VARIABLE SELECTION IN VARYING MULTI-INDEX COEFFICIENT MODELS WITH APPLICATIONS TO GENE-ENVIRONMENTAL INTERACTIONS

By

Shunjie Guan

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics—Doctor of Philosophy

2017

ABSTRACT

VARIABLE SELECTION IN VARYING MULTI-INDEX COEFFICIENT MODELS WITH APPLICATIONS TO GENE-ENVIRONMENTAL INTERACTIONS

By

Shunjie Guan

Variable selection is an important topic in modern statistics literature. And varying multi-index coefficient model(VMICM) is a promising tool to study the synergistic interaction effects between genes and multiple environmental exposures. In this dissertation, we proposed a variable selection approach for VMICM, we also generalized such approach to generalized and quantile regression settings. Their theoretical properties, simulation performance and application in genetic research were studied.

Complicated diseases have both environmental and genetic risk factors, and large amount of research have been devoted to identify gene-environment (G×E) interaction. Defined as different effect of a genotype on disease risk in persons with different environmental exposures (Ottman (1996)), we can view environmental exposures as the modulating factors in the effect of a gene. Based on this idea, we derived a three stage variable selection approach to estimate different effects of gene variables: varying, constant and zero which respectively correspond to nonlinear G×E effect, no G×E effect and no genetic effect. For multiple environmental exposure variables, we also select and estimate important environmental variables that contribute to the synergistic interaction effect. We theoretically evaluated the oracle property of the three step estimation method. We conducted simulation studies to further evaluate the finite sample performance of the method, considering both continuous and discrete predictors. Application to a real data set demonstrated the utility of the method.

In Chapter 3, we generalized such variable selection approach to binary response

setting. Instead of minimizing penalized squared error loss, we chose to maximize penalized log-likelihood function. We also theoretically evaluated the oracle property of the proposed selection approach in binary response setting. We demonstrated the performance of the model via simulation. At last, we applied our model to a Type II diabetes data set.

Compared to conditional mean regression, conditional quantile regression could provide a more comprehensive understanding of the distribution of the response variable at different quantile. Even if the center of distribution is our only interest, median regression (special case of quantile regression) could offer a more robust estimator. Hence, we extended our three stage variable selection approach to a quantile regression setting in Chapter 4. We demonstrated the finite sample performance of the model via extensive simulation. And we applied our model to a birth weight data set.

I delicate this dissertation to my parents, Bishan Lin, Ruixiong Guan and my girlfriend Lingjie Zhou.

ACKNOWLEDGMENTS

Here, I would like to offer my sincere gratitude to my advisor Dr. Yuehua Cui for his patient, support and guidance during my study and research. Dr. Cui is kind, knowledgable and humble, whenever I am stuck with a problem, he could always offer very useful idea and insights. In the genetic journal club hosted by Dr. Cui, I learned a lot from other's presentation and eventually, I was able to give a talk on my own. I really own my completion of this dissertation to Dr. Cui's guidance.

I also have to thank the members of my PhD committee, Dr. Pingshou Zhong, Dr. Hyokyoung G. Hong, and Dr Qing Lu. Their comments and suggestions are very helpful in my graduate study.

I appreciate the help I got from all the professors in the Department of Statistics and Probability. Especially, Dr. Taps Maiti, I took several of his classes, unlike other courses that focus on theory and proof, his course in linear modelling focused on interpretation and how to explain statistical concepts to non-statisticians. It is only when I began to look for a job in the past months that I realize its importance. I really own my success in finding a job after graduation to his insistence in interpreting statistics results.

My thanks also goes to other senior students in Dr. Cui's group, including Dr. Bin Gao, Dr. Honglang Wang, Dr. Tao He, Jingyi Zhang and post-doctoral researcher Xu Liu. In just a few years, many of them become successful researchers in a university or statisticians in big companies.

Last but not least, I thank my parents and my girlfriend for their support and confidence in me. They are the beacon for my journey, offered me hope and a sense of direction.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Overview

In this dissertation, we studied how genetic and environmental factors interact to affect a disease outcome by developing novel statistical methods. Ever since Gregor Mendel's famous experiments with his pea plants in the nineteenth century, researchers have been fascinated with the role of genetics played in our lives. With decades of genetics research, we knew more than ever the effect of various genes. For example, we knew mutations in the CFTR gene cause cystic fibrosis, mutations in PAH gene cause phenylketonuria. In fact, scientists have identified more than $10,000$ human disorders that are caused by mutations in single genes. However, complex diseases such as type II diabetes have various risk factors: environmental risk factors such as exercise level, body mass index, genetic risk factors such as mutations in gene TCF7L2 and ABCC8. Hence, it is of great interest to study how gene and environment interact and affect various disease traits. In this chapter, we first provided some background information and a brief review of traditional statistical models used to study gene-environment(G×E) interaction in section 1.2. To address the constrain of traditional G×E interaction models, we discussed non-parametric models in section 1.3. In section 1.4, we discussed the recent advance in variable selection via penalized regression and how we can apply variable selection in our model to select significant risk factors. We

offered a brief review of quantile regression and its benefit in section 1.5. At last, the goal and organization of this dissertation is offered in section 1.6

1.2 Gene-Environment Interaction

In recent years, more and more research suggested that gene-environment (G×E) interaction plays an important role in complex traits such as type II diabetes and birth weight. G×E interaction was defined by Ottman (1996) as "a different effect of a genotype on disease risk in persons with different environmental exposures". Traditionally, G×E interaction was investigated via linear model:

$$\boldsymbol{Y} = \beta_0 + \boldsymbol{\beta}_G * \boldsymbol{G} + \boldsymbol{\beta}_E * \boldsymbol{E}_k + \boldsymbol{\beta}_{G\times E} * \boldsymbol{G} * \boldsymbol{E}_k + \boldsymbol{\epsilon} \tag{1.1}$$

where $\boldsymbol{G}$ represents genetic factors; $\boldsymbol{E}_k$ represents an environmental factor; $\boldsymbol{\beta}_G$, $\boldsymbol{\beta}_E$ represents the genetic effects and environmental effect respectively; and $\boldsymbol{\beta}_{G\times E}$ represents the G×E effect between genetic factors $\boldsymbol{G}$ and environmental factor $\boldsymbol{E}_k$. However, such model has several drawbacks. Firstly, it assumes the G×E interaction is linear, which is often violated in many cases. Secondly, model (1.1) is only computational feasible with a single environmental factor. With the introduction of multiple environmental factors, model dimension will increase dramatically, resulting in biased and unstable estimation. Nevertheless, many epidemiological studies revealed that disease risks can be modified by simultaneously exposure to several environmental factors (Carpenter et al. (2002); Sexton and Hattis (2007)). These limitations led to the implementation of several non-parametric models in the G×E interaction studies.

1.3 Non-Parametric Models

In parametric modelling, such as linear model (1.1) or generalized linear model, we assumed we knew the model structure in advance and we estimated the model parameters based on the assumed structure and the data set. However, more often than we would prefer, such assumptions cannot be justified. Which was particularly problematic since all parameter estimation and model interpretation were made based on those assumptions. This predicament led us to consider non-parametric modelling. Non-parametric models make little to no assumptions, it lets the data decide the functional relationships between the response variable and the predictors. Due to its flexibility, non-parametric models could be applied to most data set. For example, to address the limitation of linearity in model (1.1), Ma et al. (2011) proposed to use varying coefficient(VC) model to detect non-linear gene-environmental interaction. Proposed by Hastie and Tibshirani (1993), a varying coefficient model could be of the form

$$\boldsymbol{Y} = \beta_0(X) + \sum_{k=1}^{p} \beta_k(X)\boldsymbol{G}_k + \boldsymbol{\epsilon} \tag{1.2}$$

where $\boldsymbol{Y}$ is the response; $\boldsymbol{G}_k, k = 1, \ldots, p$ is the kth genetic factor; and X is a single environmental factor and $\boldsymbol{\epsilon}$ is the random error. Representing the gene effect of $\boldsymbol{G}_k$, $\beta_k(X)$ is a smooth non-linear non-parametric functions indexed by X. Under such structure, $\beta_k(X)$ is allowed to vary as a function of the environmental factor X. It could either be a linear or a non-linear function. This is the reason why model (1.2) could be used to detect non-linear gene-environmental interaction between several genetic variants and a single environmental variant.

While VC model (1.2) alleviates the linearity constrain of model (1.1), it can only be

used to model G×E interaction with a single environmental factor. Similar to linear models, it is computational infeasible to model how a mixture of environmental factors interact with genetic variants due to dramatically increasing model dimension. To address this issue, Liu et al. (2016) proposed we could implement varying multi-index coefficient model (VMICM)

$$\boldsymbol{Y} = m_0(\boldsymbol{X}\boldsymbol{\beta}) + \sum_{k=1}^{p} m_k(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{G}_k + \boldsymbol{\epsilon} \tag{1.3}$$

where $\boldsymbol{Y}$ is the continuous response variable; $\boldsymbol{G} = (\boldsymbol{G}_1, \ldots, \boldsymbol{G}_p)$ is a p dimensional matrix representing genetic factors; $\boldsymbol{X}$ is an environmental factor matrix of dimension q; $\boldsymbol{\beta}$ is the loading parameter for environment covariates $\boldsymbol{X}$; $m_k(u)$ is a smooth non-linear non-parametric function indexed by $\boldsymbol{X}\boldsymbol{\beta}$, and it represents the gene effect of $\boldsymbol{G}_k$. One of the main advantage of model (1.3) is it considers the interaction between genetic variants $\boldsymbol{G}$ and a mixture of environmental variants $\boldsymbol{X}$ without increasing model dimension dramatically. We could interpret $m_k(\boldsymbol{X}\boldsymbol{\beta})$ as the gene effect of $\boldsymbol{G}_k$ modulated by its index $\boldsymbol{X}\boldsymbol{\beta}$. And β_d could be interpreted as the strength of interaction between the $d-th$ environmental factors X_d and $\boldsymbol{G}$. We could also observe VC model (1.2) is a special case of VMICM (1.3) when $q = 1$. Due to its unique ability to accommodate non-linear G×E interaction with a mixture of environmental variants, we decide to implement VMICM in this dissertation.

Estimation for non-parametric model such as (1.2) and (1.3) could be roughly grouped into three categories: kernel smoothing (Fan and Zhang (1999); Xia and Li (1999) ; Cai et al. (2000)), spline-based methods (Huant et al. (2004); Hoover et al. (1998); Chiang et al. (2001)) and wavelet estimation (Zhou and You (2004)). In this dissertation, we adopted the idea of B-spline approximation to estimate non-parametric function $m_k(u)$ for several reasons. Firstly, by some transformation of the B-spline basis function, we would be able to

separate the constant effect of $\boldsymbol{G}_k$ from its varying effect. This would be discussed in detail in the following chapters. Secondly, the computation algorithm of B-spline approximation is more efficient compared to kernel based methods. It is essential since we are working with high dimension genetics data. Further, it's easier to implement variable selection via penalized regression in a B-spline approximation setting.

1.4 Variable Selection

To select significant gene environmental combo from a large number of variants, we need to implement some dimension reduction technique, via either variable selection or hypothesis testing. In this dissertation, we focused on variable selection for its efficient algorithm and unique feature of simultaneous model selection and estimation. Traditional model selection techniques include backward/forward selection or information criterion based technique such as AIC or BIC. However, with the rise of big data, such methods are no longer feasible for several reasons: exponentially increasing computation time and unstable estimation due to increasing collinearity. Recently, variable selection via penalized regression has been gaining popularity. Its idea is to add a penalty term to the loss or lieklihood function. It could be of the form

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \left(||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + n \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right) \tag{1.4}$$

or

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} \left(l(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) - n \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right) \tag{1.5}$$

where $\boldsymbol{Y}$ is the dependent variable and $\boldsymbol{X}$ is the predictor; $||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$ is the squared error loss function; $l(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y})$ is the log-likelihood function; and $p_\lambda(|\beta_j|)$ is the penalty function for the j-th coordinate of $\boldsymbol{\beta}$. By adding an appropriate penalty term to the optimization function, some covariates of the penalized estimator $\hat{\boldsymbol{\beta}}$ could be shrunk to 0, therefore, achieving simultaneous model selection and estimation.

With different choices of penalty function $p_\lambda(\cdot)$, the penalized estimator of $\boldsymbol{\beta}$ could possess different properties. Fan and Li (2001) advocated three properties that a penalized estimator should possess:

(1) **Sparsity**: The penalized estimator should automatically set small coefficients to zero, therefore achieving model selection.

(2) **Approximately Unbiasedness**: The penalized estimator should be approximated unbiased, especially when the true coefficient is large.

(3) **Continuity**: The penalized estimator should be continuous in the data, therefore, reducing instability in model prediction.

To possess all three properties under squared error loss setting, Antoniadis and Fan (2001) deduced that the penalty function $p_\lambda(t)$ should satisfy: (1) $\min_{t\geq 0}\{t + p'_\lambda(t)\} > 0$; (2) $p'_\lambda(t) = 0$ for large t; (3) $arg\min_{t\geq 0}\{t + p'_\lambda(t)\} = 0$. Here, we present several popular choice of penalty functions:

(1) **Ridge Regression**: $p_\lambda(\beta_j) = \lambda|\beta_j|^2$ (Hoerl and Kennard (1970));

(2) **LASSO**: $p_\lambda(\beta_j) = \lambda|\beta_j|$ (Tibshirani (1996));

(3) **Adaptive Lasso**: $p_\lambda(\beta_j) = w_j\lambda|\beta_j|$ (Zou (2006));

(4) **SCAD**: $p'_\lambda(\beta_j) = \lambda\{I(\beta_j \leq \lambda) + \frac{(a\lambda-\beta_j)_+}{(a-1)\lambda} I(\beta_j > \lambda)\}$ for some $a > 2$ (Fan and Li (2001));

(5) **MCP**: $p_\lambda(\beta_j) = \lambda \int_0^{\beta_j} (1 - \frac{s}{\tau\lambda})_+ ds$ for some $\tau > 0$ (Zhang (2010)).

Among those penalty functions, adaptive LASSO, SCAD and MCP all possess the sparsity, approximately unbiasedness and continuity property. In this dissertation, we focused on MCP as our penalty function.

