10 weeks. Mice were sedentary and given ad libitum feeding in the control group, where they could eat as much as they wanted without doing treadmill exercises. In the AE group, mice received ad libitum feeding and ran on the treadmill every day at a speed of 0.5 mph, 1 hour per day, and 5 days a week, while mice in the PE group did the same exercise, but were given the same amount of diet as the mice in the control group. Mice in the DCR group had 20% less calorie intake than the control group, but they had the same intake of protein, vitamins, and minerals. The composition of 176 plasma neutral lipid species of interest was measured. In the current study, we only focused on diacylglycerols. In addition, the diacylglycerol lipid species that have a majority of samples lower than the detection limits were excluded so there were 31 diacylglycerols. In total, there were 31 lipid main effects and 93 lipid–environment interactions.

Using the method A1 (interep with the exchangeable working correlation) as shown in Table A.7, we identified seven lipid species that had different effects in weight control of mice (AE, PE, or DCR) on body weight compared to those of the control mice. Among them, C20:1/16:1 and C20:1/20:4 had negative interactions in AE mice, where C denotes carbon. For the lipid species of C20:1/16:1, $C_{39}H_{76}O_5N$, the regression coefficient was -2.9145 for AE mice. That is, mice with an increased amount of C20:1/16:1 tended to have a lower body weight compared to that of the control. In the AE mice, both C16:0/C16:0 and C22:6/C18:1 had strong positive associations with body weights. It is interesting that C16:0/C16:0 were negatively associated with body weight in both PE and DCR mice. C16:0 is also called palmitic acid and is one of most common saturated fatty acids. Increased consumption of palmitic acid is associated with higher risk of cardiovascular disease, type 2 diabetes, and cancer (Briggs et al. (2017)). The negative association of C16:0/16:0 and body weight in DCR and PE suggests that when the calories of the diet are restricted, the accumulation of saturated fat in the body actually decreased compared to the control. Another lipid that is negatively associated with body weight in DCR and PE mice is C18:1/16:1. The lipids that were positively associated with body weight in PE were C18:2/C16:1, C20:1/C16:1, and C22:6/C18:1. All species contain unsaturated fatty acids. Among them, C22:6 is one of the

omega-3 polyunsaturated fatty acids (PUFA). In DCR, the two lipids that were positively associated with body weight were C18:2/16:1 and C20:1/20:4. Both fatty acids C18:2 and C20:4 were PUFA. The results seem to be consistent with our previous finding that exercise with paired feeding may increase the amount of PUFA in phospholipids in mice skin (Ouyang et al. (2010)).

In addition, we adopted A4 to analyze the lipid data. A4 also had the exchangeable working correlation, but it could not conduct group level selection of the lipid–environment interactions. The identification results are tabulated in Table A.8. Note that the selection of interactions with individual dummy environment factors was not consistent with the formulation of the lipid–environment interactions. In terms of prediction, A1 had a smaller prediction error (4.04) than that of A4 (4.97).

2.4 Discussion

Investigation of the potential roles of lipids in the regulation and control of cellular function and the interactions between lipids and environmental factors are very important in the understanding of physiology and disease processes. Traditionally, the analyses mostly focus on the total amount of a particular type of lipid, such as total triglyceride, total cholesterol, and omega-3 fatty acid. With the recent advances in instrumental technology, it is feasible to analyze quantitatively a broad range of lipid species in a single platform (Bowden et al. (2017); Jiang (2012); King et al. (2015); Stegemann et al. (2014); Zhou et al. (2012)). The vast arrays of data generated in lipid profiling studies bring challenges to the statistical analysis of lipidomics data (Checa et al. (2015); Kujala and Nevalainen (2015); Wenk (2005)).

In this study, we proposed a penalized variable selection method to identify important lipid–environmental effects in longitudinal studies. Some statistical methods have already been reported for lipidomics studies, including the marginal test and variable selection methods (Checa et al. (2015); Jiang (2012); King et al. (2015); Kujala and Nevalainen (2015)); however, they cannot be directly extended to longitudinal studies. On the other hand, existing variable selection methods for longitudinal data have been predominately developed

for the identification of main effects and cannot accommodate the group level interaction structure unique to our studies. Both the simulation and case study have convincingly demonstrated the merit of the proposed interep over alternatives.

We selected tuning parameters based on cross-validation. A further investigation of different tuning criteria is interesting, but beyond the scope of this study, especially given the fact that many well known variable selection methods in longitudinal studies, such as Wang et al. (2012), have been conducted using cross-validation. To facilitate a fair cross-comparison with existing relevant studies, we believe it is reasonable to adopt cross-validation to choose tuning parameters. Note that the aforementioned stability selection analysis also partially justifies the usage of cross-validation. We acknowledge that other criteria for selecting tunings, such as double cross-validation (Filzmoser et al. (2009)), could be a potential reliable choice. However, as it is not a widely accepted tuning criterion for high-dimensional data analysis and has not been adopted in any longitudinal studies so far, we postpone the investigation to the future.

Interaction studies have been historically pursued by statisticians (Cordell (2002)). Within the high-dimensional scenario, accounting for such a complex structure, in both gene–gene (G $\times$ G) and gene–environment (G $\times$ E) interaction studies, is challenging, but also rewarding (Wu and Ma (2018)). The proposed study is among the first to investigate penalized identification of lipid–environment interactions in longitudinal studies. Both the simulation study and case study yielded interesting findings. G $\times$ G interaction is computationally more challenging than G $\times$ E interactions since both main effects involved in the interactions are of high dimensionality. Following the representative G $\times$ G interaction studies (Bien et al. (2013); Choi et al. (2010)), we can extend the proposed study to lipid–lipid interactions, which has not been investigated in longitudinal studies so far. Besides, when multi-omics measurements are available, it is also of great interest to examine interaction effects through multi-omics integration studies in the longitudinal setting (Li et al. (2019); Wu et al. (2019)).

The proposed model can also be estimated using the quadratic inference functions (QIF). GEE relies on the working correlation matrix $R(\eta)$, and it enables us to find the consistent estimator of the regression parameter if consistent estimators of the nuisance parameters η

can be obtained. However, consistent estimators of η do not always exist in some cases. QIF has been proposed to avoid explicit estimation of the nuisance parameters by assuming the inverse of the working correlation matrix $R(\eta)$ can be approximated by a linear combination of a class of base matrices (Cho and Qu (2013); Qu et al. (2000)). Thus, QIF is robust to the misspecification of the working correlation.

In this paper, we are interested in the identification of lipid-treatment (or environment) interactions through penalization. The success of set based analysis, including those for the gene set (Schaid et al. (2012)) and SNP set (Wu and Cui (2014); Wu et al. (2012)), has tremendously motivated the development of statistical methods for G $\times$ E interactions from marginal analyses (Mukherjee et al. (2012); Wu and Cui (2013)) to penalization methods (Wu et al. (2014, 2018, 2020)). Our model can be potentially extended in the following aspects. First, as data contamination and outliers have been widely observed in repeated measurements, robust variable selection methods in G $\times$ E interaction studies Wu et al. (2018, 2015); Wu and Ma (2019); Xu et al. (2018) can be extended to longitudinal settings. Second, recently, multiple Bayesian methods have been proposed for pinpointing important G $\times$ E interaction effects Ahn et al. (2013); Li et al. (2015); Ren et al. (2020). Within the framework of analyzing repeated measurements, Bayesian variable selection for interactions has not been extensively examined. Besides, test-based approaches on the analysis of longitudinal data have also been established. For example, Wang and Zhang (2010) developed a set of nonparametric tests for longitudinal DNA copy number data. Investigations of all these possible directions will be postponed to the near future.

Chapter 3

Sparse group variable selection for gene–environment interactions in the longitudinal study

3.1 Introduction

Longitudinal data have arisen in biomedical studies, clinical trials and many other areas with measurements on the same subject being taken repeatedly over time. Substantial efforts have been made to account for the correlated nature of repeated measures when modelling longitudinal data (Verbeke et al. (2014)). Recently, the importance of longitudinal design in genetic association studies has been increasingly recognized (Li et al. (2015); Sitlani et al. (2014)). As the main objective of conducting association analysis is to identify key signals associated with the disease phenotypes from a large number of genetic variants (e.g. single nucleotide polymorphisms, or SNPs) (Cordell and Clayton (2005),Wu et al. (2012)), the longitudinal design yields novel insight to elucidate the genetic control for complex disease traits over cross-sectional designs.

This study has been partially motivated by analyzing the high dimensional SNP data with longitudinal trait from the Childhood Asthma Management Program (CAMP). CAMP

has been launched in early 1990s and became the largest randomized longitudinal clinical trial developed to investigate the long term influences of Budesonide and Nedocromil, the anti-inflammatory therapy, on children with mild to moderate asthma (Childhood Asthma Management Program Research Group (1999, 2000); Covar et al. (2012)). Including placebo, the treatment thus has three levels. Our primary disease phenotype of interest is the forced expiratory volume in one second (FEV1), a repeatedly measured indicator on whether the lung growth of children has improved or not. Here, with SNPs as G factors and treatment, age and gender as environmental (E) factors, we are interested in dissecting the gene-environment (G×E) interactions under the longitudinal trait FEV1. As the number of main and interaction effects is much larger than the sample size, penalized variable selection has become a powerful tool for interaction studies (Zhou et al. (2021)).

To date, penalization methods for interaction studies have been mainly proposed under continuous disease traits, categorical status and cancer prognostic outcomes (Zhou et al. (2021)). With the longitudinal phenotype, where the response on the same subject are repeatedly measured over a set of units (e.g. time), penalized regression methods are relatively underdeveloped for interaction analyses. In fact, our limited literature search indicates that majority of the variable selection methods in longitudinal studies can only accommodate main effects. For example, Wang et al. (2012) has developed a penalized generalized estimating equation (GEE) for the identification of important main effects associated with longitudinal response. Also within the GEE framework, Ma et al. (2013) has considered an additive, partially linear model with variable selection on the main effect only. On the other hand, Cho and Qu (2013) has conducted penalized variable selection in the main effect model based on the quadratic inference function (QIF), and showed that penalized QIF outperforms penalized GEE under a variety of settings.

The relative underdevelopment of variable selection methods for longitudinal interaction studies is partially due to the challenge in accommodating structured sparsity within either the GEE or QIF framework. Consider the interaction model involving p genetic factors and q environmental factors, where the interactions are denoted by pq product terms. Such a model serves as the umbrella framework for a large number of G×E studies (Zhou et al. (2021)).

For one G factor, its main effect and interactions with the q environmental factors form a group of q+1 terms. Hence, to determine whether the genetic factor is associated with the phenotype, a group level selection should be conducted. Furthermore, if the genetic factor is associated with the phenotype, an individual level selection within the group is necessary. Overall, identification of important G×E interactions essentially amounts to a sparse group (or bi-level) variable selection problem, which becomes even more challenging when a large number of genetic factors are jointly analyzed under repeatedly measured phenotypes.

The aforementioned interaction model serves as an umbrella framework for a large number of G×E interaction studies (Zhou et al. (2021)). On a broader scope, sparse group (or bi–level) structure plays a very important role in high dimensional variable selection with structured sparsity (Breheny and Huang (2009); Friedman et al. (2010); Simon et al. (2013)). Nevertheless, the bi-level sparsity has not been examined in existing longitudinal studies by far. Our study is novel in that it is among the first to develop the sparse group regularized variable selection for high dimensional longitudinal studies. Specifically, based on the quadratic inference function (QIF), we propose a sparse group variable selection method for simultaneous selection of main and interaction effects on both the group and individual levels in G×E studies. The Minimax concave penalty (MCP) is adopted as the baseline penalty function to achieve regularized identification (Zhang (2010)).

Besides the QIF and GEE, Bayesian analysis and mixed models are also the major tools for repeated measurement studies (Fan and Li (2012); Li et al. (2015)). Our literature survey shows that the longitudinal bi-level variable selection has not been developed within the two frameworks yet. Therefore, a direct comparison is not possible. While the QIF is robust to model misspecification as well as at least a small portion of data contamination and outliers (Cho and Qu (2013); Qu and Song (2004)), the robustness of Bayesian and mixed model based high dimensional longitudinal analyses remains unanswered. For example, specifying the Bayesian hierarchical model in longitudinal studies generally involves employing a covariance structure, such as the first-order autoregressive (AR1) structure (Li et al. (2015)), when the truth is not known *a priori*. It is not clear to what extent these methods are robust to model misspecification. Besides, with the multivariate normal assumption on residual

error, Li et al. (2015) is not robust to phenotypes with long–tailed distributions. Lastly, we have implemented the proposed and alternative methods in R package *springer* (Zhou et al. (2021)). The core modules of the R package have been developed in C++ to guarantee fast computations.

3.2 Statistical Method

3.2.1 Data and Model Settings for Longitudinal G×E Studies

We consider a longitudinal scenario where there are n subjects and k_i measurements repeatedly taken over time on the ith subject ($1 \leqslant i \leqslant n$). There are correlations among measurements on the same subject, and independence is assumed for measurements between different subjects. We denote Y_{ij} as the phenotypic response of the ith subject at the jth time point ($1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant k_i$). $X_{ij} = (X_{ij1}, ..., X_{ijp})^\top$ denotes a p-dimensional vector of genetic factors and $E_{ij} = (E_{ij1}, ..., E_{ijq})^\top$ is a q-dimensional vector of environmental factors in the study. Consider the following model:

$$\begin{aligned} Y_{ij} &= \mu_{ij} + \epsilon_{ij} \\ &= \alpha_0 + \sum_{u=1}^{q} \alpha_u E_{iju} + \sum_{v=1}^{p} \gamma_v X_{ijv} + \sum_{v=1}^{p}\sum_{u=1}^{q} h_{uv} E_{iju} X_{ijv} + \epsilon_{ij} \\ &= \alpha_0 + \sum_{u=1}^{q} \alpha_u E_{iju} + \sum_{v=1}^{p} (\gamma_v + \sum_{u=1}^{q} h_{uv} E_{iju}) X_{ijv} + \epsilon_{ij} \\ &= \alpha_0 + \sum_{u=1}^{q} \alpha_u E_{iju} + \sum_{v=1}^{p} \eta_v^\top Z_{ijv} + \epsilon_{ij}, \end{aligned} \tag{3.1}$$

where α_0, α_u's, γ_v's and h_{uv}'s are the coefficients of the intercept, environmental factors, genetic factors and G×E interactions, respectively. We define $\eta_v = (\gamma_v, h_{1v}, ..., h_{qv})^\top$ and $Z_{ijv} = (X_{ijv}, E_{ij1}X_{ijv}, ..., E_{ijq}X_{ijv})^\top$. Z_{ijv} is a $(1+q)$-dimensional vector that represents the main effect of the vth genetic factor and its interactions with the q environmental factors. We assume the random error $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{ik_i})^\top \sim N_{k_i}(0, \Sigma_i)$, which is a multivariate normal distribution with Σ_i as the covariance matrix for the k_i repeated measurements of the ith

subject. Without loss of generality, we let $k_i = k$. Collectively, we can write $\alpha = (\alpha_1, ..., \alpha_q)^\top$, $\eta = (\eta_1^\top, ..., \eta_p^\top)^\top$, and $Z_{ij} = (Z_{ij1}^\top, ..., Z_{ijp}^\top)^\top$. The vector η is of length $p \times (1 + q)$. Then model (3.1) can be rewritten as:

$$Y_{ij} = \alpha_0 + E_{ij}^\top \alpha + Z_{ij}^\top \eta + \epsilon_{ij}.$$

Denote $(1 + q + p(q + 1))$-dimensional vectors $\beta = (\alpha_0, \alpha^\top, \eta^\top)^\top$ and $W_{ij} = (1, E_{ij}^\top, Z_{ij}^\top)^\top$, then model (3.1) becomes:

$$Y_{ij} = W_{ij}^\top \beta + \epsilon_{ij}.$$

While the phenotype, the G factors and E factors all have repeated measurements in the above model formulation for longitudinal G×E studies, such a formulation allows for flexible model setups. For example, it also works when only one of two types of factors is longitudinal, or neither of them have been repeatedly measured. The time-varying gene expression is a representative example of the G factor. In this study, the G factors are SNPs that do not change over time.

3.2.2 Quadratic Inference Function for Longitudinal G×E Interactions

Modeling longitudinal response Y_i is challenging, as the full likelihood function is generally difficult to specify, due to the intra-subject/cluster correlation. To overcome such an issue, LIANG and ZEGER (1986) has proposed the generalized estimating equations (GEE), where a marginal model with only the working correlation for Y_{ij} needs to be specified. The first two marginal moments of Y_{ij} are given as $\mathrm{E}(Y_{ij}) = \mu_{ij} = W_{ij}^T \beta$, and $\mathrm{Var}(Y_{ij}) = \delta(\mu_{ij})$ respectively, and $\delta(\cdot)$ is a known variance function. The score equation for GEE in the G×E setting is defined as:

$$\sum_{i=1}^{n} \frac{\partial \mu_i(\beta)}{\partial \beta} V_i^{-1} (Y_i - \mu_i(\beta)) = 0,$$

where $\mu_i(\beta) = (\mu_{i1}(\beta), ..., \mu_{ik}(\beta))^\top$. The first term in the equation, $\frac{\partial \mu_i(\beta)}{\partial \beta}$, reduces to $W_i = (W_{i1}, ..., W_{ik})^\top$. We define $Y_i = (Y_{i1}, ..., Y_{ik_i})^\top$ and $V_i = A_i^{\frac{1}{2}} R_i(\nu) A_i^{\frac{1}{2}}$ is the covariance matrix of the ith subject, with $A_i = \text{diag}\{\text{Var}(Y_{i1}), ..., \text{Var}(Y_{ik})\}$. $R_i(\nu)$ is a 'working' correlation matrix that describes the pattern of measurements and can be characterized by a finite dimensional intra–subject/cluster parameter ν. The solution of the score equation, $\hat{\beta}$, is the GEE estimator.

LIANG and ZEGER (1986) has shown that when the intra–subject parameter from the working correlation matrix can be consistently estimated, GEE yields consistent estimates of regression coefficients even if the correlation structure is misspecified. Nevertheless, the GEE estimator is not efficient under such misspecification, let alone the nonexistence of the consistent estimator for the intra–class parameter. Moreover, the GEE estimator is highly sensitive to even only one outlying observation. To overcome the drawback of GEE, Qu et al. (2000) has proposed the method of quadratic inference functions (QIF), where a direct estimation of the correlation parameter is not needed, and the corresponding estimator remains optimal even under structure misspecification. In addition, Qu and Song (2004) have further shown that QIF is more robust than GEE in the presence of outliers and data contamination, and is thus a preferable method over GEE.

In the current G×E settings, the QIF method approximates the inverse of $R(\nu)$ with a linear combination of basis matrices as $\text{R}(\nu)^{-1} \approx \sum_{t=1}^{m} b_t M_t$, where M_1 is an identity matrix, $M_2, ..., M_m$ are symmetric basis matrices with unknown coefficients $b_1, ...b_m$. Qu et al. (2000) has destribed the choice of the basis matrices $M_2, ..., M_m$ based on the working correlation. With such an approximation, the score equations become

$$\sum_{i=1}^{n} W_i^\top A_i^{-\frac{1}{2}} (b_1 M_1 + ... + b_m M_m) A_i^{-\frac{1}{2}} (Y_i - \mu_i(\beta)). \tag{3.2}$$

Within the framework of QIF, we define $\phi_i(\beta)$, the extended score vector involving the main and interaction effects for the ith subject, as

$$\phi_i(\beta) = \begin{pmatrix} W_i^\top A_i^{-\frac{1}{2}} M_1 A_i^{-\frac{1}{2}} (Y_i - \mu_i(\beta)) \\ . \\ . \\ . \\ W_i^\top A_i^{-\frac{1}{2}} M_m A_i^{-\frac{1}{2}} (Y_i - \mu_i(\beta)) \end{pmatrix}, \tag{3.3}$$

without the estimation of the coefficients $b_1, ...b_m$. Subsequently, the extended score for all subjects is $\overline{\phi_n}(\beta) = \frac{1}{n}\sum_{i=1}^n \phi_i(\beta)$.

It can be observed that the estimation functions for G×E studies (Equation (3.2)) is equivalent to a linear combination of components from the extended score vectors. Based on $\overline{\phi_n}(\beta)$, the extended score of the G×E studies, we define the corresponding quadratic inference function as

$$Q_n(\beta) = \overline{\phi_n}^\top(\beta)\overline{\Omega_n}(\beta)^{-1}\overline{\phi_n}(\beta),$$

where $\overline{\Omega_n}(\beta) = \frac{1}{n}\sum_{i=1}^n \phi_i(\beta)\phi_i(\beta)^\top$. Then the QIF estimator $\hat{\beta}$ for G×E interaction studies can be obtained as $\hat{\beta} = \underset{\beta}{\text{argmin}} Q_n(\beta)$.

3.2.3 Penalized identification of G×E interactions in longitudinal studies

In a typical G×E study, the main objective is to identify an important subset of features out of all the main and interaction effects, which is of a "large p, small n" nature. Therefore, penalized variable selection becomes a natural tool to investigate G×E interactions (Zhou et al. (2021)). With model (3.1), we propose the following penalized quadratic inference function:

$$U(\beta) = Q(\beta) + \sum_{v=1}^{p} \rho(||\eta_v||_{\Sigma_v}; \lambda_1, \gamma) + \sum_{v=1}^{p}\sum_{u=1}^{q+1} \rho(|\eta_{vu}|; \lambda_2, \gamma), \tag{3.4}$$

where the baseline penalty function $\rho(\cdot)$ is a minimax concave penalty, which is defined as $\rho(t; \lambda, \gamma) = \lambda \int_0^t (1 - \frac{x}{\gamma\lambda})_+ dx$ on $[0, \infty)$, with tuning parameter λ and regularization parameter γ (Zhang (2010)). As previously defined, η_v is a coefficient vector of length $q+1$, corre-

sponding to the main effect of the vth SNP and its interactions with the q environment factors. We denote $||\eta_v||_{\Sigma_v}$ as the empirical norm of η_v and η_{vu} as the uth component of $\eta_v (v = 1, ..., p, \text{and } u = 1, ..., q + 1)$.

Our choice of the baseline penalty function is the minimax concave penalty and the corresponding first derivative function of MCP penalty is defined as $\rho'(t; \lambda, \gamma) = (\lambda - \frac{t}{\gamma})I(t \leq \gamma\lambda)$.

Within the current longitudinal setting, identification of important G×E interactions amounts to a bi–level selection problem. In particular, selection on the group level determines whether the genetic factor is associated with the phenotypic response. If the coefficient vector η_v is 0, then the G factor does not have any contribution to the response. Otherwise, an examination on the individual level to further determine the existence of main and interaction effects is necessary. The penalized QIF function (3.4) has been formulated to accommodate individual and group level selection in longitudinal G×E studies with the sparse group MCP penalty function.

In general, the regularized loss functions of penalization problems share the form of "unregularized loss function + penalty function" (Wu and Ma (2015)). In longitudinal studies, popular choices of unregularized loss function include GEE and QIF. Our limited search suggests that existing penalization methods for longitudinal data are mostly focused on main effects, therefore only baseline penalty functions such as LASSO and SCAD are necessary (Cho and Qu (2013); Ma et al. (2013); Wang et al. (2012)). In G×E studies, the interaction structure poses a challenge to accommodate the more complicated bi-level sparsity, which has motivated the proposed study.

3.2.4 Computational Algorithms for Sparse Group QIF

Now, we outline an efficient Newton-Raphson algorithm that iteratively updates parameter estimates $\hat{\beta}$ for the penalized QIF. In particular, at the gth iteration, $\hat{\beta}^{(g+1)}$ can be obtained

based on the estimated coefficient vector $\hat{\beta}^{(g)}$ from the gth iteration as follows:

$$\hat{\beta}^{(g+1)} = \hat{\beta}^{(g)} + [V^{(g)} + nH^{(g)})]^{-1}[P^{(g)} - nH^{(g)}\hat{\beta}^{(g)}], \tag{3.5}$$

where $P^{(g)}$ and $V^{(g)}$ are the first and second order derivative functions of the score function of QIF, respectively. They are given as:

$$P^{(g)} = \frac{\partial Q(\hat{\beta}^{(g)})}{\partial \beta} = 2\frac{\partial \overline{\phi_n}^\top}{\partial \beta}\overline{\Omega_n}^{-1}\overline{\phi_n}(\hat{\beta}^{(g)}),$$

and

$$V^{(g)} = \frac{\partial^2 Q(\hat{\beta}^{(g)})}{\partial^2 \beta} = 2\frac{\partial \overline{\phi_n}^\top}{\partial \beta}\overline{\Omega_n}^{-1}\frac{\partial \overline{\phi_n}}{\partial \beta}.$$

Besides, $H^{(g)}$ is a diagonal matrix consisting of derivatives of both the individual– and group–level penalty functions, which is defined as:

$$\begin{aligned} H^{(g)} = \text{diag}(&\underbrace{0, ..., 0}_{1+q}, \underbrace{\frac{\rho'(||\hat{\eta}_1^{(g)}||_{\Sigma_1}; \sqrt{q+1}\lambda_1, \gamma)}{\epsilon + ||\hat{\eta}_1^{(g)}||_{\Sigma_1}}, ..., \frac{\rho'(||\hat{\eta}_1^{(g)}||_{\Sigma_1}; \sqrt{q+1}\lambda_1, \gamma)}{\epsilon + ||\hat{\eta}_1^{(g)}||_{\Sigma_1}}}_{1+q}, ..., \\ &\underbrace{\frac{\rho'(||\hat{\eta}_p^{(g)}||_{\Sigma_p}; \sqrt{q+1}\lambda_1, \gamma)}{\epsilon + ||\hat{\eta}_p^{(g)}||_{\Sigma_p}}, ..., \frac{\rho'(||\hat{\eta}_p^{(g)}||_{\Sigma_p}; \sqrt{q+1}\lambda_1, \gamma)}{\epsilon + ||\hat{\eta}_p^{(g)}||_{\Sigma_p}}}_{1+q}) + \text{diag}(\underbrace{0, ..., 0}_{1+q}, \\ &\underbrace{\frac{\rho'(|\hat{\eta}_{11}^{(g)}|; \lambda_2, \gamma)}{\epsilon + |\hat{\eta}_{11}^{(g)}|}, ..., \frac{\rho'(|\hat{\eta}_{1(q+1)}^{(g)}|; \lambda_2, \gamma)}{\epsilon + |\hat{\eta}_{1(q+1)}^{(g)}|}}_{1+q}, ..., \underbrace{\frac{\rho'(|\hat{\eta}_{p1}^{(g)}|; \lambda_2, \gamma)}{\epsilon + |\hat{\eta}_{p1}^{(g)}|}, ..., \frac{\rho'(|\hat{\eta}_{p(q+1)}^{(g)}|; \lambda_2, \gamma)}{\epsilon + |\hat{\eta}_{p(q+1)}^{(g)}|}}_{1+q}), \end{aligned}$$

where ϵ is a small positive number adopted to ensure that the denominator is nonzero for zero coefficients and here we set it equal to 10^{-6}. This is a common practice to avoid numerical instability in Newton–Raphson type of algorithms. The first $(1 + q)$ elements on the diagonal of matrix $H^{(g)}$ are zero, which indicates no shrinkage is added to the intercept and the coefficients of the environmental factors. Here $nH^{(g)}\hat{\beta}^{(g)}$ and $nH^{(g)}$ can be used to approximate the first and second order derivative functions of the sparse group penalty, respectively. Given an initial coefficient vector, which can be estimated by LASSO, the

proposed algorithm proceeds iteratively and update the regression parameter $\hat{\beta}^{(g+1)}$ until convergence which can be achieved when the L1 norm of the difference in coefficient vectors from adjacent iterations is less than 0.001. Our numerical experiments show that the convergence can usually be achieved in a small to moderate number of iterations.

There are usually two tuning parameters for sparse group penalty, controlling the individual and group level sparsity, respectively. In the current G×E study, for a G factor, its main effect and interactions with all the environmental factors are treated as one group. The tuning parameter λ_1 determines the amount of shrinkage on the group level, and λ_2 further tunes the shrinkage on individual effects within the group. The optimal pair of λ_1 and λ_2 are obtained through a joint search over a two-dimensional grid of (λ_1, λ_2) based on a validation approach. Specifically, the regularized estimate is computed on a training dataset, and then the prediction is evaluated on an independently generated testing dataset. Our numerical experiment shows that validation and cross validation tend to yield similar tunings, but the first one is computationally much faster.

With the nonconvex baseline penalty MCP, we will need to determine the regularization parameter γ which balances unbiasedness and concavity (Zhang (2010)). Relevant studies suggests checking with a sequence of different values, and then fixing the value. We have investigated a sequence of 1.4, 3, 4.2, 5.8, 6.9, and 10, and found that the results are not sensitive to the value of γ. Therefore, we set γ to 3. This finding is consistent with published studies (Ren et al. (2017); Wu et al. (2018)).

For fixed tuning parameters, the proposed algorithm proceeds as follows:

(a) Initialize the coefficient vector $\hat{\beta}^{(0)}$ using LASSO;

(b) At the $(g+1)$th iteration, update $\hat{\beta}^{(g+1)}$ based on equation (3.5) ;

(c) Repeat Step (b) until the convergence is achieved.

We consider three working correlation structures, exchangeable, AR(1) and independence, for the sparse group MCP based method dissecting longitudinal G×E interactions. Besides, the group MCP which ignores the within group sparsity of G×E interactions and the MCP only considering individual level main and interaction effects are included for comparison. In summary, we term the bi–level, group–level and individual–level longitudinal penalization un-

der exchangeable working correlation as sgQIF.exch, gQIF.exch and iQIF.exch, respectively. Similarly, with AR(1) correlation, the three approaches are denoted as sgQIF.ar1, gQIF.ar1 and iQIF.ar1 correspondingly. Then sgQIF.ind, gQIF.ind, and gQIF.ind are termed accordingly under independent correlation. The details of the alternative approaches are provided in Appendix B.1. We computed the CPU running time for 100 replicates of simulated gene expression data with n = 400, $p = 200$, q=5 (with a total dimension of 1206) and fixed tuning parameters on a regular laptop for the nine methods, which can be implemented using our developed package: *springer* (Zhou et al. (2021)). The average CPU running time in seconds are 34.7 (sd 4.9) (sgQIF.exch), 36.2 (sd 6.9) (gQIF.exch), 35.7 (sd 3.5) (iQIF.exch), 24.9 (sd 4.3) (sgQIF.ar1), 32.7 (sd 1.5) (gQIF.ar1), 26.5 (sd 5.3) (iQIF.ar1), 5.8 (sd 0.5)(sgQIF.ind), 6.3 (sd 0.8) (gQIF.ind) and 5.4 (sd 0.3) (iQIF.ind), respectively.

3.2.5 Unbalanced Date Implementation

In practice, due to missing data, the repeated measurements are unbalanced when cluster sizes vary among different subjects. The proposed method can still be implemented in such a case by introducing a transformation matrix to each subject (Cho and Qu (2013)). Suppose the total number of time points is denoted by k and the ith subject is repeated measured at k_i time points. Let S_i denote a $k \times k_i$ tranformation matrix for the ith subject. Then for the ith subject, the transformation matrix S_i is generated by deleting the columns of the $k \times k$ identity matrix that correspond to the time points with measurement missing. According to this strategy, transformation is performed by letting $W_i^\star = S_i W_i, Y_i^\star = S_i Y_i, \mu_i^\star(\beta) = S_i \mu_i(\beta)$ and $A_i^\star = S_i A_i S_i^\top$. Then we can replace $\phi_i(\beta)$ in equation (3.3) by the transformed extended score vector $\phi_i^\star(\beta)$, which is defined as:

$$\phi_i^\star(\beta) = \begin{pmatrix} (W_i^\star)^\top (A_i^\star)^{-\frac{1}{2}} M_1 (A_i^\star)^{-\frac{1}{2}} (Y_i^\star - \mu_i^\star(\beta)) \\ . \\ . \\ . \\ (W_i^\star)^\top (A_i^\star)^{-\frac{1}{2}} M_m (A_i^\star)^{-\frac{1}{2}} (Y_i^\star - \mu_i^\star(\beta)) \end{pmatrix},$$

and the QIF estimator can be further obtained for unbalanced data based on the transformed terms.

3.3 Simulation

The performance of the nine methods has been assessed through simulation studies to demonstrate the utility of the proposed methods. We generate the responses from model (3.1) with sample size n=400, and set the number of time points k to 5. The dimension for genetic factors is p= 200 and there are $q = 5$ environmental factors. This leads to a total dimension for all the main and interaction effects equal to 1206, which is much larger than the sample size. We have also experimented with larger dimensionality for the G factors, and found that the results are stable and consistent with the current setting as long as the total dimensionality is moderately larger than sample size. The details on scalability of the proposed method to ultra-high dimensional data is deferred to the Section of Discussion. In our simulation, the environmental factors are simulated from a multivariate normal distribution with mean 0 and AR–1 covariance matrix with marginal variance 1 and auto correlation coefficient 0.8. The first environmental factor is dichotomized at the 50th percentile and changed to a binary vector. We simulate the random error ϵ for the longitudinal response from a multivariate normal distribution by assuming 0 mean vector and an exchangeable covariance structure with parameter $\tau = 0.8$. Following all these settings, the time-independent genetic factors are simulated in four different scenarios.

In the first scenario, the genetic factors are gene expressions, which are simulated from

a multivariate normal distribution with mean 0 and AR–1 covariance matrix with marginal variance 1 and an auto correlation coefficient 0.8. In the second scenario, we consider generating SNP data by dichotomizing the gene expression values from scenario 1 at the 30th and 70th percentiles with respect to each gene, leading to the three categories (0,1,2) for genotypes (aa,Aa,AA).

In the third scenario, we simulate the SNP data using a pairwise linkage disequilibrium (LD) structure. Let q_A and q_B denote the minor allele frequencies (MAFs) for the two risk alleles A and B from two adjacent SNPs, respectively, and δ denote the LD. Then the frequencies of the four haplotypes can be derived as $p_{AB} = q_A q_B + \delta, p_{ab} = (1-q_A)(1-q_B) + \delta, p_{Ab} = q_A(1-q_B) - \delta$, and $p_{aB} = (1-q_A)q_B - \delta$. By assuming Hardy-Weinberg equilibrium, we simulate the SNP genotypes AA, Aa and aa at locus 1 from a multinomial distribution with frequencies $q_A^2, 2q_A(1-q_A)$ and $(1-q_A)^2$. Then the genotypes for SNP at locus 2 can be generated based on the conditional genotype probability matrix (Cui et al. (2008)). If the MAFs are 0.3 and pairwise correlation r is set to 0.3, we can get $\delta = r\sqrt{q_A(1-q_A)q_B(1-q_B)}$.

Next, in scenario 4, we consider a more practical approach to generate the SNP data. The first 200 SNPs from the case study have been extracted as the G factors. We randomly sample 400 subjects from the real data in each simulation replicate to generate the longitudinal responses.

The coefficients are generated from Uniform[0.3, 0.7] for 31 nonzero effects, consisting of the intercept, 5 environmental factors, and 25 genetic main effects and G×E interactions. We simulate 100 replicates for each scenario to evaluate the identification and prediction performance of all the 9 methods. The average number of true positives (TP) and false positives (FP) with the corresponding standard deviation (sd) are recorded. In addition, prediction accuracy is evaluated based on the mean squared error.

Identification results are tabulated in Tables 3.1, 3.2 in the main text, and Tables B1 and B2 in Appendix B.2. In general, the proposed sparse group G×E interactions under the exchangeable(sgQIF.exch), AR1(sgQIF.ar1) and independence (sgQIF.ind) working correlation structures outperform the alternatives focusing only on the group level effects (gQIF.exch, gQIF.ar1 and gQIF.ind) and individual level effects (iQIF.exch, iQIF.ar1 and

iQIF.ind). Table 3.1 shows the result of using gene expressions as G factors from the first scenario with $n = 400, p = 200, \tau = 0.8$. There are 25 important main and interaction effects with corresponding nonzero coefficients. Under the exchangeable working correlation, sgQIF.exch identifies 21.4 (sd 1.1) true positives, while the number of false positives, 2.6 (sd 1.5), is relatively small. On the other hand, iQIF.exch only considers the individual main and interaction effects, yielding 21.6 (sd 1.1) true positives, with 6.4 (sd 5.2) false positives. gQIF.exch identifies an FP of 14.8 (5), which is the largest number of false positives among the three under the same working correlation structure. It is also worth noting that the standard deviations associated with the alternative approaches, i.e. 5 for gQIF.exch and 5.2 for iQIF.exch, are quite larger than that of the proposed one (1.5 for sgQIF.exch). A closer look over the results shows that such all these differences mainly come from the identification of interaction effects. sgQIF.exch has the smallest FP (2.4 with sd 1.3) for the interaction effects, followed by iQIF.exch (5.4 with sd 4.6) and gQIF.exch (14.4 with sd 4.5).

Table 3.1: *Identification results for Scenario 1. TP/FP: true/false positives. mean(sd) of TP and FP based on 100 replicates.*

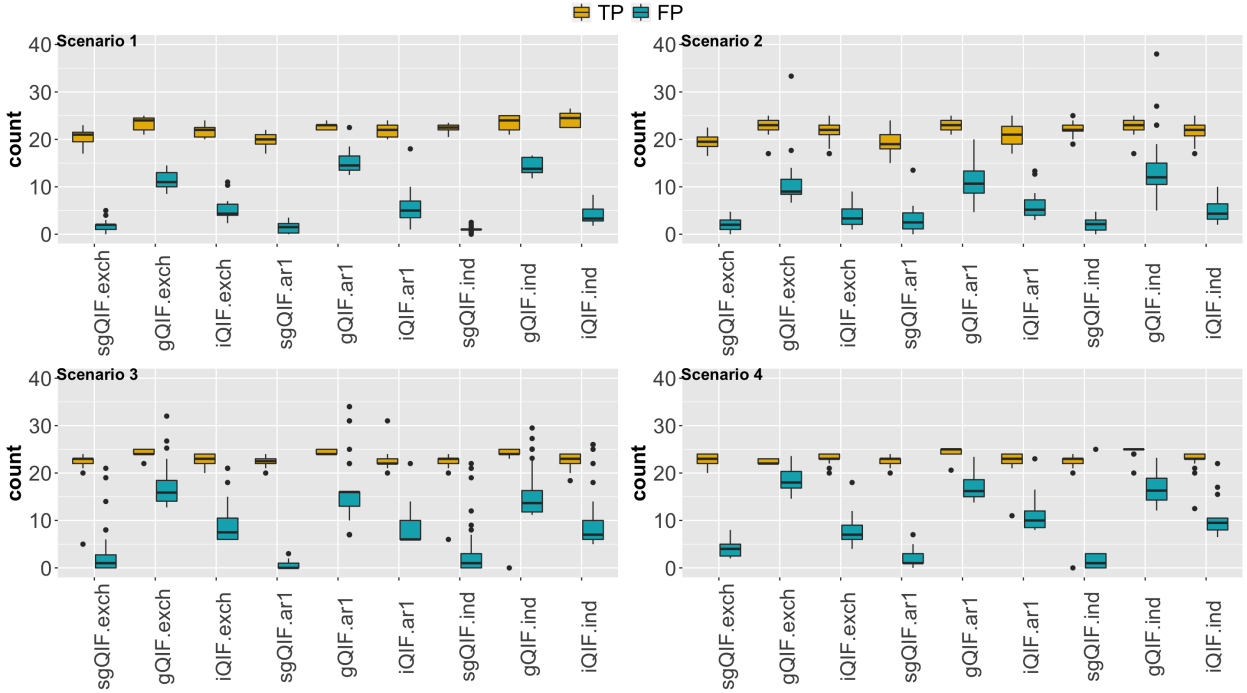
	Overall		Main		Interaction	
	TP	FP	TP	FP	TP	FP
sgQIF.exch	21.4(1.1)	2.6(1.5)	5.4(1.1)	0.2(0.4)	16.0(1.9)	2.4(1.3)
gQIF.exch	23.4(1.1)	14.8(5.0)	6.0(1.2)	0.4(0.9)	17.4(0.9)	14.4(4.5)
iQIF.exch	21.6(1.1)	6.4(5.2)	5.4(1.1)	1.0(1.7)	16.2(1.9)	5.4(4.6)
sgQIF.ar1	21.7(1.2)	3.2(1.9)	5.5(1.0)	0.3(0.5)	16.2(1.7)	2.8(1.6)
gQIF.ar1	23.7(1.2)	14.8(4.4)	6.2(1.2)	0.3(0.8)	17.5(0.8)	14.5(4)
iQIF.ar1	21.8(1.2)	6.2(4.7)	5.5(1.0)	1.0(1.5)	16.3(1.8)	5.2(4.1)
sgQIF.ind	20.7(1.0)	2.7(2.2)	4.5(1.2)	0.2(0.4)	16.2(0.8)	2.5(1.9)
gQIF.ind	22.3(1.2)	16.5(7.0)	5.5(1.0)	1.0(1.5)	16.8(0.8)	15.5(5.5)
iQIF.ind	21.0(0.9)	5.2(3.1)	4.5(1.2)	0.5(0.8)	16.5(0.8)	4.7(2.3)

Similar patterns can be observed from other settings. For instance, Table 3.2 displays the result for the simulated SNP data from Scenario 2. sgQIF.exch identifies an TP of 19.4 (sd 0.7) with 1.3 (sd 1.2) false positives. gQIF.exch has 21.5 (sd 1.9) true positives with a much larger number of false positives 13.3 (sd 4.0). The number of TP and FP pinpointed by iQIF.exch are 20.1 (sd 1.2) and 4.4 (sd 4.0), respectively. Under the same exchangeable

Table 3.2: *Identification results for Scenario 2. TP/FP: true/false positives. mean(sd) of TP and FP based on 100 replicates.*

	Overall		Main		Interaction	
	TP	FP	TP	FP	TP	FP
sgQIF.exch	19.4(0.7)	1.3(1.2)	3.3(0.7)	0.1(0.4)	16.1(0.6)	1.1(1.0)
gQIF.exch	21.5(1.9)	13.3(4.0)	4.4(1.6)	0.5(0.8)	17.1(0.6)	12.8(3.3)
iQIF.exch	20.1(1.2)	4.4(4.0)	3.3(1.0)	0.5(0.8)	16.9(0.6)	3.9(3.4)
sgQIF.ar1	19.0(0.9)	1.0(1.0)	3.3(0.6)	0.1(0.4)	15.7(0.6)	1.0(1.0)
gQIF.ar1	21.7(2.9)	12.7(4.0)	4.7(2.1)	0.3(0.6)	17.0(1.0)	12.3(3.5)
iQIF.ar1	20.7(0.6)	6.7(5.7)	3.7(0.6)	0.7(1.2)	17.0(1.0)	6.0(4.6)
sgQIF.ind	19.0(2.0)	1.8(0.7)	3.3(1.2)	0.1(0.4)	15.7(1.2)	1.6(0.7)
gQIF.ind	21.3(1.3)	15.5(8.2)	3.8(0.9)	0.8(1.8)	16.5(0.8)	14.8(7.0)
iQIF.ind	19.5(1.8)	5.3(3.6)	3.5(0.9)	1.1(1.2)	16.0(1.3)	4.1(3.0)

Figure 3.1: *Identification results under 25 important genetic main effects and G×E interactions (corresponding to 25 nonzero regression coefficients) in the 4 scenarios. TP/FP: true/false positives. mean(sd) of TP and FP based on 100 replicates.*



working correlation, while the number of identified TPs are comparable, both the average and standard deviations of alternatives are much larger than the proposed method. The identification results for the 4 scenarios are displayed in Figure 3.1, which clearly shows that

the proposed method outperforms the competing alternatives in the identification of longitudinal G×E interactions. Figure B1 summarizes the prediction results of the 4 scenarios. In Scenario 1 under the exchangeable working correlation, sgQIF.exch has a prediction error less than that of the gQIF.exch and iQIF.exch. We have similar findings in other settings as well, which indicates the proposed bi-level method has superior prediction performance over the group level and individual level based methods.

In longitudinal studies, the QIF framework is robust to the misspecification of working correlations (Qu et al. (2000)). In our simulation, although the results without misspecifying working correlation appear to be better, overall, they are comparable across different settings. Such a property is especially appealing when the ground truth on working correlation is not available. Another fold of robustness in QIF comes from its insensitivity to small portions of outlying observations and data contamination, which has been theoretically and empirically investigated in Qu and Song (2004). Meanwhile, the GEE based ones, as well as models assuming Gaussian responses and working independence among repeated outcomes, are not robust and lead to biased results given the presence of even a single outlier. A comprehensive evaluation of this fold of robustness is beyond the scope this study, and will be conducted in the near future.

3.4 Real Data Analysis

Asthma is a chronic respiratory disease with lung inflammation and reversible airflow obstruction, resulting in difficulty in breath. According to the Centers for Disease Control and Prevention (CDC), more than 25 million Americans have asthma. 7.7 percent of adults and 8.4 percent of children in the U.S. have asthma (Akinbami (2006); CDC (CDC)). Asthma is the leading chronic disease among children. We analyze the data from Childhood Asthma Management Program (CAMP) in our case study (Childhood Asthma Management Program Research Group (1999, 2000); Covar et al. (2012)). The SNP and phenotype datasets (with accession pht000701.v1.p1) from CAMP have been downloaded and pre–processed. Subjects who are 5 to 12 years and diagnosed with chronic asthma have been selected and moni-

tored over 4 years. There are three visits before treatment with each visit 1-2 weeks apart. Thirteen visits are made after treatment. The first two visits after treatment are 2 months apart and the remaining visits are 4 months apart. The twelve visits that are 4 months apart after treatment are selected in our study. Two types of treatments are given to the subjects. Treatments Budesonide and Nedocromil are assigned to 30% of the subjects, and the rest receive placebo. We consider treatment, age and gender as environmental factors. The phenotype of interest is the forced expiratory volume in one second (FEV1), which is the total volume of air expelled out of the lung in one second and it's repeatedly measured during each visit. The genotype information of each subject contains over nine hundred thousand SNPs. We match genotypes with phenotypes based on subject id's and remove the SNPs with minor allele frequency (MAF) less than 0.05 or deviation from Hardy-Weinberg equilibrium and obtain a working dataset with 438 Caucasian subjects and 447,850 SNPs.

For computational convenience in studies with ultrahigh dimensionality, such as the Genome Wide Association Studies (GWAS) and multi–omics integration studies, marginal feature prescreening needs to be conducted first so that regularization can be applied on datasets with reasonably large scale (Fan and Lv (2010); Wu et al. (2019)). For instance, Li et al. (2015), Jiang et al. (2015) and Wu et al. (2014) have adopted single SNP analysis for prescreening before applying the proposed variable selection methods in longitudinal and multivariate GWAS studies, respectively. Here, we use a marginal G×E model with FEV1 as the response to filter SNPs. The predictors of the marginal model consist of E factors, the single SNP main effect, as well as their interactions. The SNPs with at least one of the p values that correspond to G and G×E interactions in the marginal model less than a certain cutoff (0.005) are kept. 261 SNPs have passed the screening.

We apply the method sgQIF.exch under the exchangeable working correlation and analyze the data together with the alternative method iQIF.exch, which consider all the effects individually. The optimal tuning parameters are achieved through a 5-fold cross-validation. We obtain the predicted mean squared error after refitting using selected variables from the orginal data. sgQIF.exch has a smaller prediction error (0.16) than that of iQIF.ind (0.23). The identification results are tabulated in tables C1 and C2 in Appendix B.3. Methods that

consider group effects only show inferior performance in the simulation study are not adopted in the real data analysis. The proposed method sgQIF.exch identifies 130 effects in total, 34 of which are genetic main effects and the remaining 96 are interaction effects. iQIF.exch totally identifies 130 effects, with 28 genetic main effects and 102 interaction effects. sgQIF.exch and iQIF.exch commonly identify 22 genetic main effects and 62 interaction effects. There are twelve SNPs that are uniquely identified by the proposed method sgQIF.exch and they will provide some useful implications. They can be mapped to the corresponding genes and some of the genes have been found to be related to the development of asthma. For instance, sgQIF.exch identifies the main effect of the SNP rs17390967 and its interactions with the environment factors treatment and gender. The SNP rs17390967 is located within the gene SCARA5. SCARA5 is a member of the scavenger receptor A (SR-A), which is found to be protective to the lung using the ovalbumin-asthma model of lung injury (Arredouani et al. (2007)). The interaction with treatment indicates that the expression level of SCARA5 may influence the effect of medical therapy in the treatment of asthma. Another example is the SNP rs767006, which is located in the gene CYFIP2. The prposed method sgQIF.exch identifies the main effect of rs767006 and its interaction with gender. CYFIP2, together with CYFIP1 make up the CYFIP family. It has been found that there is a strong association between asthma and polymorphisms in CYFIP2 (Noguchi et al. (2005)). Method sgQIF.exch also identifies rs6914953 and its interaction with gender. The identified SNP rs6914953 is located in F13A1. F13A1 codes for the α subunit of Factor X111, which is the last enzyme generated in the blood coagulation cascade and it stabilizes blood clots with cross-linking fibrin. F13A1 has been considered as a susceptible locus for obesity and it has been found that there is a consistent link between asthma and obesity (Sharma et al. (2014)). Another identified SNP is rs4647108, that is mapped to the gene ERCC8. ERCC8 has also been found to be related with the development of asthma (Wilson et al. (2015)). The method sgQIF.exch identifies the main effect of rs4647108 and its interaction with gender. This result is consistent with previous findings that over-expression of ERCC8 is associated with a higher FEV1, which indicates a development of asthma.

3.5 Discussion

In general, regularization methods work well when the dimensionality is up to the order that is moderately larger than sample size. To handle ultra-high dimensional data, the two stage variable selection consisting of a quick marginal screening stage and a post-screening refining stage with the direct applications of regularization has been widely used (Fan and Lv (2010)), including the longitudinal GWAS (Jiang et al. (2015); Li et al. (2015)). The marginal feature screening, preferably with theoretical guarantees such as the sure independence screening (Fan and Lv (2008); Li et al. (2014); Song et al. (2014)), is necessary for reducing the ultra-high dimensionality of features to a reasonable order so regularized variable selection is applicable (Fan and Lv (2010)). By far, consensus on the optimal screening strategy with repeated measurements has not been reached yet. In this study, we have adopted a marginal G×E model to conduct screening, which is more consistent with the nature of regularization at the refining stage.

There are published studies on variable selection in varying coefficient models with repeated measurements (Wang et al. (2008), Noh and Park (2010) and Tang et al. (2013), among others). A common limitation in these studies is that the within–subject correlation has not been taken into account. From the perspective of G×E interactions, the time varying effects investigated in these studies can be viewed as nonlinear G×E interactions (Li et al. (2020); Ma and Xu (2015); Ma et al. (2011); Wu and Cui (2013); Wu et al. (2015, 2018)). In our study, the interaction effects is modeled as the product between G and E factors, which is under the linear G×E interaction assumption (Zhou et al. (2021)). To the best of our knowledge, no published studies have been developed for variable selection in G×E interaction studies with linear assumptions.

The bi–level structure plays a critical role in studies concerning the more general linear G×E interactions (Zhou et al. (2021)). The key contribution of the proposed study lies in developing sparse group regularization within the QIF framework to accommodate within–cluster correlations among repeated measurements. As a major competitor of GEE, QIF is more efficient when the working correlation is misspecified. Our work is significantly differ-

ent from Zhou et al. (2019) in that the lipid–environment interaction analysis of repeated measurements has been developed based on GEE, and, more importantly, the interaction is pursued only on a group level and does not lead to the within group sparsity. So it is not applicable to the current setting.

This study can be extended in multiple horizons. For instance, marginal regularization has been demonstrated as an effective approach to dissect G×E interactions (Lu et al. (2021); Zhang et al. (2019)). Our methods can be readily adopted to conduct marginal identification of interaction effects when the phenotypes are repeatedly measured. In addition, robust variable selection for G×E interactions have been proposed (Ren et al. (2020); Wu et al. (2018); Zhang et al. (2020)). In longitudinal G×E studies, the robustness of QIF framework to data contamination in the response can be potentially improved by modifying the weight in estimating equation to downweigh the influences of outliers. Recently, Wang et al. (2021) have revealed the benefit of accounting for network structure in large scale G×E studies. By incorporating the network constrained regularization, the proposed method can better accomodate the correlation among SNPs due to linkage disequilibrium.

Chapter 4

The Regularized Bayesian Quantile Varying Coefficient Model

4.1 Introduction

The varying coefficient model has been proposed by Hastie and Tibshirani (1993) to account for the dynamic effects of predictors on the response variable. As an extension of the linear regression model, its regression coefficients are nonparametric functions of other variables (i.e. effect modifiers termed in Hastie and Tibshirani (1993)). For example, if the effect modifier is the variable time, then the coefficients of the model are allowed to vary smoothly with the measurements on time to capture the nonparametric time-changing effects that cannot be properly modeled through linear regression. Classical estimation and inference procedures for the varying coefficient model are mainly based on the basis expansion and splines (Huang et al. (2002) and Huang et al. (2004)), the local-kernel polynomial smoothing (Fan and Zhang (2008) and Hoover et al. (1998)), and smoothing splines (Hastie and Tibshirani (1993), Hoover et al. (1998) and Chiang et al. (2001).)

The varying coefficient model enjoys wide popularity and application in a broad spectrum of scientific research areas due to its superior flexibility and interpretability over parametric models. However, as it has been developed for conditional mean regression, the varying coef-

ficient model is not robust to heavy-tailed errors and outliers in the response variable which are frequently encountered in practice. As a powerful alternative to accommodate the non-robustness, the quantile varying coefficient model has therefore received much attention. In literature, Cai and Xu (2009) and Kim (2007) have examined the quantile varying coefficient model through local polynomials and B splines, respectively. Wang et al. (2009) has studied a family of marginal semiparametric quantile models with potential varying coefficients.

With the emergence of high dimensional data, regularized variable selection has been extensively studied for varying coefficient models. Wang et al. (2008) and Wang and Xia (2009) have developed regularization procedures for the varying coefficient model based on splines and local polynomial smoothing, respectively. The selection of important varying coefficients amounts to group level selection with group SCAD and adaptive group LASSO. In addition, Huang et al. (2010) have studied variable selection for nonparametric additive models via adaptive group LASSO. In quantile regression, in addition to variable selection for linear regression models including Wu and Liu (2009) and Peng and Wang (2015), regularization for quantile varying coefficient models has also been considered in Tang et al. (2013) and Noh et al. (2012) using adaptive group LASSO and group SCAD. Tang et al. (2012) has further investigated structural identification of varying coefficients by separating the varying, nonzero constant and zero effects.

Despite the success in regularization for variable selection in quantile varying coefficient models, within the Bayesian framework, this important topic is not well studied. Li et al. (2010) has proposed a Bayesian regularized quantile regression by incorporating the Bayesian LASSO prior (Park and Casella (2008)) in the likelihood function based on asymmetric Laplace distribution. However, their study aims at linear quantile models. On the other hand, Li et al. (2015) has developed Bayesian group LASSO for varying coefficient models. Ren et al. (2020) has examined the structure identification in varying coefficient models by further considering the spike-and-slab priors. Nevertheless, these methods are vulnerable to long-tailed distributions and outliers in the response.

To the best of our knowledge, Bayesian regularized variable selection in quantile regression models with varying coefficients has not been studied. To overcome this limitation,

we develop a novel regularized Bayesian quantile varying coefficient model to identify the important covariates associated with the response. In order to shrink the coefficients of unimportant effects to exactly zero, we adopt the spike and slab priors in our model. As a comparison, Bayesian Lasso cannot shrink a posterior estimate exactly to zero. We develop an efficient MCMC algorithm for the proposed Bayesian quantile varying coefficient model. The identification of varying coefficients is equivalent to the selection of a group of basis functions and we efficiently perform Bayesian shrinkage on group level, borrowing the strength from the spike and slab priors. The simulation results have shown the superiority of the proposed method over the alternatives in terms of identification, estimation and prediction accuracy for heavy-tailed distributions. To facilitate fast computation and reproducible research, we implement the proposed and alternative methods in C++ and we will encapsulate them in a publicly available R package in the future work.

4.2 Statistical Methods

4.2.1 The Quantile Varying Coefficient Model

Let $(Y_i, \boldsymbol{X}_i, U_i, \boldsymbol{E}_i), i = 1, ..., n$, be an i.i.d. sample. $\boldsymbol{Y} = (Y_1, ..., Y_n)^\top$ is the response variable. $\boldsymbol{X}_i = (X_{i0}, X_{i1}, ..., X_{ip})^\top$ denotes the $(1+p)$-dimensional design vector of genetic factors with the first element $X_{i0} = 1$. The scalar $U_i \in \mathbb{R}^1$ is the univariate index variable. $\boldsymbol{E}_i = (E_{i1}, ..., E_{iq})^\top$ represents the q-dimensional design vector of clinical covariates. We consider the following varying coefficient model:

$$Y_i = \sum_{k=1}^{q} E_{ik}\beta_k + \sum_{j=0}^{p} \gamma_j(U_i)X_{ij} + \epsilon_i, \tag{4.1}$$

where E_{ik} denotes the kth component of $\boldsymbol{E}_i$. X_{ij} is the jth component of $\boldsymbol{X}_i$ and $\gamma_j(\cdot)$'s are unknown smooth varying-coefficient functions. The random error ϵ_i's have the θth quantile equal to 0. The covariates $\boldsymbol{E} = (\boldsymbol{E}_1, ..., \boldsymbol{E}_n)^\top$ are linearly associated with the response, but the regression coefficients of $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n)^\top$ vary with the univariate index variable

$\boldsymbol{U} = (U_1, ..., U_n)^\top$. Due to the fact that only some of the covariates in $\boldsymbol{X}$ are relevant to the response variable in practice, while the irrelevant ones have varying-coefficient functions equal to zero almost surely, the model 4.1 is proposed to identify the important relevant covariates and estimate the corresponding nonzero coefficients.

In the estimation procedure, the varying coefficient function $\gamma_j(U_i)$ is approximated using polynomial splines. Suppose the index variable $\boldsymbol{U}$ takes values from the interval $[a, b]$ with $a < b$. Let $\boldsymbol{t}_j$ denote a partition of the interval $[a, b]$, with M interior knots

$$\boldsymbol{t}_j = \{a = t_{j,0} < t_{j,1} < \cdots < t_{j,M} < t_{j,M+1} = b\}.$$

With $\boldsymbol{t}_j$ as knots for the polynomial splines, the order $O + 1$ splines functions are O-degree (or less) of polynomials on the intervals $[t_{j,h}, t_{j,h+1}), h = 0, ..., M - 1$, and $[t_{j,M}, t_{j,M+1}]$ with $O - 1$ continuous derivatives globally.

Let $\boldsymbol{\pi}_j(U_i) = (\pi_{j1}(U_i), ..., \pi_{jd}(U_i))^\top$ be a set of B-spline basis with $d = M + O + 1$. Then for $j = 0, ..., p$,

$$\boldsymbol{\gamma}_j(U_i) = \sum_{s=1}^{d} \pi_{js}(U_i) \cdot \alpha_{js} = \boldsymbol{\alpha}_j^\top \cdot \boldsymbol{\pi}_j(U_i),$$

where $\boldsymbol{\alpha}_j = (\alpha_{j1}, ..., \alpha_{jd})^\top$ is the coefficient vector. Let $\boldsymbol{Z}_{ij} = \boldsymbol{\pi}_j(U_i) \cdot X_{ij} = (\pi_{j1}(U_i)X_{ij}, ..., \pi_{jd}(U_i)X_{ij})^\top$, then the original varying coefficient model becomes

$$Y_i = \sum_{k=1}^{q} E_{ik}\beta_k + \sum_{j=0}^{p} \boldsymbol{\alpha}_j^\top \boldsymbol{Z}_{ij} + \epsilon_i. \tag{4.2}$$

The quantile regression is well known for its robustness to long-tailed distributions in response. The quantile regression estimators for quantile θ are obtained by

$$(\beta_1, ..., \beta_q, \boldsymbol{\alpha}_0, ..., \boldsymbol{\alpha}_p) = \underset{\beta_k, \alpha_j}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\theta(Y_i - \sum_{k=1}^{q} E_{ik}\beta_k - \sum_{j=0}^{p} \boldsymbol{\alpha}_j^\top \boldsymbol{Z}_{ij}),$$

where $\rho_\theta(\cdot)$ is the check loss function

$$\rho_\theta(m) = \begin{cases} \theta m & \text{if } m \geq 0 \\ -(1-\theta)m & \text{if } m < 0 \end{cases}.$$

Assume that ϵ_i's follow a skewed Laplace distribution:

$$\pi(\epsilon|\tau) = \theta(1-\theta)\tau \exp[-\tau\rho_\theta(\epsilon)] = \theta(1-\theta)\tau \begin{cases} e^{-\tau\theta\epsilon} & \text{if } \epsilon \geq 0 \\ e^{\tau(1-\theta)\epsilon} & \text{if } \epsilon < 0 \end{cases}$$

Then the joint distribution of the varying coefficient model is given as:

$$\pi(\boldsymbol{Y}|\boldsymbol{E}, \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \tau) = \theta^n(1-\theta)^n\tau^n \exp\Big(-\tau \sum_{i=1}^{n} \rho_\theta(Y_i - \sum_{k=1}^{q} E_{ik}\beta_k - \sum_{j=0}^{p} \boldsymbol{\alpha}_j^\top \boldsymbol{Z}_{ij}) \Big),$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)^\top$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^\top, ..., \boldsymbol{\alpha}_p^\top)^\top$. Thus, the previous minimization problem becomes the case of maximizing the joint likelihood. According to Li et al. (2010), assume the random variables $v \sim \text{Exp}(1)$ and $W \sim \text{N}(0,1)$. Define

$$\xi_1 = \frac{1-2\theta}{\theta(1-\theta)},$$

and

$$\xi_2 = \sqrt{\frac{2}{\theta(1-\theta)}},$$

then $\epsilon = \xi_1 v + \xi_2\sqrt{v}W$ follows a skewed Laplace distribution and the original model becomes

$$Y_i = \sum_{k=1}^{q} E_{ik}\beta_k + \sum_{j=0}^{p} \boldsymbol{\alpha}_j^\top \boldsymbol{Z}_{ij} + \tau^{-1}\xi_1 v_i + \tau^{-1}\xi_2\sqrt{v_i}W_i,$$

with $v_i \sim \text{Exp}(1)$ and $W_i \sim \text{N}(0,1)$. Let $\tilde{v}_i = \tau^{-1}v_i \sim \text{Exp}(\tau^{-1})$ and $\tilde{v} = (\tilde{v}_1, ..., \tilde{v}_n)$, then

the model can be reparameterized as:

$$Y_i = \sum_{k=1}^{q} E_{ik}\beta_k + \sum_{j=0}^{p} \boldsymbol{\alpha}_j^\top \boldsymbol{Z}_{ij} + \xi_1 \tilde{v}_i + \tau^{-\frac{1}{2}} \xi_2 \sqrt{\tilde{v}_i} W_i.$$

4.2.2 The Regularized Bayesian Quantile Varying Coefficient Model

Basis expansion results in a high-dimensional dataset. As shown in 4.2, the expanded basis are modeled on the group level, while only a small subset of the effects are associated with the disease phenotype. Therefore, the group-level penalized variable selection becomes necessary and the previous minimization problem becomes:

$$\min_{\boldsymbol{\beta},\boldsymbol{\alpha}} \sum_{i=1}^{n} \rho_\theta(\boldsymbol{Y}_i - \boldsymbol{E}_i^\top \boldsymbol{\beta} - \boldsymbol{Z}_i^\top \boldsymbol{\alpha}) + \lambda \sum_{j=1}^{p} \sqrt{d} ||\boldsymbol{\alpha}_j||_2,$$

where $\lambda > 0$ is the tuning parameter.

We have the following hierarchical model specification:

$$Y_i = \boldsymbol{E}_i^\top \boldsymbol{\beta} + \boldsymbol{Z}_i^\top \boldsymbol{\alpha} + \xi_1 \tilde{v}_i + \xi_2 \tau^{-\frac{1}{2}} \sqrt{\tilde{v}_i} W_i, i = 1, ..., n,$$

$$\frac{Y_i - \boldsymbol{E}_i^\top \boldsymbol{\beta} - \boldsymbol{Z}_i^\top \boldsymbol{\alpha} - \xi_1 \tilde{v}_i}{\xi_2 \tau^{-\frac{1}{2}} \sqrt{\tilde{v}_i}} \sim N(0,1), i = 1, ..., n,$$

$$\tilde{v_1}, ..., \tilde{v_n} \sim \prod_{i=1}^{n} \tau \exp(-\tau \tilde{v}_i), i = 1, ..., n,$$

$$W_1, ..., W_n \sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} W_i^2), i = 1, ..., n,$$

$$\boldsymbol{\alpha}_j | s_j \sim (1 - \pi_0) \mathrm{N}_d(0, s_j \mathbf{I}_d^{-1}) + \pi_0 \delta_0(\boldsymbol{\alpha}_j), j = 0, ..., p,$$

$$s_j | \eta^2 \sim (\frac{\eta^2}{2})^{\frac{d+1}{2}} s_j^{\frac{d-1}{2}} \exp(-\frac{\eta^2}{2} s_j), j = 0, ..., p,$$

$$\pi_0 \sim \mathrm{Beta}(e, f),$$

$$\tau \sim \tau^{a-1} \exp(-b\tau),$$

$$\eta^2 \sim (\eta^2)^{c-1}\exp(-m\eta^2),$$

$$\boldsymbol{\beta} \sim \mathrm{N}_q(0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}),$$

where a, b, c, e, f and m are constants. The spike-and-slab priors are imposed on the d-dimensional coefficient vectors α_j's.

4.3 The Gibbs Sampler

The joint likelihood of the unknown parameters conditional on data will be given as

$$\begin{aligned}
&\boldsymbol{\alpha}, \boldsymbol{\beta}, \tilde{v_i}, W_i, s_j, \pi_0, \tau, \eta^2 | \boldsymbol{Y} \propto \\
&\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\tau^{-1}\xi_2^2\tilde{v_i}}} \exp\{-\frac{(Y_i - \boldsymbol{E}_i^\top\boldsymbol{\beta} - \boldsymbol{Z}_i^\top\boldsymbol{\alpha} - \xi_1\tilde{v_i})^2}{\tau^{-1}\xi_2^2\tilde{v_i}}\} \\
&\times \prod_{j=0}^{p} \left((1-\pi_0) \frac{1}{\sqrt{2\pi|s_j\mathbf{I}_d^{-1}|}} \exp\Big(-\frac{1}{2}\boldsymbol{\alpha}_j^\top (s_j\mathbf{I}_d^{-1})^{-1}\boldsymbol{\alpha}_j \Big) \mathbf{I}_{(\boldsymbol{\alpha}_j \neq 0)} + \pi_0\delta_0(\boldsymbol{\alpha}_j) \right) \\
&\times \pi_0^{e-1}(1-\pi_0)^{f-1} \\
&\times \prod_{j=0}^{p} (\frac{\eta^2}{2})^{\frac{d+1}{2}} s_j^{\frac{d-1}{2}} \exp(-\frac{\eta^2}{2}s_j) \\
&\times \prod_{i=1}^{n} \tau\exp(-\tau\tilde{v_i}\Big) \\
&\times \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}W_i^2) \\
&\times \tau \sim \tau^{a-1}\exp(-b\tau) \\
&\times \eta^2 \sim (\eta^2)^{c-1}\exp(-m\eta^2) \\
&\times \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|}} \exp(-\frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}).
\end{aligned}$$

We have the full conditional distribution of $\tilde{v}_i$ listed as follows:

$$\tilde{v}_i|\text{rest} \propto (\tilde{v}_i)^{-\frac{1}{2}}\exp\bigg(-\frac{1}{2}\Big((\frac{\tau\xi_1^2}{\xi_2^2}+2\tau)\tilde{v}_i + \frac{\tau(Y_i-\boldsymbol{E}_i^\top\boldsymbol{\beta}-\boldsymbol{Z}_i^\top\boldsymbol{\alpha})^2}{\xi_2^2}\frac{1}{\tilde{v}_i}\Big)\bigg).$$

Hence, the posterior distribution of $\tilde{v}_i$ is generalized inverse Gaussian distribution.

The slab part of the full conditional distribution of α_j is given as:

$$\begin{aligned}
&\boldsymbol{\alpha}_j|\text{rest}\\
&\propto (1-\pi_0)|s_j\mathbf{I}_d^{-1}|^{-\frac{1}{2}}|\boldsymbol{\Sigma}_j|^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}\tau\xi_2^{-2}\sum_{i=1}^n\frac{1}{\tilde{v}_i}(Y_i-\boldsymbol{Z}_{i,-j}^\top\boldsymbol{\alpha}_{-j}-\boldsymbol{E}_i^\top\boldsymbol{\beta}-\xi_1\tilde{v}_i)^2\Big)\\
&\times\exp(\frac{1}{2}\boldsymbol{\mu}_j^\top\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j)\times \mathrm{N}_d(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j),
\end{aligned}$$

where variance

$$\boldsymbol{\Sigma}_j = (\tau\xi_2^{-2}\sum_{i=1}^n\frac{1}{\tilde{v}_i}\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^\top + s_j^{-1}\mathbf{I}_d)^{-1},$$

and mean

$$\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j\tau\xi_2^{-2}\sum_{i=1}^n\frac{\boldsymbol{Z}_{ij}}{\tilde{v}_i}(Y_i-\boldsymbol{Z}_{i,-j}^\top\boldsymbol{\alpha}_{-j}-\boldsymbol{E}_i^\top\boldsymbol{\beta}-\xi_1\tilde{v}_i).$$

The spike part is given as:

$$\boldsymbol{\alpha}_j|\text{rest} \propto \pi_0\exp\Big(-\frac{1}{2}\tau\xi_2^{-2}\sum_{i=1}^n\frac{1}{\tilde{v}_i}(Y_i-\boldsymbol{Z}_{i,-j}^\top\boldsymbol{\alpha}_{-j}-\boldsymbol{E}_i^\top\boldsymbol{\beta}-\xi_1\tilde{v}_i)^2\Big),$$

and the proportion of the spike part is

$$P(\alpha_j=0|\text{rest}) = \frac{\pi_0}{\pi_0+(1-\pi_0)|s_j\mathbf{I}_d^{-1}|^{-\frac{1}{2}}|\boldsymbol{\Sigma}_j|^{\frac{1}{2}}\exp(\frac{1}{2}\boldsymbol{\mu}_j^\top\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j)}.$$

The full conditional distribution of τ

$$\tau|\text{rest}$$

$$\propto \tau^{\frac{3}{2}n+a-1}\exp\bigg(-\Big(\frac{1}{2}\sum_{i=1}^{n}\frac{\tau(Y_i-\boldsymbol{Z}_i^\top\boldsymbol{\alpha}-\boldsymbol{E}_i^\top\boldsymbol{\beta}-\xi_1\tilde{v_i})^2}{\xi_2^2\tilde{v_i}}+\sum_{i=1}^{n}\tilde{v_i}+b\Big)\tau\bigg).$$

Therefore, the posterior distribution of τ is

$$\tau|\text{rest}\propto \text{Gamma}\Big(\frac{3}{2}n+a,\frac{1}{2}\sum_{i=1}^{n}\frac{\tau(Y_i-\boldsymbol{Z}_i^\top\boldsymbol{\alpha}-\boldsymbol{E}_i^\top\boldsymbol{\beta}-\xi_1\tilde{v_i})^2}{\xi_2^2\tilde{v_i}}+\sum_{i=1}^{n}\tilde{v_i}+b\Big).$$

The full conditional distribution of η^2

$$\eta^2|\text{rest}\propto(\eta^2)^{\frac{(d+1)(p+1)}{2}+c-1}\exp(-(\frac{1}{2}\sum_{j=0}^{p}s_j+m)\eta^2).$$

Therefore, the posterior distribution of η^2 is

$$\eta^2|\text{rest}\propto \text{Gamma}\Big(\frac{(d+1)(p+1)}{2}+c,\frac{1}{2}\sum_{j=0}^{p}s_j+m\Big).$$

The full conditional distribution of s_j, $j=0,...,p$

$s_j|\text{rest}$

$$\propto\bigg((1-\pi_0)\frac{1}{\sqrt{2\pi|s_j\mathbf{I}_d^{-1}|}}\exp\Big(-\frac{1}{2}\boldsymbol{\alpha}_j^\top(s_j\mathbf{I}_d^{-1})^{-1}\boldsymbol{\alpha}_j\Big)\mathbf{I}_{(\boldsymbol{\alpha}_j\neq 0)}+\pi_0\delta_0(\boldsymbol{\alpha}_j)\bigg)\times s_j^{\frac{d-1}{2}}\exp(-\frac{\eta^2}{2}s_j).$$

The slab part,

$$s_j|\text{rest}\propto s_j^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}(\eta^2 s_j+\boldsymbol{\alpha}_j^\top K\mathbf{I}_d\boldsymbol{\alpha}_j\frac{1}{s_j})\Big)$$

Therefore, the posterior distribution of s_j^{-1} is Inverse-Gaussian$(\sqrt{\frac{\eta^2}{\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j}},\eta^2)$ when $\boldsymbol{\alpha}_j\neq 0$.
The spike part,

$$s_j|\text{rest}\propto s_j^{\frac{d-1}{2}}\exp(-\frac{\eta^2}{2}s_j),$$

which is Gamma($\frac{d+1}{2}, \frac{\eta^2}{2}$). Together

$$s_j^{-1}|\text{rest} \sim \begin{cases} \text{Inverse-Gamma}(\frac{d+1}{2},\ \frac{\eta^2}{2}) & \text{if } \alpha_j = 0 \\ \text{Inverse-Gaussian}(\sqrt{\frac{\eta^2}{\boldsymbol{\alpha}_j^\top \boldsymbol{\alpha}_j}}, \eta^2) & \text{if } \alpha_j \neq 0 \end{cases}.$$

The full conditional distribution of π_0, $i = 1, ..., n$

$\pi_0|\text{rest}$

$$\propto \prod_{j=0}^{p} \left((1-\pi_0) \frac{1}{\sqrt{2\pi |s_j \mathbf{I}_d^{-1}|}} \exp\Big(-\frac{1}{2} \boldsymbol{\alpha}_j^\top (s_j \mathbf{I}_d^{-1})^{-1} \boldsymbol{\alpha}_j \Big) \mathbf{I}_{(\boldsymbol{\alpha}_j \neq 0)} + \pi_0 \delta_0(\boldsymbol{\alpha}_j) \right) \times \pi_0^{e-1} (1-\pi_0)^{f-1}.$$

Let

$$Q_j = \begin{cases} 0 & \text{if } \boldsymbol{\alpha}_j = 0 \\ & \\ 1 & \text{if } \boldsymbol{\alpha}_j \neq 0 \end{cases},$$

then the posterior distribution of π_0 becomes

$$\pi_0|\text{rest} \propto \pi_0^{1+p-\sum_{j=0}^{p} Q_j + e - 1} (1-\pi_0)^{\sum_{j=0}^{p} Q_j + f - 1},$$

which is a beta distribution. The full conditional distribution of $\boldsymbol{\beta}$ is

$\boldsymbol{\beta}|\text{rest}$

$$\propto \mathrm{N}_q \left((\sum_{i=1}^{n} \frac{\tau \boldsymbol{E}_i \boldsymbol{E}_i^\top}{\xi_2^2 \tilde{v}_i} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1} \Big(\sum_{i=1}^{n} \frac{\tau}{\xi_2^2 \tilde{v}_i} (Y_i - \boldsymbol{Z}_i^\top \boldsymbol{\alpha} - \xi_1 \tilde{v}_i) \boldsymbol{E}_i^\top \Big)^\top, (\sum_{i=1}^{n} \frac{\tau \boldsymbol{E}_i \boldsymbol{E}_i^\top}{\xi_2^2 \tilde{v}_i} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1}) \right),$$

which is a multivariate normal distribution.