To solve the optimization problem (1.4) or (1.5), there are several algorithms available. Least-angle regression (LARS) (Efron et al. (2004)) could be used to calculate the entire solution path of the LASSO problem very efficiently. Fan and Li (2001) proposed local quadratic approximation (LQA) algorithm to solve the non-concave penalized likelihood problem. It can be implemented with a number of penalty functions. Their idea was to locally approximate the optimization function with a quadratic function. Therefore, transforming problem (1.4) or (1.5) to a least square problem with a closed form solution. Even though its efficient was surpassed by recent algorithms, the idea of LQA is still of great importance. Building on the idea of LQA, Zou and Li (2008) proposed local linear approximation (LLA) algorithm. Their idea was to locally approximate the non-concave penalty function linearly. Transforming the non-concave penalty to a LASSO penalty, which could be solved using LARS. Coordinate descent algorithm (Friedman et al. (2007), Friedman et al. (2010)) is an even more efficient algorithm. With small modification, it could be implemented to a wide range of penalized optimization problems. In this dissertation, we adopted the idea of coordinate descent and LQA to solve our optimization problem.

Besides the usual penalized regression that select individual parameters, there is a group analog, call grouped penalized regression (Yuan and Lin (2006)). Instead of penalizing

individual parameter, it penalizes the L_2 norm of a group of parameters. Assume $\boldsymbol{\beta}$ is divided into q groups $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)$, and the objective function could be of the form

$$\hat{\boldsymbol{\beta}} = arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = arg\min_{\boldsymbol{\beta}} \left(||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + n\sum_{j=1}^{q} p_\lambda(||\boldsymbol{\beta}_j||_2) \right). \tag{1.6}$$

The end result is that we select non-zero parameters as a group, either a group of parameters being all zero or none of them being zero. The grouped penalized regression technique is particularly useful in this dissertation.

1.5 Quantile Regression

Another objective of this dissertation was to extend the proposed variable selection for VMICM to quantile regression setting. Quantile regression is a very important alternative to the conventional conditional mean regression. It differs from mean regression in the loss function. Instead of trying to minimize the squared error loss for linear model, quantile regression tries to minimize the quantile loss function

$$\sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{X}_i\boldsymbol{\beta}) \text{ and } \rho_\tau(u) = u\{\tau - I(u < 0)\}. \tag{1.7}$$

Quantile regression also possess several advantages. First, modelling a dataset at several different quantiles offers a far more comprehensive view of the distribution of the response variable. In many scenarios, the effect of gene $\boldsymbol{G}_k$ varies at different quantile of the distribution. Further, even when we are only interested in the center of the distribution, median regression (quantile regression with $\tau = 0.5$) can provides more robust estimator, therefore, insensitive to outliers.

1.6 Objective and Organization

To select non-linear gene-environment interaction, we proposed a three stages iterative variable selection approach for Varying Multi-Index Coefficient Model and its generalization. We also extended such approach to a quantile regression setting. One of our goal is to classify genetic variants into three categories: varying, constant and zero. Varying effect gene is the gene that interact with environmental factors. Its effect on the response varies as environmental factors changes. Constant effect gene is the gene that only has a constant effect, not being modulated by environmental factors. Zero effect gene does not have an effect at all. By selecting non-zero loading parameters $\boldsymbol{\beta}$ in model (1.3), another goal of the proposed approach is to select environmental variants that interact with genetic variants.

The rest of the dissertation is organized as follow. In chapter 2, we presented the variable selection approach for VMICM, its estimation method, and theoretical properties. We conducted extensive simulation studies to evaluate the finite sample performance of the proposed method. The utility of our model was demonstrated with a real data analysis. In chapter 3, we generalized our selection approach to a binary response generalized regression setting. We extended the model to a quantile regression setting in chapter 4, followed by conclusion and further works in chapter 5. At last, all the proofs and details of the algorithms were rendered in the Appendices.

Chapter 2

Variable Selection with Varying Multi-Index Coefficients Model for G×E Interaction

2.1 Introduction

Gene-environment (G×E) interaction study has been gaining popularity. As discussed in chapter 1, varying multi-index coefficient model (VMICM) enjoys the unique ability to model non-linear interaction between genetic variants and a mixture of environmental variants. In this chapter, we propose a variable selection model for VMICM. Consider the varying multi-index coefficient model of the form:

$$\boldsymbol{Y} = \sum_{k=0}^{p} m_k(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{G}_k + \boldsymbol{\epsilon} \tag{2.1}$$

where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ is a continuous response variable that measures certain phenotypic trait of interest; $\boldsymbol{X}$ is a q-dimensional environmental exposure variables; $\boldsymbol{G}$ is a $p+1$ dimensional genetic variables; $m_k(\cdot), k = 0, 1, \ldots, p$ is the unknown non-parametric function and $\boldsymbol{\beta}$ is a vector of unknown loading parameter of dimension q. One of the main

advantage of VMICM is that it models the effects of $\boldsymbol{G}$ on Y as functions of $\boldsymbol{X}$ without suffering the curse of dimensionality. One can interpret $m_k(\boldsymbol{X}\boldsymbol{\beta})$ as the effect of $\boldsymbol{G}_k$ on Y, modified by multiple X variables though the index $\boldsymbol{X}\boldsymbol{\beta}$. In addition, VMICM is flexible enough to cover a wide range of models. For instance, if $q = 1$ and $\beta = 1$ then it becomes an additive varying coefficient model, and if $p = 1$ and $\boldsymbol{G} = \mathbf{1}$ then it becomes a standard additive single index model.

Variable selection has been a popular statistical strategy to solve large p small n problem in a regression setup. In the past, researchers often opted for forward/backward selection, and information based criteria such as AIC and BIC for variable selection. Recently, variable selection via penalized regression is gaining more popularity and wider acceptance as it features simultaneous selection and estimation of parameters. Its idea is to add a penalty function to the loss function or log-likelihood function. Bridge regression (Frank and Friedman (1993)), least absolute shrinkage and selection operator (LASSO) (Tibshirani (1996)) and its extensions (adaptive-LASSO Zou (2006)), smoothly clipped absolute deviation (SCAD) (Fan and Li (2001)) and minimax concave penalty (MCP) (Zhang (2010)) are a few examples. To evaluate different penalized functions, Fan and Li (2001) proposed three important criteria: sparsity, unbiasedness, and continuity. They also showed that SCAD penalty possess oracle property, meaning that penalized regression featuring SCAD works as well as if the correct sub-model was known in advance. Adaptive LASSO (Zou (2006)), SCAD (Fan and Li (2001)) and MCP (Zhang (2010)) all possess oracle property. However, for adaptive LASSO, determining weights for parameters might become problematic when $p > n$. In the current work, we adopt MCP penalty function for its oracle property and fast algorithm.

Considering the nonlinear structure about the unknown non-parametric functions $m_k(\cdot)$

and its unknown parameter $\boldsymbol{\beta}$. We propose a three stage iterative variable selection strategy. Specifically, our goal is: (1) to classify the non-parametric functions $m_k(\cdot), k = 1, \ldots, p$ into three categories: varying, constant and zero; (2) to select zero and non-zero components of loading parameters $\boldsymbol{\beta}$; and (3) to estimate $m_k(\cdot), k = 0, 1, \ldots, p$ and $\boldsymbol{\beta}$. Our approach is motivated by the practical need to separate three different mechanisms in G×E interaction. The zero function of $m_k(\cdot)$ indicates no genetic effect at all; the constant function of $m_k(\cdot)$ indicates no G×E effect; while the varying function of $m_k(\cdot)$ indicates the existence of G×E effect. As shown in Liu et al. (2016), the VMICM model has the advantage to capture the joint interaction of genes with multiple exposures as a whole. Novel insights about the underlying genetic mechanism can be revealed by the proposed model. In addition to the selection of the coefficient functions, we can also select important loading parameters inside each index coefficient function, to further quantify the relative importance of individual exposure variables.

Feng and Xue (2013) proposed a variable selection approach based on VMICM by applying a group SCAD penalty on B-spline coefficients $\boldsymbol{\gamma}$ and loading parameters $\boldsymbol{\beta}$. They focused on either zero or nonzero coefficient functions $m_k(\cdot)$. We are particularly interested in the constant coefficient since it corresponds to no G×E effect and has important practical implications. Tang et al. (2012) proposed a two step variable selection approach based on an additive varying-coefficient model. They classified the non-parametric function into three categories: varying, constant or zero. However, their model is a special case of our VMICM model with the dimension of the $\boldsymbol{X}$ variable being one. No variable selection approach on VMICM has been proposed to classify unknown non-parametric functions $m_k(\cdot)$ into three categories (varying, constant or zero), while at the same time selecting non-zero loading parameter $\boldsymbol{\beta}$. Following their previous work, we use B-spline basis functions to approximate

unknown non-parametric functions $m_k(\cdot)$, then using penalized regression to classify $m_k(\cdot)$ into varying, constant or zero. In addition, we select non-zero $\boldsymbol{\beta}$ via first order approximation and penalized regression. We show that under mild regulatory conditions, our estimators possess the oracle property, indicating that our penalized estimator works as well as if the correct sub-model is known in advance.

The rest of the chapter is organized as follows. Section 2.2 introduces our proposed variable selection approach, including estimation method, iteration approach, and how to select various tuning parameters. Method on how to select initial values for $\boldsymbol{\beta}$ is also discussed. In Section 2.3, we evaluate the theoretical properties of our approach. In Section 2.4, we perform simulations to evaluate the performance of our approach in finite samples, followed by a real data application in Section 2.5 and a discussion.

2.2 Variable Selection Method

Throughout the chapter, superscript ^{T} is used to denote matrix transpose, $||.||_p$ is used to denote L_p norm, and $log(a)$ is used to denote natural logarithm of a. For the sake of simplicity, we use constant and non-zero constant interchangeably.

2.2.1 Model Setup

The varying multi-index coefficient model is set up as follows:

$$\boldsymbol{Y} = \sum_{k=0}^{p} m_k(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{G}_k + \boldsymbol{\epsilon}$$

where $\boldsymbol{Y}_{n\times 1} = (Y_1, Y_2, \ldots, Y_n)^T$ is a continuous response variable; n is the sample size. $m_k(\cdot), k = 0, 1, \ldots, p$ are $p+1$ unknown continuous functions; $\boldsymbol{X}_{n\times q} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_q)$ are continuous loading covariates; $\boldsymbol{\beta}_{q\times 1} = (\beta_1, \ldots, \beta_q)^T$ are the loading parameters; $\boldsymbol{G}_{n\times(p+1)} = (\boldsymbol{G}_0, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_p)$, $\boldsymbol{G}_0 = (1, \ldots, 1)^T$ and $\boldsymbol{G}_k = (G_{1k}, G_{2k}, \ldots, G_{nk})^T$ is a continuous or discrete vector of length n for $k = 1, 2, \ldots p$. In the model, $m_k(\boldsymbol{X}\boldsymbol{\beta})$ is the effect of $\boldsymbol{G}_k$ on Y for $k \neq 0$ and $m_0(\boldsymbol{X}\boldsymbol{\beta})$ is the intercept function which models the marginal effect of multiple $\boldsymbol{X}$ variables on Y. The error term $\boldsymbol{\epsilon}$ is an unknown random error with mean 0 and variance σ^2. We further assumed ϵ_i and ϵ_j are independent $\forall\ 1 \leq i, j \leq n$ and $i \neq j$.

2.2.2 Estimation Method

Our goal was to select and estimate unknown functions $\{m_k(\cdot)\}_{k=0,1,\ldots,p}$ and unknown loading parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$. For identifiability purpose, we assume $||\boldsymbol{\beta}||_2 = 1$ and $\beta_1 > 0$, and $m_k(\cdot)$ cannot has the form of $m_j(\boldsymbol{u}) = \boldsymbol{\alpha}^T\boldsymbol{u}\boldsymbol{\beta}^T\boldsymbol{u} + \boldsymbol{\gamma}^T\boldsymbol{u} + c$.

We approximated the unknown function $\{m_k(u)\}_{k=0,1,\ldots,p}$ using B-spline basis functions. Without loss of generality, we assumed $u \in [0, 1]$. Let K be the number of internal knots, h be the degree of the B-spline basis function. So $h = 1$ represents linear splines, $h = 2$ represents quadratic splines. Denote $u_1, u_2, ..., u_K$ be internal knots satisfying $0 = u_0 =< u_1 < u_2 < .... < u_K < u_{K+1} = 1$. Denote I_{n_t} to be left closed, right opened interval $[u_{t-1}, u_t)$ for $1 \leq t \leq K$; $I_{n_{K+1}}$ to be closed interval $[u_K, u_{K+1}]$. Denote $\mathscr{F}$ to be a collection of functions f defined on $[0, 1]$ satisfying: (i) the restriction of f to I_{n_t} is a polynomial of degree h or less for $1 \leq j \leq K + 1$; (ii) f is $h - 1$ times continuous differentiable on $[0, 1]$. Let $L = K + h + 1$, by (Schumaker (2007)), we have normalized B-spline basis function $\widetilde{\boldsymbol{B}}(u) = (\widetilde{\boldsymbol{B}}_1(u), \widetilde{\boldsymbol{B}}_2(u), \ldots, \widetilde{\boldsymbol{B}}_L(u))$ for $\mathscr{F}$. And there exists a linear transformation matrix

$\mathbf{\Pi}$, such that $\mathbf{\Pi}\widetilde{\boldsymbol{B}}(u) = (\mathbf{1}, \bar{\boldsymbol{B}}(u)) = (\mathbf{1}, \boldsymbol{B}_2(u), \boldsymbol{B}_3(u), \ldots, \boldsymbol{B}_L(u)) = \boldsymbol{B}(u)$ where each component of $\bar{\boldsymbol{B}}(u)$ is a function of u. Then for $0 \leq k \leq p$, we can estimate $m_k(u)$ by

$$m_k(u) \approx (1, B_2(u), \cdots, B_L(u)) * (\gamma_{k1}, \gamma_{k2}, \cdots, \gamma_{kL})^T = \boldsymbol{B}(u)\boldsymbol{\gamma}_k = \gamma_{k1} + \bar{\boldsymbol{B}}(u)\boldsymbol{\gamma}_{k*} \quad (2.2)$$

where $\boldsymbol{\gamma}_{k*} = (\gamma_{k2}, \gamma_{k3}, \ldots, \gamma_{kL})^T$ and $\boldsymbol{\gamma}_k = (\gamma_{k1}, \boldsymbol{\gamma}_{k*}^T)^T$. With B-spline approximation, (2.1) can be rewritten as

$$\boldsymbol{Y} = \sum_{k=0}^{p} \{\gamma_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{\gamma}_{k*}\}\boldsymbol{G}_k + \boldsymbol{\epsilon}. \quad (2.3)$$

Thus, the original estimation problem can be transformed to estimate $\{\gamma_{k1}, \boldsymbol{\gamma}_{k*}\}_{k=0,1,\ldots,p}$ and $\boldsymbol{\beta}$. Note: the transformation $\mathbf{\Pi}$ enable us to separate the constant effect of $\boldsymbol{G}_k$ on $\boldsymbol{Y}$ from its jointly effect with $\boldsymbol{X}$ on $\boldsymbol{Y}$. That is: (1) if $||\boldsymbol{\gamma}_{k*}||_2 = (\sum_{l=2}^{L} \gamma_{kl}^2)^{1/2} \neq 0$, then there exists interaction between $\boldsymbol{G}_k$ and multiple $\boldsymbol{X}$; (2) if $||\boldsymbol{\gamma}_{k*}||_2 = 0$ and $|\gamma_{k1}| \neq 0$, then $\boldsymbol{G}_k$ has a constant effect on $\boldsymbol{Y}$, i.e., no G×E interaction effect; and (3) if further $||\boldsymbol{\gamma}_{k*}||_2 = 0$ and $|\gamma_{k1}| = 0$ then $\boldsymbol{G}_k$ has no effect on $\boldsymbol{Y}$ at all.