Gibbs Sampler for the alternative methods are attached in Appendix C.4.

4.4 Simulation

We compare the performance of the proposed method, bayesian quantile regression with group lasso penalty and spike and slab priors, which is termed as BQRVCSS, with five alternative methods: bayesian quantile regression with group lasso penalty (BQRVC), bayesian group Lasso penalty and spike and slab priors (BVCSS), bayesian group Lasso (BVC), frequentist quantile varying coefficient model with adaptive group lasso penalty (QRVC-adp) and frequentist varying coefficient model with adaptive group lasso penalty (VC-adp). The coefficients of the d basis functions of the varying coefficient γ_j are treated as one group and are subject to selection at the group level.The three alternative methods, BQRVC, BVCSS and BVC, are compared with the proposed method to evaluate the strength of the spike-and-slab prior and the necessity of fitting quantile regression.

We comprehensively evaluate the proposed and alternative methods through simulation studies at quantiles 0.3, 0.5 and 0.7. The responses are generated according to model 4.1 with sample size n=200 and p=100 genetic factors. When the number of basis function d=5, the total dimension of the regression coefficient is 505, which forms 101 groups with group size equal to 5 and the first group corresponds to the varying intercept. The coefficients are set as $\gamma_0(U) = 2 + 2\sin(2\pi U), \gamma_1(U) = 2\exp(2U-1), \gamma_2(U) = -6U(1-U), \gamma_3(U) = -4U^3$. The rest of the coefficients are set to 0. The genetic factors are simulated in two different scenarios. In the first scenario, the genetic factors are simulated as gene expressions from a multivariate normal distribution with mean 0 and an AR-1 covariance matrix with marginal mean 0 and correlation coefficient 0.5. In the second scenario, we generate the genetic factors as SNP data by dichotomizing the gene expressoin values of each gene in Scenario 1 at the 1st and 3rd quartiles, leading to the 3-level categories (0,1,2) for genotypes (aa, Aa, AA).

We consider 5 choices of error distribution for ϵ_i's in model 4.1: N(μ, 1)(Error 1), 80%N(μ,1) + 20%Normal(μ, 3) (Error 2), Laplace(μ,b) with the scale parameter b = 1 (Error 3), LogNormal(μ,1) (Error 4), t(2) with mean=μ (Error 5). All of them are heavy-tailed distributions but Error 1. For each error, μ is chosen so that the θth quantile is 0. We also consider the case of heterogeneous random errors by replacing the i.i.d random errors

in model 4.1 by $(1 + X_{i2})\epsilon_i$ and in this case the responses are generated as:

$$Y_i = \sum_{k=1}^{q} E_{ik}\beta_k + \sum_{j=0}^{p} \gamma_j(U_i)X_{ij} + (1 + X_{i2})\epsilon_i,$$

where X_{i2} corresponds to the second genetic factor.

Then we evaluate the performance of each method using identification and estimation accuracy. Identification performance is evaluated based on the proportion of times a method underselecting (U), overselecting (O) and correctly selecting (C) the covariates with nonzero coefficients. We use the integrated mean squared error (IMSE) to assess the estimation accuracy of each method on nonlinear effects. Let $\hat{\alpha}_j(U)$ denote the estimated nonparametric function $\alpha_j(U)$ and $(U_1, ..., U_{n_{grid}})$ be the grid of points where α_j is evaluated. Then the IMSE of $\hat{\alpha}_j(U)$ is given as $\text{IMSE}(\hat{\alpha}_j(U)) = \frac{1}{n_{grid}} \sum_{t=1}^{n_{grid}} \left(\hat{\alpha}_j(U_t) - \alpha_j(U_t)\right)^2$. We use the total mean squared error (TMSE), which is the sum of all the IMSE's, to denote the overall estimation accuracy. Prediction performance is assessed based on the mean prediction errors, which are calculated as check loss for quantile regression and squared loss for the rest of the methods, on an independently generated testing dataset with the same settings. Besides, the mean absolute prediction errors are also calculated. The simulation performance is evaluated based on 100 replicates.

We have collected the posterior samples from the Gibbs sampler running 10,000 iterations in wich the first 5,000 samples as burn-ins. The Bayesian estimates are calculated using the posterior medians. As methods BQRVCSS and BVCSS incorporate spike–and–slab priors, we consider the median probability model (MPM) to identify the important effects that are associated with the response. We define the indicator $\phi_j^{(m)} = 1$ if the jth predictor is included in the model in the mth iteration. Suppose M posterior samples are collected from the MCMC after burn-ins. Then the posterior probability of including the jth predictor in the final model is given as

$$prob_j = \hat{\pi}(\phi_j = 1|y) = \frac{1}{M} \sum_{m=1}^{M} \phi_j^{(m)}, j = 1, ..., p.$$

A higher posterior inclusion probability p_j indicates a stronger empirical evidence that the corresponding predictor has a non-zero coefficient and is associated with the response variable. The MPM model is defined as the model consisting of predictors with at least $\frac{1}{2}$posterior inclusion probability. Barbieri and Berger recommend using MPM because of its optimal prediction performance when the goal is to select a single model. For methods without spike–and–slab priors, the 95% credible interval (95%CI) is adopted to select important varying effects.

The simulation results of the 4 methods are tabulated in Table 4.1, Table 4.2 and Table C1 to Table C6 in the Appendix. In general, the proposed method, BQRVCSS, does a better job in terms of identification, estimation and prediction accuracy compared with the alternative methods on the heavy-tailed distributions. For example Table 4.1 and Table 4.2 contain the result for the simulated gene expression data with i.i.d. random errors. At each setting, BQRVCSS outperforms its alternatives. For example, at quantile 0.5 with the t(2) error distribution, BQRVCSS correctly selects the exact model 97% of the times, while the percentage for BQRVC is 18%, 50% for BVCSS, 4% for BVC, 88% for QRVC-adp and 46% for VC-adp. The TMSE's for the six methods are 0.33 (sd 0.23), 4.35 (sd 0.78), 2.04 (sd 1.48), 6.82 (sd 6.51), 0.76 (sd 0.99) and 2.05 (sd 4.32), respectively. BQRVCSS has the smallest TMSE among them, which indicates it has the highest estimation accuracy. Besides, BQRVCSS has the smallest prediction error, 0.17 (sd 0.04), which is also smaller than those of its alternatives. The superior performance of BQRVCSS mainly lies in the robustness to skewed error distribution and spike and slab priors for achieving sparsity. It turns out all the six methods have a better performance at quantile 0.5 for all the five error distributions, except that the better performance occurs at quantile 0.3 for the lognormal error distribution, which is positively skewed, while the the other error distributions are all symmetric. For instance, when the error distribution is t(2), BQRVCSS selects the correct model 90% of the times at quantiles 0.3 and 0.7, which is less than that for quantile 0.5. BQRVCSS has TMSE and prediction error at quantile 0.3 equal to 0.44 (sd 0.24) and 0.22 (sd 0.06), respectively, which are less than those for quantile 0.5. However, for the lognormal error distribution, BQRVCSS has the best identification performance at quantile 0.3, with

a correct selection percentage of 99%, which is greater than the 97% for quantile 0.5. The percentage at quantile 0.7, which is 71%, is even lower. At quantile 0.3, the TMSE and prediction error for BQRVCSS are 0.11 (sd 0.05) and 0.09 (sd 0.01), respectively, while both tend to get larger as the quantile increases. The TMSE and prediction error at quantile 0.5 for BQRVCSS are 0.25 (sd 0.19) and 0.15 (sd 0.04), respectively, and they increase to 0.71 (sd 0.45) and 0.30 (sd 0.11) at quantile 0.7. Table C1 and Table C2 tabulate the simulation result for the gene expression data with heterogeneous errors. There's no difference for the quantile methods between the i.i.d. and heterogeneous errors, which is due to the property of robustness. However, the non-quantile methods perform worse when the random errors are heterogeneous. For example, BVCSS has a correct selection percentage of 50% at quantile 0.5 with i.i.d. t(2) error. The TMSE and prediction error are 2.04 (sd 1.48) and 0.28 (sd 1.33), respectively. But for the heterogeneous t(2) error, those terms for BVCSS at quantile 0.5 are 28%, 3.33 (sd 3.15) and 0.39(sd 2.16), respectively, which suggests worse performance. Table C3 to Table C6 have the results for the simulated SNP data with the same settings for the simulated gene expression data and we get similar findings.

We also made plots for the varying coefficients in the simulation study. Continue using Error 2 as an example, Figure 4.1 shows the estimated varying coefficients from the proposed method (BQRVCSS) fit the underlying trend of varying coefficients relatively well. We assess the convergence of the MCMC chains using the potential scale reduction factor (PSRF) (Gelman and Rubin (1992),Brooks and Gelman (1998)) following the work of Li et al. (2015). It implies that the chains converge to a stationary distribution if PSRF values are close to 1. According to Gelman et al. (2013), we adopt PSRF$\leqslant$1.1 as the cutoff threshold for convergence. Then we compute the PSRF for each parameter in our study and it turns out all chains converge after burn-ins. Figure 4.2 clearly shows the PSRF of the five estimated spline coefficients of each varying coefficient function below the threshold, indicating convergence of the Gibbs sampler.

We demonstrate the sensitivity of the proposed method BQRVCSS for variable selection to the choice of the hyperparameters for π_0 and η in the Appendix and tabulate the results from Table C7 to Table C10. These results suggest that the MPM model is insensitive to

different choices of the hyperparameters. We also conduct senstitivity analysis on whether the smoothness specification of the parameters in the B spline will impact the variable selection. The sensitivity analysis results are shown in Table C11 to Table C16 in the Appendix. It is evident that the proposed method is insensitive to the number of spline basis in smoothness specification. Based on this finding, we set the degree $O = 2$ and the number of interior knots $M = 2$ for the B spline basis, which leads to $d = 5$ basis functions.

Table 4.1: *Identification results for i.i.d. errors based on 100 replicates. C: correct-fitting proportion; O: overfitting proportion; U: underfitting proportion.*

θ			BQRVCSS	BQRVC	BVCSS	BVC	QRVC-adp	VC-adp
$\theta = 0.3$	Normal	C	0.96	0.70	0.90	0.34	0.90	0.83
		O	0.04	0.08	0.10	0.44	0.10	0.15
		U	0	0.22	0	0.22	0	0.03
	NormalMix	C	0.89	0.38	0.86	0.24	0.82	0.80
		O	0.11	0.16	0.14	0.50	0.18	0.20
		U	0	0.46	0	0.26	0	0
	Laplace	C	0.90	0.68	0.86	0.32	0.90	0.85
		O	0.10	0.12	0.14	0.58	0.10	0.15
		U	0	0.2	0	0.1	0	0
	Lognormal	C	0.99	0.36	0.70	0.10	0.82	0.56
		O	0.01	0.04	0.18	0.48	0.16	0.42
		U	0	0.62	0.12	0.42	0.02	0.02
	t(2)	C	0.90	0.18	0.26	0.10	0.83	0.30
		O	0.10	0.02	0.42	0.24	0.18	0.50
		U	0	0.80	0.32	0.66	0	0.20
$\theta = 0.5$	Normal	C	0.98	0.70	0.98	0.36	0.90	0.87
		O	0.02	0.16	0.02	0.6	0.10	0.13
		U	0	0.14	0	0.04	0	0
	NormalMix	C	0.96	0.42	0.86	0.12	0.90	0.84
		O	0.04	0.10	0.14	0.74	0.10	0.16
		U	0	0.48	0	0.14	0	0
	Laplace	C	0.94	0.7	0.90	0.4	0.90	0.85
		O	0.06	0.12	0.10	0.58	0.10	0.15
		U	0	0.18	0	0.02	0	0
	Lognormal	C	0.97	0.32	0.60	0.14	0.86	0.62
		O	0.03	0.08	0.32	0.42	0.10	0.38
		U	0	0.60	0.08	0.44	0.04	0
	t(2)	C	0.96	0.18	0.50	0.04	0.88	0.46
		O	0.02	0.02	0.26	0.42	0.08	0.50
		U	0.02	0.80	0.24	0.54	0.04	0.04
$\theta = 0.7$	Normal	C	0.96	0.70	0.96	0.32	0.90	0.90
		O	0.04	0.16	0.04	0.64	0.10	0.10
		U	0	0.14	0	0.04	0	0
	NormalMix	C	0.90	0.36	0.86	0.16	0.82	0.84
		O	0.10	0.14	0.12	0.66	0.18	0.16
		U	0	0.50	0.02	0.18	0	0
	Laplace	C	0.90	0.56	0.89	0.32	0.88	0.70
		O	0.10	0.20	0.11	0.58	0.12	0.30
		U	0	0.24	0	0.10	0	0
	Lognormal	C	0.68	0.2	0.64	0.12	0.56	0.62
		O	0.30	0.14	0.20	0.40	0.40	0.34
		U	0.02	0.66	0.16	0.48	0.04	0.04
	t(2)	C	0.90	0.18	0.42	0.12	0.85	0.32
		O	0.10	0.02	0.28	0.30	0.04	0.58
		U	0	0.80	0.30	0.58	0.08	0.10

Table 4.2: *Estimation and prediction results for i.i.d. errors based on 100 replicates. TMSE: total mean squared equared error. pred: prediction error (check loss for quantile methods and squared loss for non-quantile methods). pred.mad: mean absolute prediction error.*

θ			BQRVCSS	BQRVC	BVCSS	BVC	QRVC-adp	VC-adp
$\theta = 0.3$	Normal	TMSE	0.23(0.10)	2.28(0.35)	0.45(0.09)	1.56(0.16)	0.25(0.10)	0.70(0.09)
		pred	0.14(0.03)	0.31(0.02)	0.21(0.08)	1.08(0.11)	0.17(0.03)	0.23(0.02)
		pred.mad	0.29(0.05)	0.72(0.04)	0.57(0.06)	0.83(0.04)	0.34(0.06)	0.66(0.04)
	NormalMix	TMSE	0.34(0.19)	3.9(0.62)	0.76(0.27)	3.04(0.43)	0.45(0.23)	0.92(0.16)
		pred	0.18(0.05)	0.44(0.05)	0.23(0.21)	2.22(0.30)	0.22(0.05)	0.27(0.03)
		pred.mad	0.35(0.07)	0.92(0.06)	0.67(0.11)	1.08(0.06)	0.44(0.08)	0.77(0.07)
	Laplace	TMSE	0.27(0.13)	2.97(0.45)	0.47(0.15)	2.12(0.31)	0.26(0.11)	0.71(0.11)
		pred	0.17(0.05)	0.38(0.04)	0.39(0.11)	1.44(0.20)	0.18(0.05)	0.25(0.02)
		pred.mad	0.31(0.06)	0.78(0.05)	0.52(0.08)	0.90(0.06)	0.34(0.07)	0.67(0.05)
	Lognormal	TMSE	0.11(0.05)	3.38(0.55)	1.14(0.85)	5.84(1.92)	0.18(0.41)	1.22(2.45)
		pred	0.09(0.01)	0.34(0.03)	0.24(0.61)	4.42(1.50)	0.10(0.05)	0.36(0.10)
		pred.mad	0.19(0.03)	0.78(0.07)	1.12(0.16)	1.18(0.15)	0.24(0.10)	1.13(0.26)
	t(2)	TMSE	0.44(0.24)	5.01(1.16)	2.63(5.24)	8.35(9.76)	0.84(0.98)	2.58(3.22)
		pred	0.22(0.06)	0.58(0.09)	0.32(3.27)	8.90(4.58)	0.29(0.14)	0.42(0.18)
		pred.mad	0.39(0.08)	1.05(0.11)	0.99(0.49)	1.35(0.25)	0.54(0.23)	1.05(0.36)
$\theta = 0.5$	Normal	TMSE	0.21(0.06)	2.42(0.36)	0.40(0.06)	1.57(0.16)	0.21(0.07)	0.62(0.11)
		pred	0.14(0.02)	0.36(0.02)	0.23(0.04)	1.80(0.08)	0.16(0.03)	0.28(0.02)
		pred.mad	0.28(0.04)	0.71(0.04)	0.52(0.04)	0.72(0.04)	0.31(0.05)	0.55(0.04)
	NormalMix	TMSE	0.31(0.17)	3.75(0.6)	0.74(0.24)	2.71(0.49)	0.35(0.16)	0.92(0.11)
		pred	0.16(0.03)	0.46(0.04)	0.27(0.12)	1.87(0.33)	0.19(0.03)	0.31(0.03)
		pred.mad	0.32(0.07)	0.92(0.08)	0.39(0.08)	0.95(0.08)	0.38(0.06)	0.61(0.06)
	Laplace	TMSE	0.22(0.06)	3.07(0.48)	0.46(0.08)	1.83(0.28)	0.22(0.09)	0.70(0.08)
		pred	0.15(0.02)	0.39(0.03)	0.17(0.05)	1.21(0.20)	0.15(0.03)	0.29(0.02)
		pred.mad	0.30(0.04)	0.78(0.06)	0.31(0.05)	0.79(0.06)	0.30(0.06)	0.58(0.04)
	Lognormal	TMSE	0.25(0.19)	4.59(0.94)	1.18(1.69)	5.09(2.28)	0.40(0.56)	1.26(0.68)
		pred	0.15(0.04)	0.46(0.06)	0.38(1.13)	3.95(1.80)	0.18(0.06)	0.43(0.08)
		pred.mad	0.30(0.08)	0.92(0.11)	0.83(0.22)	1.03(0.14)	0.36(0.12)	0.85(0.16)
	t(2)	TMSE	0.33(0.23)	4.35(0.78)	2.04(1.48)	6.82(6.51)	0.76(0.99)	2.05(4.32)
		pred	0.17(0.04)	0.49(0.04)	0.28(1.33)	5.05(4.55)	0.27(0.14)	0.34(0.26)
		pred.mad	0.35(0.08)	0.99(0.09)	0.89(0.26)	1.20(0.19)	0.52(0.23)	0.91(0.56)
$\theta = 0.7$	Normal	TMSE	0.21(0.08)	2.53(0.41)	0.41(0.08)	1.58(0.18)	0.23(0.1)	0.71(0.10)
		pred	0.15(0.03)	0.32(0.02)	0.36(0.07)	1.08(0.12)	0.16(0.03)	0.23(0.02)
		pred.mad	0.29(0.04)	0.74(0.05)	0.53(0.07)	0.83(0.05)	0.33(0.06)	0.65(0.05)
	NormalMix	TMSE	0.33(0.14)	3.84(0.58)	0.78(0.3)	3.03(0.53)	0.45(0.26)	0.92(0.18)
		pred	0.19(0.04)	0.44(0.03)	0.65(0.21)	2.25(0.35)	0.22(0.04)	0.26(0.02)
		pred.mad	0.36(0.06)	0.93(0.06)	0.68(0.11)	1.10(0.07)	0.42(0.07)	0.73(0.06)
	Laplace	TMSE	0.29(0.11)	3.22(0.49)	0.49(0.16)	2.18(0.34)	0.3(0.17)	0.73(0.12)
		pred	0.18(0.04)	0.39(0.03)	0.23(0.12)	1.50(0.20)	0.18(0.04)	0.24(0.02)
		pred.mad	0.33(0.06)	0.8(0.05)	0.52(0.08)	0.92(0.05)	0.35(0.07)	0.66(0.05)
	Lognormal	TMSE	0.71(0.45)	5.44(1.52)	0.99(0.9)	4.19(2.07)	0.96(0.95)	1.35(3.65)
		pred	0.30(0.11)	0.60(0.15)	0.35(0.59)	2.87(1.39)	0.33(0.16)	0.36(0.23)
		pred.mad	0.50(0.15)	0.99(0.18)	0.55(0.2)	1.07(0.11)	0.6(0.23)	0.73(0.46)
	t(2)	TMSE	0.42(0.35)	5.07(1.21)	2.65(3.35)	9.10(11.24)	0.97(1.42)	2.02(1.75)
		pred	0.22(0.07)	0.58(0.11)	0.35(2.31)	7.08(9.51)	0.30(0.18)	0.38(0.17)
		pred.mad	0.39(0.10)	1.07(0.12)	0.95(0.38)	1.39(0.30)	0.57(0.29)	0.96(0.32)

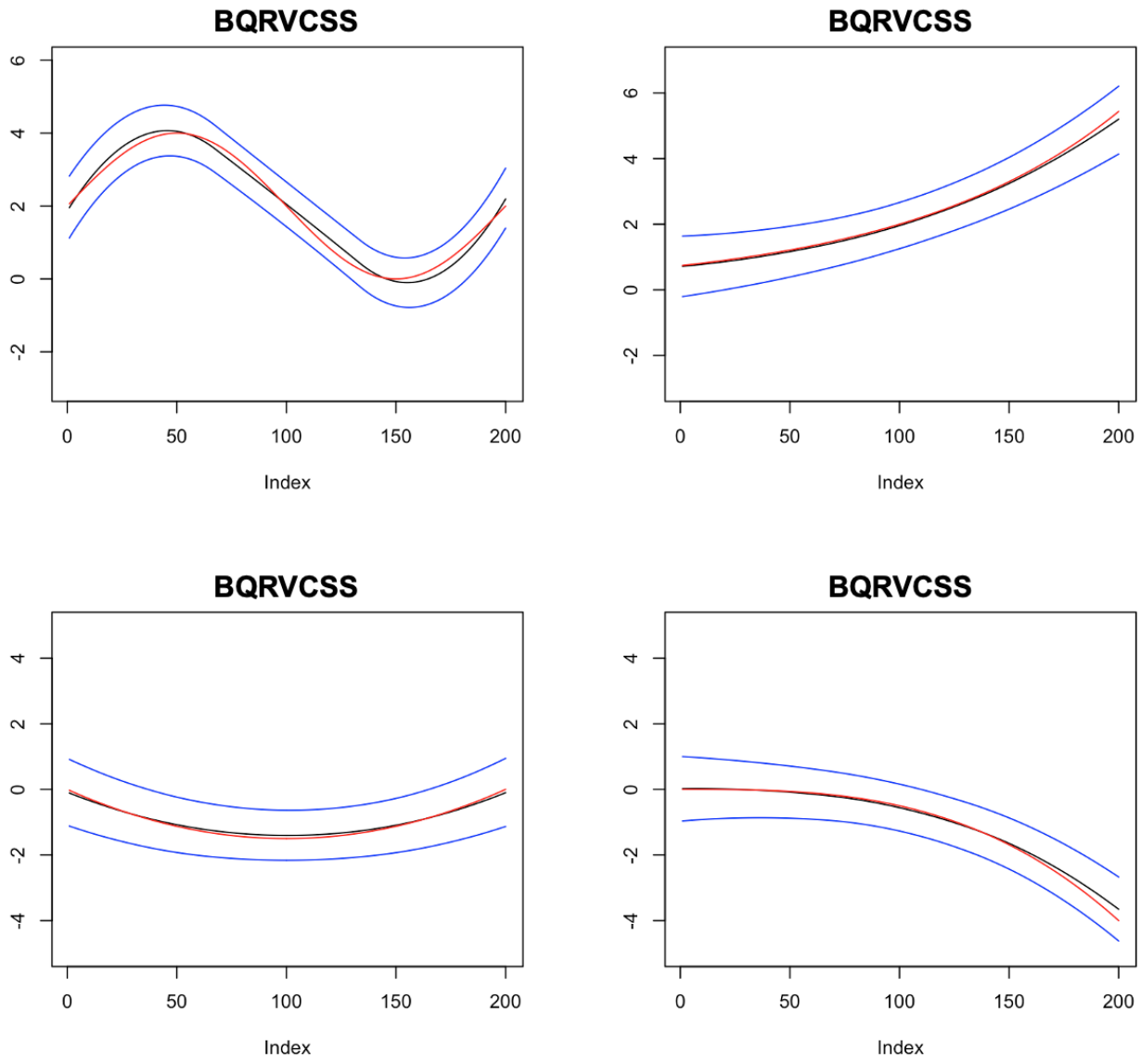


Figure 4.1: *Simulation study for Error using the proposed method (BQRVCSS). Red line: true parameter values. Black line: median estimates of varying coefficients from BQRVCSS. Blue lines: 95% credible intervals for the estimated varying coefficients.*

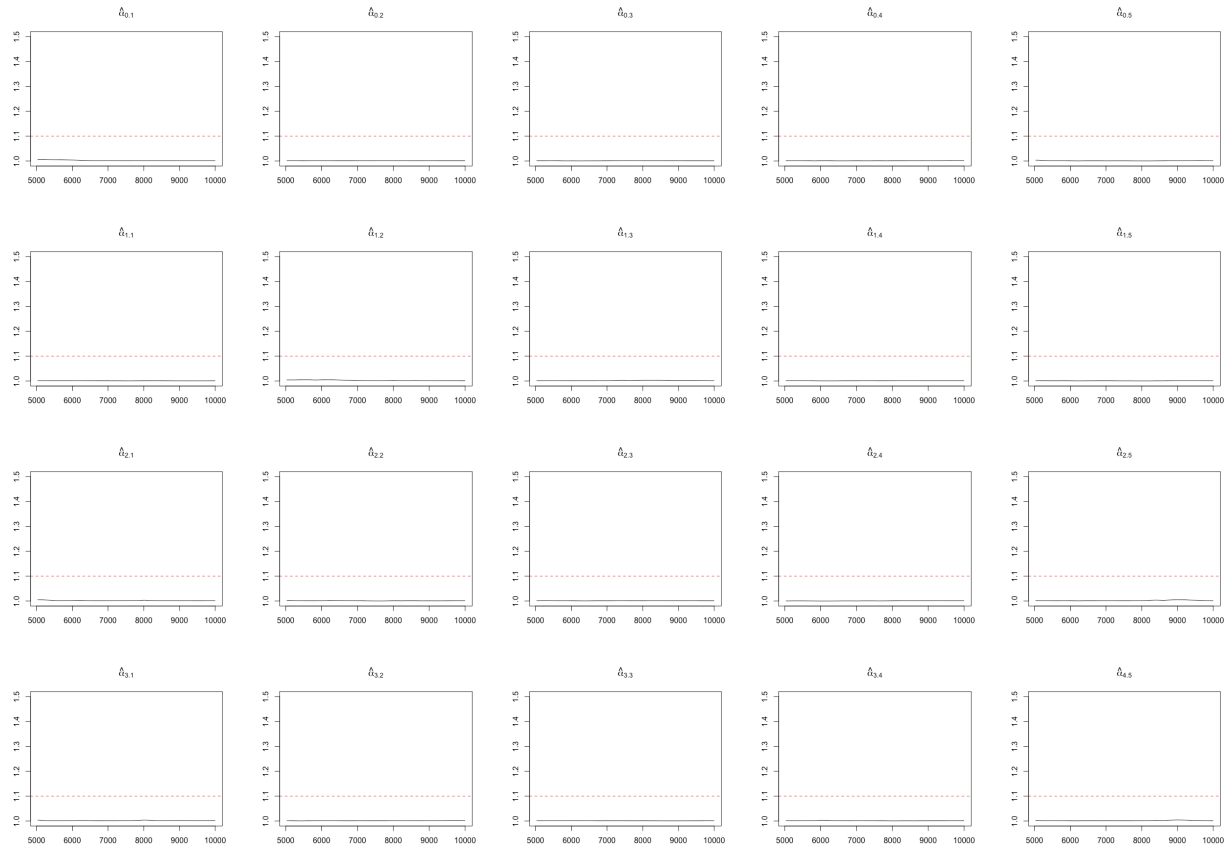


Figure 4.2: *Potential scale reduction factor (PSRF) versus iteractions for the varying functions in Figure 4.1. Black line: PSRF. Red line: the threshold of 1.1.* $\hat{\alpha}_{j1}$ *to* $\hat{\alpha}_{j5}(j = 0, ..., 3)$ *represent the five estimated spline coefficients for the varying coefficient function* γ_j*, respectively.*

4.5 Real Data Analysis

We analyze the dataset from the Nurse's Health Study (NHS). The body mass index (BMI), which can quantify the obesity leve, is set as the response. We focus on SNPs on chromosome 2. We consider age as the environment factor since it is known to be associated with the variations in the obesity level. Besides, three clinical covariates are included: total physical activity, trans fat intake and cereal fiber intake. Only the health subjects in the NHS are selected in our case study. We clean the data by keeping subjects with matched phenotypes

and genotypes, removing SNPs with minor allele frequency (MAF) less than 0.05 or deviation from Hardy-Weinberg equilibrium and obtain a working dataset that contains 1716 subjects with 53,408 SNPs. We impute the missing data using fastPHASE (Scheet and Stephens (2006)).

We reduce the feature space through prescreening to make it more attainable for variable selection. For example, Li et al. (2015) perform the single SNP analysis to filter SNPs in a GWA study before downstream analysis. In our case study, we screen the SNPs using the established procedure as described by Ma et al. (2011) and Wu and Cui (2013). Here, we conduct three statistical tests to evaluate the penetrance effect of a variant under the environmental stimuli to test whether the interaction effects are nonlinear, linear, constant, or zero. We keep the SNPs with p-values less than a certain cutoff (0.005, for instance) from any of the tests with BMI as the response. 300 SNPs pass the screening.

We analyze the screened data using the proposed method BQRVCSS at the median and the alternative BVCSS. Other methods, such as BQRVC and BVC are not considered since they have inferior performance in the simulation studies. BQRVCSS identifies 11 SNPs while BVCSS identifies 9 SNPs. The identification results are displayed in Figures 4.3 and 4.4. We can see 6 SNPs are commonly identified by both methods. Besides, the proposed method uniquely identified 5 other SNPs and the genes where the SNPs are located have been found to be associated with body weight change. For example, BQRVCSS identifies the SNP rs17783776, which is located in the gene ALK. ALK (anaplastic lymphoma kinase) has been identified as a thinness gene which suggests it could be the target gene for obesity treatment (Orthofer et al. (2020)). As a comparison, the alternative method BVCSS misses this important gene. The proposed method also identifies rs 41349646, a SNP that is mapped to the gene NPAS2. NPAS2 has been found to play an essential role in the regulation of peripheral circadian response and hepatic metabolism, therefore affects weight change (O'Neil et al. (2013)). The SNP rs10933420 is also uniquely identified by our proposed method and it is located in the gene NGEF. Kim et al. (2015) has found NGEF associated with intra-abdominal fat accumulation. Besides, our proposed method BQRVCSS identifies rs4854071 as well. The SNP rs4854071 is located within the gene NDUFA10 (NADH:Ubiquinone Ox-

idoreductase Subunit A10), which has been found to be involved in the NAFLD pathway regulating weight loss together with 10 other genes (Mirhashemi et al. (2021)).

We also applied the proposed method BQRVCSS to the screened data at other quantiles, such as 0.3 and 0.7. BQRVCSS identifies 11 SNPs at quantile 0.3. Compared with the identification result of BVCSS, BQRVCSS uniquely identifies 10 SNPs and the remaining 1 SNP is also identified by BVCSS. Looking into this difference, we obtain some interesting findings. For example, BQRVCSS identifies the SNP rs10084365 at quantile 0.3, while BVCSS doesn't. We locate the SNP rs10084365 to the gene GPR39, which is a constitutively active 7TM receptor, and its deficiency has been found to be associated with obesity (Petersen et al. (2011)). BQRVCSS also identifies the SNP rs11885893, which is mapped to the gene PLEKHH2. Benton et al. (2015) has found PLEKHH2 associated with weight loss through the regulation of DNA methylation. At the quantile 0.7, BQRVCSS identifies 10 SNPs, 3 of which are commonly identified by BVCSS. Interestingly, BQRVCSS also identifies the SNP rs4854071 as it does at the median. BQRVCSS uniquely identifies the SNP rs752833 at quantile 0.7, while BVCSS misses this SNP. SNP rs752833 is located to the gene ACOXL, which a member of the acyl-CoA oxidase family involved in lipid metabolism and therefore associated with obesity (Vuillaume et al. (2014)). Besides, BQRVCSS uniquely identifies the SNP rs17533992, which is located within the gene SPRED2. Ohkura et al. (2019) has uncovered that SPRED2 regulates high fat diet-induced obesity negatively. BQRVCSS also identifies rs4894108. The identified SNP rs4894108 is located in ZNF385B, which has been found to be associated with obesity (Kim et al. (2012))

It is difficult to evaluate the selection performance with real data objectively. The prediction performance is evaluated as it may provide partial information on the relative performance of different methods. We refit the selected model of each method by Bayesian LASSO following the methods in Li et al. (2015) and Yan and Huang (2012). The prediction mean squared errors (PMSEs) and prediction mean absolute deviations (PMADs) are computed based on the posterior median estimates. The proposed method BQRVCSS has the PMSE and PMAD equal to 13.13 and 1.34, respectively, while the PMSE and PMAD for BVCSS are 15.04 and 3.05, which are both larger than the counterparts of BQRVCSS. Therefore,

the proposed method has better performance.