To select and estimate the parameters $\{\boldsymbol{\gamma}_k\}_{k=0,1,\ldots,p}$ and $\boldsymbol{\beta}$, we adopted the penalized regression idea and minimized the following objective function:

$$\begin{aligned} Q(\boldsymbol{\beta}, \boldsymbol{\gamma}) = & \sum_{i=1}^{n} g(Y_i - \sum_{k=0}^{p} [\gamma_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{\gamma}_{k*}] G_{ik}) + n \sum_{k=1}^{p} p_{\lambda_1}(||\boldsymbol{\gamma}_{k*}||_2) \\ & + n \sum_{k=1}^{p} p_{\lambda_2}(|\gamma_{k1}|) I(||\boldsymbol{\gamma}_{k*}||_2 = 0) + n \sum_{d=2}^{q} p_{\lambda_3}(|\beta_d|) \end{aligned} \quad (2.4)$$

where $g(\cdot)$ is a loss function; $p_{\lambda_1}(\cdot), p_{\lambda_2}(\cdot), p_{\lambda_3}(\cdot)$ are penalty function of the corresponding parameters; and $I(\cdot)$ is an indicator function.

Remarks: (1) From the construction of the penalty function, we penalized γ_{k1} only if

$||\boldsymbol{\gamma}_{k*}||_2 = 0$. If $||\boldsymbol{\gamma}_{k*}||_2 \neq 0$, it implies that the function is varying and no need to penalize the constant part.

(2) No penalty was applied to the intercept function $m_0(\cdot)$ as both $\boldsymbol{\gamma}_{0*}$ and γ_{01} was not involved in the penalty term. There is no practical motivation of penalizing the marginal intercept function.

(3) No penalty was applied to the first loading parameter β_1 in $\boldsymbol{\beta}$ due to the constraint.

For the penalty function, we used MCP penalty proposed by C. Zhang (Zhang (2010)), $p(x, \lambda) = \lambda \int_0^x (1 - \frac{s}{\tau\lambda})_+ ds$ with regularization parameters $\tau > 0$ and $\lambda > 0$. For the loss function, we set $g(\cdot)$ to be squared error loss, (2.4) can be further rewritten as :

$$\begin{aligned} Q(\boldsymbol{\beta}, \boldsymbol{\gamma}) = & \sum_{i=1}^{n} \left\{ Y_i - \sum_{k=0}^{p} [\gamma_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{\gamma}_{k*}] G_{ik} \right\}^2 + n \sum_{k=1}^{p} p_{\lambda_1}(||\boldsymbol{\gamma}_{k*}||_1) \\ & + n \sum_{k=1}^{p} p_{\lambda_2}(|\gamma_{k1}|) I(||\boldsymbol{\gamma}_{k*}||_1 = 0) + n \sum_{d=2}^{q} p_{\lambda_3}(|\beta_d|) \end{aligned} \tag{2.5}$$

where $p_{\lambda_1}(\cdot), p_{\lambda_2}(\cdot), p_{\lambda_3}(\cdot)$ are the MCP penalty function defined above.

2.2.3 Iterative Approach

Our modeling purpose was to separate the $m_k(\cdot)$ function into three different categories: varying, constat or zero, denoted by V, C and Z respectively. Following the ideas by Feng and Xue (2013) and Tang et al. (2012), we adopted a three step iterative approach to obtain our penalized estimator.

Step 1: For given initial values of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}^{(0)}$, we obtained our 1st step estimation

of $\boldsymbol{\gamma}$, denoted by $\hat{\boldsymbol{\gamma}}^{(1)} = \{\hat{\gamma}_{k1}^{(1)}, \hat{\boldsymbol{\gamma}}_{k*}^{(1)T}\}_{k=0,1,\ldots,p}^{T}$ by following a group penalized regression,

$$\hat{\boldsymbol{\gamma}}^{(1)} = \arg\min_{\boldsymbol{\gamma}} Q_1(\boldsymbol{\gamma}|\lambda_1, \hat{\boldsymbol{\beta}}^{(0)})$$

where

$$Q_1(\boldsymbol{\gamma}|\lambda_1, \hat{\boldsymbol{\beta}}^{(0)}) = \sum_{i=1}^{n}\{Y_i - \sum_{k=0}^{p}[\gamma_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})\boldsymbol{\gamma}_{k*}]G_{ik}\}^2 + n\sum_{k=1}^{p} p_{\lambda_1}(||\boldsymbol{\gamma}_{k*}||_2).$$

Note that instead of penalizing each coordinate of $\boldsymbol{\gamma}_{k*} = (\gamma_{k2}, \ldots, \gamma_{kL})^T$ separately, we penalized the L_2 norm of $\boldsymbol{\gamma}_{k*}$ because we want to assess the presence of joint varying effect of $\boldsymbol{X}$ and $\boldsymbol{G}_k$ on $\boldsymbol{Y}$. No penalty was applied to $\boldsymbol{\gamma}_{0*}$ (the B-spline coefficients for the intercept function $m_0(\cdot)$). Step 1 separates $m_k(\cdot), k = 1, \ldots, p$ into two categories: varying(V) or non-varying(NV), and $m_k(\cdot) \in V$ if $||\hat{\boldsymbol{\gamma}}_{k*}^{(1)}||_2 > 0$ and $m_k(\cdot) \in NV$ if $||\hat{\boldsymbol{\gamma}}_{k*}^{(1)}||_2 = 0$.

Step 2: The aim of this step was to select γ_{k1} given $\hat{\boldsymbol{\gamma}}_{k*}^{(1)} = 0$. From the non-varying functions obtained from step 1, we would like to further select the variables with constant effects, and classify the non-parametric functions into constant(C) and zero (0). We penalized γ_{k1} only when $||\hat{\boldsymbol{\gamma}}_{k*}^{(1)}||_2 = 0,\ k = 1, 2, \ldots, p$, and no penalty is applied to γ_{01}.

We obtained our step 2 estimator $\hat{\boldsymbol{\gamma}}^{(2)} = \{(\hat{\gamma}_{k1}^{(2)}, \hat{\boldsymbol{\gamma}}_{k*}^{(2)})_{k\in V}, (\hat{\gamma}_{k1}^{(2)})_{k\in NV}\}$ via penalized regression

$$\hat{\boldsymbol{\gamma}}^{(2)} = \arg\min_{\boldsymbol{\gamma}} Q_2(\boldsymbol{\gamma}|\lambda_2, \boldsymbol{\beta}^{(0)}, \hat{\boldsymbol{\gamma}}^{(1)})$$

where

$$Q_2(\boldsymbol{\gamma}|\lambda_2, \boldsymbol{\beta}^{(0)}, \hat{\boldsymbol{\gamma}}^{(1)}) = \sum_{i=1}^{n}\{Y_i - \sum_{k\in V}[\gamma_{k1}^{(2)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta}^{(0)})\boldsymbol{\gamma}_{k*}^{(2)}]G_k - \sum_{k\in NV}\gamma_{k1}^{(2)}G_k\}^2 + n\sum_{k=1}^{p} p_{\lambda_2}(|\gamma_{k1}^{(2)}|)I(||\hat{\boldsymbol{\gamma}}_{k*}^{(1)}||_2 = 0).$$

After Step 1 and 2, we had obtained the estimator of the B-spline coefficients $\boldsymbol{\gamma}$ and classify $m_k(\cdot)$ $k = 1, \ldots, p$ into V, C or 0. The next step is to select loading parameter $\boldsymbol{\beta}$ given $\hat{\boldsymbol{\gamma}}^{(2)}$.

Step 3: We obtained $\hat{\boldsymbol{\beta}}$ via penalized regression

$$\hat{\boldsymbol{\beta}} = \arg\min_{||\boldsymbol{\beta}||_2=1} Q_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}^{(2)})$$

where

$$Q_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}^{(2)}) = \sum_{i=1}^{n}(Y_i - \sum_{k=0}^{p}[\hat{\gamma}_{k1}^{(2)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}_{k*}^{(2)}]G_k)^2 + n\sum_{d=2}^{q} p_{\lambda_3}(|\beta_d|).$$

Then we replaced $\hat{\boldsymbol{\beta}}^{(0)}$ by $\hat{\boldsymbol{\beta}}$, and iterate step 1 to 3 until convergence. The algorithms used in each step would be discussed in Appendices.

2.2.4 Selection of tuning parameters

We proposed the following three step tuning parameters selection process based on Bayesian Information Criterion (BIC) (Schwarz (1978)).

Step 1: We took λ_1 as the minimizer of

$$BIC(\lambda_1) = log \sum_{i=1}^{n} (Y_i - \sum_{k=0}^{p} [\hat{\gamma}_{k1}^{(\lambda_1)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})\hat{\boldsymbol{\gamma}}_{k*}^{(\lambda_1)}]G_k)^2 + \frac{log(n)}{n} * df_{\lambda_1}$$

where $\{\hat{\gamma}_{k1}^{(\lambda_1)}, \hat{\boldsymbol{\gamma}}_{k*}^{(\lambda_1)}\}_{k=0,1,...,p}$ are the minimizers of $Q_1(\boldsymbol{\gamma}|\lambda_1, \hat{\boldsymbol{\beta}}^{(0)})$ defined above; $\hat{\boldsymbol{\beta}}^{(0)}$ is chosen as the estimator from previous iteration; and df_{λ_1} is defined as the total number of non zero coefficients if λ_1 is the penalized parameter.

Step 2: We took λ_2 as the minimizer of

$$BIC(\lambda_2) = log \sum_{i=1}^{n} (Y_i - \sum_{k=0}^{p} [\hat{\gamma}_{k1}^{(\lambda_2)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})\hat{\boldsymbol{\gamma}}_{k*}^{(\lambda_2)}]G_k)^2 + \frac{log(n)}{n} * df_{\lambda_2}$$

where $\{\hat{\gamma}_{k1}^{(\lambda_2)}, \hat{\boldsymbol{\gamma}}_{k*}^{(\lambda_2)}\}_{k=0,1,...,p}$ are the minimizers of $Q_2(\boldsymbol{\gamma}|\lambda_2, \hat{\boldsymbol{\beta}}^{(0)})$ defined above; $\hat{\boldsymbol{\beta}}^{(0)}$ is chosen as the estimator from previous iteration; and df_{λ_2} is defined as the total number of non zero coefficients if λ_2 is the penalized parameter.

Step 3: We took λ_3 as the minimizer of

$$BIC(\lambda_3) = log \sum_{i=1}^{n} (Y_i - \sum_{k=0}^{p} [\hat{\gamma}_{k1}^{(\lambda_2)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda_3)})^T\hat{\boldsymbol{\gamma}}_{k*}^{(\lambda_2)}]G_k)^2 + \frac{logn}{n} * df_{\lambda_3}$$

where $\hat{\boldsymbol{\beta}}^{(\lambda_3)}$ is the minimizer of $Q_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}^{(2)})$ defined above; $\{\hat{\gamma}_{k1}^{(\lambda_2)}, \hat{\boldsymbol{\gamma}}_{k*}^{(\lambda_2)}\}_{k=0,1,...,p}$ are minimizer of the B-spline coefficient from Step 2; and df_{λ_3} is defined as the total number of non zero $\boldsymbol{\beta}$ if λ_3 is the penalized parameter. We searched the optimal of $\lambda_1, \lambda_2, \lambda_3$ over a grid of 100 exponentially decreasing values with the minimum to be 1E-3, and the maximum of $\lambda_1, \lambda_2, \lambda_3$ were set to be the minimum value such that all of the penalized estimators are 0.

2.2.5 Selection of the order h and the number of internal knots K

Since h is the order of the B-spline basis function, higher degree corresponds to more complicated interactions and is less interpretable in practice. For instance, $h = 2$ implies quadratic splines while $h = 3$ implies cubic splines. Hence, we searched optimal h over the set $h \in \{2, 3, 4\}$. As for the selection of K, since only when $K = O_p(n^{\frac{1}{2r+1}})$ (n is the number of samples and r is the smoothness of our nonparametric function $m_k(\cdot)$ and $r > 2$), our selection approach possesses oracle properties. So we can search optimal K in the neighborhood of $n^{\frac{1}{2r+1}}$, which is denoted by $\mathscr{K}$. In our simulation, $\mathscr{K} = \{2, 3, 4, 5\}$.