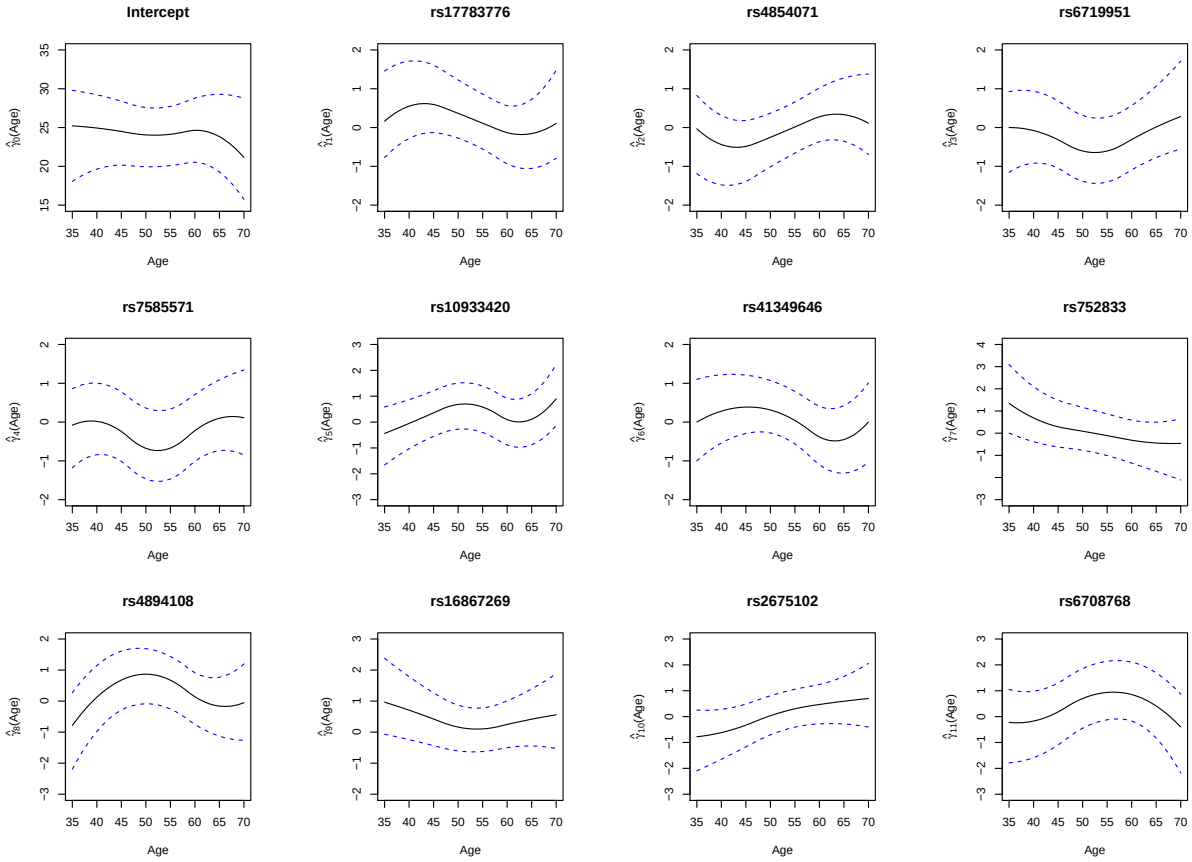


Figure 4.3: *Real data analysis using the proposed method (BQRVCSS). Black line: median estimates of varying coefficients for BQRVCSS. Blue dashed lines: 95% credible intervals for the estimated varying coefficients.*

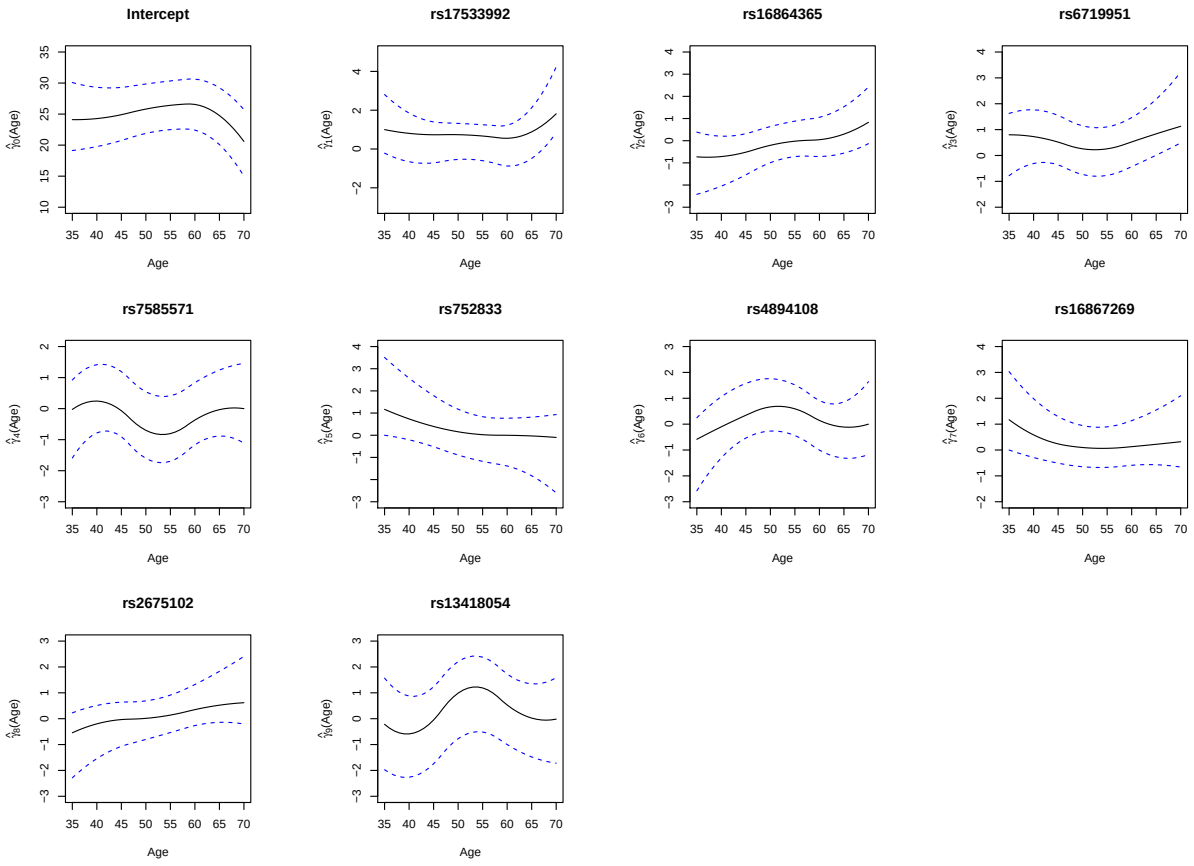


Figure 4.4: *Real data analysis using the alternative method (BVCSS). Black line: median estimates of varying coefficients for BVCSS. Blue dashed lines: 95% credible intervals for the estimated varying coefficients.*

4.6 Discussion

High-dimensional data, which have the "large p, small n" nature, frequently occur in biomedical studies, such as genomewide association studies and clinical research. As only a subset of the covariates is associated with the response variable while the rest are irrelevant, penalized variable selection methods have been developed to overcome "the curse of dimensionality". Besides, in practice, parametric models are not sufficient enough to capture the true underlying relationship between the response variabe and the covariates when the dynamic changes of the regression coefficients exist. This brings obstacles to identify the important covariates

that are associated with the response.

In particular, we are interested in the varying coefficient model (Hastie and Tibshirani (1993)) where the regression coefficients depend on some covariate through a nonparametric function, thus the VC model provides more flexibility than the linear models and reduces biases. Although the asymptotic theory for the varying coefficient model has been well developed, a Bayesian approach has also been established. Biller and Fahrmeir (2001) proposed a Bayesian B-spline basis function approach to varying coefficient models with adaptive knot selection. Reich et al. (2010) introduced a Bayesian variable selection procedure for multivariate spatially varying coefficient regression. Li et al. (2015) incorporated Bayesian group lasso algorithm to the high-dimensional varying coefficient model, which is applied to functional genome-wide association studies.

Although the penalized linear squares approach has become an useful tool in variable selection, there is a drawback in that it summarized the average relationship between the response variable and covariates based on the conditional mean function, which only provides a partial view of the relationship. It is possible that a certain covariate may not have a significant effect on the mean of the response but have a greater influence at other segments of the conditional distribution. Quantile regression provides the capability of describing the relationship at different points in the conditional distribution of the response variable.

Quantile regression has become more and more popular in recent years as it is robust to non-normal errors and outliers while the ordinary least squares methods is inefficient. Quantile regression also provides richer information of the data than the classic mean regression. The development of regularized variable selection methods allows us to build a regularized quantile regression model. Koenker (2004) penalized the random effects in a mixed-effect quantile regression model and shrank the random effects towards zero. Li and Zhu (2008) incorporated the Lasso penalty to quantile regression and developed its piecewise linear solution path. Wu and Liu (2009) demonstrated the oracle properties of the SCAD and adaptive lasso penalties in regularized quantile regression. Li et al. (2010) developed regularized Bayesian quantile regression using Lasso, elastic net and group Lasso penalties. Noh et al. (2012) developed a penalized variable selection method for varying coefficient

models in quantile regression. There's a limitation in these studies that Bayesian regularized variable selection in quantile models with varying coefficients has not been well established. Therefore, we propose a novel Bayesian regularized quantile varying coefficient model to identify the important genetic covariates that are associated with the phenotype. Besides, we develop a C++ based R package, which incorporates the proposed and alternative Bayesian methods in this project and the package will be submitted to CRAN soon.

Chapter 5

Summary

In this dissertation, we aim at the development of data-driven penalized variable selection methods that enable efficient variable selection on longitudinal and nonlinear gene-environment interactions within both frequentist and Bayesian frameworks. The correlation nature within each cluster of repeated measurements on the response brings challenges to the methodology development, while existing penalized methods in longitudinal studies mainly focus on the identification of main effects only (Wang et al. (2012), Cho and Qu (2013)., Ma et al. (2013)). In Chapter 2, a novel Newton-Raphson based penalized variable selection method has been proposed to identify important lipid-environment interactions within the GEE framework in a longitudinal lipidomics study. Our method significantly advances the existing ones by considering the interaction effects. Simultaneous selection of both the main and interaction effects can be accommodated by the incorporation of the group structure within GEE. The paper associated with this study has been published at the Genes (Zhou et al. (2019)). As penalized variable selection has become a powerful tool in longitudinal interaction studies, we move on and develop a sparse group penalization method to carry out a bi-level selection on G×E interactions for the repeatedly measured phenotype. The penalized QIF framework is adopted as it has better performance compared with penalized GEE under a variety of settings. The proposed method enables a simultaneous identification of main and interaction effects on both the group and individual level. Simulation studies

and a case study have demonstrated that the proposed method outperforms the competing alternatives in terms of both identification and prediction with satisfactory computation speeds. In the last chapter of this dissertation, we have proposed a regularized Bayesian quantile varying coefficient model to identify non-linear G×E interactions. This method provides the capability of describing the relationship between the response and predictors at different quantiles of the response variable while effectively accommodating robustness to heavy-tailed errors and outliers in the response variable within the Bayesian framework. Moreover, this method accounts for sparsity in the identification of the non-linear G×E interactions. We have developed open-source R packages with core modules written in C++ for each project to facilitate fast computation. We have published the R packages interep and springer, which correspond to the first and second projects respectively, on CRAN. The R package associated with the third project will be publicly available in the near future.

Bibliography

Ahn, J., B. Mukherjee, S. B. Gruber, and M. Ghosh (2013). Bayesian semiparametric analysis for two-phase studies of gene-environment interaction. *Ann. Appl. Stat. 7*(1), 543–569.

Akinbami, O. J. (2006). The state of childhood asthma: United states, 1980-2005. (381).

Arredouani, M. S., F. Franco, A. Imrich, et al. (2007). Scavenger receptors SR-AI/II and MARCO limit pulmonary dendritic cell migration and allergic airway inflammation. *The Journal of Immunology 178*(9), 5912–5920.

Bandyopadhyay, S., B. Ganguli, and A. Chatterjee (2011). A review of multivariate longitudinal data analysis. *Statistical Methods in Medical Research 20*(4), 299–330.

Barona, T., R. Byrne, T. Pettitt, M. Wakelam, B. Larijani, and D. Poccia (2005). Diacylglycerol induces fusion of nuclear envelope membrane precursor vesicles. *J. Biol. Chem. 280*, 41171–41177.

Benton, M. C., A. Johnstone, D. Eccles, B. Harmon, M. T. Hayes, R. A. Lea, L. Griffiths, E. P. Hoffman, R. S. Stubbs, and D. Macartney-Coxson (2015). An analysis of DNA methylation in human adipose tissue reveals differential modification of obesity genes before and after gastric bypass and weight loss. *Genome Biology 16*(1).

Berridge, M. J. (1987). Inositol trisphosphate and diacylglycerol: Two interacting second messengers. *Annual Review of Biochemistry 56*(1), 159–193.

Bien, J., J. Taylor, and R. Tibshirani (2013). A lasso for hierarchical interactions. *Ann. Statist. 41*(3), 1111–1141.

Biller, C. and L. Fahrmeir (2001). Bayesian varying-coefficient models using adaptive regression splines. *Statistical Modelling 1*(3), 195–211.

Bowden, J. A., A. Heckert, C. Z. Ulmer, C. M. Jones, J. P. Koelmel, L. Abdullah, L. Ahonen, Y. Alnouti, A. M. Armando, J. M. Asara, T. Bamba, J. R. Barr, J. Bergquist, C. H. Borchers, J. Brandsma, S. B. Breitkopf, T. Cajka, A. Cazenave-Gassiot, A. Checa, M. A. Cinel, R. A. Colas, S. Cremers, E. A. Dennis, J. E. Evans, A. Fauland, O. Fiehn, M. S. Gardner, T. J. Garrett, K. H. Gotlinger, J. Han, Y. Huang, A. H. Neo, T. Hyötyläinen, Y. Izumi, H. Jiang, H. Jiang, J. Jiang, M. Kachman, R. Kiyonami, K. Klavins, C. Klose, H. C. Köfeler, J. Kolmert, T. Koal, G. Koster, Z. Kuklenyik, I. J. Kurland, M. Leadley, K. Lin, K. R. Maddipati, D. McDougall, P. J. Meikle, N. A. Mellett, C. Monnin, M. A. Moseley, R. Nandakumar, M. Oresic, R. Patterson, D. Peake, J. S. Pierce, M. Post, A. D. Postle, R. Pugh, Y. Qiu, O. Quehenberger, P. Ramrup, J. Rees, B. Rembiesa, D. Reynaud, M. R. Roth, S. Sales, K. Schuhmann, M. L. Schwartzman, C. N. Serhan, A. Shevchenko, S. E. Somerville, L. St John-Williams, M. A. Surma, H. Takeda, R. Thakare, J. W. Thompson, F. Torta, A. Triebl, M. Trötzmüller, S. J. K. Ubhayasekera, D. Vuckovic, J. M. Weir, R. Welti, M. R. Wenk, C. E. Wheelock, L. Yao, M. Yuan, X. H. Zhao, and S. Zhou (2017). Harmonizing lipidomics: Nist interlaboratory comparison exercise for lipidomics using srm 1950-metabolites in frozen human plasma. *Journal of Lipid Research 58*(12), 2275–2288.

Breheny, P. and J. Huang (2009). Penalized methods for bi-level variable selection. *Statistics and its Interface 2*(3), 369–380.

Briggs, M. A., K. S. Petersen, and P. M. Kris-Etherton (2017). Saturated fatty acids and cardiovascular disease: replacements for saturated fat to reduce cardiovascular risk. In *Healthcare*, Volume 5, pp. 29. Multidisciplinary Digital Publishing Institute.

Brooks, S. P. and A. Gelman (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics 7*(4), 434–455.

Cai, Z. and X. Xu (2009). Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association 104*(485), 371–383.

CDC. Data, statistics, and surveillance: asthma surveillance data [Internet]. Atlanta (GA): Centers for Disease Control and Prevention; 2020 [cited 2020 jan 23].

Checa, A., C. Bedia, and J. Jaumot (2015). Lipidomic data analysis: Tutorial, practical guidelines and applications. *Analytica Chimica Acta 885*, 1 – 16.

Chiang, C.-T., J. A. Rice, and C. O. Wu (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association 96*(454), 605–619.

Childhood Asthma Management Program Research Group (1999). The childhood asthma management program (CAMP): design, rationale, and methods. *Controlled clinical trials 20*(1), 91–120.

Childhood Asthma Management Program Research Group (2000). Long-term effects of budesonide or nedocromil in children with asthma. *New England Journal of Medicine 343*(15), 1054–1063.

Cho, H. and A. Qu (2013). Model selection for correlated data with diverging number of parameters. *Statistica Sinica 23*(2), 901–927. Full publication date: April 2013.

Choi, N. H., W. Li, and J. Zhu (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association 105*(489), 354–364.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics 11*(20), 2463–2468.

Cordell, H. J. and D. G. Clayton (2005). Genetic association studies. *The Lancet 366*(9491), 1121–1131.

Cornelis, M. C. and F. B. Hu (2012). Gene-environment interactions in the development of type 2 diabetes: Recent progress and continuing challenges. *Annual Review of Nutrition 32*(1), 245–259. PMID: 22540253.

Cornelis, M. C., E. J. Tchetgen Tchetgen, L. Liang, L. Qi, N. Chatterjee, F. B. Hu, and P. Kraft (2012). Gene-environment interactions in genome-wide association studies: A comparative study of tests applied to empirical studies of type 2 diabetes. *American Journal of Epidemiology 175*(3), 191–202.

Covar, R. A., A. L. Fuhlbrigge, P. Williams, and H. W. Kelly (2012). The childhood asthma management program (camp): contributions to the understanding of therapy and the natural history of childhood asthma. *Current Respiratory Care Reports 1*(4), 243–250.

Cui, Y., G. Kang, K. Sun, M. Qian, R. Romero, and W. Fu (2008). Gene-centric genomewide association study via entropy. *Genetics 179*(1), 637–650.

Dempfle, A., A. Scherag, R. Hein, L. Beckmann, J. Chang-Claude, and H. Schäfer (2008). Gene–environment interactions for complex traits: definitions, methodological requirements and challenges. *European Journal of Human Genetics 16*(10), 1164–1172.

Du, Y., K. Fan, X. Lu, and C. Wu (2021). Integrating multi–omics data for gene-environment interactions. *BioTech 10*(1), 3.

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

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

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

Fan, J. and W. Zhang (2008). Statistical methods with varying coefficient models. *Statistics and its Interface 1*(1), 179.

Fan, Y. and R. Li (2012). Variable selection in linear mixed effects models. *Annals of Statistics 40*(4), 2043.

Fan, Y., G. Qin, and Z. Zhu (2012). Variable selection in robust regression models for longitudinal data. *Journal of Multivariate Analysis 109*, 156 – 167.

Filzmoser, P., B. Liebmann, and K. Varmuza (2009). Repeated double cross validation. *Journal of Chemometrics 23*(4), 160–171.

Flowers, E., E. S. Froelicher, and B. E. Aouizerat (2012). Gene-environment interactions in cardiovascular disease. *European journal of cardiovascular nursing : journal of the Working Group on Cardiovascular Nursing of the European Society of Cardiology 11*(4), 472–478. 21684212[pmid].

Friedman, J., T. Hastie, and R. Tibshirani (2010). A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis*. CRC press.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4).

George, E. I. and R. E. McCulloch (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association 88*(423), 881–889.

Goñi, F. M. and A. Alonso (1999). Structure and functional properties of diacylglycerols in membranes1this work is dedicated to professor vittorio luzzati on occasion of his 75th birthday.1. *Progress in Lipid Research 38*(1), 1–48.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological) 55*(4), 757–779.

Hirschhorn, J. N., K. Lohmueller, E. Byrne, and K. Hirschhorn (2002). A comprehensive review of genetic association studies. *Genetics in Medicine 4*(2), 45–61.

Hoover, D. R., J. A. Rice, C. O. Wu, and L.-P. Yang (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85*(4), 809–822.

Huang, H. and Y. Liang (2019). A novel cox proportional hazards model for high-dimensional genomic data in cancer prognosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Huang, H.-H., J.-G. Dai, and Y. Liang (2018). Clinical drug response prediction by using a lq penalized network-constrained logistic regression method. *Cellular Physiology and Biochemistry 51*(5), 2073–2084.

Huang, H.-H. and Y. Liang (2018). Hybrid l1/2+ 2 method for gene selection in the cox proportional hazards model. *Computer Methods and Programs in Biomedicine 164*, 65–73.

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

Huang, J., J. L. Horowitz, and F. Wei (2010). Variable selection in nonparametric additive models. *Annals of Statistics 38*(4), 2282.

Huang, J. Z., C. O. Wu, and L. Zhou (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika 89*(1), 111–128.

Huang, J. Z., C. O. Wu, and L. Zhou (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, 763–788.

Hunter, D. J. (2005). Gene–environment interactions in human diseases. *Nature Reviews Genetics 6*(4), 287–298.

Jiang, L., J. Liu, X. Zhu, M. Ye, L. Sun, X. Lacaze, and R. Wu (2015). 2HiGWAS: a unifying high-dimensional platform to infer the global genetic architecture of trait development. *Briefings in Bioinformatics 16*(6), 905–911.

Jiang, Y. (2012). IGF-1 mediates exercise-induced phospholipid alteration in the murine skin tissues. *Journal of Nutrition & Food Sciences S2*(01).

Kim, H.-J., J.-H. Park, S. Lee, H.-Y. Son, J. Hwang, J. Chae, J. M. Yun, H. Kwon, J.-I. Kim, and B. Cho (2015). A common variant of ngef is associated with abdominal visceral fat in korean men. *PLOS ONE 10*(9), 1–11.

Kim, J., T. Lee, T.-H. Kim, K.-T. Lee, and H. Kim (2012). An integrated approach of comparative genomics and heritability analysis of pig and human on obesity trait: evidence for candidate genes on human chromosome 2. *BMC Genomics 13*(1), 711.

Kim, M.-O. (2007). Quantile regression with varying coefficients. *The Annals of Statistics*, 92–108.

King, B. S., L. Lu, M. Yu, Y. Jiang, J. Standard, X. Su, Z. Zhao, and W. Wang (2015). Lipidomic profiling of di- and tri-acylglycerol species in weight-controlled mice. *PLOS ONE 10*(2), 1–12.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis 91*(1), 74 – 89. Special Issue on Semiparametric and Nonparametric Mixed Models.

Kujala, M. and J. Nevalainen (2015). A case study of normalization, missing data and variable selection methods in lipidomics. *Statistics in Medicine 34*(1), 59–73.

Kuo, L. and B. Mallick (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 65–81.

Kyung, M., J. Gill, M. Ghosh, and G. Casella (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Anal. 5*(2), 369–411.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist. 44*(3), 907–927.

Li, J., Q. Lu, and Y. Wen (2019). Multi-kernel linear mixed model with adaptive lasso for prediction analysis on high-dimensional multi-omics data. *Bioinformatics 36*(6), 1785–1794.

Li, J., Z. Wang, R. Li, and R. Wu (2015). Bayesian group lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *Ann. Appl. Stat. 9*(2), 640–664.

Li, J., W. Zhong, R. Li, and R. Wu (2014). A fast algorithm for detecting gene–gene interactions in genome-wide association studies. *The Annals of Applied Statistics 8*(4), 2292.

Li, Q., R. Xi, and N. Lin (2010). Bayesian regularized quantile regression. *Bayesian Analysis 5*(3), 533–556.

Li, Y., F. Wang, R. Li, and Y. Sun (2020). Semiparametric integrative interaction analysis for non-small-cell lung cancer. *Statistical Methods in Medical Research 29*(10), 2865–2880.

Li, Y. and J. Zhu (2008). L1-norm quantile regression. *Journal of Computational and Graphical Statistics 17*(1), 163–185.

LIANG, K.-Y. and S. L. ZEGER (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Lockhart, R., J. Taylor, R. J. Tibshirani, and R. Tibshirani (2014). A significance test for the lasso. *Ann. Statist. 42*(2), 413–468.

Lu, X., K. Fan, J. Ren, and C. Wu (2021). Identifying gene-environment interactions with robust marginal bayesian variable selection. *Frontiers in Genetics (accepted)*.

Lunetta, K. L. (2008). Genetic association studies. *Circulation 118*(1), 96–101.

Ma, S., Q. Song, and L. Wang (2013). Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustered data. *Bernoulli 19*(1), 252–274.

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

Ma, S., L. Yang, R. Romero, and Y. Cui (2011). Varying coefficient model for gene–environment interaction: a non-linear look. *Bioinformatics 27*(15), 2119–2126.

Markgraf, D., H. Al-Hasani, and S. Lehr (2016). Lipidomics—reshaping the analysis and perception of type 2 diabetes. *Int. J. Mol. Sci. 17*, 1841.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 417–473.

Mirhashemi, M. E., R. V. Shah, R. R. Kitchen, J. Rong, A. Spahillari, A. R. Pico, O. Vitseva, D. Levy, D. Demarco, S. Shah, M. D. Iafrati, M. G. Larson, K. Tanriverdi, and J. E. Freedman (2021). The dynamic platelet transcriptome in obesity and weight loss. *Arteriosclerosis, Thrombosis, and Vascular Biology 41*(2), 854–864.

Mitchell, T. J. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*(404), 1023–1032.

Mukherjee, B., J. Ahn, S. B. Gruber, and N. Chatterjee (2012). Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *American journal of epidemiology 175*(3), 177–190.

Murcray, C. E., J. P. Lewinger, and W. J. Gauderman (2009). Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology 169*(2), 219–226.

Noguchi, E., Y. Yokouchi, J. Zhang, et al. (2005). Positional identification of an asthma susceptibility gene on human chromosome 5q33. *American Journal of Respiratory and Critical Care Medicine 172*(2), 183–188.

Noh, H., K. Chung, and I. V. Keilegom (2012). Variable selection of varying coefficient models in quantile regression. *Electronic Journal of Statistics 6*(0), 1220–1238.

Noh, H. S. and B. U. Park (2010). Sparse varying coefficient models for longitudinal data. *Statistica Sinica 20*(3), 1183–1202.

O'Hara, R. B. and M. J. Sillanpaa (2009). A review of bayesian variable selection methods: what, how and which. *Bayesian Anal. 4*(1), 85–117.

Ohkura, T., T. Yoshimura, M. Fujisawa, T. Ohara, R. Marutani, K. Usami, and A. Matsukawa (2019). Spred2 regulates high fat diet-induced adipose tissue inflammation, and metabolic abnormalities in mice. *Frontiers in Immunology 10*.

O'Neil, D., H. Mendez-Figueroa, T.-A. Mistretta, C. Su, R. H. Lane, and K. M. Aagaard (2013). Dysregulation of npas2 leads to altered metabolic pathways in a murine knockout model. *Molecular Genetics and Metabolism 110*(3), 378–387.

Orthofer, M., A. Valsesia, R. Mägi, Q.-P. Wang, J. Kaczanowska, I. Kozieradzki, A. Leopoldi, D. Cikes, L. M. Zopf, E. O. Tretiakov, E. Demetz, R. Hilbe, A. Boehm, M. Ticevic, M. Nõukas, A. Jais, K. Spirk, T. Clark, S. Amann, M. Lepamets, C. Neumayr, C. Arnold, Z. Dou, V. Kuhn, M. Novatchkova, S. J. Cronin, U. J. Tietge, S. Müller, J. A. Pospisilik, V. Nagy, C.-C. Hui, J. Lazovic, H. Esterbauer, A. Hagelkruys, I. Tancevski, F. W. Kiefer, T. Harkany, W. Haubensak, G. G. Neely, A. Metspalu, J. Hager, N. Gheldof, and J. M. Penninger (2020). Identification of ALK in thinness. *Cell 181*(6), 1246–1262.e22.

Ottman, R. (1996). Gene–environment interaction: Definitions and study design. *Preventive Medicine 25*(6), 764 – 770.

Ouyang, P., Y. Jiang, H. M. Doan, L. Xie, D. Vasquez, R. Welti, X. Su, N. Lu, B. Herndon, S.-S. Yang, R. Jeannotte, and W. Wang (2010). Weight loss via exercise with controlled dietary intake may affect phospholipid profile for cancer prevention in murine skin tissues. *Cancer prevention research (Philadelphia, Pa.) 3*(4), 466–477. 20233900[pmid].

Park, T. and G. Casella (2008). The bayesian lasso. *Journal of the American Statistical Association 103*(482), 681–686.

Peng, B. and L. Wang (2015). An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics 24*(3), 676–694.

Petersen, P. S., C. Jin, A. N. Madsen, M. Rasmussen, R. Kuhre, K. L. Egerod, L. B. Nielsen, T. W. Schwartz, and B. Holst (2011). Deficiency of the GPR39 receptor is associated with obesity and altered adipocyte metabolism. *The FASEB Journal 25*(11), 3803–3814.

Qu, A., B. G. Lindsay, and B. Li (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika 87*(4), 823–836.

Qu, A. and P. X.-K. Song (2004). Assessing robustness of generalised estimating equations and quadratic inference functions. *Biometrika 91*(2), 447–459.

Reich, B. J., M. Fuentes, A. H. Herring, and K. R. Evenson (2010). Bayesian variable selection for multivariate spatially varying coefficient regression. *Biometrics 66*(3), 772–782.

Ren, J., T. He, Y. Li, S. Liu, Y. Du, and C. Wu (2017). Network-based regularization for high dimensional SNP data in the case–control study of type 2 diabetes. *BMC Genetics 18*(1).

Ren, J., F. Zhou, X. Li, Q. Chen, H. Zhang, S. Ma, Y. Jiang, and C. Wu (2020). Semiparametric bayesian variable selection for gene-environment interactions. *Statistics in Medicine 39*(5), 617–638.

Ren, J., F. Zhou, X. Li, S. Ma, Y. Jiang, and C. Wu (2020). Robust bayesian variable selection for gene-environment interactions. *Biometrics (Revision Invited)*.

Ročková, V. and E. I. George (2018). The spike-and-slab lasso. *Journal of the American Statistical Association 113*(521), 431–444.

Schaid, D. J., J. P. Sinnwell, G. D. Jenkins, S. K. McDonnell, J. N. Ingle, M. Kubo, P. E. Goss, J. P. Costantino, D. L. Wickerham, and R. M. Weinshilboum (2012). Using the gene

ontology to scan multilevel gene sets for associations in genome wide association studies. *Genetic Epidemiology 36*(1), 3–16.

Scheet, P. and M. Stephens (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics 78*(4), 629–644.

Sharma, S., X. Zhou, D. M. Thibault, et al. (2014). A genome-wide survey of cd4+ lymphocyte regulatory genetic variants identifies novel asthma genes. *Journal of Allergy and Clinical Immunology 134*(5), 1153–1162.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics 22*(2), 231–245.

Simonds, N. I., A. A. Ghazarian, C. B. Pimentel, S. D. Schully, G. L. Ellison, E. M. Gillanders, and L. E. Mechanic (2016). Review of the gene-environment interaction literature in cancer: what do we know? *Genetic Epidemiology 40*(5), 356–365.

Sitlani, C. M., K. M. Rice, T. Lumley, B. McKnight, L. A. Cupples, C. L. Avery, R. Noordam, B. H. Stricker, E. A. Whitsel, and B. M. Psaty (2014). Generalized estimating equations for genome-wide association studies using longitudinal phenotype data. *Statistics in Medicine 34*(1), 118–130.

Song, R., W. Lu, S. Ma, and X. Jessie Jeng (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika 101*(4), 799–814.

Stegemann, C., R. Pechlaner, P. Willeit, S. R. Langley, M. Mangino, U. Mayr, C. Menni, A. Moayyeri, P. Santer, G. Rungger, T. D. Spector, J. Willeit, S. Kiechl, and M. Mayr (2014). Lipidomics profiling and risk of cardiovascular disease in the prospective population-based bruneck study. *Circulation 129*(18), 1821–1831.

Stephenson, D. J., L. A. Hoeferlin, and C. E. Chalfant (2017). Lipidomics in translational research and the clinical significance of lipid-based biomarkers. *Translational Research 189*, 13–29.

Tang, Y., H. J. Wang, and Z. Zhu (2013). Variable selection in quantile varying coefficient models with longitudinal data. *Computational Statistics and Data Analysis 57*(1), 435–449.

Tang, Y., H. J. Wang, Z. Zhu, and X. Song (2012). A unified variable selection approach for varying coefficient models. *Statistica Sinica 22*(2).

Taylor, J. and R. J. Tibshirani (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences 112*(25), 7629–7634.

Thiam, A., J. Farese, R.V., and T. Walther (2013). The biophysics and cell biology of lipid droplets. *Nat. Rev. Mol. Cell Biol. 14*, 775–786.

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

Verbeke, G., S. Fieuws, G. Molenberghs, and M. Davidian (2014). The analysis of multivariate longitudinal data: a review. *Statistical Methods in Medical Research 23*(1), 42–59.

Vuillaume, M.-L., S. Naudion, G. Banneau, G. Diene, A. Cartault, D. Cailley, J. Bouron, J. Toutain, G. Bourrouillou, A. Vigouroux, L. Bouneau, F. Nacka, I. Kieffer, B. Arveiler, A. Knoll-Gellida, P. J. Babin, E. Bieth, B. Jouret, S. Julia, P. Sarda, D. Geneviève, L. Faivre, D. Lacombe, P. Barat, M. Tauber, M.-A. Delrue, and C. Rooryck (2014). New candidate loci identified by array-CGH in a cohort of 100 children presenting with syndromic obesity. *American Journal of Medical Genetics Part A 164*(8), 1965–1975.

Wang, H. and Y. Xia (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association 104*(486), 747–757.

Wang, H., M. Ye, Y. Fu, A. Dong, M. Zhang, L. Feng, X. Zhu, W. Bo, L. Jiang, C. H. Griffin, D. Liang, and R. Wu (2021). Modeling genome-wide by environment interactions through omnigenic interactome networks. *Cell Reports 35*(6), 109114.

Wang, H. and K. Zhang (2010). Nonparametric tests for longitudinal dna copy number data. *Statistics and Its Interface 3*(2), 211–221.

Wang, H. J., Z. Zhu, and J. Zhou (2009). Quantile regression in partially linear varying coefficient models. *The Annals of Statistics*, 3841–3866.

Wang, L., H. Li, and J. Z. Huang (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association 103*(484), 1556–1569.

Wang, L., J. Zhou, and A. Qu (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics 68*(2), 353–360.

Wenk, M. R. (2005). The emerging field of lipidomics. *Nature Reviews Drug Discovery 4*(7), 594–610.

Wilson, B. T., Z. Stark, R. E. Sutton, et al. (2015). The cockayne syndrome natural history (CoSyNH) study: clinical findings in 102 individuals and recommendations for care. *Genetics in Medicine 18*(5), 483–493.

Winham, S. J. and J. M. Biernacka (2013). Gene–environment interactions in genome-wide association studies: current approaches and new directions. *Journal of Child Psychology and Psychiatry 54*(10), 1120–1134.

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

Wu, C. and Y. Cui (2014). Boosting signals in gene-based association studies via efficient snp selection. *Briefings in Bioinformatics 15*(2), 279–291.

Wu, C., Y. Cui, and S. Ma (2014). Integrative analysis of gene–environment interactions under a multi-response partially linear varying coefficient model. *Statistics in Medicine 33*(28), 4988–4998.

Wu, C., Y. Jiang, J. Ren, Y. Cui, and S. Ma (2018). Dissecting gene-environment interactions: A penalized robust approach accounting for hierarchical structures. *Statistics in Medicine 37*(3), 437–456.

Wu, C., S. Li, and Y. Cui (2012). Genetic association studies: an information content perspective. *Current Genomics 13*(7), 566–573.

Wu, C. and S. Ma (2015). A selective review of robust variable selection with applications in bioinformatics. *Briefings in Bioinformatics 16*(5), 873–883.

Wu, C., X. Shi, Y. Cui, and S. Ma (2015). A penalized robust semiparametric approach for gene–environment interactions. *Statistics in Medicine 34*(30), 4016–4030.

Wu, C., Q. Zhang, Y. Jiang, and S. Ma (2018). Robust network-based analysis of the associations between (epi)genetic measurements. *Journal of Multivariate Analysis 168*, 119–130.

Wu, C., P. Zhong, and Y. Cui (2018). Additive varying-coefficient model for nonlinear gene-environment interactions. *Stat. Appl. Genet. Mol. Biol. 17*.

Wu, C., F. Zhou, J. Ren, X. Li, Y. Jiang, and S. Ma (2019). A selective review of multi-level omics data integration using variable selection. *High-throughput 8*(1), 4.

Wu, M. and S. Ma (2018). Robust genetic interaction analysis. *Briefings in Bioinformatics 20*(2), 624–637.

Wu, M. and S. Ma (2019). Robust semiparametric gene-environment interaction analysis using sparse boosting. *Statistics in Medicine 38*(23), 4625–4641.

Wu, M., Q. Zhang, and S. Ma (2020). Structured gene-environment interaction analysis. *Biometrics 76*(1), 23–35.

Wu, Y. and Y. Liu (2009). Variable selection in quantile regression. *Statistica Sinica*, 801–817.

Xu, X. and M. Ghosh (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Anal. 10*(4), 909–936.

Xu, Y., M. Wu, S. Ma, and S. E. Ahmed (2018). Robust gene–environment interaction analysis using penalized trimmed regression. *Journal of Statistical Computation and Simulation 88*(18), 3502–3528.

Yan, J. and J. Huang (2012). Model selection for cox models with time-varying coefficients. *Biometrics 68*(2), 419–428.

Yuan, M. and Y. Lin (2005). Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association 100*(472), 1215–1225.

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

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*(2), 894–942.

Zhang, H., X. Huang, J. Gan, W. Karmaus, and T. Sabo-Attwood (2016). A two-component g-prior for variable selection. *Bayesian Analysis 11*(2), 353–380.

Zhang, Q., H. Chai, and S. Ma (2020). Robust identification of gene-environment interactions under high-dimensional accelerated failure time models. *arXiv preprint arXiv:2003.02580*.

Zhang, S., Y. Xue, Q. Zhang, C. Ma, M. Wu, and S. Ma (2019). Identification of gene–environment interactions with marginal penalization. *Genetic Epidemiology 44*(2), 159–196.

Zhou, F., X. Lu, J. Ren, and C. Wu (2021). Package 'springer': sparse group variable selection for gene-environment interactions in the longitudinal study (R package version 0.1.2.).

Zhou, F., J. Ren, G. Li, Y. Jiang, X. Li, W. Wang, and C. Wu (2019). Penalized variable selection for lipid–environment interactions in a longitudinal lipidomics study. *Genes 10*(12), 1002.

Zhou, F., J. Ren, X. Li, C. Wu, and Y. Jiang (2020). interep: Interaction analysis of repeated measure data (R package version 0.3.1.).

Zhou, F., J. Ren, X. Lu, S. Ma, and C. Wu (2021). Gene–environment interaction: A variable selection perspective. *Epistasis: Methods and Protocols, Springer*, 191–223.

Zhou, X., J. Mao, Y. Ai, J.; Deng, C. Roth, M.R.; Pound, J. Henegar, R. Welti, and S. Bigler (2012). Identification of plasma lipid biomarkers for prostate cancer by lipidomics and bioinformatics. *PLoS ONE 7*, e48889.

Appendix A

Appendices for Chapter 2

Table A.1: *Identification results for $n = 250$, $p = 75$ with an actual dimension of 304. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

$n = 250$	$p = 75$	Overall		Main		Interaction	
		TP	FP	TP	FP	TP	FP
ρ=0.5	A1	14.5(1.9)	4.8(3.1)	7.2(0.8)	1.7(1.2)	7.4(1.5)	3.1(2.6)
	A2	14.7(1.8)	5.0(3.2)	7.2(0.9)	1.7(1.3)	7.5(1.4)	3.2(2.6)
	A3	14.7(1.7)	5.0(3.3)	7.2(0.8)	1.8(1.4)	7.6(1.3)	3.2(2.6)
	A4	13.3(1.5)	6.6(4.2)	7.2(0.7)	1.6(1.4)	6.1(1.1)	5.1(3.3)
	A5	13.3(1.5)	6.8(4.4)	7.2(0.8)	1.7(1.4)	6.1(1.1)	5.2(3.5)
	A6	13.3(1.5)	7.3(4.7)	7.2(0.8)	1.8(1.5)	6.1(1.1)	5.5(3.7)
ρ=0.8	A1	13.7(2.3)	4.1(2.8)	7.2(0.8)	1.5(1.0)	6.5(2.1)	2.7(2.4)
	A2	13.9(2.4)	4.1(2.8)	7.2(0.8)	1.5(1.0)	6.6(2.1)	2.7(2.4)
	A3	14.2(2.3)	4.5(2.9)	7.2(0.7)	1.6(1.0)	7.0(2.2)	2.9(2.5)
	A4	12.9(1.9)	5.5(2.7)	7.2(0.7)	1.1(1.0)	5.6(1.6)	4.5(2.3)
	A5	12.9(1.9)	5.8(2.9)	7.2(0.7)	1.1(0.9)	5.7(1.6)	4.7(2.5)
	A6	13.0(1.8)	6.5(3.5)	7.2(0.7)	1.2(0.9)	5.8(1.4)	5.5(3.2)

Table A.2: *Identification results for $n = 250$, $p = 150$ with an actual dimension of 604. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

$n = 250$	$p = 150$	Overall		Main		Interaction	
		TP	FP	TP	FP	TP	FP
ρ=0.5	A1	13.9(2.3)	5.0(3.0)	7.2(0.7)	1.7(1.1)	6.7(2.0)	3.3(2.6)
	A2	14.0(2.2)	5.0(3.0)	7.2(0.7)	1.7(1.1)	6.8(1.9)	3.3(2.6)
	A3	14.4(2.2)	5.1(3.2)	7.3(0.7)	1.8(1.2)	7.1(1.9)	3.3(2.8)
	A4	12.9(1.9)	5.7(2.5)	7.3(0.8)	1.4(0.9)	5.6(1.5)	4.4(2.3)
	A5	13.0(1.8)	5.9(2.6)	7.2(0.8)	1.4(0.9)	5.7(1.4)	4.5(2.3)
	A6	13.0(1.8)	6.4(2.7)	7.2(0.8)	1.4(1.0)	5.8(1.5)	5.0(2.5)
ρ=0.8	A1	13.5(2.0)	5.3(3.0)	7.2(0.9)	2.1(1.2)	6.3(1.9)	3.2(2.4)
	A2	13.5(2.0)	5.4(3.2)	7.2(0.9)	2.2(1.3)	6.3(1.9)	3.2(2.5)
	A3	13.4(2.1)	6.0(3.0)	7.1(0.9)	2.4(1.3)	6.2(1.9)	3.6(2.7)
	A4	12.5(1.9)	7.6(3.3)	7.3(0.7)	1.8(1.2)	5.2(1.7)	5.7(2.7)
	A5	12.6(1.8)	7.8(3.4)	7.3(0.7)	1.9(1.2)	5.3(1.6)	5.9(2.8)
	A6	12.6(1.8)	8.4(4.1)	7.3(0.8)	1.9(1.2)	5.4(1.7)	6.5(3.6)

Table A.3: *Identification results for $n = 500$, $p = 150$ with an actual dimension of 604. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

$n = 500$	$p = 150$	Overall		Main		Interaction	
		TP	FP	TP	FP	TP	FP
ρ=0.5	A1	15.7(1.4)	2.7(1.9)	7.7(0.5)	1.3(0.7)	8.0(1.4)	1.4(1.7)
	A2	15.8(1.3)	2.7(2)	7.7(0.5)	1.3(0.7)	8.1(1.3)	1.3(1.8)
	A3	16.2(1.2)	2.7(1.9)	7.8(0.4)	1.3(0.8)	8.4(1.2)	1.3(1.6)
	A4	14.7(1.0)	2.5(1.7)	7.8(0.4)	0.9(0.8)	6.9(1.0)	1.6(1.4)
	A5	14.7(1.1)	2.6(1.7)	7.8(0.4)	0.9(0.7)	6.9(1.0)	1.7(1.4)
	A6	14.9(1.0)	2.7(2.0)	7.8(0.4)	0.8(0.7)	7.0(0.9)	1.8(1.6)
ρ=0.8	A1	15.5(1.7)	3.0(2.9)	7.7(0.6)	1.1(0.8)	7.9(1.5)	1.9(2.2)
	A2	15.4(1.7)	2.9(2.8)	7.7(0.6)	1.1(0.8)	7.8(1.5)	1.8(2.2)
	A3	15.7(1.6)	2.6(2.6)	7.7(0.5)	1.2(0.9)	8.0(1.4)	1.4(2.1)
	A4	14.8(1.4)	3.7(1.8)	7.5(0.6)	1.2(0.7)	7.2(1.2)	2.5(1.5)
	A5	14.7(1.3)	3.6(1.9)	7.5(0.5)	1.1(0.7)	7.2(1.2)	2.5(1.5)
	A6	15.0(1.3)	3.8(1.9)	7.7(0.6)	1.1(0.7)	7.4(1.1)	2.7(1.6)

Table A.4: *Identification results for $n = 500$, $p = 300$ with an actual dimension of 1204. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

$n = 500$	$p = 300$	Overall		Main		Interaction	
		TP	FP	TP	FP	TP	FP
ρ=0.5	A1	16.1(1.2)	3.2(2.4)	7.6(0.6)	1.4(0.8)	8.5(1.0)	1.8(2.2)
	A2	16.3(1.1)	3.2(2.4)	7.7(0.5)	1.4(0.8)	8.5(0.9)	1.8(2.2)
	A3	16.3(1)	2.9(2.2)	7.8(0.5)	1.4(0.8)	8.6(0.8)	1.5(1.9)
	A4	14.8(0.8)	2.9(2.1)	7.8(0.4)	1.0(0.8)	7.0(0.8)	1.9(1.7)
	A5	14.8(0.9)	3.1(2.3)	7.8(0.4)	1.0(0.8)	7.0(0.8)	2.0(1.9)
	A6	14.9(0.9)	3.3(2.6)	7.8(0.4)	1.0(0.8)	7.1(0.9)	2.3(2.1)
ρ=0.8	A1	15.9(1.2)	3(2.6)	7.6(0.5)	1.5(0.8)	8.3(1.1)	1.5(2.2)
	A2	15.9(1.3)	3.0(2.7)	7.6(0.5)	1.5(0.9)	8.2(1.1)	1.5(2.2)
	A3	15.8(1.4)	3.1(2.8)	7.7(0.5)	1.6(1.0)	8.1(1.2)	1.6(2.2)
	A4	14.5(1.2)	4.5(3.0)	7.8(0.6)	1.0(0.7)	6.8(1.0)	3.5(2.6)
	A5	14.5(1.2)	4.7(3.3)	7.8(0.6)	1.1(0.8)	6.7(0.9)	3.6(2.9)
	A6	14.5(1.1)	4.9(3.6)	7.8(0.6)	1.0(0.8)	6.7(0.8)	3.8(3.3)

Figure A.1: *Plot of the identification results for* $n = 250$. $p = 75$ *with an actual dimension of 304.* $p = 150$ *with an actual dimension of 604. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

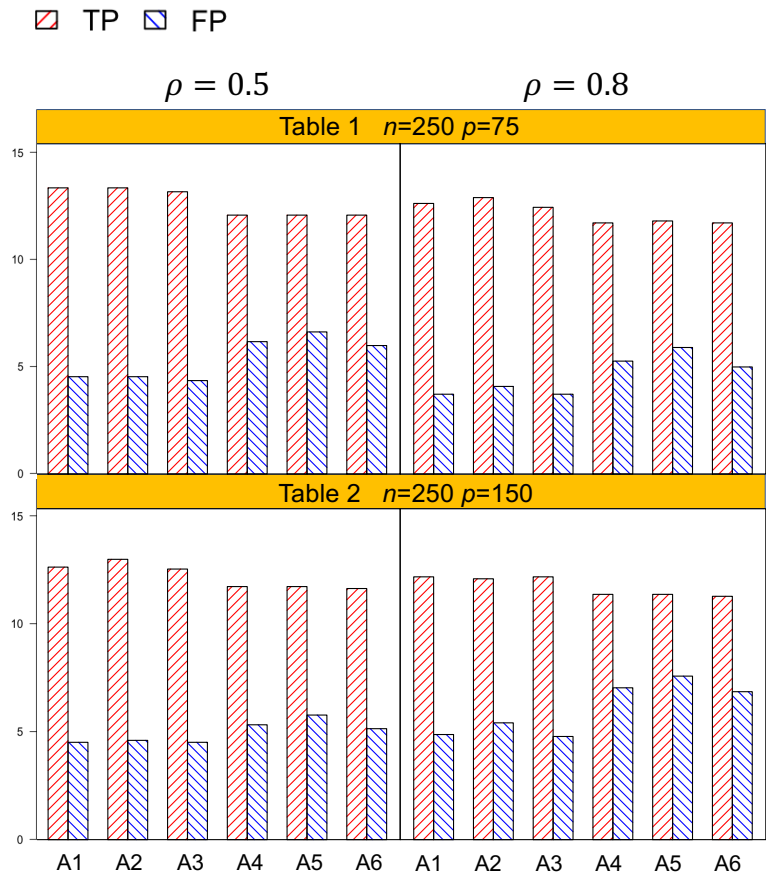


Figure A.2: *Plot of the identification results for* $n = 500$. $p = 150$ *with an actual dimension 604.* $p = 300$ *with an actual dimension of 1204. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

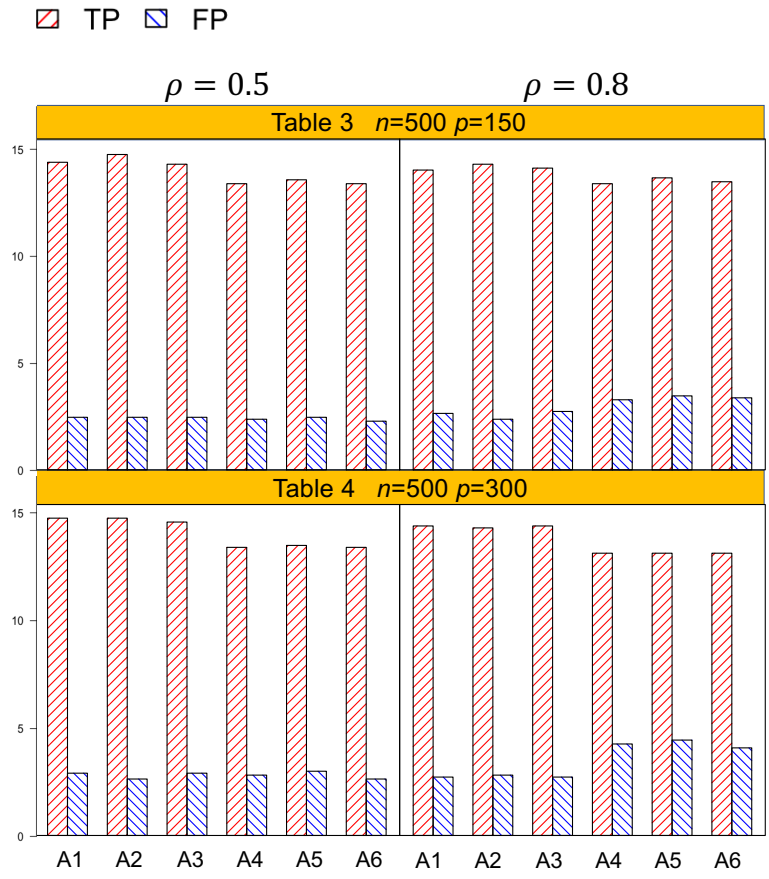


Table A.5: *Estimation accuracy results for $n = 250$. $p = 75$ with an actual dimension of 304. $p = 150$ with an actual dimension of 604. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

		$n = 250$					
		$p = 75$			$p = 150$		
		MSE	NMSE	TMSE	MSE	NMSE	TMSE
ρ=0.5	A1	0.1055	0.0026	0.0043	0.1264	0.0045	0.0072
	A2	0.1042	0.0026	0.0042	0.1259	0.0045	0.0072
	A3	0.1030	0.0026	0.0042	0.1174	0.0041	0.0066
	A4	0.2321	0.0018	0.0056	0.2435	0.0032	0.0084
	A5	0.2304	0.0018	0.0055	0.2402	0.0031	0.0082
	A6	0.2288	0.0018	0.0055	0.2346	0.0030	0.0080
ρ=0.8	A1	0.1187	0.0087	0.0135	0.129	0.0048	0.0075
	A2	0.1163	0.0085	0.0132	0.1295	0.0048	0.0075
	A3	0.1066	0.0075	0.0118	0.1319	0.0049	0.0077
	A4	0.2410	0.0060	0.0162	0.2531	0.0038	0.0092
	A5	0.2426	0.0060	0.0162	0.2487	0.0038	0.0091
	A6	0.2335	0.0058	0.0157	0.2431	0.0037	0.0089

Table A.6: *Estimation accuracy results for $n = 500$. $p = 150$ with an actual dimension of 604. $p = 300$ with an actual dimension of 1204. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

		$n = 500$					
		$p = 150$			$p = 300$		
		MSE	NMSE	TMSE	MSE	NMSE	TMSE
ρ=0.5	A1	0.0754	0.0026	0.0042	0.0660	0.0010	0.0017
	A2	0.0731	0.0026	0.0041	0.0659	0.0010	0.0017
	A3	0.0648	0.0022	0.0035	0.0663	0.0010	0.0017
	A4	0.1872	0.0015	0.0055	0.1635	0.0007	0.0024
	A5	0.1837	0.0015	0.0054	0.1612	0.0007	0.0024
	A6	0.1792	0.0013	0.0052	0.1603	0.0007	0.0024
ρ=0.8	A1	0.0708	0.0023	0.0037	0.0688	0.0010	0.0018
	A2	0.0716	0.0023	0.0038	0.0688	0.0011	0.0018
	A3	0.0704	0.0025	0.0039	0.0718	0.0012	0.0020
	A4	0.1480	0.0013	0.0049	0.1949	0.0007	0.0028
	A5	0.1492	0.0013	0.0045	0.1945	0.0007	0.0028
	A6	0.1479	0.0012	0.0044	0.1899	0.0007	0.0027

Table A.7: *Real data analysis result from Method A1 (method accommodating the lipid–environment interactions with exchangeable working correlation).*

	Lipid	AE	PE	DCR
C16:0/16:1	0	0.0117	-0.0239	-0.0057
C18:2/16:1	0	0.1544	3.3322	0.3924
C18:1/16:1	0	0.4857	-0.6299	-0.5559
C20:1/16:1	0.5966	-2.9145	0.1299	-1.4836
C16:0/16:0	0	1.3742	-0.8817	-1.8070
C20:6/16:0	0.0369	0	0	0
C20:0/18:3	-1.3628	0	0	0
C18:0/18:2	-1.6154	0	0	0
C22:6/18:1	1.1717	1.7526	0.2287	-0.4079
C18:2/20:4	1.1497	0	0	0
C18:1/20:4	0.8490	0	0	0
C20:1/20:4	0	-0.2169	-0.6096	3.0537

Table A.8: *Real data analysis result from Method A4 (method not accommodating the lipid–environment interactions with exchangeable working correlation).*

	Lipid	AE	DCR	PE
C16:0/16:1	0	0	-0.0024	0
C18:2/16:1	-2.1856	0	3.2306	0
C18:1/16:1	0	0	-1.4641	-2.3563
C20:1/16:1	0.0042	-2.6768	0	-1.7757
C16:0/16:0	0	2.8757	-0.9389	-2.6791
C18:2/16:0	0	0	0	-1.7688
C20:6/16:0	0.1481	-0.1276	0	0
C18:1/18:3	0	0	1.2917	0
C20:0/18:3	-1.6171	0	0	0
C18:0/18:2	-1.7695	0	0	0
C22:6/18:1	0.8851	3.4714	0.4809	0
C18:1/18:0	0	-1.2901	0	0
C22:7/18:0	0	-0.9839	0	0
C18:2/20:4	2.5871	0.6150	0	1.9327
C18:1/20:4	0	0	-0.0031	0
C20:1/20:4	0.7542	-1.1147	0	3.5396

Table A.9: *Identification results for $n = 60$, $p = 30$ with an actual dimension of 124. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

$n = 60$	$p = 30$	Overall		Main		Interaction	
		TP	FP	TP	FP	TP	FP
ρ=0.5	A1	13.6(2.5)	4.7(2.7)	7.4(0.8)	2.1(1.6)	6.2(2.1)	2.5(2.6)
	A2	13.6(2.5)	4.8(2.8)	7.3(0.8)	2.2(1.6)	6.2(2.1)	2.6(2.6)
	A3	13.7(2.5)	4.9(3.0)	7.4(0.7)	2.1(1.6)	6.3(2.1)	2.7(2.7)
	A4	11.1(2.6)	5.4(2.8)	6.4(1.1)	1.1(1.0)	4.6(1.9)	4.3(2.3)
	A5	11.1(2.6)	5.4(2.8)	6.4(1.1)	1.1(1.0)	4.6(1.9)	4.3(2.3)
	A6	11.1(2.5)	5.5(2.8)	6.5(1.2)	1.1(1.0)	4.7(1.8)	4.4(2.3)
ρ=0.8	A1	13.2(2.2)	4.4(2.9)	7.5(0.6)	2.4(1.7)	5.7(2.1)	1.9(2.1)
	A2	13.2(2.2)	4.4(2.9)	7.5(0.6)	2.4(1.7)	5.7(2.1)	2.0(2.1)
	A3	13.4(2.0)	4.4(3.0)	7.5(0.6)	2.4(1.7)	5.9(1.9)	2.0(2.1)
	A4	11.0(2.4)	5.5(2.5)	6.5(1.4)	1.3(1.2)	4.5(1.8)	4.2(2.1)
	A5	11.0(2.4)	5.6(2.6)	6.5(1.4)	1.3(1.2)	4.5(1.8)	4.2(2.2)
	A6	11.1(2.4)	5.8(2.7)	6.5(1.4)	1.4(1.3)	4.5(1.8)	4.3(2.2)

Table A.10: *Estimation accuracy results for $n = 60$, $p = 30$ with an actual dimension of 124. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

	$n = 60, p = 30$					
	$\rho = 0.5$			$\rho = 0.8$		
	MSE	NMSE	TMSE	MSE	NMSE	TMSE
A1	0.9352	0.1928	0.2732	0.9820	0.2108	0.2944
A2	0.9387	0.1924	0.2733	0.9809	0.2105	0.2940
A3	0.9324	0.1914	0.2717	1.0098	0.2063	0.2933
A4	1.9732	0.1560	0.3528	1.9910	0.1488	0.3484
A5	1.9709	0.1556	0.3523	1.9887	0.1487	0.348
A6	1.9629	0.1543	0.3502	1.9795	0.1474	0.3458

Table A.11: *Data simulated based upon the underlying main effect only model. Identification results for $n = 250, p = 75, \rho = 0.8$ with an actual dimension of 304. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

	Overall		Main		Interaction				
	TP	FP	TP	FP	TP	FP	MSE	NMSE	TMSE
A1	7.7(0.9)	0.7(1.7)	7.7(0.9)	0.0(0.0)	0.0(0.0)	0.7(1.7)	0.1025	0.0000	0.0014
A2	7.8(0.6)	0.4(1.3)	7.8(0.6)	0.0(0.2)	0.0(0.0)	0.4(1.3)	0.0730	0.0000	0.0010
A3	7.9(0.3)	0.5(1.2)	7.9(0.3)	0.3(0.7)	0.0(0.0)	0.2(0.8)	0.0288	0.0000	0.0004
A4	7.3(1.1)	0.8(0.9)	7.3(1.1)	0.0(0.0)	0.0(0.0)	0.8(0.9)	0.2530	0.0000	0.0034
A5	7.2(1.1)	0.9(1.1)	7.2(1.1)	0.0(0.0)	0.0(0.0)	0.9(1.1)	0.2273	0.0001	0.0031
A6	7.5(0.7)	1.2(1.1)	7.5(0.7)	0.0(0.2)	0.0(0.0)	1.2(1.1)	0.1932	0.0001	0.0027

Table A.12: *Null models. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

	$n = 250$				$n = 500$			
	$p = 75$		$p = 150$		$p = 150$		$p = 300$	
	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$
A1	0.00(0.00)	0.03(0.18)	0.03(0.18)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
A2	0.03(0.10)	0.03(0.18)	0.30(0.70)	0.10(0.31)	0.00(0.00)	0.00(0.00)	0.03(0.18)	0.00(0.00)
A3	0.13(0.51)	0.17(0.44)	0.97(1.47)	0.77(0.81)	0.10(0.40)	0.50(0.20)	0.10(0.31)	0.10(0.25)
A4	0.00(0.00)	0.03(0.18)	0.03(0.18)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
A5	0.03(0.10)	0.03(0.18)	0.30(0.70)	0.10(0.31)	0.00(0.00)	0.00(0.00)	0.03(0.18)	0.00(0.00)
A6	0.13(0.51)	0.17(0.44)	0.97(1.47)	0.77(0.81)	0.10(0.40)	0.50(0.20)	0.10(0.31)	0.10(0.25)

Table A.13: *Stability selection percentages for all the 17 true effects in the simulated data when $n = 250$, $p = 75$, $\rho = 0.8$ with an actual dimension of 304. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

true effect	A1	A2	A3	A4	A5	A6
1	1	1	1	1	1	1
2	0.73	1	1	0.82	0.98	1
3	1	0.80	1	1	1	1
4	1	1	1	1	1	1
5	1	0.45	1	1	0.93	0.98
6	0.13	0.14	0.38	0.65	0.98	0.98
7	0.58	0.65	1	0.99	1	0.92
8	0.61	0.25	0.45	0.89	1	1
9	1	0.84	1	0.46	0.02	0.10
10	1	0.86	1	0.07	0.01	0.10
11	1	0.83	1	0.70	0.66	0.84
12	0.77	0.91	0.72	0.36	0.87	0.01
13	0.77	0.91	0.73	0.39	0.94	0.45
14	0.75	0.94	0.77	0.48	1	0.98
15	0.81	0.82	0.98	0.30	0.55	1
16	0.80	0.86	0.99	0.98	0.75	0.99
17	0.80	0.87	0.99	0.66	0.93	1

Table A.14: *Validation methods. Identification results for $n = 250$, $p = 75$ with an actual dimension of 304. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

$n = 250$	$p = 75$	Overall		Main		Interaction	
		TP	FP	TP	FP	TP	FP
ρ=0.5	A1	14.1(2.1)	4.6(3.1)	7.0(0.8)	1.1(0.8)	7.0(1.8)	3.5(2.9)
	A2	14.2(2.1)	4.7(3.1)	7.0(0.9)	1.1(0.9)	7.1(1.8)	3.6(2.8)
	A3	14.4(1.7)	4.6(3.2)	7.1(0.8)	1.1(0.9)	7.2(1.5)	3.5(3.0)
	A4	13.1(1.1)	6.1(2.8)	6.9(0.8)	1.0(0.8)	6.1(0.9)	5.3(2.6)
	A5	13.1(1.1)	6.4(2.8)	6.9(0.8)	1.0(0.8)	6.1(0.9)	5.6(2.5)
	A6	13.0(1.2)	6.7(3.1)	6.9(0.8)	1.0(0.8)	6.1(1.0)	5.9(2.9)
ρ=0.8	A1	13.7(2.6)	4.7(2.9)	7.2(0.8)	1.4(0.9)	6.5(2.3)	3.2(2.5)
	A2	13.8(2.6)	4.6(3.1)	7.3(0.8)	1.4(1.0)	6.6(2.3)	3.1(2.6)
	A3	13.8(2.5)	5.1(3.0)	7.3(0.7)	1.5(0.8)	6.5(2.1)	3.6(2.9)
	A4	12.9(2.1)	5.7(2.5)	7.3(0.8)	1.3(0.9)	5.6(1.6)	4.5(2.1)
	A5	12.9(2.1)	5.8(2.6)	7.3(0.8)	1.3(1.0)	5.6(1.6)	4.5(2.2)
	A6	12.9(2.2)	6.8(2.7)	7.3(0.7)	1.4(0.9)	5.6(1.8)	5.5(2.5)

Table A.15: *Validation methods. Estimation accuracy results for $n = 250$, $p = 75$ with an actual dimension of 304. mean(sd) based on 100 replicates. A1–A3: methods accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively. A4–A6: methods not accommodating the lipid–environment interactions with exchangeable, AR(1) and independence working correlations, respectively.*

$n = 250, p = 75$						
	$\rho = 0.5$			$\rho = 0.8$		
	MSE	NMSE	TMSE	MSE	NMSE	TMSE
A1	0.1126	0.0074	0.0120	0.1205	0.0085	0.0134
A2	0.1095	0.0071	0.0115	0.1200	0.0085	0.0133
A3	0.1082	0.0071	0.0115	0.1245	0.0090	0.0140
A4	0.2344	0.0051	0.0150	0.2610	0.0060	0.0171
A5	0.2335	0.0050	0.0149	0.2627	0.0060	0.0171
A6	0.2302	0.0048	0.0146	0.2565	0.0058	0.0166

Appendix B

Appendices for Chapter 3

B.1 Derivations of Alternative Methods

The alternative methods fall into the following two categories: (1) gQIF.exch, gQIF.ar1 and gQIF.ind only conduct penalized identification on the group level, corresponding to the penalized group QIF, and (2) iQIF.exch, iQIF.ar1 and iQIF.ind ignore the group level effects, and only focus on the individual level effects (penalized QIF).

B.1.1 Penalized Group QIF

The penalized group QIF methods considered in this study (gQIF.exch, gQIF.ar1 and gQIF.ind) can only identify the main and interaction effects on a group–in/group–out basis. The corresponding score equation is defined as

$$U(\beta) = Q(\beta) + \sum_{v=1}^{p} \rho(||\eta_v||_{\Sigma_v}; \sqrt{q+1}\lambda_1, \gamma),$$

where ρ denotes MCP penalty with tuning parameter λ_1 and regularization parameter γ. As defined in Section 3.2.2, the coefficient vector β corresponds to all the main and interaction effects. η_v, the vector of length q+1 in β, represents the main effect of the vth G factor as well as its interactions with the q environment factors. The penalty is imposed on $||\eta_v||_{\Sigma_v}$,

the empirical norm of η_v. Thus the penalized identification can merely performed on group level.

We have developed a Newton-Raphson based algorithm to obtain the penalized QIF estimate $\hat{\beta}$. The estimate $\hat{\beta}^{(g+1)}$ in the $(g+1)$th iteration can be solved based on the previous coefficient vector $\hat{\beta}^{(g)}$ in the gth iteration:

$$\hat{\beta}^{(g+1)} = \hat{\beta}^{(g)} + [V^{(g)} + nH^{(g)})]^{-1}[P^{(g)} - nH^{(g)}\hat{\beta}^{(g)}],$$

with $P^{(g)}$ and $V^{(g)}$ as the first and second order derivative of the score function of QIF, respectively. They are defined as:

$$P^{(g)} = \frac{\partial Q(\hat{\beta}^{(g)})}{\partial \beta} = 2\frac{\partial \overline{\phi_n}^\top}{\partial \beta}\overline{\Omega_n}^{-1}\overline{\phi_n}(\hat{\beta}^{(g)}),$$

$$V^{(g)} = \frac{\partial^2 Q(\hat{\beta}^{(g)})}{\partial^2 \beta} = 2\frac{\partial \overline{\phi_n}^\top}{\partial \beta}\overline{\Omega_n}^{-1}\frac{\partial \overline{\phi_n}}{\partial \beta}.$$

$H^{(g)}$ is a diagonal matrix containing the derivatives of the penalty function and it's defined as:

$$H^{(g)} = \text{diag}(\underbrace{0, ..., 0}_{1+q}, \underbrace{\frac{\rho'(||\hat{\eta}_1^{(g)}||_{\Sigma_1}; \sqrt{q+1}\lambda_1, \gamma)}{\epsilon + ||\hat{\eta}_1^{(g)}||_{\Sigma_1}}, ..., \frac{\rho'(||\hat{\eta}_1^{(g)}||_{\Sigma_1}; \sqrt{q+1}\lambda_1, \gamma)}{\epsilon + ||\hat{\eta}_1^{(g)}||_{\Sigma_1}}}_{1+q}, ...,$$
$$\underbrace{\frac{\rho'(||\hat{\eta}_p^{(g)}||_{\Sigma_p}; \sqrt{q+1}\lambda_1, \gamma)}{\epsilon + ||\hat{\eta}_p^{(g)}||_{\Sigma_p}}, ..., \frac{\rho'(||\hat{\eta}_p^{(g)}||_{\Sigma_p}; \sqrt{q+1}\lambda_1, \gamma)}{\epsilon + ||\hat{\eta}_p^{(g)}||_{\Sigma_p}}}_{1+q}),$$

where λ_1 is the tuning parameter of genetic effects and gene-environment interactions and γ is the regularization parameter. The first $(1+q)$ elements on the diagonal of matrix H are set to zero, since there is no shrinkage imposed on the intercept and the coefficients of the environmental factors. We can use $nH\hat{\beta}$ and nH to approximate the first and second drivative functions of the the group MCP penalty. Starting with an inital coefficient vector, we can repeat the proposed algorithm and update the regression parameter $\hat{\beta}^{(g+1)}$ through iterations. We set the stop criterion $\text{mean}(|\hat{\beta}^{(g+1)} - \hat{\beta}^{(g)}|) < 0.001$ and convergence can usually be achieved in a small to moderate number of iterations.

B.1.2 Penalized QIF

iQIF.exch, iQIF.ar1 and iQIF.ind are the second category of alternative methods considering only the individual level effects. The derivations for the three methods proceeds in a similar fashion. We have the penalized score function as:

$$U(\beta) = Q(\beta) + \sum_{v=1}^{p}\sum_{u=1}^{q+1}\rho(|\eta_{vu}|;\lambda_1,\gamma),$$

where η_{vu} denotes the uth element of η_v. The Newton-Raphson update of $\hat{\beta}$ can be obtained as:

$$\hat{\beta}^{(g+1)} = \hat{\beta}^{(g)} + [V^{(g)} + nH^{(g)})]^{-1}[P^{(g)} - nH^{(g)}\hat{\beta}^{(g)}],$$

where $P^{(g)}$ and $V^{(g)}$ are given as the corresponding first and second order derivatives of the score function of QIF as follows:

$$P^{(g)} = \frac{\partial Q(\hat{\beta}^{(g)})}{\partial\beta} = 2\frac{\partial\overline{\phi_n}^{\top}}{\partial\beta}\overline{\Omega_n}^{-1}\overline{\phi_n}(\hat{\beta}^{(g)}),$$

$$V^{(g)} = \frac{\partial^2 Q(\hat{\beta}^{(g)})}{\partial^2\beta} = 2\frac{\partial\overline{\phi_n}^{\top}}{\partial\beta}\overline{\Omega_n}^{-1}\frac{\partial\overline{\phi_n}}{\partial\beta}.$$

The main diagonal of the diagonal matrix $H^{(g)}$ consists of the first order derivative of MCP:

$$H^{(g)} = \text{diag}(\underbrace{0,...,0}_{1+q}, \underbrace{\frac{\rho'(|\hat{\eta}_{11}^{(g)}|;\lambda_2,\gamma)}{\epsilon+|\hat{\eta}_{11}^{(g)}|},...,\frac{\rho'(|\hat{\eta}_{1(q+1)}^{(g)}|;\lambda_2,\gamma)}{\epsilon+|\hat{\eta}_{1(q+1)}^{(g)}|}}_{1+q},...,$$
$$\underbrace{\frac{\rho'(|\hat{\eta}_{p1}^{(g)}|;\lambda_2,\gamma)}{\epsilon+|\hat{\eta}_{p1}^{(g)}|},...,\frac{\rho'(|\hat{\eta}_{p(q+1)}^{(g)}|;\lambda_2,\gamma)}{\epsilon+|\hat{\eta}_{p(q+1)}^{(g)}|}}_{1+q}),$$

where λ_2 and γ are the tuning and regularization parameters, respectively. There is no shrinkage on the intercept and the coefficients of the environmental factors. Hence the first $(1+q)$ elements on the diagonal of matrix H are set to zero. Here $nH\hat{\beta}$ and nH can also be used to approximate the first and second drivative functions of the the MCP penalty. The iterative update of $\hat{\beta}$ can be conducted till convergence.