In theory, we can select the optimal order and the interior knots for all the nonparametric functions. However, this is practically infeasible due to the large search space and the computational cost. Thus, we assumed that all the nonparametric functions share common h and K, and fit the following intercept only model to select the optimal h and K,

$$Y = m_0(X\boldsymbol{\beta}) + \boldsymbol{\epsilon}. \tag{2.6}$$

We searched the optimal K and h over a grid $K \in \mathscr{K}$ and $h \in \{2, 3, 4\}$. The optimal K and h is defined as the K, h that minimize $log \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2 + \frac{log(n)}{n}(K + h + 1)$, where $\hat{Y}_i$ is the prediction of the ith subject under model (2.6).

2.2.6 Selection of Initial values for $\boldsymbol{\beta}$

The algorithm described above needs a reasonable initial value for $\boldsymbol{\beta}$ (denoted by $\boldsymbol{\beta}^{initial}$) to start the iteration. Based on the optimal K and h selected, we fit the intercept only model (2.6) via B-spline approximation and Newton-Raplson algorithm. $\boldsymbol{\beta}^{initial}$ was set to be the estimator for β in (2.6).

2.3 Theoretical Properties

Let $\boldsymbol{\beta}^0$ and $m_k^0(\cdot),\ k = 0, 1, \ldots, p$ be the true value of $\boldsymbol{\beta}$ and $m_k(\cdot)$, respectively, and denote $\boldsymbol{\gamma}^0$ be the true value of the B-spline coefficient $\boldsymbol{\gamma}$. Without loss of generality, we assumed $\beta_l^0 \neq 0$ for $l = 1, \ldots s$, $\beta_l^0 = 0$ for $l = s+1, \ldots q$; $m_k^0(\cdot)$ is varying for $k = 0, 1, \ldots, v$, $m_k^0(\cdot)$ is non-zero constant for $k = v+1, \ldots, c$ and $m_k^0(\cdot)$ is zero for $k = c+1, \ldots, p$. The following theorems established the consistency of the penalized least square estimators.

Theorem 2.3.1. *Assume the regulatory conditions (A1)-(A7) in Appendices hold and the number of knots* $K = O_p(n^{1/(2r+1)})$. *Then*

(i) $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0|| = O_p(n^{-r/(2r+1)} + a_n)$;

(ii) $||\hat{m}_k(\cdot) - m_k^0(\cdot)|| = O_p(n^{-r/(2r+1)} + a_n),\ k = 1, \ldots, q$

where

$$a_n = \max_{k,j,l}\{p'_{\lambda_1}(||\boldsymbol{\gamma}_{k*}^0||_2), p'_{\lambda_2}(|\gamma_{j1}^0|), p'_{\lambda_3}(|\beta_l^0|), \boldsymbol{\gamma}_{k*}^0 \neq 0, \gamma_{j1}^0 \neq 0, \beta_l^0 \neq 0\}$$

and $k, j = 1, \ldots, p, l = 2, \ldots, q$, *and* r *is defined in Appendices.*

Theorem 2.3.2. *Assume the regularity conditions (A1)-(A7) the Appendices hold and the number of knots* $K = O_p(n^{1/(2r+1)})$. *Let*

$$\lambda_{max} = \max\{\lambda_1, \lambda_2, \lambda_3\}, \quad \lambda_{min} = \min\{\lambda_1, \lambda_2, \lambda_3\}.$$

Suppose $\lambda_{max} \to 0$ *and* $n^{r/(2r+1)}\lambda_{min} \to \infty$ *as* $n \to \infty$. *Then with probability approaching 1,* $\hat{\boldsymbol{\beta}}$ *and* $\hat{m}_k(\cdot)$ *must satisfy:*

(i) $\hat{\beta}_j = 0$ *for* $j = s+1, \ldots, q$

(ii) $\hat{m}_k(\cdot) = c_k$ *for* $k = v+1, \ldots, c$ *where* c_k *is some non-zero constant*

(iii) $\hat{m}_k(\cdot) = 0$ *for* $k = c+1, \ldots, p$

Theorem 2.3.1 and 2.3.2 show that our proposed variable selection approach is consistent and possesses oracle property.

2.4 Simulation

We conducted extensive simulation to evaluate the performance of our proposed approach. The performance is measured in several ways: (1) classification accuracy for function $m(\cdot)$, denoted as oracle percentage; (2) IMSE of the estimated $m(\cdot)$ function; (3) selection accuracy of $\boldsymbol{\beta}$; and (4) estimation accuracy of $\boldsymbol{\beta}$(MSE). Denote R as the total number of simulations.

Oracle percentage of $m(\cdot)$ is defined as the percentage of correct classification out of a total of R simulations. For example, if $m_k(\cdot) \in V$, and out of R simulations, $m_k(\cdot)$ is classified as varying for g times, then the oracle percentage of $m_k(\cdot)$ is $\frac{g}{R} \times 100\%$.

IMSE of $m_k(\cdot)$ is defined as

$$IMSE = \frac{1}{R}\sum_{r=1}^{R}[\frac{1}{n_{grid}}\sum_{j=1}^{n_{grid}}(\hat{\gamma}_{k1}^{(r)} + \bar{\boldsymbol{B}}(u_j)\hat{\boldsymbol{\gamma}}_{k*}^{(r)} - m_k(u_j))^2]$$

where n_{grid} is the number of points that we want to estimate the MSE of the predicted function; $\hat{\boldsymbol{\gamma}}_{k*}^{(r)}$ and $\hat{\gamma}_{k1}^{(r)}$ are the estimators of the B-spline coefficients for the rth simulation using the proposed estimation approach; $\hat{\boldsymbol{\beta}}^{(r)}$ is the estimator of the loading parameter $\boldsymbol{\beta}$ for the rth simulation; and u_j is taken at the $j/n_{grid} \times 100\%$ quantile among the range of $\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(r)}$. For our simulations, n_{grid} is set to be 100.

Oracle percentage of $\boldsymbol{\beta}$ is defined as the percentage of correct selection of $\boldsymbol{\beta}$ out of R simulations. For example, if $\beta_d \neq 0$ and out of R simulations, β_d is selected to be non-zero

for g times, then the oracle percentage of β_d is $\frac{g}{R} \times 100\%$.

MSE of β_d is calculated as $\frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_d^{(r)} - \beta_d)^2$ where $\hat{\beta}_d^{(r)}$ is the estimator for β_d in the rth simulation.

2.4.1 Simulation Setting

The simulation data was generated according to the following model,

$$\boldsymbol{Y} = m_0(\boldsymbol{X}\boldsymbol{\beta}) + \sum_{k=1}^{p} m_k(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{G}_k + \boldsymbol{\epsilon}$$

where $\boldsymbol{X}$ was generated from a $Unif(0,1)$ distribution; $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)^T$; $\beta_1 = \beta_2 = \frac{1}{\sqrt{2}}$; and the rest $\beta_j's$ are zeros. We evaluated the performance of the proposed approach with both continuous and discrete predictors $\boldsymbol{G}$. For continuous $\boldsymbol{G}$ variables, they can be gene expressions. For discrete $\boldsymbol{G}$ variables, they can be SNP variants. In either case, the dimension of $\boldsymbol{G}$ can be large.

2.4.2 The Continuous Cases

In the continuous case, the non-parametric functions $m_k(u)$ are defined as follows: $m_0(u) = 2sin(2\pi u)$, $m_1(u) = 2cos(\pi u) + 2$ and $m_2(u) = sin(2\pi u) + cos(\pi u) + 1$ are varying functions; $m_3(u) = 2$ and $m_4(u) = 2.5$ are non-zero constants; $m_k(u) = 0$ for $k = 5, \ldots, p$ are zeros. The number of loading parameters is set as $q = 5$ and $\beta_1 = \beta_2 = \frac{1}{\sqrt{2}}$, $\beta_3 = \beta_4 = \beta_5 = 0$. $\boldsymbol{G}$ was generated randomly from a $N(0,1)$ distribution, $\boldsymbol{\epsilon} \sim N(0,1)$. We ran 1000 simulations (R = 1000) to evaluate the performance of the proposed model under p = 50, 100.

Table 1 demonstrates the selection and estimation accuracy for non-parametric functions with continuous $\boldsymbol{G}$. The left and right penal corresponds to the case where p = 50 and

100 respectively. For all the cases, the selection accuracy (oracle %) is very closed to 100%. IMSE for varying functions ($m_0(\cdot), m_1(\cdot)$ and $m_2(\cdot)$) is in the order of -2, and the IMSE for constant functions ($m_3(\cdot)$ and $m_4(\cdot)$) is in the order of -3. All of the model IMSE and oracle IMSE are in the same order. Overall, the differences in oracle percentage and IMSE are negligible between the case $p = 50$ and the case $p = 100$. These observations suggest that our proposed variable selection approach possesses reasonable selection and prediction accuracy for non-parametric function $m_k(\cdot)$.

Table 1: Selection and prediction accuracy of $m_k(\cdot)$ for continuous $\boldsymbol{G}$

		p = 50			p = 100		
			IMSE			IMSE	
		Oracle %	Model	Oracle	Oracle %	Model	Oracle
$n = 500$	$m_0(\cdot)$	100.0%	3.87E-02	4.27E-02	100.0%	3.77E-02	4.51E-02
	$m_1(\cdot)$	99.6%	1.58E-02	2.42E-02	99.9%	1.57E-02	3.14E-02
	$m_2(\cdot)$	99.9%	2.33E-02	2.58E-02	99.9%	2.26E-02	2.96E-02
	$m_3(\cdot)$	100.0%	2.09E-03	2.11E-03	100.0%	1.90E-03	1.97E-03
	$m_4(\cdot)$	100.0%	2.04E-03	2.06E-03	100.0%	2.07E-03	2.12E-03
	Zero	99.7%	1.94E-05	0	99.9%	1.12E-05	0
$n = 1000$	$m_0(\cdot)$	100.0%	3.23E-02	3.40E-02	100.0%	3.31E-02	3.47E-02
	$m_1(\cdot)$	100.0%	7.17E-03	1.21E-02	100.0%	7.07E-03	1.17E-02
	$m_2(\cdot)$	100.0%	1.46E-02	1.59E-02	100.0%	1.46E-02	1.64E-02
	$m_3(\cdot)$	100.0%	1.02E-03	1.02E-03	100.0%	9.60E-04	9.55E-04
	$m_4(\cdot)$	100.0%	1.09E-03	1.09E-03	100.0%	1.06E-03	1.07E-03
	Zero	99.8%	8.50E-06	0	99.9%	3.46E-06	0

Table 2 presents the selection and prediction accuracy for loading parameter $\boldsymbol{\beta}$. The left and the right penal correspond to the case where p = 50 and 100 respectively. It shows that the selection accuracy for all β is reasonably good ($> 98\%$ in all cases). For most of the β, the MSE is in the order of -4 or lower, except for β_2, which was -3 for both $p = 50$ and $p = 100$ when $n = 500$. The order of the model estimation for $\boldsymbol{\beta}$ is at least the same as that of the oracle model if not lower. And we did not observe a difference in performance between

the case $p = 50$ and the case $p = 100$. These results indicate that our model possesses good selection and prediction accuracy for loading parameters $\boldsymbol{\beta}$.

Table 2: Prediction accuracy of $\boldsymbol{\beta}$ for continuous $\boldsymbol{G}$ ($\beta_1 = \beta_2 = \frac{1}{\sqrt{2}}$, $\beta_3 = \beta_4 = \beta_5 = 0$)

		p = 50			p = 100		
			MSE			MSE	
		Oracle %	Model	Oracle	Oracle %	Model	Oracle
$n = 500$	β_1	100.0%	1.15E-04	1.07E-04	100.0%	1.17E-04	1.30E-04
	β_2	100.0%	8.04E-03	4.12E-03	100.0%	2.26E-03	7.62E-03
	β_3	98.1%	9.98E-05	0	98.2%	3.64E-05	0
	β_4	98.8%	2.99E-05	0	99.1%	3.13E-05	0
	β_5	98.6%	1.00E-04	0	98.5%	7.73E-05	0
$n = 1000$	β_1	100.0%	5.30E-05	5.52E-05	100.0%	5.00E-05	5.49E-05
	β_2	100.0%	5.34E-05	1.86E-03	100.0%	5.04E-05	1.79E-03
	β_3	98.9%	9.36E-06	0	98.8%	1.16E-05	0
	β_4	99.4%	6.30E-06	0	99.5%	5.49E-06	0
	β_5	99.1%	7.17E-06	0	99.0%	6.93E-06	0

2.4.3 For discrete G

We further evaluated how our proposed model performs with the discrete $\boldsymbol{G}$. In this simulation, each G variable was simulated from a multinomial distributions with minor allele frequency (MAF) denoted as P_a. The G variable took values $0, 1$, and 2 corresponding to the genotype aa, Aa, and AA where a is the minor allele. The frequencies corresponding to the three genotypes (aa, Aa, and AA) are P_a^2, $2P_a(1 - P_a)$ and $(1 - P_a)^2$. We varied the MAF for different G variables to evaluate the impact of the MAF on the selection performance. Specifically, for $k = 1, 2, 7$, $P_a = 0.5$; for $k = 3, 4, 8$, $P_a = 0.3$; for $k = 5, 6, 9$, $P_a = 0.1$, and for $k = 10, 11, \ldots, p$, $P_a \sim Unif(0.05, 0.5)$. For the non-parametric functions, we set $m_0(u) = 2sin(2\pi u)$, $m_1(u) = m_3(u) = m_5(u) = 2cos(\pi u) + 2$, $m_2(u) = m_4(u) = m_6(u) = sin(2\pi u) + cos(\pi u) + 1$; $m_7(u) = m_8(u) = m_9(u) = 2$; and $m_k(u) = 0$ for

$k = 10, 11, \ldots, p$. Under the setup, we have both varying and constant effect with different minor allele frequencies. For the zero effects, the MAF for $\boldsymbol{G}_k$ ranged uniformly from 0.05 to 0.5. $\boldsymbol{X}$ was generated from $Unif(0, 1)$ and $\boldsymbol{\epsilon}$ was generated from $N(0, 1)$. Finally, $\boldsymbol{Y}$ was generated according to model (2.1). We evaluated the performance of our proposed model via 1000 simulations under $p = 50, 100$ and $n = 500, 1000$.