B.2 Other Simulation Results

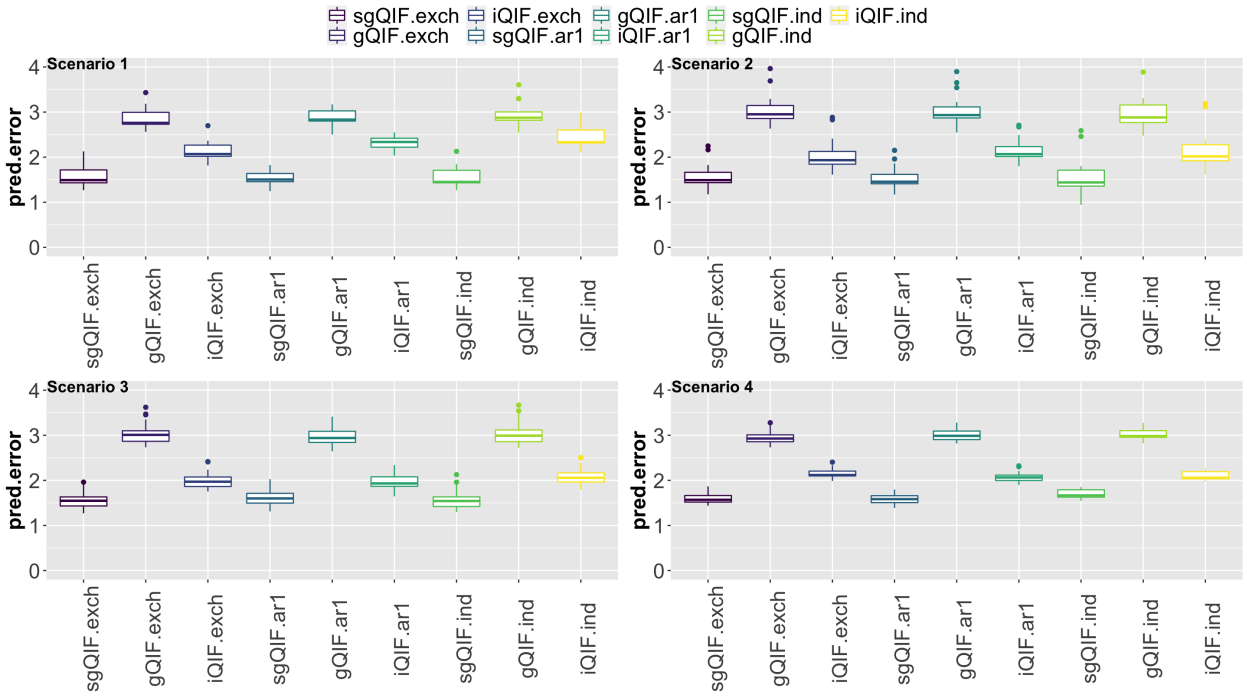
Table B1: *Identification results for Scenario 3. TP/FP: true/false positives. mean(sd) of TP and FP based on 100 replicates.*

	Overall		Main		Interaction	
	TP	FP	TP	FP	TP	FP
sgQIF.exch	19.4(1.0)	2.1(1.1)	3.1(1.1)	0.9(0.7)	16.3(0.8)	1.3(0.8)
gQIF.exch	22.1(1.6)	19.6(6.0)	4.9(0.9)	1.1(1.1)	17.3(1.0)	18.4(5.2)
iQIF.exch	19.7(1.4)	9.0(4.8)	3.1(1.2)	2.0(1.2)	16.6(1.1)	7.0(4.0)
sgQIF.ar1	19.6(1.3)	2.8(1.3)	3.2(1)	0.6(0.7)	16.5(0.8)	2.2(1.4)
gQIF.ar1	22.1(1.4)	18.5(5.3)	4.6(0.9)	0.9(1)	17.5(0.8)	17.6(4.5)
iQIF.ar1	20.0(1.5)	9.2(4.4)	3.5(1.3)	1.6(1.3)	16.5(0.9)	7.5(3.6)
sgQIF.ind	19.3(1.5)	3.0(2.6)	3.0(1.0)	0.3(0.6)	16.3(1.5)	2.7(2.1)
gQIF.ind	21.7(1.2)	16.3(5.1)	4.3(1.2)	1.7(1.2)	17.3(1.2)	14.7(4.0)
iQIF.ind	20.0(1.0)	8.0(4.4)	3.0(1.0)	1.0(1.0)	17.0(1.0)	7.0(3.5)

Table B2: *Identification results for Scenario 4. TP/FP: true/false positives. mean(sd) of TP and FP based on 100 replicates.*

	Overall		Main		Interaction	
	TP	FP	TP	FP	TP	FP
sgQIF.exch	21.9(1.6)	4.7(2.4)	6.6(0.5)	0.2(0.1)	15.3(1.5)	4.5(2.4)
gQIF.exch	21.1(2.7)	19.2(4.4)	6.4(0.9)	3.1(1.6)	14.7(2.0)	16.1(3.1)
iQIF.exch	22.3(1.3)	7.1(2.3)	6.8(0.5)	0.1(0.1)	15.5(1.2)	7.1(2.3)
sgQIF.ar1	22.3(2.1)	4.0(1.0)	6.7(0.6)	0.1(0.1)	15.7(1.5)	4.0(1.0)
gQIF.ar1	22.4(1.9)	17.1(7.4)	6.9(0.4)	2.4(2.2)	15.6(1.8)	14.7(5.5)
iQIF.ar1	23.3(0.6)	10.0(3.5)	7.0(0.6)	0.3(0.1)	16.3(0.6)	10.0(3.5)
sgQIF.ind	20.3(1.0)	3.5(1.3)	5.8(0.5)	0.1(0.1)	14.5(0.6)	3.5(1.3)
gQIF.ind	22.5(0.7)	16.5(6.2)	7(0.3)	2.5(2.1)	15.5(0.7)	14.0(4.0)
iQIF.ind	21.8(0.5)	9.5(1.3)	6.8(0.5)	0.1(0.1)	15.0(0.8)	9.5(1.3)

Figure B1: *Prediction (MSE) results of the 4 scenarios. mean(sd) of prediction error based on 100 replicates.*



B.3 Real Data Analysis

Table C1: Identification results on CAMP data using the bi-level selection method under the exchangeable working correlation (sgQIF.exch). The identified SNPs and the corresponding genes are listed in the first two columns. The third column contains the coefficients of the main effects for each SNP. The last three columns correspond to the interactions between the SNPs and environmental factors.

SNP	Gene		trt	age	gender
rs1276888	FAM46A	0	0.116	0	0
rs10139964	AKAP6	0	0	0.125	0
rs10852830	AC005703.2	0	0	0.111	0

rs10995722	RP11-170M17.1	-0.398	0.103	0	0
rs329614	NDUFAF2	0	0	0	0.598
rs17431749	DKK2	0	-0.246	0	0.325
rs2453021	TNFRSF9	0	0	0	-0.123
rs1922134	RP11-170M17.1	-0.143	0	0	0
rs290505	NDUFAF2	-0.155	0	0	0.300
rs4969059	SLC39A11	-0.212	0	0	0
rs4730738	CAV2	0	0	0	0.145
rs162240	NDUFAF2	0.198	0	0	0
rs6869332	ELOVL7	-0.246	-0.221	0	0
rs167912	NDUFAF2	0	0	-0.282	0
rs158928	ERCC8	0	-0.151	0	0.214
rs131815	NCAPH2	0	0	0.129	0
rs4280657	AC144521.1	0	0	0	-0.274
rs11778333	TOX	0	0	-0.299	0
rs11803207	KCND3	0	0	0	-0.139
rs12299421	rs12299421	0	0	-0.105	0
rs11257102	PFKFB3	-0.582	0	0	-0.353
rs8141896	MICAL3	0.218	0	0	-0.152
rs162231	NDUFAF2	0	-0.347	0	0
rs10857493	RP11-123B3.2	0.468	-0.123	-0.495	0
rs11257103	PFKFB3	-0.508	0	0	-0.192
rs1251577	ST6GALNAC3	0	0	0.112	0
rs4897284	LAMA2	0	0	0	-0.177
rs10491881	RP11-202G18.1	0.128	-0.342	0.339	0
rs566979	CAT	0	0	-0.105	0
rs4904516	FOXN3	0	0.178	0	0

rs681561	PCCA	0	0	-0.286	0.258
rs1618870	CATSPERB	0	0.432	0	-0.489
rs17010079	RP11-123B3.2	-0.214	0	0.364	0
rs11031570	RCN1	0.291	0	-0.504	0
rs909768	RPS6KA2	0	0	0	0.304
rs9891809	SLC39A11	0	0.158	0	0
rs8079240	SLC39A11	0	0.158	0	0
rs7951816	SYT9	0	0	0.141	0
rs1180286	CAV2	0	0	0	-0.152
rs17813724	RP11-202G18.1	0	0.209	0	0.192
rs17241424	TOX	-0.147	0	0.270	0
rs11708933	AC144521.1	0	0	0	0.423
rs197394	FAM212B	0	0	-0.276	0
rs6008813	CELSR1	0	0.142	0	-0.119
rs742267	RPS6KA2	0	0	-0.194	0
rs7712473	ELOVL7	0	0	0	-0.154
rs1704630	CATSPERB	0	0.638	0	-0.493
rs10995701	RP11-170M17.1	0	0	-0.312	0
rs4647078	ERCC8	-0.105	0.515	0	-0.115
rs6877849	ELOVL7	0	0	0.405	0
rs7029556	RP11-63P12.6	0	-0.119	0	0
rs6449502	ELOVL7	0	0	-0.266	0
rs12101359	UNC13C	0.107	0	0	0
rs4716370	RP1-137D17.1	-0.227	0	0.215	0
rs12060403	SLC35F3	0	0	-0.139	0
rs12071173	SLC35F3	0	0	-0.139	0
rs513555	SPRR2G	0	-0.289	0.291	0

rs767006	CYFIP2	0.198	0	0	-0.164
rs4700398	ELOVL7	0	0.205	0	-0.480
rs197380	FAM212B	0.254	0	-0.479	0
rs6914953	F13A1	-0.318	0	0	-0.227
rs264356	NRG2	0	0	0.189	0
rs10972815	CLTA	0	-0.108	0	0.253
rs4700392	ELOVL7	-0.279	-0.916	0	0
rs13194966	F13A1	-0.426	0	0.664	0
rs1119266	SPRR2B	0	0.186	0	0
rs11031563	RCN1	0.423	0	-0.549	0
rs12101884	UNC13C	-0.225	0	0.436	0
rs4647108	ERCC8	0.239	0	0	-0.607
rs7718320	IQGAP2	-0.192	0	0.561	0
rs2303921	TAF1B	0	0	0	-0.174
rs1136062	CCNF	0	-0.125	0.191	0
rs17390967	SCARA5	0.107	-0.196	0	-0.101
rs7243734	ZBTB7C	-0.544	0	0	0
rs17023415	AFF3	-0.305	0.383	0	0
rs10995687	RP11-170M17.1	0.169	0	-0.294	0
rs13265701	MYOM2	-0.218	0.263	0	0
rs4940195	ZBTB7C	-0.51	0	0	0
rs2918528	ZNF717	0	0	0.290	-0.148
rs17819589	RP11-392P7.6	0	0	-0.157	0
rs1360176	RP11-82L2.1	-0.268	0	0	0.213
rs17660456	MYO5B	-0.138	0.157	0	0
rs10871386	RP11-525K10.3	0	0	-0.105	0.201

Table C2: Identification results on CAMP data using the individual-level selection method under the exchangeable working correlation (iQIF.exch). The identified SNPs and the corresponding genes are listed in the first two columns. The third column contains the coefficients of the main effects for each SNP. The last three columns correspond to the interactions between the SNPs and environmental factors.

SNP	Gene		trt	age	gender
rs10050758	SLC36A2	0	-0.136	0.135	0
rs1276888	FAM46A	0	0	0	-0.168
rs10852830	AC005703.2	0	-0.138	0	0
rs10995722	RP11-170M17.1	-0.419	0	0.413	0
rs329614	NDUFAF2	0	0.153	0	0
rs1922134	RP11-170M17.1	0.178	0	0.332	0
rs290505	NDUFAF2	0	0	0	0.429
rs4969059	SLC39A11	0	0.257	0	0
rs4730738	CAV2	0	0	-0.132	0
rs162240	NDUFAF2	-0.374	0	0	0.817
rs6869332	ELOVL7	-0.498	0.696	0	0
rs167912	NDUFAF2	-0.210	0	0	0.378
rs158928	ERCC8	0	0.314	0	0
rs131815	NCAPH2	0	0	0.146	0
rs4280657	AC144521.1	0	0.222	-0.250	0
rs11778333	TOX	0	-0.235	0	0.255
rs11803207	KCND3	0	0	-0.241	0.175
rs11257102	PFKFB3	-0.238	0	0	-0.233
rs8141896	MICAL3	0.278	0	-0.370	0
rs162231	NDUFAF2	0	-0.262	0	-0.218

rs10857493	RP11-123B3.2	0.451	0	-0.536	0
rs10796011	CCDC3	0	0	0	0.145
rs11257103	PFKFB3	-0.170	0	0	0
rs1251577	ST6GALNAC3	0	0	0.133	0
rs4897284	LAMA2	0	-0.273	0.290	0
rs10491881	RP11-202G18.1	0	-0.169	0	0
rs4904516	FOXN3	0	0.247	0	0
rs681561	PCCA	0	0	-0.258	0.263
rs1618870	CATSPERB	0	-0.16	0.686	-0.551
rs17010079	RP11-123B3.2	-0.227	0	0.428	0
rs11031570	RCN1	0.410	-0.670	0	0
rs909768	RPS6KA2	0	0	0	0.142
rs7951816	SYT9	0	0	0.133	0
rs1180286	CAV2	0	0	0	-0.221
rs17241424	TOX	0	0	0	0.210
rs17044664	AC144521.1	0	0.248	0	-0.159
rs11708933	AC144521.1	0	0	0	0.163
rs197394	FAM212B	0	0.270	0	-0.224
rs6008813	CELSR1	0	0.156	0	0
rs742267	RPS6KA2	0	0	-0.371	0
rs742269	RPS6KA2	0	-0.144	0	0
rs7712473	ELOVL7	-0.271	-0.567	0	0.722
rs1704630	CATSPERB	0	0	0.569	-0.575
rs17015079	ROBO2	-0.357	0.548	0	0
rs10995701	RP11-170M17.1	0	0	0	-0.348
rs4647078	ERCC8	-0.214	0	0	0.335
rs6877849	ELOVL7	0	0	0.993	-0.414

rs7029556	RP11-63P12.6	0	0	0.315	0
rs6449502	ELOVL7	0	0	0.491	-0.867
rs12101359	UNC13C	0.133	0	0	0
rs34673	TNPO1	0	0.198	0	0
rs12060403	SLC35F3	0	0	-0.344	0
rs12071173	SLC35F3	0	0	-0.344	0
rs12073596	SLC35F3	0	0	-0.159	0
rs12085211	SLC35F3	0	0	-0.159	0
rs1545854	LINC00880	0	0.139	0	-0.149
rs513555	SPRR2G	0	-0.290	0	0
rs4700398	ELOVL7	-0.507	0	0	0.160
rs197380	FAM212B	0.224	0	-0.273	0
rs6914953	F13A1	0	0	0	-0.275
rs264356	NRG2	0	0	0.270	0
rs463221	CTD-2193G5.1	-0.156	0.154	0	0
rs4700392	ELOVL7	0.365	-0.982	0	0
rs13194966	F13A1	-0.178	0	0.391	0
rs1119266	SPRR2B	0	0.291	-0.264	0
rs11031563	RCN1	0.542	-0.474	0	0
rs12101884	UNC13C	-0.176	0	0.364	0
rs4647108	ERCC8	0	0.319	0	-0.389
rs719628	TASP1	0	0.154	0	0
rs7718320	IQGAP2	0	0	0.355	-0.409
rs1136062	CCNF	0	-0.131	0.135	0
rs7243734	ZBTB7C	-0.703	0	0	0
rs17023415	AFF3	-0.225	0.353	0	0
rs17128269	SH2D4A	0	0	-0.342	0.196

rs10995687	RP11-170M17.1	0	0	-0.173	0.171
rs10734883	SLC2A14	-0.133	0.173	0	0
rs13265701	MYOM2	-0.245	0.157	0	0
rs4940195	ZBTB7C	-0.708	0	0	0
rs2918528	ZNF717	0	0	0.218	0
rs1360176	RP11-82L2.1	0	0	0	0.246
rs17660456	MYO5B	0.229	0	-0.398	0
rs10871386	RP11-525K10.3	0	0	-0.199	0.184

Appendix C

Appendices for Chapter 4

C.1 Other simulation results

Table C1: *Identification results for heterogeneous errors based on 100 replicates. C: correct-fitting proportion; O: overfitting proportion; U: underfitting proportion.*

θ			BQRVCSS	BQRVC	BVCSS	BVC	QRVC-adp	VC-adp
$\theta = 0.3$	Normal	C	0.96	0.72	0.94	0.32	0.88	0.86
		O	0.04	0.14	0.06	0.68	0.12	0.14
		U	0	0.14	0	0	0	0
	NormalMix	C	0.92	0.20	0.64	0.06	0.92	0.59
		O	0.08	0.04	0.2	0.38	0.08	0.41
		U	0	0.76	0.16	0.56	0	0
	Laplace	C	0.94	0.50	0.76	0.12	0.91	0.80
		O	0.06	0.06	0.20	0.76	0.09	0.21
		U	0	0.44	0.04	0.12	0	0
	Lognormal	C	0.93	0.2	0.28	0.08	0.87	0.26
		O	0.07	0	0.46	0.32	0.10	0.7
		U	0	0.8	0.26	0.60	0.08	0.04
	t(2)	C	0.93	0.16	0.24	0.08	0.89	0.20
		O	0.07	0.04	0.3	0.28	0.07	0.64
		U	0	0.8	0.46	0.64	0.04	0.16
$\theta = 0.5$	Normal	C	0.97	0.54	0.89	0.24	0.80	0.87
		O	0.03	0.14	0.11	0.64	0.20	0.13
		U	0	0.32	0	0.12	0	0
	NormalMix	C	0.96	0.22	0.58	0.08	0.90	0.56
		O	0.04	0.04	0.26	0.34	0.10	0.24
		U	0	0.74	0.16	0.58	0	0.2
	Laplace	C	0.95	0.44	0.74	0.20	0.88	0.76
		O	0.05	0.06	0.26	0.54	0.12	0.24
		U	0	0.50	0	0.26	0	0
	Lognormal	C	0.97	0.14	0.40	0.06	0.92	0.38
		O	0.03	0.02	0.26	0.38	0.04	0.58
		U	0	0.84	0.34	0.56	0.04	0.04
	t(2)	C	0.96	0.20	0.28	0.08	0.93	0.22
		O	0.04	0.08	0.24	0.08	0.05	0.64
		U	0	0.72	0.48	0.84	0.02	0.14
$\theta = 0.7$	Normal	C	0.95	0.60	0.92	0.36	0.92	0.86
		O	0.05	0.16	0.08	0.56	0.08	0.14
		U	0	0.24	0	0.08	0	0
	NormalMix	C	0.92	0.14	0.64	0.06	0.94	0.54
		O	0.08	0	0.18	0.24	0.05	0.42
		U	0	0.86	0.18	0.70	0.01	0.04
	Laplace	C	0.90	0.4	0.76	0.14	0.82	0.74
		O	0.10	0.04	0.24	0.42	0.18	0.26
		U	0	0.56	0	0.44	0	0
	Lognormal	C	0.94	0.10	0.36	0.1	0.93	0.38
		O	0.06	0.08	0.3	0.16	0.03	0.56
		U	0	0.82	0.34	0.74	0.03	0.06
	t(2)	C	0.88	0.16	0.22	0.04	0.83	0.13
		O	0.06	0.04	0.18	0.14	0.10	0.55
		U	0.06	0.80	0.60	0.82	0.07	0.32

Table C2: *Estimation and prediction results for heterogeneous errors based on 100 replicates. TMSE: total mean squared equared error. pred: prediction error (check loss or squared loss). pred.mad: mean absolute prediction error.*

θ			BQRVCSS	BQRVC	BVCSS	BVC	QRVC-adp	VC-adp
$\theta = 0.3$	Normal	TMSE	0.35(0.15)	3.44(0.54)	0.94(0.30)	2.82(0.37)	0.37(0.20)	0.95(0.17)
		pred	0.20(0.03)	0.43(0.04)	0.24(0.28)	2.08(0.36)	0.21(0.05)	0.29(0.02)
		pred.mad	0.39(0.06)	0.89(0.07)	0.7(0.12)	0.97(0.08)	0.4(0.08)	0.76(0.07)
	NormalMix	TMSE	0.50(0.24)	5.05(0.99)	1.04(1.20)	5.79(1.7)	0.45(0.23)	1.62(0.61)
		pred	0.26(0.05)	0.59(0.06)	0.49(0.70)	4.13(1.07)	0.22(0.05)	0.57(0.03)
		pred.mad	0.46(0.09)	1.11(0.09)	0.71(0.23)	1.27(0.13)	0.44(0.08)	0.77(0.07)
	Laplace	TMSE	0.35(0.15)	4.04(0.79)	1.03(0.67)	3.57(0.9)	0.41(0.21)	0.94(0.27)
		pred	0.21(0.05)	0.49(0.05)	0.81(0.47)	2.59(0.66)	0.22(0.05)	0.29(0.04)
		pred.mad	0.37(0.07)	0.93(0.07)	0.66(0.16)	1.02(0.09)	0.41(0.08)	0.75(0.11)
	Lognormal	TMSE	0.20(0.09)	4.18(0.93)	2.55(2.57)	9.84(4.87)	0.37(0.54)	3.59(2.03)
		pred	0.15(0.03)	0.47(0.06)	0.36(1.91)	7.81(3.61)	0.17(0.09)	0.51(0.13)
		pred.mad	0.29(0.07)	0.95(0.11)	1.37(0.27)	1.35(0.22)	0.34(0.18)	1.38(0.27)
	t(2)	TMSE	0.64(0.39)	5.87(1.29)	2.99(2.83)	10.94(6.72)	1.37(1.59)	3.27(1.27)
		pred	0.29(0.08)	0.73(0.12)	0.52(2.01)	8.82(6.62)	0.34(0.13)	0.55(0.22)
		pred.mad	0.51(0.12)	1.25(0.15)	1.16(0.34)	1.58(0.27)	0.65(0.21)	1.30(0.38)
$\theta = 0.5$	Normal	TMSE	0.27(0.21)	3.38(0.53)	0.93(0.17)	2.21(0.36)	0.28(0.16)	0.96(0.16)
		pred	0.15(0.04)	0.4(0.03)	0.23(0.19)	1.48(0.23)	0.17(0.04)	0.30(0.03)
		pred.mad	0.30(0.09)	0.8(0.06)	0.65(0.09)	0.82(0.06)	0.33(0.07)	0.76(0.05)
	NormalMix	TMSE	0.29(0.12)	4.61(0.82)	1.12(0.94)	5.2(1.48)	0.35(0.16)	1.62(0.61)
		pred	0.17(0.03)	0.5(0.05)	0.28(0.41)	3.74(0.98)	0.19(0.03)	0.31(0.03)
		pred.mad	0.33(0.07)	1.00(0.09)	0.54(0.27)	1.14(0.11)	0.38(0.06)	0.61(0.06)
	Laplace	TMSE	0.21(0.1)	3.84(0.67)	0.98(0.41)	3.18(0.72)	0.21(0.12)	1.06(0.33)
		pred	0.14(0.03)	0.44(0.04)	0.28(0.43)	2.34(0.52)	0.14(0.03)	0.31(0.03)
		pred.mad	0.28(0.06)	0.87(0.07)	0.56(0.10)	0.93(0.08)	0.28(0.07)	0.63(0.06)
	Lognormal	TMSE	0.29(0.16)	4.36(0.95)	2.09(2.13)	8.26(3.61)	0.40(0.48)	2.45(2.17)
		pred	0.18(0.04)	0.46(0.06)	0.54(1.91)	6.33(3.33)	0.18(0.09)	0.53(0.15)
		pred.mad	0.33(0.08)	0.91(0.12)	0.98(0.28)	1.19(0.18)	0.37(0.17)	1.05(0.3)
	t(2)	TMSE	0.38(0.22)	5.31(1.12)	3.33(3.15)	11.94(15.06)	1.16(2.2)	3.92(5.56)
		pred	0.18(0.04)	0.54(0.05)	0.39(2.16)	10.45(6.85)	0.26(0.13)	0.52(0.29)
		pred.mad	0.36(0.08)	1.09(0.09)	1.02(0.68)	1.40(0.28)	0.53(0.26)	1.19(0.56)
$\theta = 0.7$	Normal	TMSE	0.33(0.11)	3.65(0.59)	0.85(0.25)	2.71(0.47)	0.38(0.16)	1.06(0.27)
		pred	0.20(0.04)	0.44(0.04)	0.23(0.23)	2.00(0.37)	0.21(0.05)	0.30(0.03)
		pred.mad	0.37(0.06)	0.90(0.07)	0.66(0.11)	0.96(0.07)	0.41(0.08)	0.78(0.1)
	NormalMix	TMSE	0.51(0.22)	5.32(0.89)	1.22(1.04)	5.91(1.57)	0.78(0.56)	1.65(0.61)
		pred	0.25(0.05)	0.61(0.08)	0.35(0.91)	4.45(1.09)	0.30(0.10)	0.36(0.06)
		pred.mad	0.47(0.09)	1.17(0.10)	0.80(0.23)	1.31(0.12)	0.55(0.17)	0.93(0.15)
	Laplace	TMSE	0.42(0.22)	4.25(0.73)	0.93(0.42)	3.37(0.72)	0.42(0.24)	1.1(0.39)
		pred	0.23(0.05)	0.51(0.06)	0.79(0.35)	2.55(0.61)	0.21(0.06)	0.3(0.04)
		pred.mad	0.41(0.08)	0.98(0.10)	0.66(0.14)	1.03(0.12)	0.41(0.1)	0.77(0.12)
	Lognormal	TMSE	0.80(0.58)	6.85(1.71)	2.47(8.41)	7.98(6.94)	2.72(6.07)	2.54(3.39)
		pred	0.33(0.13)	0.80(0.17)	0.35(6.38)	6.04(6.79)	0.32(0.30)	0.49(0.28)
		pred.mad	0.58(0.17)	1.30(0.21)	0.86(0.38)	1.31(0.18)	0.78(0.56)	1.00(0.52)
	t(2)	TMSE	0.62(0.29)	6.44(1.31)	5.37(4.67)	13.41(12.08)	1.27(1.13)	3.32(3.06)
		pred	0.28(0.06)	0.79(0.12)	0.62(3.36)	11.13(10.49)	0.32(0.13)	0.59(0.28)
		pred.mad	0.50(0.10)	1.38(0.16)	1.42(0.44)	1.71(0.32)	0.63(0.25)	1.26(0.44)

Table C3: *Identification results for simulated SNPs with i.i.d. errors based on 100 replicates. C: correct-fitting proportion; O: overfitting proportion; U: underfitting proportion.*

θ			BQRVCSS	BQRVC	BVCSS	BVC	QRVC-adp	VC-adp
$\theta = 0.3$	Normal	C	0.96	0.78	0.94	0.46	0.92	0.89
		O	0.04	0.16	0.06	0.54	0.08	0.11
		U	0	0.06	0	0	0	0
	NormalMix	C	0.88	0.62	0.89	0.18	0.84	0.83
		O	0.12	0.18	0.11	0.72	0.14	0.15
		U	0	0.2	0	0.1	0.02	0.02
	Laplace	C	0.91	0.7	0.87	0.3	0.88	0.83
		O	0.09	0.18	0.13	0.68	0.12	0.17
		U	0	0.12	0	0.02	0	0
	Lognormal	C	0.99	0.62	0.8	0.06	0.89	0.76
		O	0.01	0.02	0.18	0.76	0.05	0.14
		U	0	0.36	0.02	0.18	0.06	0.1
	t(2)	C	0.92	0.34	0.5	0.1	0.84	0.38
		O	0.08	0.06	0.24	0.42	0.08	0.32
		U	0	0.6	0.26	0.48	0.08	0.3
$\theta = 0.5$	Normal	C	0.98	0.88	0.96	0.5	0.94	0.91
		O	0.02	0.1	0.04	0.5	0.06	0.09
		U	0	0.02	0	0	0	0
	NormalMix	C	0.94	0.44	0.86	0.18	0.88	0.84
		O	0.06	0.26	0.14	0.72	0.1	0.14
		U	0	0.3	0	0.1	0.02	0.02
	Laplace	C	0.96	0.76	0.88	0.24	0.91	0.82
		O	0.04	0.16	0.12	0.74	0.09	0.16
		U	0	0.08	0	0.02	0	0.02
	Lognormal	C	0.98	0.4	0.79	0.06	0.91	0.72
		O	0.02	0.04	0.11	0.48	0.06	0.2
		U	0	0.56	0.1	0.46	0.03	0.08
	t(2)	C	0.97	0.3	0.54	0.12	0.91	0.4
		O	0.03	0.06	0.22	0.36	0.09	0.28
		U	0	0.64	0.24	0.52	0	0.32
$\theta = 0.7$	Normal	C	0.97	0.82	0.95	0.3	0.93	0.88
		O	0.04	0.08	0.05	0.68	0.07	0.12
		U	0	0.1	0	0.02	0	0
	NormalMix	C	0.87	0.6	0.84	0.2	0.82	0.79
		O	0.13	0.22	0.16	0.72	0.16	0.18
		U	0	0.18	0	0.08	0.02	0.04
	Laplace	C	0.88	0.78	0.87	0.32	0.88	0.79
		O	0.12	0.16	0.13	0.66	0.12	0.21
		U	0	0.06	0	0.02	0	0
	Lognormal	C	0.65	0.24	0.76	0.16	0.54	0.78
		O	0.35	0.36	0.12	0.56	0.34	0.16
		U	0	0.4	0.12	0.28	0.12	0.06
	t(2)	C	0.88	0.18	0.64	0.18	0.82	0.4
		O	0.11	0.1	0.1	0.38	0.12	0.4
		U	0.01	0.72	0.26	0.44	0.06	0.2

Table C4: *Estimation and prediction results for simulated SNPs with i.i.d. errors based on 100 replicates. TMSE: total mean squared equared error. pred: prediction error (check loss or squared loss). pred.mad: mean absolute prediction error.*

θ			BQRVCSS	BQRVC	BVCSS	BVC	QRVC-adp	VC-adp
$\theta = 0.3$	Normal	TMSE	0.23(0.1)	2.32(0.4)	0.45(0.12)	1.51(0.19)	0.28(0.11)	0.79(0.14)
		pred	0.15(0.03)	0.32(0.02)	0.4(0.11)	1.09(0.14)	0.17(0.03)	0.25(0.02)
		pred.mad	0.31(0.06)	0.74(0.05)	0.55(0.08)	0.84(0.06)	0.34(0.05)	0.68(0.05)
	NormalMix	TMSE	0.34(0.17)	3.47(0.59)	0.76(0.23)	2.92(0.46)	0.53(0.35)	0.98(0.27)
		pred	0.19(0.04)	0.44(0.04)	0.68(0.19)	2.29(0.4)	0.23(0.06)	0.26(0.03)
		pred.mad	0.37(0.07)	0.94(0.07)	0.69(0.1)	1.1(0.07)	0.45(0.1)	0.75(0.08)
	Laplace	TMSE	0.26(0.1)	2.91(0.53)	0.45(0.12)	2.06(0.32)	0.34(0.15)	0.8(0.11)
		pred	0.19(0.04)	0.38(0.04)	0.39(0.11)	1.48(0.23)	0.19(0.05)	0.25(0.02)
		pred.mad	0.33(0.05)	0.8(0.06)	0.52(0.08)	0.91(0.07)	0.36(0.07)	0.69(0.05)
	Lognormal	TMSE	0.11(0.07)	3.23(0.61)	1.76(0.64)	4.7(1.38)	0.28(0.51)	1.45(0.76)
		pred	0.09(0.01)	0.34(0.03)	1.52(0.47)	3.85(1.23)	0.12(0.06)	0.35(0.05)
		pred.mad	0.21(0.04)	0.81(0.07)	1.06(0.13)	1.13(0.12)	0.26(0.11)	1.11(0.16)
	t(2)	TMSE	0.38(0.17)	4.7(1.07)	1.99(1.66)	7.91(9.55)	1.3(1.3)	1.54(1.52)
		pred	0.22(0.06)	0.57(0.08)	1.5(1.17)	6.19(7.8)	0.34(0.17)	0.38(0.2)
		pred.mad	0.39(0.08)	1.06(0.1)	0.91(0.32)	1.34(0.2)	0.63(0.27)	0.99(0.39)
$\theta = 0.5$	Normal	TMSE	0.19(0.07)	2.14(0.38)	0.41(0.09)	1.21(0.14)	0.28(0.12)	0.76(0.1)
		pred	0.14(0.02)	0.35(0.02)	0.41(0.10)	0.8(0.09)	0.16(0.03)	0.28(0.02)
		pred.mad	0.29(0.05)	0.71(0.04)	0.56(0.05)	0.72(0.04)	0.33(0.05)	0.67(0.03)
	NormalMix	TMSE	0.27(0.12)	3.67(0.58)	0.73(0.16)	2.65(0.43)	0.49(0.37)	1.03(0.32)
		pred	0.17(0.03)	0.47(0.03)	0.66(0.15)	1.93(0.32)	0.21(0.05)	0.31(0.04)
		pred.mad	0.33(0.05)	0.94(0.06)	0.63(0.09)	0.97(0.06)	0.41(0.09)	0.72(0.08)
	Laplace	TMSE	0.16(0.05)	2.88(0.43)	0.45(0.09)	1.87(0.35)	0.28(0.19)	0.78(0.23)
		pred	0.13(0.02)	0.39(0.03)	0.48(0.11)	1.28(0.22)	0.15(0.03)	0.3(0.03)
		pred.mad	0.26(0.04)	0.79(0.05)	0.52(0.07)	0.8(0.06)	0.31(0.07)	0.65(0.06)
	Lognormal	TMSE	0.23(0.13)	4.16(0.83)	1.55(1.14)	5.3(2.43)	0.44(0.45)	1.43(0.66)
		pred	0.15(0.03)	0.46(0.05)	1.23(0.73)	4.22(1.75)	0.19(0.06)	0.37(0.08)
		pred.mad	0.31(0.06)	0.93(0.11)	0.83(0.2)	1.06(0.12)	0.37(0.12)	0.95(0.17)
	t(2)	TMSE	0.31(0.18)	4.17(0.83)	1.94(1.63)	7.49(7.61)	1.25(1.23)	2.14(1.9)
		pred	0.17(0.03)	0.5(0.05)	1.53(1.09)	6.09(6.96)	0.29(0.15)	0.36(0.18)
		pred.mad	0.34(0.06)	1(0.1)	0.76(0.44)	1.21(0.18)	0.59(0.3)	0.93(0.36)
$\theta = 0.7$	Normal	TMSE	0.19(0.07)	2.37(0.46)	0.41(0.1)	1.5(0.18)	0.3(0.16)	0.78(0.12)
		pred	0.15(0.04)	0.32(0.02)	0.37(0.09)	1.06(0.12)	0.17(0.03)	0.23(0.02)
		pred.mad	0.3(0.06)	0.73(0.04)	0.53(0.08)	0.83(0.05)	0.34(0.06)	0.65(0.05)
	NormalMix	TMSE	0.35(0.15)	3.49(0.53)	0.7(0.19)	2.94(0.45)	0.52(0.3)	1.11(0.39)
		pred	0.21(0.05)	0.45(0.03)	0.61(0.15)	2.28(0.35)	0.22(0.06)	0.26(0.03)
		pred.mad	0.39(0.08)	0.94(0.06)	0.65(0.08)	1.11(0.08)	0.44(0.1)	0.76(0.09)
	Laplace	TMSE	0.25(0.13)	2.76(0.46)	0.46(0.13)	1.99(0.27)	0.36(0.16)	0.86(0.19)
		pred	0.17(0.04)	0.38(0.03)	0.4(0.11)	1.49(0.22)	0.18(0.05)	0.25(0.02)
		pred.mad	0.32(0.07)	0.79(0.06)	0.53(0.08)	0.91(0.06)	0.36(0.07)	0.66(0.06)
	Lognormal	TMSE	0.78(0.79)	5.24(1.38)	1.06(1.07)	4.21(1.91)	1.05(0.88)	0.49(0.77)
		pred	0.34(0.13)	0.63(0.14)	0.71(0.63)	3.29(1.66)	0.32(0.12)	0.33(0.08)
		pred.mad	0.55(0.18)	1.04(0.16)	0.58(0.22)	1.11(0.12)	0.59(0.17)	0.98(0.13)
	t(2)	TMSE	0.46(0.38)	4.83(1.35)	1.9(1.67)	7.59(7.54)	1.13(1.01)	1.77(1)
		pred	0.23(0.07)	0.6(0.13)	1.46(1.13)	6.03(5.75)	0.31(0.13)	0.34(0.1)
		pred.mad	0.42(0.1)	1.1(0.15)	0.9(0.28)	1.38(0.22)	0.57(0.2)	0.89(0.18)

Table C5: *Identification results for simulated SNPs with heterogeneous errors based on 100 replicates. C: correct-fitting proportion; O: overfitting proportion; U: underfitting proportion.*

θ			BQRVCSS	BQRVC	BVCSS	BVC	QRVC-adp	VC-adp
$\theta = 0.3$	Normal	C	0.96	0.62	0.9	0.14	0.88	0.82
		O	0.04	0.28	0.1	0.8	0.12	0.18
		U	0	0.1	0	0.06	0	0
	NormalMix	C	0.91	0.34	0.8	0.16	0.87	0.74
		O	0.09	0.04	0.14	0.42	0.07	0.22
		U	0	0.62	0.06	0.42	0.06	0.04
	Laplace	C	0.93	0.52	0.88	0.22	0.86	0.84
		O	0.07	0.18	0.12	0.68	0.1	0.12
		U	0	0.3	0	0.1	0.04	0.04
	Lognormal	C	0.97	0.44	0.74	0.04	0.94	0.38
		O	0.03	0.02	0.16	0.58	0.06	0.34
		U	0	0.54	0.1	0.38	0	0.28
	t(2)	C	0.95	0.16	0.4	0.1	0.89	0.24
		O	0.03	0.04	0.1	0.22	0.08	0.4
		U	0.02	0.8	0.5	0.68	0.03	0.36
$\theta = 0.5$	Normal	C	0.98	0.74	0.81	0.22	0.85	0.83
		O	0.02	0.18	0.19	0.78	0.15	0.17
		U	0	0.08	0	0	0	0
	NormalMix	C	0.95	0.46	0.86	0.06	0.76	0.74
		O	0.05	0.06	0.06	0.66	0.18	0.14
		U	0	0.48	0.08	0.28	0.06	0.12
	Laplace	C	0.98	0.46	0.9	0.18	0.87	0.78
		O	0.02	0.16	0.1	0.64	0.13	0.18
		U	0	0.38	0	0.18	0	0.04
	Lognormal	C	0.94	0.26	0.72	0.04	0.86	0.36
		O	0.06	0.06	0.12	0.44	0.12	0.4
		U	0	0.68	0.16	0.52	0.02	0.24
	t(2)	C	0.96	0.14	0.38	0.02	0.92	0.24
		O	0.04	0.02	0.18	0.26	0.06	0.32
		U	0	0.84	0.44	0.72	0.02	0.44
$\theta = 0.7$	Normal	C	0.94	0.5	0.87	0.2	0.82	0.79
		O	0.06	0.28	0.13	0.72	0.18	0.19
		U	0	0.22	0	0.08	0	0.02
	NormalMix	C	0.89	0.24	0.76	0.16	0.78	0.75
		O	0.11	0.1	0.22	0.46	0.14	0.17
		U	0	0.66	0.02	0.38	0.08	0.08
	Laplace	C	0.91	0.46	0.9	0.16	0.85	0.76
		O	0.09	0.08	0.1	0.64	0.09	0.14
		U	0	0.46	0	0.2	0.06	0.1
	Lognormal	C	0.93	0.18	0.74	0.08	0.85	0.4
		O	0.07	0.02	0.23	0.28	0.13	0.32
		U	0	0.8	0.03	0.64	0.02	0.28
	t(2)	C	0.91	0.06	0.38	0	0.86	0.29
		O	0.09	0	0.1	0.1	0.12	0.28
		U	0	0.94	0.52	0.9	0.02	0.43

Table C6: *Estimation and prediction results for simulated SNPs with heterogeneous errors based on 100 replicates. TMSE: total mean squared equared error. pred: prediction error (check loss or squared loss). pred.mad: mean absolute prediction error.*

θ			BQRVCSS	BQRVC	BVCSS	BVC	QRVC-adp	VC-adp
$\theta = 0.3$	Normal	TMSE	0.26(0.11)	3.17(0.57)	0.83(0.24)	2.71(0.44)	0.35(0.24)	1.13(0.3)
		pred	0.2(0.04)	0.43(0.03)	0.74(0.21)	2.05(0.33)	0.2(0.05)	0.3(0.03)
		pred.mad	0.35(0.06)	0.91(0.06)	0.65(0.09)	0.94(0.07)	0.38(0.08)	0.79(0.09)
	NormalMix	TMSE	0.4(0.2)	4.59(0.79)	1.72(0.86)	5.66(1.35)	0.63(0.54)	1.63(0.57)
		pred	0.23(0.06)	0.59(0.07)	1.4(0.61)	4.42(1.02)	0.27(0.1)	0.35(0.05)
		pred.mad	0.42(0.1)	1.13(0.09)	0.9(0.2)	1.28(0.12)	0.48(0.15)	0.92(0.13)
	Laplace	TMSE	0.3(0.14)	3.72(0.68)	0.9(0.36)	3.74(0.8)	0.42(0.45)	1.17(0.49)
		pred	0.22(0.06)	0.51(0.07)	0.76(0.31)	2.93(0.63)	0.21(0.08)	0.3(0.04)
		pred.mad	0.37(0.08)	0.98(0.09)	0.66(0.13)	1.05(0.09)	0.39(0.13)	0.77(0.1)
	Lognormal	TMSE	0.17(0.08)	3.54(0.67)	3.7(2.01)	8.86(3.66)	0.72(0.98)	4.32(3)
		pred	0.16(0.03)	0.48(0.04)	2.99(1.2)	7.24(3.2)	0.21(0.13)	0.52(0.15)
		pred.mad	0.29(0.06)	0.98(0.09)	1.3(0.24)	1.3(0.19)	0.4(0.23)	1.41(0.29)
	t(2)	TMSE	0.66(0.65)	5.92(1.54)	4.64(5.71)	16.16(23.82)	2.09(4.68)	3.78(4.41)
		pred	0.31(0.12)	0.76(0.15)	3.36(4.59)	12.58(18.66)	0.38(0.2)	0.51(0.2)
		pred.mad	0.53(0.16)	1.31(0.17)	1.26(0.68)	1.65(0.47)	0.67(0.31)	1.24(0.38)
$\theta = 0.5$	Normal	TMSE	0.17(0.08)	3.11(0.48)	0.82(0.21)	2.1(0.29)	0.25(0.18)	1.09(0.31)
		pred	0.13(0.03)	0.4(0.03)	0.72(0.19)	1.49(0.2)	0.15(0.04)	0.3(0.02)
		pred.mad	0.26(0.06)	0.8(0.06)	0.63(0.07)	0.81(0.05)	0.3(0.08)	0.80(0.094)
	NormalMix	TMSE	0.25(0.12)	4.2(0.71)	1.66(0.63)	4.56(1.02)	0.68(0.73)	1.74(0.71)
		pred	0.16(0.04)	0.49(0.04)	1.26(0.58)	3.49(0.74)	0.21(0.09)	0.39(0.07)
		pred.mad	0.32(0.08)	0.99(0.07)	0.55(0.16)	1.11(0.1)	0.43(0.19)	0.97(0.14)
	Laplace	TMSE	0.18(0.12)	3.78(0.63)	0.46(0.26)	3.18(0.55)	0.23(0.16)	0.85(0.45)
		pred	0.13(0.03)	0.44(0.04)	0.35(0.17)	2.44(0.48)	0.14(0.04)	0.32(0.05)
		pred.mad	0.27(0.07)	0.88(0.09)	0.42(0.09)	0.93(0.09)	0.28(0.09)	0.65(0.1)
	Lognormal	TMSE	0.17(0.08)	4.2(0.74)	2.88(4.66)	9.86(9.94)	0.7(1.14)	2.79(3.26)
		pred	0.13(0.03)	0.46(0.05)	2.03(2.83)	7.64(7.7)	0.19(0.12)	0.52(0.17)
		pred.mad	0.26(0.05)	0.92(0.1)	0.94(0.36)	1.18(0.24)	0.39(0.24)	1.04(0.33)
	t(2)	TMSE	0.3(0.16)	4.75(0.66)	3.19(4.68)	12.78(12.71)	1.55(1.46)	3.78(3.64)
		pred	0.17(0.04)	0.53(0.04)	2.23(3.35)	10.21(10.03)	0.31(0.15)	0.51(0.18)
		pred.mad	0.34(0.08)	1.06(0.08)	0.99(0.57)	1.43(0.31)	0.62(0.3)	1.31(0.36)
$\theta = 0.7$	Normal	TMSE	0.25(0.11)	3.4(0.59)	0.8(0.22)	2.63(0.39)	0.3(0.13)	1.12(0.29)
		pred	0.19(0.05)	0.45(0.03)	0.73(0.2)	2.04(0.31)	0.19(0.04)	0.31(0.03)
		pred.mad	0.35(0.07)	0.94(0.06)	0.66(0.1)	0.95(0.08)	0.36(0.06)	0.8(0.07)
	NormalMix	TMSE	0.39(0.17)	4.77(0.76)	1.35(0.49)	4.76(0.97)	0.94(1.08)	1.85(0.77)
		pred	0.25(0.05)	0.62(0.06)	1.18(0.44)	3.94(0.81)	0.3(0.13)	0.38(0.08)
		pred.mad	0.43(0.06)	1.2(0.09)	0.82(0.16)	1.25(0.1)	0.55(0.23)	0.96(0.18)
	Laplace	TMSE	0.25(0.11)	4.14(0.7)	0.88(0.25)	3.57(0.64)	0.43(0.53)	1.3(0.5)
		pred	0.21(0.04)	0.53(0.05)	0.76(0.23)	2.82(0.54)	0.2(0.08)	0.33(0.05)
		pred.mad	0.36(0.06)	1.02(0.08)	0.66(0.1)	1.05(0.09)	0.36(0.14)	0.82(0.11)
	Lognormal	TMSE	0.58(0.23)	6.55(1.35)	5.32(22.78)	9.11(11.68)	1.26(1.18)	2.15(2.68)
		pred	0.29(0.08)	0.81(0.15)	3.26(13.26)	7.85(12.78)	0.35(0.17)	0.42(0.17)
		pred.mad	0.52(0.1)	1.32(0.17)	0.84(0.32)	1.32(0.18)	0.63(0.26)	0.87(0.32)
	t(2)	TMSE	0.49(0.25)	6.08(0.99)	5.98(9.18)	18.73(22.33)	3.2(3.84)	4.94(5.77)
		pred	0.28(0.06)	0.8(0.12)	5.12(8.74)	15.91(20.98)	0.54(0.31)	0.65(0.27)
		pred.mad	0.48(0.1)	1.39(0.14)	1.48(0.58)	1.72(0.36)	0.98(0.55)	1.45(0.58)

C.2 Hyper-parameters sensitivity analysis

Table C7: *Sensitivity analysis on the choice of the hyperparameter for π_0 using different Beta priors for the Laplace error dirstribution for the 30% quantile.*

	C	O	U	TMSE	pred	pred.mad
Beta(0.5,0.5)	0.9	0.1	0	0.27(0.12)	0.19(0.08)	0.33(0.06)
Beta(1,1)	0.9	0.1	0	0.28(0.12)	0.19(0.08)	0.33(0.06)
Beta(2,2)	0.9	0.1	0	0.28(0.11)	0.19(0.08)	0.33(0.06)
Beta(1,5)	0.9	0.1	0	0.27(0.11)	0.19(0.07)	0.33(0.06)
Beta(5,1)	0.9	0.1	0	0.27(0.11)	0.19(0.07)	0.33(0.06)

Table C8: *Sensitivity analysis on the choice of the hyperparameter for η using different Gamma priors for the Laplace error dirstribution for the 30% quantile.*

	C	O	U	TMSE	pred	pred.mad
Gamma(0.1,1)	0.9	0.1	0	0.29(0.17)	0.2(0.09)	0.33(0.06)
Gamma(1,1)	0.9	0.1	0	0.29(0.16)	0.2(0.09)	0.33(0.06)
Gamma(1,5)	0.9	0.1	0	0.3(0.16)	0.2(0.09)	0.33(0.06)
Gamma(2,5)	0.88	0.12	0	0.3(0.16)	0.2(0.09)	0.33(0.06)
Gamma(5,1)	0.9	0.1	0	0.29(0.16)	0.2(0.09)	0.33(0.06)

Table C9: *Sensitivity analysis on the choice of the hyperparameter for π_0 using different Beta priors for the Laplace error dirstribution for the 50% quantile.*

	C	O	U	TMSE	pred	pred.mad
Beta(0.5,0.5)	0.92	0.08	0	0.22(0.05)	0.16(0.03)	0.29(0.04)
Beta(1,1)	0.94	0.06	0	0.22(0.06)	0.14(0.03)	0.29(0.04)
Beta(2,2)	0.94	0.06	0	0.22(0.06)	0.14(0.03)	0.29(0.04)
Beta(1,5)	0.94	0.06	0	0.22(0.06)	0.14(0.03)	0.29(0.04)
Beta(5,1)	0.92	0.08	0	0.22(0.06)	0.15(0.03)	0.29(0.04)

Table C10: *Sensitivity analysis on the choice of the hyperparameter for η using different Gamma priors for the Laplace error dirstribution for the 50% quantile.*

	C	O	U	TMSE	pred	pred.mad
Gamma(0.1,1)	0.96	0.04	0	0.22(0.05)	0.15(0.03)	0.29(0.04)
Gamma(1,1)	0.94	0.06	0	0.22(0.05)	0.15(0.03)	0.29(0.04)
Gamma(1,5)	0.94	0.06	0	0.23(0.05)	0.16(0.03)	0.29(0.04)
Gamma(2,5)	0.94	0.06	0	0.22(0.06)	0.15(0.03)	0.29(0.04)
Gamma(5,1)	0.94	0.06	0	0.22(0.05)	0.15(0.03)	0.29(0.04)

C.3 Sensitivity analysis on smoothness specification

Let O denote the degree of B spline basis and K denote the number of interior knots. For quadratic and cubic splines corresponding to O=2 and O=3 respectively, we conduct a sensitivity analysis for the proposed model.

Table C11: *Sensitivity analysis on smoothness specification for the Laplace error dirstribution for the 30% quantile.*

O=2	K	1	2	3	4	5
Laplace	C	0.88	0.90	0.92	0.89	0.91
	O	0.12	0.1	0.08	0.11	0.09
	U	0	0	0	0	0
	TMSE	0.33(0.19)	0.28(0.12)	0.31(0.14)	0.24(0.12)	0.25(0.15)
	pred	0.18(0.05)	0.17(0.05)	0.21(0.05)	0.19(0.04)	0.20(0.05)
	pred.mad	0.32(0.07)	0.32(0.06)	0.30(0.07)	0.35(0.06)	0.34(0.07)
O=3	K	1	2	3	4	5
Laplace	C	0.89	0.90	0.92	0.86	0.88
	O	0.11	0.10	0.08	0.14	0.12
	U	0	0	0	0	0
	TMSE	0.25(0.11)	0.28(0.12)	0.28(0.15)	0.26(0.19)	0.25(0.16)
	pred	0.17(0.04)	0.21(0.05)	0.19(0.04)	0.23(0.05)	0.22(0.04)
	pred.mad	0.30(0.06)	0.38(0.08)	0.35(0.06)	0.34(0.08)	0.33(0.06)

Table C12: *Sensitivity analysis on smoothness specification for the Normal error dirstribution for the 30% quantile.*

O=2	K	1	2	3	4	5
Normal	C	0.97	0.96	0.98	0.95	0.94
	O	0.03	0.04	0.04	0.05	0.06
	U	0	0	0	0	0
	TMSE	0.26(0.12)	0.22(0.09)	0.29(0.16)	0.23(0.12)	0.22(0.18)
	pred	0.15(0.04)	0.14(0.03)	0.17(0.04)	0.17(0.03)	0.16(0.03)
	pred.mad	0.30(0.06)	0.29(0.05)	0.26(0.06)	0.26(0.06)	0.29(0.07)
O=3	K	1	2	3	4	5
Normal	C	0.96	0.94	0.97	0.94	0.95
	O	0.04	0.06	0.03	0.06	0.05
	U	0	0	0	0	0
	TMSE	0.24(0.09)	0.26(0.14)	0.21(0.10)	0.25(0.19)	0.24(0.12)
	pred	0.15(0.03)	0.16(0.03)	0.16(0.03)	0.18(0.03)	0.18(0.03)
	pred.mad	0.29(0.05)	0.28(0.06)	0.30(0.05)	0.28(0.06)	0.26(0.04)

Table C13: *Sensitivity analysis on smoothness specification for the Laplace error dirstribution for the 50% quantile.*

O=2	K	1	2	3	4	5
Laplace	C	0.96	0.94	0.92	0.95	0.96
	O	0.04	0.06	0.08	0.05	0.04
	U	0	0	0	0	0
	TMSE	0.25(0.11)	0.21(0.09)	0.29(0.16)	0.28(0.11)	0.25(0.19)
	pred	0.14(0.03)	0.14(0.03)	0.17(0.03)	0.16(0.02)	0.17(0.03)
	pred.mad	0.29(0.06)	0.28(0.05)	0.33(0.06)	0.31(0.05)	0.37(0.07)
O=3	K	1	2	3	4	5
Laplace	C	0.95	0.93	0.94	0.96	0.93
	O	0.05	0.07	0.06	0.04	0.07
	U	0	0	0	0	0
	TMSE	0.24(0.07)	0.31(0.14)	0.26(0.12)	0.22(0.16)	0.26(0.13)
	pred	0.13(0.02)	0.15(0.03)	0.15(0.03)	0.17(0.03)	0.17(0.03)
	pred.mad	0.29(0.05)	0.31(0.07)	0.30(0.05)	0.35(0.06)	0.34(0.05)

Table C14: *Sensitivity analysis on smoothness specification for the Normal error dirstribution for the 50% quantile.*

O=2	K	1	2	3	4	5
Normal	C	0.97	0.98	0.96	0.99	0.98
	O	0.03	0.02	0.04	0.01	0.02
	U	0	0	0	0	0
	TMSE	0.21(0.06)	0.23(0.13)	0.22(0.07)	0.24(0.14)	0.22(0.09)
	pred	0.12(0.03)	0.11(0.04)	0.13(0.03)	0.14(0.04)	0.13(0.04)
	pred.mad	0.30(0.06)	0.28(0.08)	0.28(0.06)	0.29(0.08)	0.29(0.07)
O=3	K	1	2	3	4	5
Normal	C	0.98	0.96	0.98	0.98	0.97
	O	0.02	0.04	0.02	0.02	0.03
	U	0	0	0	0	0
	TMSE	0.19(0.07)	0.29(0.11)	0.25(0.07)	0.24(0.14)	0.23(0.08)
	pred	0.13(0.02)	0.15(0.02)	0.15(0.02)	0.12(0.02)	0.14(0.02)
	pred.mad	0.27(0.04)	0.30(0.04)	0.29(0.04)	0.26(0.04)	0.27(0.03)

Table C15: *Sensitivity analysis on smoothness specification for BVCSS with the Normal error dirstribution for the 30% quantile.*

O=2	K	1	2	3	4	5
Normal	C	0.90	0.94	0.93	0.94	0.92
	O	0.10	0.06	0.07	0.06	0.08
	U	0	0	0	0	0
	TMSE	0.55(0.13)	0.47(0.11)	0.46(0.15)	0.43(0.12)	0.49(0.22)
	pred	0.23(0.10)	0.21(0.10)	0.25(0.10)	0.23(0.10)	0.22(0.11)
	pred.mad	0.57(0.08)	0.56(0.08)	0.61(0.07)	0.57(0.08)	0.54(0.07)
O=3	K	1	2	3	4	5
Normal	C	0.92	0.91	0.90	0.94	0.93
	O	0.08	0.09	0.10	0.06	0.07
	U	0	0	0	0	0
	TMSE	0.42(0.08)	0.48(0.18)	0.48(0.09)	0.51(0.25)	0.56(0.1)
	pred	0.27(0.07)	0.27(0.09)	0.24(0.08)	0.25(0.1)	0.23(0.08)
	pred.mad	0.53(0.06)	0.57(0.06)	0.55(0.06)	0.61(0.06)	0.56(0.06)

Table C16: *Sensitivity analysis on smoothness specification for BVCSS with the Normal error dirstribution for the 50% quantile.*

O=2	K	1	2	3	4	5
Normal	C	0.92	0.90	0.94	0.96	0.98
	O	0.08	0.10	0.06	0.04	0.02
	U	0	0	0	0	0
	TMSE	0.42(0.12)	0.41(0.05)	0.47(0.19)	0.43(0.07)	0.46(0.25)
	pred	0.24(0.04)	0.21(0.03)	0.23(0.07)	0.24(0.04)	0.31(0.09)
	pred.mad	0.57(0.05)	0.55(0.03)	0.55(0.06)	0.52(0.04)	0.51(0.06)
O=3	K	1	2	3	4	5
Normal	C	0.96	0.96	0.92	0.95	0.93
	O	0.04	0.04	0.08	0.05	0.07
	U	0	0	0	0	0
	TMSE	0.45(0.05)	0.42(0.15)	0.41(0.06)	0.45(0.22)	0.43(0.10)
	pred	0.21(0.03)	0.21(0.06)	0.23(0.03)	0.29(0.08)	0.26(0.04)
	pred.mad	0.54(0.04)	0.53(0.06)	0.52(0.04)	0.53(0.06)	0.53(0.04)

C.4 Posterior inference

C.4.1 Posterior inference for BQRVCSS

Priors

$$\boldsymbol{Y}_i = \boldsymbol{E}_i^\top \boldsymbol{\beta} + \boldsymbol{Z}_i^\top \boldsymbol{\alpha} + \xi_1 \tilde{v_i} + \xi_2 \tau^{-\frac{1}{2}} \sqrt{\tilde{v_i}} W_i, i = 1, ..., n,$$

$$\tilde{v_1}, ..., \tilde{v_n} \sim \prod_{i=1}^{n} \tau \exp(-\tau \tilde{v_i}), i = 1, ..., n,$$

$$W_1, ..., W_n \sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} W_i^2), i = 1, ..., n,$$

$$\boldsymbol{\alpha}_j | s_j \sim (1 - \pi_0) \mathrm{N}_d(0, s_j \mathbf{I}_d^{-1}) + \pi_0 \delta_0(\boldsymbol{\alpha}_j), j = 0, ..., p,$$

$$s_j|\eta^2 \sim (\frac{\eta^2}{2})^{\frac{d+1}{2}} s_j^{\frac{d-1}{2}} \exp(-\frac{\eta^2}{2} s_j), j = 0, ..., p,$$

$$\pi_0 \sim \text{Beta}(e, f),$$

$$\tau \sim \tau^{a-1} \exp(-b\tau),$$

$$\eta^2 \sim (\eta^2)^{c-1} \exp(-m\eta^2),$$

$$\boldsymbol{\beta} \sim \text{N}_q(0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}).$$

Gibbs Sampler

• The full conditional distribution of $\tilde{v}_i$, $i = 1, ..., n$

$$\begin{aligned}
&\tilde{v}_i|\text{rest} \\
&\propto \frac{1}{\sqrt{2\pi\tau^{-1}\xi_2^2\tilde{v}_i}} \exp\Big(-\frac{1}{2}\frac{(\boldsymbol{Y}_i - \boldsymbol{E}_i^\top\boldsymbol{\beta} - \boldsymbol{Z}_i^\top\boldsymbol{\alpha} - \xi_1\tilde{v}_i)^2}{\xi_2^2\tau^{-1}\tilde{v}_i}) \tau \exp(-\tau\tilde{v}_i\Big) \\
&\propto (\tilde{v}_i)^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}\frac{(\boldsymbol{Y}_i - \boldsymbol{E}_i^\top\boldsymbol{\beta} - \boldsymbol{Z}_i^\top\boldsymbol{\alpha})}{\xi_2^2\tau^{-1}\tilde{v}_i} - \frac{1}{2}\frac{\xi_1^2\tilde{v}_i}{\tau^{-1}\xi_2^2} - \tau\tilde{v}_i\Big) \\
&\propto (\tilde{v}_i)^{-\frac{1}{2}} \exp\Bigg(-\frac{1}{2}\Big((\frac{\tau\xi_1^2}{\xi_2^2} + 2\tau)\tilde{v}_i + \frac{\tau(\boldsymbol{Y}_i - \boldsymbol{E}_i^\top\boldsymbol{\beta} - \boldsymbol{Z}_i^\top\boldsymbol{\alpha})^2}{\xi_2^2}\frac{1}{\tilde{v}_i}\Big)\Bigg).
\end{aligned}$$

Hence, the full conditional distribution of $\tilde{v}_i$ is generalized inverse Gaussian distribution.

• The full conditional distribution of $\boldsymbol{\alpha}_j$, $j = 0, ..., p$

$$\begin{aligned}
&\boldsymbol{\alpha}_j|\text{rest} \\
&\propto \prod_{i=1}^{n} \exp\Big(-\frac{\tau}{2\xi_2^2\tilde{v}_i}(\boldsymbol{Y}_i - \boldsymbol{Z}_{i,-j}^\top\boldsymbol{\alpha}_{-j} - \boldsymbol{Z}_{ij}^\top\boldsymbol{\alpha}_j - \boldsymbol{E}_i^\top\boldsymbol{\beta} - \xi_1\tilde{v}_i)^2\Big) \\
&\times \Bigg((1-\pi_0)\frac{1}{\sqrt{2\pi|s_j\mathbf{I}_d^{-1}|}} \exp\Big(-\frac{1}{2}\boldsymbol{\alpha}_j^\top(s_j\mathbf{I}_d^{-1})^{-1}\boldsymbol{\alpha}_j\Big)\mathbf{I}_{(\boldsymbol{\alpha}_j \neq 0)} + \pi_0\delta_0(\boldsymbol{\alpha}_j)\Bigg).
\end{aligned}$$

The slab part,

$$
\begin{aligned}
&\boldsymbol{\alpha}_j|\text{rest} \\
&\propto \prod_{i=1}^{n} \exp\Big(-\frac{\tau}{2\xi_2^2 \tilde{v}_i}(\boldsymbol{Y}_i - \boldsymbol{Z}_{i,-j}^\top \boldsymbol{\alpha}_{-j} - \boldsymbol{Z}_{ij}^\top \boldsymbol{\alpha}_j - \boldsymbol{E}_i^\top \boldsymbol{\beta} - \xi_1 \tilde{v}_i)^2 \Big) \\
&\times (1-\pi_0) \frac{1}{\sqrt{2\pi |s_j \mathbf{I}_d^{-1}|}} \exp\Big(-\frac{1}{2} \boldsymbol{\alpha}_j^\top (s_j \mathbf{I}_d^{-1})^{-1} \boldsymbol{\alpha}_j \Big) \\
&\propto (1-\pi_0) \frac{1}{\sqrt{2\pi |s_j \mathbf{I}_d^{-1}|}} \exp\Big(-\frac{1}{2} \tau \xi_2^{-2} \sum_{i=1}^{n} \frac{1}{\tilde{v}_i} (\boldsymbol{Y}_i - \boldsymbol{Z}_{i,-j}^\top \boldsymbol{\alpha}_{-j} - \boldsymbol{E}_i^\top \boldsymbol{\beta} - \xi_1 \tilde{v}_i)^2 \Big) \\
&\times \exp\Bigg(-\frac{1}{2} \Big(\boldsymbol{\alpha}_j^\top (\tau \xi_2^{-2} \sum_{i=1}^{n} \frac{\boldsymbol{Z}_{ij} \boldsymbol{Z}_{ij}^\top}{\tilde{v}_i} + s_j^{-1} \mathbf{I}_d) \boldsymbol{\alpha}_j \\
&- 2\tau \xi_2^{-2} \sum_{i=1}^{n} \frac{1}{\tilde{v}_i} (\boldsymbol{Y}_i - \boldsymbol{Z}_{i,-j}^\top \boldsymbol{\alpha}_{-j} - \boldsymbol{E}_i^\top \boldsymbol{\beta} - \xi_1 \tilde{v}_i) Z_{ij}^\top \boldsymbol{\alpha}_j \Big) \Bigg).
\end{aligned}
$$

Let the variance

$$
\boldsymbol{\Sigma}_j = (\tau \xi_2^{-2} \sum_{i=1}^{n} \frac{1}{\tilde{v}_i} \boldsymbol{Z}_{ij} \boldsymbol{Z}_{ij}^\top + s_j^{-1} \boldsymbol{K}_j)^{-1}
$$

and the mean

$$
\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j \tau \xi_2^{-2} \sum_{i=1}^{n} \frac{\boldsymbol{Z}_{ij}}{\tilde{v}_i} (\boldsymbol{Y}_i - \boldsymbol{Z}_{i,-j}^\top \boldsymbol{\alpha}_{-j} - \boldsymbol{E}_i^\top \boldsymbol{\beta} - \xi_1 \tilde{v}_i),
$$

then

$$
\begin{aligned}
&\boldsymbol{\alpha}_j|\text{rest} \\
&\propto (1-\pi_0) |s_j \mathbf{I}_d^{-1}|^{-\frac{1}{2}} |\boldsymbol{\Sigma}_j|^{-\frac{1}{2}} \exp\Big(-\frac{1}{2} \tau \xi_2^{-2} \sum_{i=1}^{n} \frac{1}{\tilde{v}_i} (\boldsymbol{Y}_i - \boldsymbol{Z}_{i,-j}^\top \boldsymbol{\alpha}_{-j} - \boldsymbol{E}_i^\top \boldsymbol{\beta} - \xi_1 \tilde{v}_i)^2 \Big) \\
&\times \exp(\frac{1}{2} \boldsymbol{\mu}_j^\top \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j) \times \mathrm{N}_d(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j).
\end{aligned}
$$

The spike part,

$$\boldsymbol{\alpha}_j|\text{rest}$$

$$\propto \pi_0 \exp\Big(-\frac{1}{2}\tau\xi_2^{-2}\sum_{i=1}^{n}\frac{1}{\tilde{v}_i}(\boldsymbol{Y}_i - \boldsymbol{Z}_{i,-j}^{\top}\boldsymbol{\alpha}_{-j} - \boldsymbol{E}_i^{\top}\boldsymbol{\beta} - \xi_1\tilde{v}_i)^2\Big).$$

Proportion of the spike part

$$P(\boldsymbol{\alpha}_j = 0|\text{rest}) = \frac{\pi_0}{\pi_0 + (1-\pi_0)|s_j\mathbf{I}_d^{-1}|^{-\frac{1}{2}}|\boldsymbol{\Sigma}_j|^{\frac{1}{2}}\exp(\frac{1}{2}\boldsymbol{\mu}_j^{\top}\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j)}.$$

- The full conditional distribution of τ is

$\tau|\text{rest}$

$$\propto \prod_{i=1}^{n}\sqrt{\tau}\exp\Big(-\frac{\tau(\boldsymbol{Y}_i - \boldsymbol{Z}_i^{\top}\boldsymbol{\alpha} - \boldsymbol{E}_i^{\top}\boldsymbol{\beta} - \xi_1\tilde{v}_i)^2}{\xi_2^2\tilde{v}_i}\Big) \times \prod_{i=1}^{n}\tau\exp(-\tau\tilde{v}_i) \times \tau^{a-1}\exp(-b\tau)$$

$$\propto \tau^{\frac{3}{2}n+a-1}\exp\left(-\Big(\frac{1}{2}\sum_{i=1}^{n}\frac{\tau(\boldsymbol{Y}_i - \boldsymbol{Z}_i^{\top}\boldsymbol{\alpha} - \boldsymbol{E}_i^{\top}\boldsymbol{\beta} - \xi_1\tilde{v}_i)^2}{\xi_2^2\tilde{v}_i} + \sum_{i=1}^{n}\tilde{v}_i + b\Big)\tau\right).$$

Therefore, the posterior distribution of τ is

$$\tau|\text{rest} \propto \text{Gamma}\Big(\frac{3}{2}n + a, \frac{1}{2}\sum_{i=1}^{n}\frac{\tau(\boldsymbol{Y}_i - \boldsymbol{Z}_i^{\top}\boldsymbol{\alpha} - \boldsymbol{E}_i^{\top}\boldsymbol{\beta} - \xi_1\tilde{v}_i)^2}{\xi_2^2\tilde{v}_i} + \sum_{i=1}^{n}\tilde{v}_i + b\Big).$$

- The full conditional distribution of η^2 is

$$\eta^2|\text{rest}$$

$$\propto \prod_{j=0}^{p}(\frac{\eta^2}{2})^{\frac{d+1}{2}}\exp(-\frac{\eta^2}{2}s_j) \times (\eta^2)^{c-1}\exp(-m\eta^2)$$

$$\propto (\eta^2)^{\frac{(d+1)(p+1)}{2}+c-1}\exp(-(\frac{1}{2}\sum_{j=0}^{p}s_j + m)\eta^2).$$

Therefore, the posterior distribution of η^2 is

$$\eta^2|\text{rest} \propto \text{Gamma}\Big(\frac{(d+1)(p+1)}{2}+c, \frac{1}{2}\sum_{j=0}^{p} s_j + m\Big).$$

- The full conditional distribution of s_j $(j = 0, ..., p)$ is

$s_j|\text{rest}$

$$\propto \left((1-\pi_0)\frac{1}{\sqrt{2\pi|s_j\mathbf{I}_d^{-1}|}}\exp\Big(-\frac{1}{2}\boldsymbol{\alpha}_j^\top(s_j\mathbf{I}_d^{-1})^{-1}\boldsymbol{\alpha}_j\Big)\mathbf{I}_{(\boldsymbol{\alpha}_j\neq 0)} + \pi_0\delta_0(\boldsymbol{\alpha}_j)\right) \times s_j^{\frac{d-1}{2}}\exp(-\frac{\eta^2}{2}s_j).$$

The slab part,

$$s_j|\text{rest}$$

$$\propto s_j^{-\frac{d}{2}}\exp\Big(-\frac{1}{2}(\eta^2 s_j + \boldsymbol{\alpha}_j^\top\mathbf{I}_d\boldsymbol{\alpha}_j\frac{1}{s_j})\Big).$$

Therefore, the posterior distribution of s_j is

$$s_j^{-1}|\text{rest} \propto \text{invGamma}(\sqrt{\frac{\eta^2}{\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j}}, \eta^2), \text{if } \boldsymbol{\alpha}_j \neq 0.$$

The spike part,

$$s_j|\text{rest} \propto s_j^{\frac{d-1}{2}}\exp(-\frac{\eta^2}{2}s_j),$$

which is Gamma$(\frac{d+1}{2}, \frac{\eta^2}{2})$. Together

$$s_j^{-1}|\text{rest} \sim \begin{cases} \text{Inverse-Gamma}(\frac{d+1}{2},\ \frac{\eta^2}{2}) & \text{if } \boldsymbol{\alpha}_j = 0 \\ \text{Inverse-Gaussian}(\sqrt{\frac{\eta^2}{\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j}},\eta^2) & \text{if } \boldsymbol{\alpha}_j \neq 0 \end{cases}.$$

- The full conditional distribution of π_0, $i = 1, ..., n$

$\pi_0|\text{rest}$

$$\propto \prod_{j=0}^{p} \left((1-\pi_0) \frac{1}{\sqrt{2\pi|s_j \mathbf{I}_d^{-1}|}} \exp\Big(-\frac{1}{2} \boldsymbol{\alpha}_j^\top (s_j \mathbf{I}_d^{-1})^{-1} \boldsymbol{\alpha}_j \Big) \mathbf{I}_{(\boldsymbol{\alpha}_j \neq 0)} + \pi_0 \delta_0(\boldsymbol{\alpha}_j) \right) \times \pi_0^{e-1} (1-\pi_0)^{f-1}$$

Let

$$Q_j = \begin{cases} 0 & \text{if } \boldsymbol{\alpha}_j = 0 \\ & \\ 1 & \text{if } \boldsymbol{\alpha}_j \neq 0 \end{cases},$$

then the posterior distribution of π_0 becomes

$$\pi_0|\text{rest} \propto \pi_0^{1+p-\sum_{j=0}^{p} Q_j + e - 1} (1-\pi_0)^{\sum_{j=0}^{p} Q_j + f - 1},$$

which is Beta$(1 + p - \sum_{j=0}^{p} Q_j + e - 1, \sum_{j=0}^{p} Q_j + f)$.