Table 3 and figure 1 present the selection and estimation accuracy of non-parametric function $m_k(\cdot)$ for discrete $\boldsymbol{G}$. We observed that the oracle percentage is very high ($> 99\%$) for all cases, indicating our proposed model can correctly select the coefficient functions with high accuracy. The IMSE for varying functions was of the order -1 or -2, while the IMSE for constant functions was of the order -2 or-3. Moreover, the IMSE of the proposed model was in the same order of the IMSE of the oracle model. With the decrease of minor allele frequency P_a of G_k (from 0.5 to 0.1), we observed an increase in both model IMSE and oracle IMSE. This is consistent with our expectation since SNP with lower MAF provides less information. We did not observe a difference in oracle percentage and IMSE between the case $p = 50$ and $p = 100$. This suggests that our model performs reasonably well in both cases. The case $n = 1000$ performed slightly better than the case $n = 500$, and it is consistent with the asymptotic property of the model. Overall, the proposed model performs reasonably well in the selection and estimation of non-parametric functions.

Table 3: Selection and prediction accuracy of $m_k(\cdot)$ for discrete $\boldsymbol{G}$.

		p = 50			p = 100		
			IMSE			IMSE	
		Oracle %	Model	Oracle	Oracle %	Model	Oracle
n = 500	$m_0(\cdot)$	100.0%	4.67E-02	4.34E-02	100.0%	5.06E-02	5.22E-02
	$m_1(\cdot)$	99.7%	3.71E-02	3.54E-02	99.6%	4.02E-02	4.53E-02
	$m_2(\cdot)$	99.7%	4.56E-02	4.26E-02	99.4%	4.92E-02	4.77E-02
	$m_3(\cdot)$	99.7%	4.39E-02	4.19E-02	99.4%	4.95E-02	5.44E-02
	$m_4(\cdot)$	99.6%	5.13E-02	4.93E-02	99.4%	5.68E-02	5.45E-02
	$m_5(\cdot)$	99.6%	1.15E-01	1.53E-01	99.2%	1.33E-01	2.02E-01
	$m_6(\cdot)$	99.6%	1.29E-01	1.24E-01	99.3%	1.33E-01	1.30E-01
	$m_7(\cdot)$	100.0%	4.53E-03	4.50E-03	99.9%	4.74E-03	4.56E-03
	$m_8(\cdot)$	100.0%	5.34E-03	5.30E-03	99.8%	6.26E-03	6.03E-03
	$m_9(\cdot)$	100.0%	1.27E-02	1.26E-02	99.9%	1.21E-02	1.24E-02
	Zero	99.6%	1.82E-04	0	99.6%	1.28E-04	0
n = 1000	$m_0(\cdot)$	100.0%	3.11E-02	3.31E-02	100.0%	3.11E-02	3.20E-02
	$m_1(\cdot)$	99.9%	1.44E-02	1.96E-02	100.0%	1.44E-02	1.93E-02
	$m_2(\cdot)$	99.9%	2.16E-02	2.37E-02	100.0%	2.13E-02	2.34E-02
	$m_3(\cdot)$	99.9%	1.69E-02	2.19E-02	100.0%	1.74E-02	2.20E-02
	$m_4(\cdot)$	99.9%	2.35E-02	2.50E-02	100.0%	2.40E-02	2.56E-02
	$m_5(\cdot)$	99.9%	4.10E-02	4.55E-02	100.0%	4.25E-02	4.74E-02
	$m_6(\cdot)$	99.9%	4.62E-02	5.01E-02	100.0%	4.65E-02	4.85E-02
	$m_7(\cdot)$	100.0%	2.29E-03	2.34E-03	100.0%	1.89E-03	1.89E-03
	$m_8(\cdot)$	100.0%	2.67E-03	2.66E-03	100.0%	2.62E-03	2.63E-03
	$m_9(\cdot)$	100.0%	5.61E-03	5.69E-03	100.0%	6.03E-03	6.08E-03
	Zero	99.8%	3.10E-05	0	99.9%	1.15E-05	0

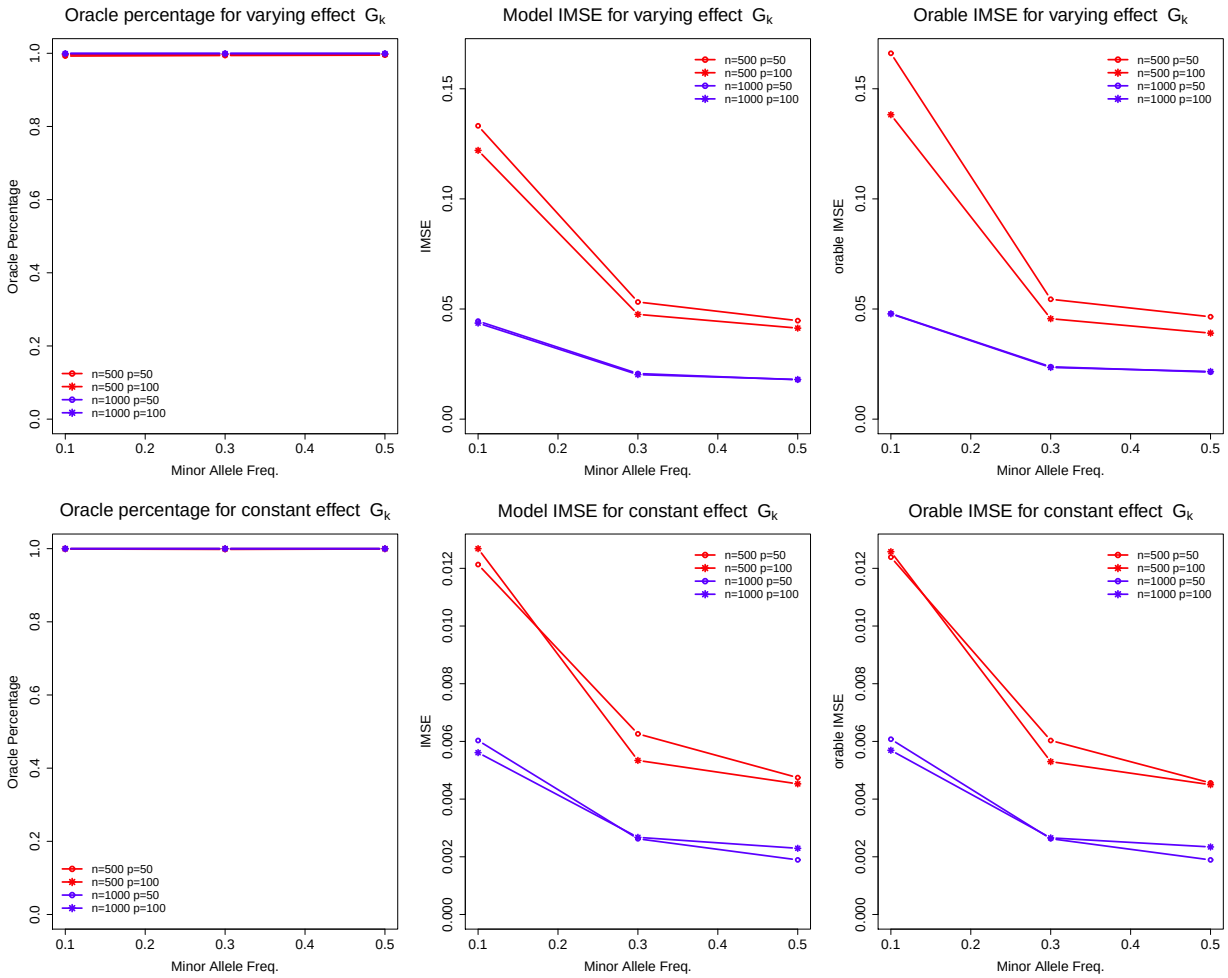

Figure 1: Selection and estimation accuracy of $m_k(\cdot)$ for discrete $\boldsymbol{G}$

Table 4 presents the selection and estimation result of the loading parameters $\boldsymbol{\beta}$. The left and right panel represent the case where $p = 50$ and $p = 100$, respectively. We observed that the oracle percentage in all the cases is above 97% and the MSE for the estimation of $\boldsymbol{\beta}$ is of the order -3 or lower in the proposed and oracle model. With the increase of model dimension p (from 50 to 100), we observed a slight increase model MSE. This is expected since model performance usually decreases as complexity increases. Further, the case where $n = 1000$ performed slightly better than the case $n = 500$, and it is consistent with the asymptotic theory of the model. Overall, the proposed variable selection approach

can correctly select and estimate the loading parameters with high accuracy.

Table 4: Prediction accuracy of $\boldsymbol{\beta}$ for discrete G ($\beta_1 = \beta_2 = \frac{1}{\sqrt{2}}$, $\beta_3 = \beta_4 = \beta_5 = 0$)

		p = 50			p = 100		
			MSE			MSE	
		Oracle %	Model	Oracle	Oracle %	Model	Oracle
n = 500	β_1	100.0%	2.86E-04	1.04E-04	100.0%	6.36E-04	2.13E-04
	β_2	99.8%	2.84E-03	1.04E-04	99.7%	6.89E-03	6.13E-03
	β_3	97.4%	9.38E-05	0	97.0%	2.15E-04	0
	β_4	99.1%	4.08E-05	0	98.6%	2.21E-04	0
	β_5	97.3%	7.04E-05	0	98.2%	1.68E-04	0
n = 1000	β_1	100.0%	5.45E-05	5.03E-05	100.0%	5.20E-05	6.03E-05
	β_2	100.0%	1.92E-03	2.09E-03	100.0%	5.20E-05	1.74E-03
	β_3	99.3%	5.19E-06	0	99.0%	1.10E-05	0
	β_4	99.3%	4.21E-06	0	98.9%	9.69E-06	0
	β_5	99.2%	7.51E-06	0	99.5%	2.64E-06	0

To summarize, the proposed method possesses reasonable selection and estimation accuracy for non-parametric functions $m_k(\cdot)$ and their loading parameters $\boldsymbol{\beta}$ in all cases. With the increase of the sample size n, we observed a small increase in oracle percentage for both non-parametric functions $m_k(\cdot)$ and their loading parameters $\boldsymbol{\beta}$. We also observed a decrease in IMSE of $m_k(\cdot)$ and MSE of $\boldsymbol{\beta}$. These coincide with the asymptotic theory of our model.

2.5 Real Data Application

We demonstrated the utility of the model with a human liver cohort (HLC) data set. The data set is consisted of genotype (SNPs), gene expressions, and phenotypes (activity of several liver enzymes). The data set can be downloaded from `www.synapse.org` using synapse ID: syn4499. For more details regarding the data set, please refer to Schadt et al. (2008) and

Yang et al. (2010). In the HLC data set, the phenotypes are enzyme activity measurements of Cytochrom P450. There are a total of nine P450 enzymes (CYP1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1, and 3A4). However, only CYP2E1 passed Shapiro-Wilk normality test (p-value > 0.1) after log transformation. Hence, we focused the analysis on CYP2E1 activity ($\boldsymbol{Y}$). For the environmental variable ($\boldsymbol{X}$), we set $\boldsymbol{X}_1$ = Age, $\boldsymbol{X}_2$ = Aldehyde Oxydase, and $\boldsymbol{X}_3$ = Liver Triglyceride, then transform $\boldsymbol{X}_i,\ i = 1, 2, 3$ to range [0,1] with $X'_{ij} = \frac{X_{ij} - \min_{j=1,2,\ldots,n}(X_{ij})}{\max_{j=1,2,\ldots,n}(X_{ij}) - \min_{j=1,2,\ldots,n}(X_{ij})}$. Where X_{ij} denotes the jth observation of $\boldsymbol{X}_i$, $j = 1, \ldots n$. In this analysis, we focused on gene expressions and treated them as the $\boldsymbol{G}$ variable. After data cleaning, we had $n = 394$ (sample size) and $N = 19,172$ (number of gene expressions).

We implemented the proposed selection approach to pathway hsa00510 N-Glycan biosynthesis. Based on the KEGG pathway database (http://www.genome.jp/kegg/pathway.html), pathway hsa00510 is mapped to 44 gene expressions in our data. Due to model constrain of VMICM, the first loading parameter must be a non-zero positive number ($\beta_1 > 0$). Hence, we fit the proposed model three times with age, aldehyde oxydase, and liver triglyceride being the first loading covariate. Gene expression $PGGT1B$ was selected as varying coefficient predictor in all three models. Gene expression $B4GALT3$ and $B3GNT3$ were selected as constant coefficient predictors in two models (age and aldehyde oxydase being the first loading covariate).

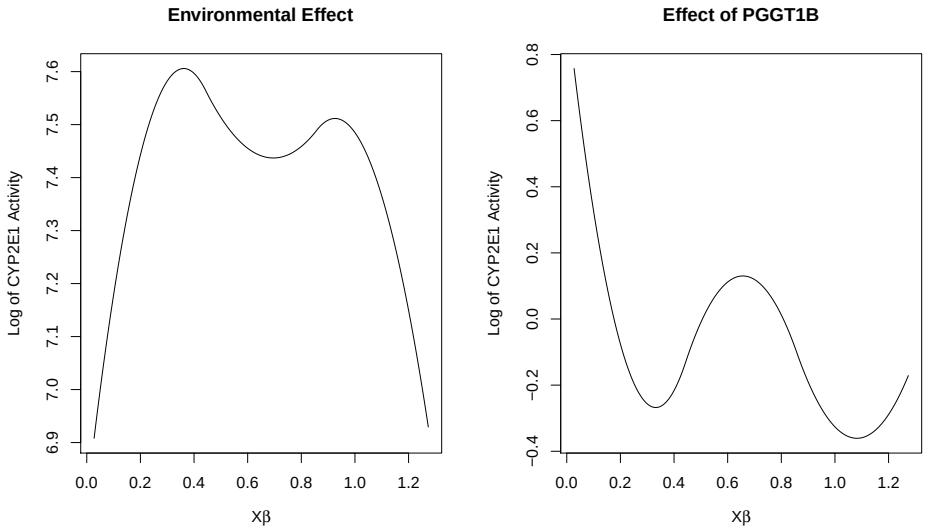

Figure 2: Plot of the varying coefficient effect for gene expression $PGGT1B$.

Table 5: The estimated loading parameters

Aldehyde Oxydase	Age	Liver Triglyceride
0.879	0.477	0

Figure 2 presents the plot of the marginal environmental effect and the effect of gene $PGGT1B$ on log of CYP2E1 activity with aldehyde oxydase being the first loading covariate. With the increase of loading index $\boldsymbol{X\beta}$, marginal environmental effect ($m_0(\cdot)$) first increased from 6.9 to 7.6, then it fluctuated around 7.5 before decreasing sharply to 7.0. For gene expression $PGGT1B$, its effect on the log of $CYP2E1$ activity first decreased sharply from 0.8 to -0.2, then it fluctuated around -0.1 for the remainder of the index. This suggests that the G×E effect of gene expression $PGGT1B$ mainly exists in the lower quantile of aldehyde oxydase and age. With aldehyde oxydase being the first loading covariate, the constant effects of gene expression $B4GALT3$ and $B3GNT3$ were -0.0618 and -0.0624 respectively. These suggest that the effect of gene expression $B4GALT3$ and $B3GNT3$ on the log of

$CYP2E1$ activity does not interact with environmental factors.

Table 5 presents the selection result of environmental factors with aldehyde oxydase being the first loading covariate. The model selected aldehyde oxydase and age as significant environmental factors, their estimation were 0.879 and 0.477 respectively. Base on their value, aldehyde oxydase have a stronger effect than age.

This result demonstrates the unique ability of the proposed method to capture the non-linear interaction between a mixture of environmental variants and genetic variables. However, further biological investigation is needed to confirm this finding.

2.6 Discussion

VMICM is a promising candidate to model non-linear interaction between multiple genes and multiple environments as a whole. It combines multiple exposure variables $\boldsymbol{X}$ into a single index $\boldsymbol{X\beta}$, hence can reduce model dimension and alleviate the curse of dimensionality. In this paper, we developed a three stage variable selection approach for VMICM model. Our goal was to identify varying, constant and zero effect that interacted with a gene. In the meantime, we also selected important exposure variables. Rather than modeling the G×E effect for each X variable separately, our approach can model the joint effect of the environmental factors ($\boldsymbol{X}$) as a whole, then identify how different genes ($\boldsymbol{G}$) interacted with the environmental mixture to affect the phenotype Y. Biologically speaking, our approach is attractive since it offers an alternative strategy to look for G×E interaction, and our model is flexible to detect any non-linear interactions. In addition, we theoretically evaluated the selection consistency of the variable selection. Both simulation and real data analysis demonstrated the utility of the proposed method.

In our model setup, the covariates $\boldsymbol{X}$ were assumed to be continuous. This is due to the fact that the index $\boldsymbol{X\beta}$ has to be continuous in order to model the nonlinear function. In real applications, environmental variables can be discrete such as smoking status, gender and ethnicity. To accommodate the presence of discrete factors, the model VMICM could be generalized to partial-linear VMICM: $\boldsymbol{Y} = \boldsymbol{Z\alpha} + \sum_{k=0}^{p} m_k(\boldsymbol{X\beta})\boldsymbol{G}_k + \boldsymbol{\epsilon}$ where $\boldsymbol{Z}$ is the discrete covariates of dimension r and $\boldsymbol{\alpha}$ is its effect on the response. Our variable selection approach can be modified slightly to perform selection of non-parametric function and the parametric component simultaneously. More specifically, the objected function (2.5) is modified as

$$\begin{aligned} Q(\boldsymbol{\beta},\boldsymbol{\gamma}) =& \sum_{i=1}^{n}\{Y_i - \boldsymbol{Z}_i\boldsymbol{\alpha} - \sum_{k=0}^{p}[\gamma_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{X\beta})\boldsymbol{\gamma}_{k*}]G_{ik}\}^2 + n\sum_{k=1}^{p} p_{\lambda_1}(||\boldsymbol{\gamma}_{k*}||_1) \\ &+ n\sum_{j=1}^{r} p_{\lambda_z}(|\alpha_j|) + n\sum_{k=1}^{p} p_{\lambda_2}(|\gamma_{k1}|)I(||\boldsymbol{\gamma}_{k*}||_1 = 0) + n\sum_{d=2}^{q} p_{\lambda_3}(|\beta_d|). \end{aligned}$$

And the objective function in step 1 is modified as

$$\begin{aligned} Q_1(\boldsymbol{\gamma}|\lambda_z,\lambda_1,\hat{\boldsymbol{\beta}}^{(0)}) =& \sum_{i=1}^{n}\{Y_i - \boldsymbol{Z}_i\boldsymbol{\alpha} - \sum_{k=0}^{p}[\gamma_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})\boldsymbol{\gamma}_{k*}]G_{ik}\}^2 \\ &+ n\sum_{k=1}^{p} p_{\lambda_1}(||\boldsymbol{\gamma}_{k*}||_2) + n\sum_{j=1}^{r} p_{\lambda_z}(|\alpha_j|). \end{aligned}$$

In this paper, we discussed the variable selection approach for VMICM with a continuous response variable. However, many response variables are measured on a binary scale such as the presence of a certain disease in humans. It is natural to extend the current selection approach to a generalized VIMCM framework, which will be investigated in chapter 3.

In our model formulation, we assumed different index coefficients share the common loading parameters, i.e., $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_p = \boldsymbol{\beta}$. From a practical point of view, allowing different loading parameters makes perfect sense. However, such a model imposes theoretical challenges when evaluating the theoretical properties such as the selection consistency. This is because that the loading coefficients for the kth index coefficients are not identifiable when $m_k(u) \notin V$. When a coefficient function does not vary, $\boldsymbol{\beta}_k$ does not exists. Thus, the selection consistency for $\boldsymbol{\beta}_k$ does not exists. For this reason, we imposed the same loading parameters for all the index coefficient functions.

In addition to the application to G×E study, our model has many applications in other fields where a potential nonlinear varying effect exists. Our method enriches the catalog of variable selection. It contributes to the methodology development of variable selection in theory and to the application of G×E study in practice.

Chapter 3

Variable Selection for Generalized VMICM

3.1 Introduction

There has been a growing interest in identifying gene-environment (G×E) interaction in scientific literatures. Ottman(1996) defined gene-environment interaction as "a different effect of a genotype on disease risk in persons with different environmental exposures". Traditionally, G×E interactions were investigated based on a single environmental factor, since introduction of several environmental factors will increase model dimension exponentially, which could potentially lead to biased estimation and large standard error (curse of dimensionality). However, more and more epidemiological studies revealed that disease risk can be modified by simultaneously exposure to several environmental factors (Carpenter et al. (2002), Sexton and Hattis (2007)). Further, little was known about how multiple environmental factors as a whole could interact with genetic factors to affect the response variable. Any investigation in this area could shed some light into the disease etiology and offer prospects for future disease prevention.

In chapter 2, we proposed to use varying multi-index coefficient model of the form $\boldsymbol{Y} = \sum_{k=0}^{p} m_k(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{G}_k + \boldsymbol{\epsilon}$ to model the non-linear gene-environmental interaction. However,

VMICM can only model continuous phenotype. In this chapter, we generalized VMICM to model data with binary response variables. Consider the generalized varying multi-index coefficient model (gVMICM)

$$log\Big(\frac{P(\boldsymbol{Y}=1|\boldsymbol{X},\boldsymbol{G})}{P(\boldsymbol{Y}=0|\boldsymbol{X},\boldsymbol{G})}\Big) = \sum_{k=0}^{p} m_k(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{G}_k \tag{3.1}$$

where $m_k(\cdot), k = 0, 1, \ldots, p$ represents the gene effect of $\boldsymbol{G}_k$, and is modeled as a smoothed non-linear non-parametric function modulated by the loading index $\boldsymbol{X}\boldsymbol{\beta}$. Its unique structure allows us to capture the interaction effect between genetic factors and a mixture of environmental factors. Further, model (3.1) is very flexible to cover a wide range of models. For instance, if $q = 1$ and $\beta = 1$ then it reduces to a generalized varying coefficient mode; if $p = 1$ and $\boldsymbol{G} = \mathbf{1}$ then it becomes a standard generalized single index model.

Variable selection has been a popular topic in modern statistics literature. In the past, people often implemented hypothesis testing, forward/backward selection combined with AIC or BIC. Nevertheless, with the rise of big data, we were able to collect hundreds of thousands of variables at the same time. Traditional methods are no longer feasible because of exponentially increasing computation time and unstable estimation due to increasing collinearity. Recently, variable selection via penalized regression has become fairly popular. The idea is to add a penalty term to the optimization function. With different penalty functions, the penalized estimator could possess different properties. Fan and Li (2001) proposed three important criteria for penalized estimator: sparsity, unbiasedness and continuity. They also characterized oracle property meaning the model performs as well as if the true sub-model is known in advance. It has become the standard for new penalized estimator. A few examples of the penalized function would be Bridge regression

(Frank and Friedman (1993)), least absolute shrinkage and selection operator (LASSO) (Tibshirani (1996)), its extension adaptive LASSO (Zou (2006)), smoothly clipped absolute deviation (SCAD) (Fan and Li (2001)), and minimax concave penalty (MCP) (Zhang (2010)). Although LASSO enjoys simple formulation and efficient algorithm (LARS), it does not possess oracle property. On the other side, Adaptive LASSO, SCAD, MCP all posses oracle property and we decided to implement MCP in our model.

Due to the complexity of model (3.1), variable selection presents unique challenge, specifically, the nonlinear and non-parametric structure of function $m_k(\cdot)$ and its unknown loading parameter $\boldsymbol{\beta}$. In chapter 2, we proposed a three stage variable selection approach for VMICM. The model classified the non-parametric gene effect into varying, constant and zero. It also selected non-zero loading parameters. In this chapter, we extended such approach to gVMICM. In stead of penalizing squared error loss for VMICM, we decided to implement the penalized log-likelihood method in this model.

The rest of this chapter is organized as follows: section 3.2 introduced the proposed variable selection method, formulation of the penalized log-likelihood, three step iterative optimization approach, and selection of tuning parameters. In section 3.3, we discussed the asymptotic properties of the proposed method. And we evaluated the performance of our method via several simulations in section 3.4. Utility of our model was demonstrated with a Type II diabetes data set in section 3.5, followed by discussion and future work.

3.2 Variable Selection for gVMICM

3.2.1 Model Setup

Consider the following gVMICM model

$$log(\frac{P(\boldsymbol{Y}=1|\boldsymbol{X},\boldsymbol{G})}{P(\boldsymbol{Y}=0|\boldsymbol{X},\boldsymbol{G})}) = \sum_{k=0}^{p} m_k(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{G}_k$$

where $\boldsymbol{Y}_{n\times 1} = (Y_1, Y_2, \ldots, Y_n)^T$ is a binary response variable measured over n subjects; $m_k(\cdot), k = 0, 1, \ldots, p$ are $p+1$ unknown smooth non-linear functions; $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_q)$ is a matrix of dimension $n\times q$ representing continuous environmental variables; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$ is the loading parameter of dimension q; $\boldsymbol{G}_{n\times(p+1)} = (\boldsymbol{G}_0, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_p)$, $\boldsymbol{G}_0 = (1, \ldots, 1)^T$ and $\boldsymbol{G}_k = (G_{1k}, G_{2k}, \ldots, G_{nk})^T$ is the $k-th$ genetic factor. For $k \neq 0$, $m_k(\boldsymbol{X}\boldsymbol{\beta})$ is the effect of $\boldsymbol{G}_k$ on the response on a log odda scale; $m_0(\boldsymbol{X}\boldsymbol{\beta})$ is the intercept term measuring the marginal environmental effect.

For ease of notation, we denoted $\boldsymbol{\mu} = \sum_{k=0}^{p} m_k(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{G}_k$. Thus, model (3.1) can be rewritten as

$$log(\frac{P(\boldsymbol{Y}=1|\boldsymbol{X},\boldsymbol{G})}{P(\boldsymbol{Y}=0|\boldsymbol{X},\boldsymbol{G})}) = \boldsymbol{\mu}. \tag{3.2}$$

3.2.2 Estimation Method

Our goal was to select and estimate unknown functions $\{m_k(\cdot)\}_{k=0,1,\ldots,p}$ and their unknown loading parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$. For the sake of identifiability, we assumed $||\boldsymbol{\beta}||_2 = 1$ and $\beta_1 > 0$, and $m_k(\cdot)$ cannot has the form of $m_j(\boldsymbol{u}) = \boldsymbol{\alpha}^T\boldsymbol{u}\boldsymbol{\beta}^T\boldsymbol{u} + \boldsymbol{\gamma}^T\boldsymbol{u} + c$.

We first approximated the non-parametric function $m_k(u)$ with B-spline basis functions. Without loss of generality, we assumed $u \in [0, 1]$, and denoted K to be the number of internal

knots and h to be the degree of B-spline basis function. So $h = 1$ represents linear splines; $h = 2$ represents quadratic splines; and $h = 3$ represents cubic splines. From standard B-spline theory, we denoted $u_1, u_2, ..., u_K$ to be the interior knots satisfying $0 = u_0 =< u_1 < u_2 < .... < u_K < u_{K+1} = 1$. Let I_{n_t} be left closed, right opened interval $[u_{t-1}, u_t)$ for $1 \leq t \leq K$, and $I_{n_{K+1}}$ be closed interval $[u_K, u_{K+1}]$. Let $\mathscr{F}$ to be a collection of functions f defined on $[0, 1]$ satisfying: (i) the restriction of f to I_{n_t} is a polynomial of degree h or less for $1 \leq j \leq K + 1$; (ii) f is $h - 1$ times continuously differentiable on $[0, 1]$. Let $L = K + h + 1$, then by Schumaker (1981)(Schumaker (2007)), we normalized B-spline basis function $\widetilde{\boldsymbol{B}}(u) = (\widetilde{\boldsymbol{B}}_1(u), \widetilde{\boldsymbol{B}}_2(u), \ldots, \widetilde{\boldsymbol{B}}_L(u))$ for $\mathscr{F}$. And there exists a linear transformation matrix $\boldsymbol{\Pi}$, such that $\boldsymbol{\Pi}\widetilde{\boldsymbol{B}}(u) = (\mathbf{1}, \bar{\boldsymbol{B}}(u)) = (\mathbf{1}, \boldsymbol{B}_2(u), \boldsymbol{B}_3(u), \ldots, \boldsymbol{B}_L(u)) = \boldsymbol{B}(u)$ where each component of $\bar{\boldsymbol{B}}(u)$ is a function of u. Hence, for $0 \leq k \leq p$, we approximated $m_k(u)$ by

$$m_k(u) \approx (1, B_2(u), \cdots, B_L(u)) * (\gamma_{k1}, \gamma_{k2}, \cdots, \gamma_{kL})^T = \boldsymbol{B}(u)\boldsymbol{\gamma}_k = \gamma_{k1} + \bar{\boldsymbol{B}}(u)\boldsymbol{\gamma}_{k*} \quad (3.3)$$

where $\boldsymbol{\gamma}_{k*} = (\gamma_{k2}, \gamma_{k3}, \ldots, \gamma_{kL})^T$ and $\boldsymbol{\gamma}_k = (\gamma_{k1}, \boldsymbol{\gamma}_{k*}^T)^T$. We approximated $\boldsymbol{\mu}$ with $\boldsymbol{\mu}^B$:

$$\boldsymbol{\mu}^B = \sum_{k=0}^{p} [\gamma_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{\gamma}_{k*}]\boldsymbol{G}_k \quad (3.4)$$

where γ_{k1}, $k \geq 1$ represents the main genetic effect for the kth variant and $\boldsymbol{\gamma}_{k*}$, $k \geq 1$ represents the G×E interaction effect between the kth variant and a mixture of the environmental variables.