- The full conditional distribution of $\boldsymbol{\beta}$

$\boldsymbol{\beta}|\text{rest}$

$$\propto \prod_{i=1}^{n} \exp\Big(-\frac{\tau}{2\xi_2^2 \tilde{v}_i} (\boldsymbol{Y}_i - \boldsymbol{Z}_i^\top \boldsymbol{\alpha} - \boldsymbol{E}_i^\top \boldsymbol{\beta} - \xi_1 \tilde{v}_i)^2 \Big) \exp(-\frac{1}{2} \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta})$$

$$\propto \exp\left(-\frac{1}{2} \Big(\boldsymbol{\beta}^\top (\sum_{i=1}^{n} \frac{\tau \boldsymbol{E}_i \boldsymbol{E}_i^\top}{\xi_2^2 \tilde{v}_i} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}) \boldsymbol{\beta} - 2 \sum_{i=1}^{n} \frac{\tau}{\xi_2^2 \tilde{v}_i} (\boldsymbol{Y}_i - \boldsymbol{Z}_i^\top \boldsymbol{\alpha} - \xi_1 \tilde{v}_i) \boldsymbol{E}_i^\top \boldsymbol{\beta} \Big) \right)$$

$$\propto \mathrm{N}_q \left((\sum_{i=1}^{n} \frac{\tau \boldsymbol{E}_i \boldsymbol{E}_i^\top}{\xi_2^2 \tilde{v}_i} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1} \Big(\sum_{i=1}^{n} \frac{\tau}{\xi_2^2 \tilde{v}_i} (\boldsymbol{Y}_i - \boldsymbol{Z}_i^\top \boldsymbol{\alpha} - \xi_1 \tilde{v}_i) \boldsymbol{E}_i^\top \Big)^\top, (\sum_{i=1}^{n} \frac{\tau \boldsymbol{E}_i \boldsymbol{E}_i^\top}{\xi_2^2 \tilde{v}_i} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1} \right),$$

which is a multivariate normal distribution.

C.4.2 Posterior inference for BQRVC

Priors

$$Y_i = \boldsymbol{E}_i^\top \boldsymbol{\beta} + Z_i^\top \boldsymbol{\alpha} + \xi_1 \tilde{v_i} + \xi_2 \tau^{-\frac{1}{2}} \sqrt{\tilde{v_i}} W_i, i = 1, ..., n,$$

$$\tilde{v_1}, ..., \tilde{v_n} \sim \prod_{i=1}^{n} \tau \exp(-\tau \tilde{v_i}), i = 1, ..., n,$$

$$W_1, ..., W_n \sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} W_i^2), i = 1, ..., n,$$

$$\boldsymbol{\alpha}_j | s_j \sim \mathrm{N}_d(0, s_j \mathbf{I}_d^{-1}), j = 0, ..., p,$$

$$s_j | \eta^2 \sim (\frac{\eta^2}{2})^{\frac{d+1}{2}} s_j^{\frac{d-1}{2}} \exp(-\frac{\eta^2}{2} s_j), j = 0, ..., p,$$

$$\pi_0 \sim \mathrm{Beta}(e, f),$$

$$\tau \sim \tau^{a-1} \exp(-b\tau),$$

$$\eta^2 \sim (\eta^2)^{c-1} \exp(-m\eta^2),$$

$$\boldsymbol{\beta} \sim \mathrm{N}_q(0, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}).$$

Gibbs Sampler

The full conditional distribution of $\tilde{v_i}$, $i = 1, ..., n$

$$\begin{aligned}
&\tilde{v_i} | \text{rest} \\
&\propto \frac{1}{\sqrt{2\pi\tau^{-1}\xi_2^2 \tilde{v_i}}} \exp\Big(-\frac{1}{2} \frac{(Y_i - \boldsymbol{E}_i^\top \boldsymbol{\beta} - \boldsymbol{Z}_i^\top \boldsymbol{\alpha} - \xi_1 \tilde{v_i})^2}{\xi_2^2 \tau^{-1} \tilde{v_i}}) \tau \exp(-\tau \tilde{v_i}\Big) \\
&\propto (\tilde{v_i})^{-\frac{1}{2}} \exp\Big(-\frac{1}{2} \frac{(Y_i - \boldsymbol{E}_i^\top \boldsymbol{\beta} - \boldsymbol{Z}_i^\top \boldsymbol{\alpha})}{\xi_2^2 \tau^{-1} \tilde{v_i}} - \frac{1}{2} \frac{\xi_1^2 \tilde{v_i}}{\tau^{-1} \xi_2^2} - \tau \tilde{v_i}\Big) \\
&\propto (\tilde{v_i})^{-\frac{1}{2}} \exp\Bigg(-\frac{1}{2}\Big((\frac{\tau \xi_1^2}{\xi_2^2} + 2\tau) \tilde{v_i} + \frac{\tau (Y_i - \boldsymbol{E}_i^\top \boldsymbol{\beta} - Z_i^\top \boldsymbol{\alpha})^2}{\xi_2^2} \frac{1}{\tilde{v_i}}\Big)\Bigg)
\end{aligned}$$

Therefore, the full conditional distribution of $\tilde{v_i}$ is generalized inverse Gaussian distribution.

The full conditional distribution of $s_j (j = 0, ..., p)$ is

$$
\begin{aligned}
& s_j|\text{rest} \\
& \propto \frac{1}{\sqrt{2\pi|s_j \mathbf{I}_d^{-1}|}} \exp\Big(-\frac{1}{2} \boldsymbol{\alpha}_j^\top (s_j \mathbf{I}_j^{-1})^{-1} \boldsymbol{\alpha}_j \Big) \times s_j^{\frac{d-1}{2}} \exp(-\frac{\eta^2}{2} s_j) \\
& \propto s_j^{-\frac{1}{2}} \exp\Big(-\frac{1}{2} (\eta^2 s_j + \boldsymbol{\alpha}_j^\top \boldsymbol{\alpha}_j \frac{1}{s_j}) \Big)
\end{aligned}
$$

Therefore, the posterior distribution of s_j is

$$
s_j^{-1}|\text{rest} \propto \text{invGaussian}(\sqrt{\frac{\eta^2}{\boldsymbol{\alpha}_j^\top \boldsymbol{\alpha}_j}}, \eta^2).
$$

The full conditional distribution of $\boldsymbol{\alpha}_j$, $j = 0, ..., p$

$\boldsymbol{\alpha}_j|\text{rest}$

$$
\begin{aligned}
& \propto \prod_{i=1}^{n} \exp\Big(-\frac{\tau}{2\xi_2^2 \tilde{v}_i} (Y_i - \boldsymbol{Z}_{i,-j}^\top \boldsymbol{\alpha}_{-j} - \boldsymbol{Z}_{ij}^\top \boldsymbol{\alpha}_j - \boldsymbol{E}_i^\top \boldsymbol{\beta} - \xi_1 \tilde{v}_i)^2 \Big) \\
& \times \frac{1}{\sqrt{2\pi|s_j \mathbf{I}_d^{-1}|}} \exp\Big(-\frac{1}{2} \boldsymbol{\alpha}_j^\top (s_j \mathbf{I}_j^{-1})^{-1} \boldsymbol{\alpha}_j \Big) \\
& \propto \exp\Big(-\frac{1}{2} \tau \xi_2^{-2} \sum_{i=1}^{n} \frac{1}{\tilde{v}_i} (Y_i - \boldsymbol{Z}_{i,-j}^\top \boldsymbol{\alpha}_{-j} - \boldsymbol{E}_i^\top \boldsymbol{\beta} - \xi_1 \tilde{v}_i)^2 \Big) \\
& \times \exp\Bigg(-\frac{1}{2} \Big(\alpha_j^\top (\tau \xi_2^{-2} \sum_{i=1}^{n} \frac{\boldsymbol{Z}_{ij} \boldsymbol{Z}_{ij}^\top}{\tilde{v}_i} + s_j^{-1} \mathbf{I}_d) \boldsymbol{\alpha}_j - 2\tau \xi_2^{-2} \sum_{i=1}^{n} \frac{1}{\tilde{v}_i} (Y_i - \boldsymbol{Z}_{i,-j}^\top \boldsymbol{\alpha}_{-j} - \boldsymbol{E}_i^\top \boldsymbol{\beta} - \xi_1 \tilde{v}_i) \boldsymbol{Z}_{ij}^\top \boldsymbol{\alpha}_j \Big) \Bigg)
\end{aligned}
$$

Denote the variance

$$
\boldsymbol{\Sigma}_j = (\tau \xi_2^{-2} \sum_{i=1}^{n} \frac{1}{\tilde{v}_i} \boldsymbol{Z}_{ij} \boldsymbol{Z}_{ij}^\top + s_j^{-1} \mathbf{I}_j)^{-1}
$$

and the mean

$$
\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j \tau \xi_2^{-2} \sum_{i=1}^{n} \frac{\boldsymbol{Z}_{ij}}{\tilde{v}_i} (Y_i - \boldsymbol{Z}_{i,-j}^\top \boldsymbol{\alpha}_{-j} - \boldsymbol{E}_i^\top \boldsymbol{\beta} - \xi_1 \tilde{v}_i),
$$

then the posterior distribution $\boldsymbol{\alpha}_j|\text{rest}$ is

$$\boldsymbol{\alpha}_j|\text{rest} \propto \mathrm{N}_d(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j).$$

The full conditional distribution of τ

$$\begin{aligned}
&\tau|\text{rest} \\
&\propto \prod_{i=1}^{n} \sqrt{\tau}\exp\Big(- \frac{\tau(Y_i - \boldsymbol{Z}_i^\top\boldsymbol{\alpha} - \boldsymbol{E}_i^\top\boldsymbol{\beta} - \xi_1\tilde{v}_i)^2}{\xi_2^2\tilde{v}_i}\Big) \times \prod_{i=1}^{n} \tau\exp(-\tau\tilde{v}_i) \times \tau^{a-1}\exp(-b\tau) \\
&\propto \tau^{\frac{3}{2}n+a-1}\exp\Bigg(- \Big(\frac{1}{2}\sum_{i=1}^{n} \frac{\tau(Y_i - \boldsymbol{Z}_i^\top\boldsymbol{\alpha} - \boldsymbol{E}_i^\top\boldsymbol{\beta} - \xi_1\tilde{v}_i)^2}{\xi_2^2\tilde{v}_i} + \sum_{i=1}^{n}\tilde{v}_i + b\Big)\tau\Bigg)
\end{aligned}$$

Therefore, the posterior distribution of τ is

$$\tau|\text{rest} \propto \text{Gamma}\Big(\frac{3}{2}n + a, \frac{1}{2}\sum_{i=1}^{n} \frac{\tau(Y_i - Z_i^\top\alpha - E_i^\top\beta - \xi_1\tilde{v}_i)^2}{\xi_2^2\tilde{v}_i} + \sum_{i=1}^{n}\tilde{v}_i + b\Big).$$

The full conditional distribution of η^2

$$\begin{aligned}
&\eta^2|\text{rest} \\
&\propto \prod_{j=0}^{p} (\frac{\eta^2}{2})^{\frac{d+1}{2}}\exp(-\frac{\eta^2}{2}s_j) \times (\eta^2)^{c-1}\exp(-m\eta^2) \\
&\propto (\eta^2)^{\frac{(d+1)(p+1)}{2}+c-1}\exp(-(\frac{1}{2}\sum_{j=0}^{p} s_j + m)\eta^2)
\end{aligned}$$

Therefore, the posterior distribution of η^2 is

$$\eta^2|\text{rest} \propto \text{Gamma}\Big(\frac{(d+1)(p+1)}{2} + c, \frac{1}{2}\sum_{j=0}^{p} s_j + m\Big).$$

The full conditional distribution of $\boldsymbol{\beta}$

$\boldsymbol{\beta}|\text{rest}$

$$\propto \prod_{i=1}^{n} \exp\Big(-\frac{\tau}{2\xi_2^2\tilde{v}_i}(Y_i - \boldsymbol{Z}_i^\top\boldsymbol{\alpha} - \boldsymbol{E}_i^\top\boldsymbol{\beta} - \xi_1\tilde{v}_i)^2\Big)\exp(-\frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta})$$

$$\propto \exp\Bigg(-\frac{1}{2}\Big(\boldsymbol{\beta}^\top(\sum_{i=1}^{n}\frac{\tau\boldsymbol{E}_i\boldsymbol{E}_i^\top}{\xi_2^2\tilde{v}_i} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})\boldsymbol{\beta} - 2\sum_{i=1}^{n}\frac{\tau}{\xi_2^2\tilde{v}_i}(Y_i - \boldsymbol{Z}_i^\top\boldsymbol{\alpha} - \xi_1\tilde{v}_i)\boldsymbol{E}_i^\top\boldsymbol{\beta}\Big)\Bigg)$$

$$\propto \mathrm{N}_q\Bigg((\sum_{i=1}^{n}\frac{\tau\boldsymbol{E}_i\boldsymbol{E}_i^\top}{\xi_2^2\tilde{v}_i} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1}\Big(\sum_{i=1}^{n}\frac{\tau}{\xi_2^2\tilde{v}_i}(Y_i - \boldsymbol{Z}_i^\top\boldsymbol{\alpha} - \xi_1\tilde{v}_i)\boldsymbol{E}_i^\top\Big)^\top, (\sum_{i=1}^{n}\frac{\tau\boldsymbol{E}_i\boldsymbol{E}_i^\top}{\xi_2^2\tilde{v}_i} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1})\Bigg)$$

which is a multivariate normal distribution.

C.4.3 Posterior inference for BVCSS

Priors

$$\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2,\tau_j^2 \sim \mathrm{N}_n(\boldsymbol{E}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), i = 1,...,n; j = 0,...,p,$$

$$\boldsymbol{\alpha}_j|\tau_j^2,\sigma^2 \sim (1-\pi_0)\mathrm{N}_d(0,\sigma^2\tau_j^2\mathbf{I}_d) + \pi_0\delta_0(\boldsymbol{\alpha}_j), j = 0,...,p,$$

$$\tau_j^2|\lambda^2 \sim \Gamma(\frac{d+1}{2}, \frac{\lambda^2}{2}), j = 0,...,p,$$

$$\pi_0 \sim \mathrm{Beta}(a,b),$$

$$\sigma^2 \sim \mathrm{invGamma}(s,h),$$

$$\lambda^2 \sim \Gamma(t,\theta),$$

$$\boldsymbol{\beta} \sim \mathrm{N}_q(0,\boldsymbol{\Sigma}_{\boldsymbol{\beta}}).$$

Gibbs Sampler

The full conditional distribution of $\boldsymbol{\alpha}_j$, $j = 0, ..., p$

$$\boldsymbol{\alpha}_j|\text{rest}$$

$$\propto \exp(-\frac{1}{2\sigma^2}||\boldsymbol{Y} - \boldsymbol{Z}_{-j}\boldsymbol{\alpha}_{-j} - \boldsymbol{Z}_j\alpha_j - \boldsymbol{E}\boldsymbol{\beta}||^2)$$

$$\times \left((1-\pi_0)\frac{1}{\sqrt{2\pi|\sigma^2\tau_j^2\mathbf{I}_d|}}\exp\Big(-\frac{1}{2}\boldsymbol{\alpha}_j^\top(\sigma^2\tau_j^2\mathbf{I}_d)^{-1}\boldsymbol{\alpha}_j\Big)\mathbf{I}_{(\boldsymbol{\alpha}_j\neq 0)} + \pi_0\delta_0(\boldsymbol{\alpha}_j)\right)$$

The slab part,

$\boldsymbol{\alpha}_j|\text{rest}$

$$\propto \exp\left(-\frac{1}{2\sigma^2}\Big(\boldsymbol{\alpha}_j^\top\boldsymbol{Z}_j^\top\boldsymbol{Z}_j\boldsymbol{\alpha}_j - 2\boldsymbol{\alpha}_j\boldsymbol{Z}_j^\top(\boldsymbol{Y} - \boldsymbol{E}\boldsymbol{\beta} - \boldsymbol{Z}_{-j}\boldsymbol{\alpha}_{-j})\Big)\right)\exp(-\frac{1}{2\sigma^2}||\boldsymbol{Y} - \boldsymbol{Z}_{-j}\boldsymbol{\alpha}_{-j} - \boldsymbol{E}\boldsymbol{\beta}||^2)$$

$$\times (1-\pi_0)(2\pi)^{-\frac{d}{2}}(\sigma^2\tau_j^2)^{-\frac{d}{2}}\exp\Big(-\frac{1}{2}\boldsymbol{\alpha}_j^\top(\sigma^2\tau_j^2\mathbf{I}_d)^{-1}\boldsymbol{\alpha}_j\Big)$$

$$\propto (1-\pi_0)(2\pi)^{-\frac{d}{2}}(\sigma^2\tau_j^2)^{-\frac{d}{2}}\exp(-\frac{1}{2\sigma^2}||\boldsymbol{Y} - \boldsymbol{Z}_{-j}\boldsymbol{\alpha}_{-j} - \boldsymbol{E}\boldsymbol{\beta}||^2)$$

$$\times \exp\left(-\frac{1}{2\sigma^2}\Big(\boldsymbol{\alpha}_j^\top(\boldsymbol{Z}_j^\top\boldsymbol{Z}_j + \tau_j^{-2}\mathbf{I}_d)\boldsymbol{\alpha}_j - 2\boldsymbol{\alpha}_j\boldsymbol{Z}_j^\top(\boldsymbol{Y} - \boldsymbol{E}\boldsymbol{\beta} - \boldsymbol{Z}_{-j}\boldsymbol{\alpha}_{-j})\Big)\right)$$

Denote the variance

$$\boldsymbol{\Sigma}_j = (\boldsymbol{Z}_j^\top\boldsymbol{Z}_j + \tau_j^{-2}\mathbf{I}_d)^{-1}$$

and the mean

$$\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j\boldsymbol{Z}_j^\top(\boldsymbol{Y} - \boldsymbol{E}\boldsymbol{\beta} - \boldsymbol{Z}_{-j}\boldsymbol{\alpha}_{-j}),$$

then when $\boldsymbol{\alpha}_j \neq 0$, the posterior distribution of $\boldsymbol{\alpha}_j$ becomes

$$\begin{aligned}
&\boldsymbol{\alpha}_j|\text{rest} \\
&\propto (1-\pi_0)(\tau_j^2)^{-\frac{d}{2}}\sqrt{|\boldsymbol{\Sigma}_j|}\exp(-\frac{1}{2\sigma^2}||\boldsymbol{Y}-\boldsymbol{Z}_{-j}\boldsymbol{\alpha}_{-j}-\boldsymbol{E}\boldsymbol{\beta}||^2) \\
&\times \exp\Big(-\frac{1}{2}\boldsymbol{\mu}_j^\top(\sigma^2\boldsymbol{\Sigma}_j)^{-1}\boldsymbol{\mu}_j\Big)\times \mathrm{N}_d(\boldsymbol{\mu}_j,\sigma^2\boldsymbol{\Sigma}_j)
\end{aligned}$$

The spike part

$$\begin{aligned}
&\boldsymbol{\alpha}_j|\text{rest} \\
&\propto \pi_0\exp(-\frac{1}{2\sigma^2}||\boldsymbol{Y}-\boldsymbol{Z}_{-j}\boldsymbol{\alpha}_{-j}-\boldsymbol{E}\boldsymbol{\beta}||^2)
\end{aligned}$$

Proportion of the spike part

$$P(\boldsymbol{\alpha}_j=0|\text{rest})=\frac{\pi_0}{\pi_0+(1-\pi_0)(\tau_j^2)^{-\frac{d}{2}}\sqrt{|\boldsymbol{\Sigma}_j|}\exp\Big(-\frac{1}{2}\boldsymbol{\mu}_j^\top(\sigma^2\boldsymbol{\Sigma}_j)^{-1}\boldsymbol{\mu}_j\Big)}$$

The full conditional distribution of σ^2

$$\begin{aligned}
&\sigma^2|\text{rest} \\
&\propto (\sigma^2)^{-\frac{n}{2}}\exp(-\frac{1}{2\sigma^2}||\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\alpha}-\boldsymbol{E}\boldsymbol{\beta}||^2)\times(\frac{1}{\sigma^2})^{s+1}\exp(-\frac{h}{\sigma^2}) \\
&\times\prod_{j=0}^{p}\left((1-\pi_0)\frac{1}{\sqrt{2\pi|\sigma^2\tau_j^2\mathbf{I}_d|}}\exp\Big(-\frac{1}{2}\alpha_j^\top(\sigma^2\tau_j^2\mathbf{I}_d)^{-1}\boldsymbol{\alpha}_j\Big)\mathbf{I}_{(\boldsymbol{\alpha}_j\neq 0)}+\pi_0\delta_0(\boldsymbol{\alpha}_j)\right)
\end{aligned}$$

Let

$$Q_j=\begin{cases}0 & \text{if } \boldsymbol{\alpha}_j=0\\ 1 & \text{if } \boldsymbol{\alpha}_j\neq 0\end{cases}$$

then the posterior distribution of σ^2 becomes

$$
\begin{aligned}
&\sigma^2|\text{rest} \\
&\propto (\sigma^2)^{-\frac{n}{2}}\exp(-\frac{1}{2\sigma^2}||\boldsymbol{Y}-\boldsymbol{Z\alpha}-\boldsymbol{E\beta}||^2)\times(\frac{1}{\sigma^2})^{s+1}\exp(-\frac{h}{\sigma^2}) \\
&\times\prod_{j=0}^{p}(1-\pi_0)^{Q_j}(\sigma^2)^{-\frac{d}{2}\sum_{j=0}^{p}Q_j}\prod_{j=0}^{p}\pi_0^{1-Q_j}\exp\Big(-\frac{1}{\sigma^2}\cdot\frac{1}{2}\sum_{j=0}^{p}(\tau_j^2)^{-1}\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j\Big) \\
&\propto (\sigma^2)^{-\frac{n}{2}-\frac{d}{2}\sum_{j=0}^{p}Q_j-s-1}\exp\Big(-\frac{1}{\sigma^2}(\frac{1}{2}||\boldsymbol{Y}-\boldsymbol{Z\alpha}-\boldsymbol{E\beta}||^2+\frac{1}{2}\sum_{j=0}^{p}(\tau_j^2)^{-1}\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j+h)\Big)
\end{aligned}
$$

Therefore, the posterior distribution of σ^2 is

$$
\sigma^2|\text{rest}\propto\text{invGamma}\Big(\frac{n}{2}+\frac{d}{2}\sum_{j=0}^{p}Q_j+s,\frac{1}{2}||\boldsymbol{Y}-\boldsymbol{Z\alpha}-\boldsymbol{E\beta}||^2+\frac{1}{2}\sum_{j=0}^{p}(\tau_j^2)^{-1}\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j+h\Big).
$$

The full conditional distribution of τ_j^2, $j=0,...,p$

$$
\begin{aligned}
&\tau_j^2|\text{rest} \\
&\propto\prod_{j=0}^{p}\left((1-\pi_0)\frac{1}{\sqrt{2\pi|\sigma^2\tau_j^2\mathbf{I}_d|}}\exp\Big(-\frac{1}{2}\boldsymbol{\alpha}_j^\top(\sigma^2\tau_j^2\mathbf{I}_d)^{-1}\boldsymbol{\alpha}_j\Big)\mathbf{I}_{(\boldsymbol{\alpha}_j\neq 0)}+\pi_0\delta_0(\boldsymbol{\alpha}_j)\right) \\
&\times(\tau_j^2)^{\frac{d-1}{2}}\exp(-\frac{\lambda^2}{2}\tau_j^2)
\end{aligned}
$$

The slab part

$$
\begin{aligned}
&\tau_j^2|\text{rest} \\
&\propto(\tau_j^2)^{-\frac{1}{2}}\exp(-\frac{1}{2}(\frac{\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j}{\sigma^2}\frac{1}{\tau_j^2}+\lambda^2\tau_j^2))
\end{aligned}
$$

therefore $(\tau_j^2)^{-1}\propto\text{invGaussian}(\sqrt{\frac{\sigma^2\lambda^2}{\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j}},\lambda^2)$

The spike part

$$\begin{aligned}
&\tau_j^2|\text{rest} \\
&\propto (\tau_j^2)^{\frac{d-1}{2}}\exp(-\frac{\lambda^2}{2}\tau_j^2) \\
&\propto \Gamma(\frac{d+1}{2}, \frac{\lambda^2}{2})
\end{aligned}$$

Together

$$(\tau_j^2)^{-1}|\text{rest} \sim \begin{cases} \text{Inverse-Gamma}(\frac{d+1}{2},\ \frac{\lambda^2}{2}) & \text{if } \boldsymbol{\alpha}_j = 0 \\ \text{Inverse-Gaussian}(\sqrt{\frac{\sigma^2\lambda^2}{\boldsymbol{\alpha}_j^\top \boldsymbol{\alpha}_j}}, \lambda^2) & \text{if } \boldsymbol{\alpha}_j \neq 0 \end{cases}$$

The full conditional distribution of λ^2

$$\begin{aligned}
&\lambda^2|\text{rest} \\
&\propto \prod_{j=0}^{p}\left((\frac{\lambda^2}{2})^{\frac{d+1}{2}}\exp(-\frac{\lambda^2}{2}\tau_j^2)\right) \times (\lambda^2)^{t-1}\exp(-\theta\lambda^2) \\
&\propto (\lambda^2)^{\frac{1}{2}(d+1)(p+1)+t-1}\exp(-(\frac{1}{2}\sum_{j=0}^{p}\tau_j^2 + \theta)\lambda^2) \\
&\propto \Gamma(\frac{1}{2}(d+1)(p+1)+t, \frac{1}{2}\sum_{j=0}^{p}\tau_j^2 + \theta)
\end{aligned}$$

which is a gamma distribution.

The full conditional distribution of π_0

$$
\begin{aligned}
&\pi_0|\text{rest} \\
&\propto \prod_{j=0}^{p}\left((1-\pi_0)\frac{1}{\sqrt{2\pi|\sigma^2\tau_j^2\mathbf{I}_d|}}\exp\Big(-\frac{1}{2}\boldsymbol{\alpha}_j^\top(\sigma^2\tau_j^2\mathbf{I}_d)^{-1}\boldsymbol{\alpha}_j\Big)\mathbf{I}_{(\boldsymbol{\alpha}_j\neq 0)}+\pi_0\delta_0(\boldsymbol{\alpha}_j)\right) \\
&\times \pi_0^{a-1}(1-\pi_0)^{b-1} \\
&\propto \pi_0^{1+p+a-\sum_{j=0}^{p}Q_j-1}(1-\pi_0)^{b+\sum_{j=0}^{p}Q_j-1} \\
&\propto \text{Beta}(1+p+a-\sum_{j=0}^{p}Q_j, b+\sum_{j=0}^{p}Q_j)
\end{aligned}
$$

which is a Beta distribution.

The full conditional distribution of $\boldsymbol{\beta}$

$$
\begin{aligned}
&\boldsymbol{\beta}|\text{rest} \\
&\propto \exp(-\frac{1}{2\sigma^2}||\boldsymbol{Y}-\boldsymbol{E}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{\alpha}||^2)\times\exp(-\frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}) \\
&\propto \exp\left(-\frac{1}{2}\Big(\boldsymbol{\beta}^\top(\frac{\boldsymbol{E}^\top\boldsymbol{E}}{\sigma^2}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})\boldsymbol{\beta}-2\cdot\frac{1}{\sigma^2}(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\alpha})^\top\boldsymbol{E}\boldsymbol{\beta}\Big)\right) \\
&\propto \text{N}_q((\frac{\boldsymbol{E}^\top\boldsymbol{E}}{\sigma^2}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1}(\frac{1}{\sigma^2}(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\alpha})^\top\boldsymbol{E})^\top, (\frac{\boldsymbol{E}^\top\boldsymbol{E}}{\sigma^2}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1})
\end{aligned}
$$

which is a multivariate normal distribution.

C.4.4 Posterior inference for BVC

Priors

$$\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2,\tau_j^2 \sim \text{N}_n(\boldsymbol{E}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{\beta},\sigma^2\mathbf{I}_n), i=1,...,n; j=0,...,p,$$

$$\boldsymbol{\alpha}_j|\tau_j^2,\sigma^2 \sim \text{N}_d(0,\sigma^2\tau_j^2\mathbf{I}_d), j=0,...,p,$$

$$\tau_j^2|\lambda^2 \sim \Gamma(\frac{d+1}{2}, \frac{\lambda^2}{2}), j = 0, ..., p,$$

$$\sigma^2 \sim \text{invGamma}(s, h),$$

$$\lambda^2 \sim \Gamma(t, \theta),$$

$$\beta \sim \text{N}_q(0, \Sigma_\beta).$$

Gibbs Sampler

The full conditional distribution of $\boldsymbol{\alpha}_j$, $j = 0, ..., p$

$\boldsymbol{\alpha}_j|\text{rest}$

$$\propto \exp(-\frac{1}{2\sigma^2}||\boldsymbol{Y} - \boldsymbol{Z\alpha} - \boldsymbol{E\beta}||^2)\exp\Big(-\frac{1}{2}\boldsymbol{\alpha}_j^\top(\sigma^2\tau_j^2\mathbf{I}_d)^{-1}\boldsymbol{\alpha}_j\Big)$$

$$\propto \exp\bigg(-\frac{1}{2\sigma^2}\Big(\boldsymbol{\alpha}_j^\top\boldsymbol{Z}_j^\top\boldsymbol{Z}_j\boldsymbol{\alpha}_j - 2\boldsymbol{\alpha}_j\boldsymbol{Z}_j^\top(\boldsymbol{Y} - \boldsymbol{E\beta} - \boldsymbol{Z}_{-j}\boldsymbol{\alpha}_{-j})\Big)\bigg)\exp\Big(-\frac{1}{2}\boldsymbol{\alpha}_j^\top(\sigma^2\tau_j^2\mathbf{I}_d)^{-1}\boldsymbol{\alpha}_j\Big)$$

$$\propto \exp\bigg(-\frac{1}{2\sigma^2}\Big(\boldsymbol{\alpha}_j^\top(\boldsymbol{Z}_j^\top\boldsymbol{Z}_j + \tau_j^{-2}\mathbf{I}_d)\boldsymbol{\alpha}_j - 2\boldsymbol{\alpha}_j\boldsymbol{Z}_j^\top(\boldsymbol{Y} - \boldsymbol{E\beta} - \boldsymbol{Z}_{-j}\boldsymbol{\alpha}_{-j})\Big)\bigg)$$

Denote the variance

$$\boldsymbol{\Sigma}_j = (\boldsymbol{Z}_j^\top\boldsymbol{Z}_j + \tau_j^{-2}\mathbf{I}_d)^{-1}$$

and the mean

$$\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j\boldsymbol{Z}_j^\top(\boldsymbol{Y} - \boldsymbol{E\beta} - \boldsymbol{Z}_{-j}\boldsymbol{\alpha}_{-j}),$$

then the posterior distribution of $\boldsymbol{\alpha}_j$ is

$$\boldsymbol{\alpha}_j|\text{rest} \propto \text{N}_d(\boldsymbol{\mu}_j, \sigma^2\boldsymbol{\Sigma}_j).$$

The full conditional distribution of $\boldsymbol{\beta}$

$$
\begin{aligned}
&\boldsymbol{\beta}|\text{rest} \\
&\propto \exp(-\frac{1}{2\sigma^2}||\boldsymbol{Y}-\boldsymbol{E}\beta-\boldsymbol{Z}\boldsymbol{\alpha}||^2)\times\exp(-\frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}) \\
&\propto \exp\bigg(-\frac{1}{2}\Big(\beta^\top(\frac{\boldsymbol{E}^\top\boldsymbol{E}}{\sigma^2}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})\beta-2\cdot\frac{1}{\sigma^2}(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\alpha})^\top\boldsymbol{E}\boldsymbol{\beta}\Big)\bigg) \\
&\propto \mathrm{N}_q((\frac{\boldsymbol{E}^\top\boldsymbol{E}}{\sigma^2}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1}(\frac{1}{\sigma^2}(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\alpha})^\top\boldsymbol{E})^\top,(\frac{\boldsymbol{E}^\top\boldsymbol{E}}{\sigma^2}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1})
\end{aligned}
$$

which is a multivariate normal distribution.

The full conditional distribution of τ_j^2, $j=0,...,p$

$$
\begin{aligned}
&\tau_j^2|\text{rest} \\
&\propto \frac{1}{\sqrt{2\pi|\sigma^2\tau_j^2\mathbf{I}_d|}}\exp\Big(-\frac{1}{2}\boldsymbol{\alpha}_j^\top(\sigma^2\tau_j^2\mathbf{I}_d)^{-1}\boldsymbol{\alpha}_j\Big) \\
&\propto (\tau_j^2)^{-\frac{1}{2}}\exp(-\frac{1}{2}(\frac{\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j}{\sigma^2}\frac{1}{\tau_j^2}+\lambda^2\tau_j^2))
\end{aligned}
$$

therefore $(\tau_j^2)^{-1}\propto \text{invGaussian}(\sqrt{\frac{\sigma^2\lambda^2}{\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j}},\lambda^2)$

The full conditional distribution of λ^2

$$
\begin{aligned}
&\lambda^2|\text{rest} \\
&\propto \prod_{j=0}^{p}\Big((\frac{\lambda^2}{2})^{\frac{d+1}{2}}\exp(-\frac{\lambda^2}{2}\tau_j^2)\Big)\times(\lambda^2)^{t-1}\exp(-\theta\lambda^2) \\
&\propto (\lambda^2)^{\frac{1}{2}(d+1)(p+1)+t-1}\exp(-(\frac{1}{2}\sum_{j=0}^{p}\tau_j^2+\theta)\lambda^2) \\
&\propto \Gamma(\frac{1}{2}(d+1)(p+1)+t,\frac{1}{2}\sum_{j=0}^{p}\tau_j^2+\theta)
\end{aligned}
$$

which is a gamma distribution.

The full conditional distribution of σ^2

$\sigma^2|\text{rest}$

$$\propto (\sigma^2)^{-\frac{n}{2}}\exp(-\frac{1}{2\sigma^2}||\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\alpha}-\boldsymbol{E}\boldsymbol{\beta}||^2)\times(\frac{1}{\sigma^2})^{s+1}\exp(-\frac{h}{\sigma^2})$$

$$\times\prod_{j=0}^{p}\frac{1}{\sqrt{2\pi|\sigma^2\tau_j^2\mathbf{I}_d|}}\exp\Big(-\frac{1}{2}\boldsymbol{\alpha}_j^\top(\sigma^2\tau_j^2\mathbf{I}_d)^{-1}\boldsymbol{\alpha}_j\Big)$$

$$\propto (\sigma^2)^{-\frac{n}{2}-\frac{d(p+1)}{2}-s-1}\exp\Big(-\frac{1}{\sigma^2}(\frac{1}{2}||\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\alpha}-\boldsymbol{E}\boldsymbol{\beta}||^2+\frac{1}{2}\sum_{j=0}^{p}(\tau_j^2)^{-1}\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j+h)\Big)$$

Therefore, the posterior distribution of σ^2 is invGamma$\Big(\frac{n}{2}+\frac{d(p+1)}{2}+s, \frac{1}{2}||\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\alpha}-\boldsymbol{E}\boldsymbol{\beta}||^2+\frac{1}{2}\sum_{j=0}^{p}(\tau_j^2)^{-1}\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j+h\Big)$.