With this new representation, the selection of the non-parametric function $m_k(\cdot)$ was transformed to the selection of its B-splines coefficients $\boldsymbol{\gamma} = \{\gamma_{k1}, \boldsymbol{\gamma}_{k*}\}_{k=0,1,...,p}$. The transformation $\boldsymbol{\Pi}$ allows us to separate constant effect of $\boldsymbol{G}_k$ from its varying effect. More

specifically, (1) if $||\boldsymbol{\gamma}_{k*}||_2 \neq 0$, then there exists G×E effect; (2) if $||\boldsymbol{\gamma}_{k*}||_2 = 0$ and $|\gamma_{k1}| \neq 0$, then there only exists main genetic effect; and (3) if $||\boldsymbol{\gamma}_{k*}||_2 = 0$ and $|\gamma_{k1}| = 0$, then $\boldsymbol{G}_k$ has no effect at all. This transformation is the key step to separate different genetic effects.

Given the binary response, the log-likelihood function is defined as:

$$l(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i \mu_i^B - log(1 + e^{\mu_i^B}))$$

where μ_i^B is the ith subject of $\boldsymbol{\mu}^B$. We defined the following penalized log-likelihood objective function,

$$\begin{aligned} M(\boldsymbol{\beta}, \boldsymbol{\gamma}) = & \sum_{i=1}^{n} (Y_i \mu_i^B - log(1 + e^{\mu_i^B})) - n \sum_{k=1}^{p} p_{\lambda_1}(||\boldsymbol{\gamma}_{k*}||_2) \\ & - n \sum_{k=1}^{p} p_{\lambda_2}(|\gamma_{k1}|) I(||\boldsymbol{\gamma}_{k*}||_2 = 0) - n \sum_{d=2}^{q} p_{\lambda_3}(|\beta_d|) \end{aligned} \tag{3.5}$$

where $p_{\lambda_1}(\cdot), p_{\lambda_2}(\cdot), p_{\lambda_3}(\cdot)$ are the penalty functions for $\boldsymbol{\gamma}_{k*}$, γ_{k1} and $\boldsymbol{\beta}$, respectively; $I(\cdot)$ is an indicator function. The construction of the penalty functions in (3.5) implies that there is a natural order when selecting the effect of $\boldsymbol{G}_k$. First, the model selects $\boldsymbol{G}_k$ with varying effect. If $\boldsymbol{G}_k$ does not have a varying effect, the model penalizes the constant effects of $\boldsymbol{G}_k$. We did not penalize β_1 due to the model constraint. For the penalty function, we adopted the MCP penalty function proposed by Zhang (Zhang (2010)), i.e., $p(x, \lambda) = \lambda \int_0^x (1 - \frac{s}{\tau\lambda})_+ ds$ with regularization parameters $\tau > 0$ and $\lambda > 0$.

3.2.3 Computational algorithm

To optimize function (3.5), we followed the idea proposed in chapter 2 and adopted a 3 step interactive approach.

Step 1: Given an initial value for $\boldsymbol{\beta}$, denoted as $\hat{\boldsymbol{\beta}}^{(0)}$, we obtained the 1st step estimation of $\boldsymbol{\gamma}$, denoted as $\hat{\boldsymbol{\gamma}}^{(1)} = \{\hat{\gamma}_{k1}^{(1)}, \hat{\boldsymbol{\gamma}}_{k*}^{(1)T}\}_{k=0,1,\ldots,p}^{T}$ by following a group penalized regression, i.e.,

$$\hat{\boldsymbol{\gamma}}^{(1)} = \arg\max_{\boldsymbol{\gamma}} M_1(\boldsymbol{\gamma}|\lambda_1, \hat{\boldsymbol{\beta}}^{(0)})$$

where

$$M_1(\boldsymbol{\gamma}|\lambda_1, \hat{\boldsymbol{\beta}}^{(0)}) = \sum_{i=1}^{n}(Y_i\mu_i^B - log(1 + e^{\mu_i^B})) - n\sum_{k=1}^{p} p_{\lambda_1}(||\boldsymbol{\gamma}_{k*}||_2).$$

Step 1 classified $m_k(\cdot), k = 1, \ldots, p$ into two categories: varying(V) or non-varying(NV). That is, $m_k(\cdot) \in V$ if $||\hat{\boldsymbol{\gamma}}_{k*}^{(1)}||_2 > 0$ and $m_k(\cdot) \in NV$ if $||\hat{\boldsymbol{\gamma}}_{k*}^{(1)}||_2 = 0$.

Step 2: In this step, our interest was to select γ_{k1} given $\hat{\boldsymbol{\gamma}}_{k*}^{(1)} = 0$. We further selected constant coefficient from the non-varying coefficient obtained in step 1. We penalized γ_{k1} only when $||\hat{\boldsymbol{\gamma}}_{k*}^{(1)}||_2 = 0,\ k = 1, 2, \ldots, p$, and no penalty was applied to γ_{01}. We excluded $\boldsymbol{\gamma}_{k*}$ from the model if $||\hat{\boldsymbol{\gamma}}_{k*}^{(1)}||_2 = 0$ in step 1, i.e. $\hat{\boldsymbol{\gamma}}_{k*}^{(2)} = \mathbf{0}$. We obtained our step 2 estimator $\hat{\boldsymbol{\gamma}}^{(2)} = \{(\hat{\gamma}_{k1}^{(2)}, \hat{\boldsymbol{\gamma}}_{k*}^{(2)})_{k\in V}, (\hat{\gamma}_{k1}^{(2)})_{k\in NV}\}$ via penalized regression

$$\hat{\boldsymbol{\gamma}}^{(2)} = \arg\max_{\boldsymbol{\gamma}} M_2(\boldsymbol{\gamma}|\lambda_2, \boldsymbol{\beta}^{(0)}, \hat{\boldsymbol{\gamma}}^{(1)})$$

where

$$M_2(\boldsymbol{\gamma}|\lambda_2, \boldsymbol{\beta}^{(0)}, \hat{\boldsymbol{\gamma}}^{(1)}) = \sum_{i=1}^{n}(Y_i\mu_i^{B^{(2)}} - log(1 + e^{\mu_i^{B^{(2)}}})) - n\sum_{k=1}^{p} p_{\lambda_2}(|\gamma_{k1}^{(2)}|)I(||\hat{\boldsymbol{\gamma}}_{k*}^{(1)}||_2 = 0).$$

$\mu_i^{B^{(2)}}$ is the ith element of $\boldsymbol{\mu}^{B^{(2)}}$ with $\boldsymbol{\mu}^{B^{(2)}} = \sum_{k\in V}[\gamma_{k1}^{(2)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta}^{(0)})\boldsymbol{\gamma}_{k*}^{(2)}]\boldsymbol{G}_k + \sum_{k\in NV}\gamma_{k1}^{(2)}\boldsymbol{G}_k$.

After step 2, we obtained our estimators of the B-splines coefficients $\boldsymbol{\gamma}$ for given $\hat{\boldsymbol{\beta}}^{(0)}$

and classified $m_k(\cdot)$ $k = 1, \ldots, p$ into varying, constant or zero effects. The next step is to update and select the loading parameter $\boldsymbol{\beta}$ given $\hat{\boldsymbol{\gamma}}^{(2)}$.

Step 3: We obtained $\hat{\boldsymbol{\beta}}$ via penalized regression

$$\hat{\boldsymbol{\beta}} = \arg \max_{||\boldsymbol{\beta}||_2=1} M_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}^{(2)})$$

where

$$M_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}^{(2)}) = \sum_{i=1}^{n}(Y_i \mu_i^{B^{(3)}} - log(1 + e^{\mu_i^{B^{(3)}}})) - n \sum_{d=2}^{q} p_{\lambda_3}(|\beta_d|).$$

$\mu_i^{B^{(3)}}$ is the ith element of $\boldsymbol{\mu}^{B^{(3)}}$ with $\boldsymbol{\mu}^{B^{(3)}} = \sum_{k=0}^{p}[\hat{\gamma}_{k1}^{(2)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}_{k*}^{(2)}]\boldsymbol{G}_k$.

Once we have the updated $\hat{\boldsymbol{\beta}}$, we set $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}$, then iterate step 1 through 3 until convergence.

Remark: For step 1 and 2, we implemented the block coordinate descent algorithm for group penalty. For step 3, we developed our algorithm based on the idea of the local quadratic approximation (LQA) proposed by Fan and Li(2001). For more detail, please refer to the Appendices. Next we discuss details about selecting the tuning parameters $\lambda_1, \lambda_2, \lambda_3$, order h and the number of interior knots K for the B-spline approximation, as well as a reasonable initial value for $\boldsymbol{\beta}$.

3.2.4 Selection of Parameters

3.2.4.1 Selection of tuning parameters $\lambda_1, \lambda_2, \lambda_3$

We proposed to use Bayesian Information Criterion (BIC) (Schwarz (1978)) to select the tuning parameters.

Step 1: We selected λ_1 as the minimizer of

$$BIC(\lambda_1) = -2l(\hat{\boldsymbol{\gamma}}^{(1)}_{\lambda_1}, \hat{\boldsymbol{\beta}}^{(0)}) + log(n) * df_{\lambda_1}$$

where $\hat{\boldsymbol{\gamma}}^{(1)}_{\lambda_1}$ is the minimizer of $M_1(\boldsymbol{\gamma}|\lambda_1, \hat{\boldsymbol{\beta}}^{(0)})$ defined above; $\hat{\boldsymbol{\beta}}^{(0)}$ is chosen as the estimator from the previous iteration; and df_{λ_1} is defined as the total number of non-zero coefficients if λ_1 is the penalized parameter.

Step 2: We selected λ_2 as the minimizer of

$$BIC(\lambda_2) = -2l(\hat{\boldsymbol{\gamma}}^{(2)}_{\lambda_2}, \hat{\boldsymbol{\beta}}^{(0)}) + log(n) * df_{\lambda_2}$$

where $\hat{\boldsymbol{\gamma}}^{(2)}_{\lambda_2}$ is the minimizer of $M_2(\boldsymbol{\gamma}|\lambda_2, \boldsymbol{\beta}^{(0)}, \hat{\boldsymbol{\gamma}}^{(1)})$ defined above; $\hat{\boldsymbol{\beta}}^{(0)}$ is chosen as the estimator from the previous iteration; and df_{λ_2} is defined as the total number of non-zero coefficients if λ_2 is the penalized parameter.

Step 3: We selected λ_3 as the minimizer of

$$BIC(\lambda_3) = -2l(\hat{\boldsymbol{\gamma}}^{(2)}, \hat{\boldsymbol{\beta}}_{\lambda_3}) + log(n) * df_{\lambda_3}$$

where $\hat{\boldsymbol{\beta}}_{\lambda_3}$ is the minimizer of $M_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}^{(2)})$ defined above; $\hat{\boldsymbol{\gamma}}^{(2)}$ is the minimizer of the B-spline coefficient from step 2; and df_{λ_3} is defined as the total number of non-zero $\boldsymbol{\beta}$ if λ_3 is the penalized parameter.

The penalized parameter $\lambda_1, \lambda_2, \lambda_3$ were chosen over a grid of exponentially decreasing values with the minimum to be 1E-3. The maximum of $\lambda_1, \lambda_2, \lambda_3$ were set to be the minimum values such that all of the penalized estimators are zeros. The number of grid to be searched was set as 100.

3.2.4.2 Selection of order h and number of interior knots K

In theory, it might be appealing to allow different K and h for different non-parametric functions $m_k(\cdot)$. In practice, it would be computationally infeasible. For computational purpose, we let all $m_k(\cdot)$ share the same K and h. Since h is the order of the B-spline basis function, higher degree implies more complicated interactions between environmental factors and genetic predictors, thus, more difficult to interpret. For this reason, we searched optimal h over the set $h \in \{2, 3, 4\}$. For the interior knots K, only when $K = O_p(n^{\frac{1}{2r+1}})$ (n is the sample size and r is the smoothness of the nonparametric function $m_k(\cdot)$ and $r > 2$), our selection approach possesses oracle properties. Thus, we searched optimal K in the neighborhood of $n^{\frac{1}{2r+1}}$, denoted by $\mathscr{K}$. In our simulation, we chosed $\mathscr{K} = \{2, 3, 4, 5\}$.

The knots K and order h were then selected by fitting the following intercept only model with the B-spline approximation and Newton-Raphson algorithm.

$$log(\frac{P(\boldsymbol{Y} = 1|\boldsymbol{X}, \boldsymbol{G})}{P(\boldsymbol{Y} = 0|\boldsymbol{X}, \boldsymbol{G})}) = m_0(\boldsymbol{X}\boldsymbol{\beta}). \tag{3.6}$$

Denote the estimated spline coefficients as $(\hat{\gamma}_{01}, \hat{\boldsymbol{\gamma}}_{0*})$ and the loading parameters as $\hat{\boldsymbol{\beta}}$, and let $\hat{m}_0(\boldsymbol{X}\hat{\boldsymbol{\beta}}) = \hat{\gamma}_{01} + \bar{B}(\boldsymbol{X}\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_{0*}$. We searched the optimal K and h over a grid $K \in \mathscr{K}$ and $h \in \{2, 3, 4\}$. Optimal K and h are defined as the K and h that minimized $log\big(\boldsymbol{Y} * \hat{m}_0(\boldsymbol{X}\hat{\boldsymbol{\beta}}) - log(1 + e^{\hat{m}_0(\boldsymbol{X}\hat{\boldsymbol{\beta}})})\big) + \frac{log(n)}{n}(K + h + 1)$.

3.2.5 Choosing the initial values

To start the iterative optimization process described above, a reasonably good initial value of $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{initial}$ is essential. In many literature, $\boldsymbol{\beta}^{initial}$ is set to be $(1, 0, \ldots, 0)^T$ or $(\frac{1}{\sqrt{q}}, \ldots, \frac{1}{\sqrt{q}})^T$. However, neither works well in our simulations. Hence, we set the initial value of $\boldsymbol{\beta}$ as the

byproduct of selecting the optimal K and h by fitting the intercept only model in (3.6).

3.3 Theoretical Properties

Let $\boldsymbol{\beta}^0$ and $m_k^0(\cdot),\ k = 0, 1, \ldots, p$ be the true value of $\boldsymbol{\beta}$ and $m_k(\cdot)$, respectively. Denote $\boldsymbol{\gamma}^0$ to be the true value of the B-spline coefficients $\boldsymbol{\gamma}$. Without loss of generality, we assumed $\beta_l^0 \neq 0$ for $l = 1, \ldots s$, $\beta_l^0 = 0$ for $l = s+1, \ldots q$; $m_k^0(\cdot)$ has varying coefficients for $k = 0, 1, \ldots, v$, $m_k^0(\cdot)$ has non-zero constant coefficients for $k = v+1, \ldots, c$ and $m_k^0(\cdot)$ is zero for $k = c+1, \ldots, p$. The following theorems give the consistency of the penalized log-likelihood estimators.

Theorem 3.3.1. *Assume the regulatory conditions (A1)-(A7) in Appendices hold and the number of knots* $K = O_p(n^{1/(2r+1)})$. *Then*

(i) $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0|| = O_p(n^{-r/(2r+1)} + a_n)$;

(ii) $||\hat{m}_k(\cdot) - m_k^0(\cdot)|| = O_p(n^{-r/(2r+1)} + a_n),\ k = 1, \ldots, q$

where

$$a_n = \max_{k,j,l}\{p'_{\lambda_1}(||\boldsymbol{\gamma}_{k*}^0||_2), p'_{\lambda_2}(|\boldsymbol{\gamma}_{j1}^0|), p'_{\lambda_3}(|\beta_l^0|), \boldsymbol{\gamma}_{k*}^0 \neq 0, \gamma_{j1} \neq 0, \beta_l^0 \neq 0\}.$$

$k, j = 1, \ldots, p, l = 2, \ldots, q$. r *is defined in condition (A2) in Appendices.* $p'_\lambda(\cdot)$ *denotes the first order derivative of the penalty function* $p_\lambda(\cdot)$.

Theorem 3.3.2. *Assume the regularity conditions(A1)-(A7) in Appendices hold and the number of knots* $K = O_p(n^{1/(2r+1)})$. *Let*

$$\lambda_{max} = \max\{\lambda_1, \lambda_2, \lambda_3\}, \quad \lambda_{min} = \min\{\lambda_1, \lambda_2, \lambda_3\}.$$

Suppose $\lambda_{max} \to 0$ *and* $n^{r/(2r+1)}\lambda_{min} \to \infty$ *when* $n \to \infty$. *Then with probability approaching 1,* $\hat{\boldsymbol{\beta}}$ *and* $\hat{m}_k(\cdot)$ *must satisfy:*

(i) $\hat{\beta}_j = 0$ *for* $j = s+1, \ldots, q$

(ii) $\hat{m}_k(\cdot) = c_k$ *for* $k = v+1, \ldots, c$ *where* c_k *is some non-zero constant*

(iii) $\hat{m}_k(\cdot) = 0$ *for* $k = c+1, \ldots, p$

Theorem 3.3.1 and 3.3.2 show that our penalized likelihood estimators are consistent and possess oracle properties.

3.4 Simulation

We evaluated the finite sample performance of our model via simulations. Its performance was evaluated in several ways: (1) classification accuracy of the $\boldsymbol{m}(\cdot)$ denoted as oracle percentage; (2) IMSE of the estimated m-functions; (3) selection accuracy of $\boldsymbol{\beta}$; and (4) estimation accuracy of $\boldsymbol{\beta}$ (MSE). For all cases, we ran a total of R simulations.

The oracle percentage of $\boldsymbol{m}(\cdot)$ is defined as the percentage of correct classification for varying, constant and zero effects. For instance, if $m_k(\cdot)$ is a varying function and $m_k(\cdot)$ is classified as varying for g times, then the oracle percentage of $m_k(\cdot)$ is calculated as $\frac{g}{R} \times 100\%$. IMSE of $m_k(\cdot)$ is defined as

$$\frac{1}{R}\sum_{r=1}^{R}[\frac{1}{n_{grid}}\sum_{j=1}^{n_{grid}}(\hat{\gamma}_{k1}^{(r)} + \bar{\boldsymbol{B}}(u_j)\hat{\boldsymbol{\gamma}}_{k*}^{(r)} - m_k(u_j))^2]$$

where n_{grid} is the number of grid points; $\hat{\boldsymbol{\gamma}}_{k*}^{(r)}$ and $\hat{\gamma}_{k1}^{(r)}$ are the estimators of the B-spline coefficients for the rth simulation using the proposed estimation approach; $\hat{\boldsymbol{\beta}}^{(r)}$ is the estimator of the loading parameter $\boldsymbol{\beta}$ for the rth simulation; and u_j is taken at the

$j/n_{grid} \times 100\%$ quantile among the range of $\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(r)}$. In the simulations, n_{grid} was set to be 100.

The oracle percentage of $\boldsymbol{\beta}$ is defined as the percentage of correct selection of $\boldsymbol{\beta}$ out of R simulations. For example, if $\beta_d \neq 0$, and β_d is selected as non-zero for g times, then the oracle percentage of β_d is calculated as $\frac{g}{R} \times 100\%$. MSE of β_d is calculated as $\frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_d^{(r)} - \beta_d)^2$, where $\hat{\beta}_d^{(r)}$ is the estimator for β_d in the rth simulation.

The simulation data were generated according to model (3.1). The index matrix $\boldsymbol{X}$ was generated from a $Unif(0,1)$ distribution. For the loading parameter $\boldsymbol{\beta}$, $\beta_1 = \beta_2 = \frac{1}{\sqrt{2}}$ and the rest $\beta_j's$ are zeros. We evaluated the performance of the proposed approach with both continuous and discrete predictors $\boldsymbol{G}$. The continuous $\boldsymbol{G}$ can be gene expressions and the discrete $\boldsymbol{G}$ can be SNP variants.

3.4.1 For continuous G

In the continuous case, the non-parametric functions $m_k(u)$ are defined as follows: $m_0(u) = 2sin(2\pi u)$, $m_1(u) = 2cos(\pi u) + 2$ and $m_2(u) = sin(2\pi u) + cos(\pi u) + 1$ are varying coefficient functions. $m_3(u) = 2$ and $m_4(u) = 2.5$ are non-zero constant coefficients. $m_k(u) = 0$ for $k = 5, \ldots, p$ are zeros. $\boldsymbol{G} \sim N(0,1)$. We conducted 1000 simulations ($R = 1000$) to evaluate the performance of the proposed model under $p = 50, 100$, $q = 5$ and $n = 1000, 2000$.

Table 6 demonstrates the selection and estimation accuracy for $m_k(\cdot)$ with continuous predictors. The left and the right penal corresponds to the case where $p = 50$ and 100, respectively. The upper and lower panel corresponds to the case where $n = 1000$ and 2000, respectively. In all cases, the selection accuracy was closed to 100% for varying, constant and zero effect coefficients. The IMSE of the proposed model was in the order of -1 or -2

for varying and constant effect predictors. With the increase of model dimension p (from 50 to 100), we observed a small increase in model IMSE. With the increase of the sample size n (from 1000 to 2000), there were decreases in both model IMSE and oracle IMSE, which is consistent with the asymptotic property of the proposed model. These suggest that the proposed variable selection approach performs reasonably well in selection and estimation accuracy for the non-parametric functions $m_k(\cdot)$.

Table 6: Selection and prediction accuracy of $m_k(\cdot)$ for continuous $\boldsymbol{G}$

		p = 50			p = 100		
			IMSE			IMSE	
		Oracle %	Model	Oracle	Oracle %	Model	Oracle
n = 1000	$m_0(.)$	100.0%	1.50E-01	1.47E-01	100.0%	1.48E-01	1.31E-01
	$m_1(.)$	99.2%	1.44E-01	2.14E-01	99.7%	1.48E-01	1.66E-01
	$m_2(.)$	99.4%	1.34E-01	1.61E-01	99.7%	1.37E-01	1.43E-01
	$m_3(.)$	100.0%	3.95E-02	3.85E-02	100.0%	4.19E-02	3.33E-02
	$m_4(.)$	99.9%	5.58E-02	5.53E-02	100.0%	5.93E-02	4.86E-02
	Zero	99.1%	1.03E-03	0	99.0%	1.16E-03	0
n = 2000	$m_0(.)$	100.0%	6.85E-02	6.88E-02	100.0%	7.00E-02	7.05E-02
	$m_1(.)$	100.0%	5.17E-02	6.00E-02	100.0%	5.43E-02	6.34E-02
	$m_2(.)$	100.0%	5.46E-02	5.99E-02	100.0%	5.40E-02	5.88E-02
	$m_3(.)$	100.0%	1.40E-02	1.46E-02	100.0%	1.46E-02	1.48E-02
	$m_4(.)$	100.0%	1.79E-02	1.93E-02	100.0%	1.89E-02	1.87E-02
	Zero	99.4%	3.33E-04	0	99.4%	3.14E-04	0

Table 7 presents the selection and prediction accuracy for loading parameter $\boldsymbol{\beta}$. The left and right penal corresponds to the case where $p = 50$ and 100, respectively. The upper and lower panel corresponds to the case where $n = 1000$ and 2000, respectively. For all cases, the selection accuracy for non-zero loading parameters (β_1, β_2) was close to 100%. Their MSE were in order of -2 to -4. For zero loading parameters $(\beta_3, \beta_4, \beta_5)$, the selection accuracy for the case $n = 1000$ was around 97%. When the sample size increases to $n = 2000$, their oracle percentages increased to 99%. Their MSE were in the order of -4 to -5. These

suggest that the algorithm could not shrink estimators to 0 in around 4% of the cases when $n = 1000$, which is a common drawback of the LQA algorithm.

Table 7: Prediction accuracy of $\boldsymbol{\beta}$ for continuous $\boldsymbol{G}$ ($\beta_1 = \beta_2 = \frac{1}{\sqrt{2}}$, $\beta_3 = \beta_4 = \beta_5 = 0$)

		p = 50			p = 100		
			MSE			MSE	
		Oracle %	Model	Oracle	Oracle %	Model	Oracle
n = 1000	β_1	100.0%	7.26E-04	5.91E-04	100.0%	6.75E-04	5.75E-04
	β_2	100.0%	1.36E-02	1.11E-02	100.0%	3.28E-03	5.78E-04
	β_3	97.3%	2.34E-04	0	96.7%	6.22E-04	0
	β_4	97.5%	1.90E-04	0	96.7%	2.72E-04	0
	β_5	96.8%	9.01E-04	0	96.6%	2.95E-04	0
n = 2000	β_1	100.0%	2.76E-04	2.68E-04	100.0%	2.77E-04	2.75E-04
	β_2	100.0%	2.71E-04	2.66E-04	100.0%	2.76E-04	2.74E-04
	β_3	99.1%	5.49E-05	0	99.6%	1.57E-05	0
	β_4	99.6%	2.68E-05	0	99.4%	3.16E-05	0
	β_5	99.3%	4.58E-05	0	99.4%	2.22E-05	0

3.4.2 For discrete G

For the discrete case, each $\boldsymbol{G}$ was simulated from a multinomial distribution assuming a minor allele frequency (MAF) P_a. $\boldsymbol{G}$ took values as $0, 1, 2$ corresponds to aa, Aa, and AA genotype, where a is the minor allele. $\boldsymbol{G}$ was simulated via the following probability distribution function:

$$P(G_{ij} = 0) = P_a^2, P(G_{ij} = 1) = 2 * P_a(1 - P_a), P(G_{ij} = 2) = (1 - P_a)^2$$

where G_{ij} is the j-th variable of the i-th subject, for $i = 1, \ldots, n$, $j = 1, \ldots, p$. The following lists the choice of the coefficient functions $m_k(u)$.

Table 8: Setup for $m_k(u)$

Function	P_a of $\boldsymbol{G}_k$
$m_0(u) = 2sin(2\pi u)$	NA
$m_1(u) = 2cos(\pi u) + 2$	0.5
$m_2(u) = sin(2\pi u) + cos(\pi u) + 1$	0.5
$m_3(u) = 2cos(\pi u) + 2$	0.3
$m_4(u) = sin(2\pi u) + cos(\pi u) + 1$	0.3
$m_5(u) = 2cos(\pi u) + 2$	0.1
$m_6(u) = sin(2\pi u) + cos(\pi u) + 1$	0.1
$m_7(u) = 2$	0.5
$m_8(u) = 2$	0.3
$m_9(u) = 2$	0.1
$m_k(u) = 0, k > 9$	$Unif(0.05, 0.5)$

In this setup, there were varying and constant coefficient functions corresponding to different MAF. The purpose of setting varying MAFs is to check the selection and estimation performance under different MAFs in $\boldsymbol{G}$. For zero coefficients, P_a ranged uniformly from 0.05 to 0.5. $\boldsymbol{X}$ was simulated from a $Unif(0, 1)$ distribution. $\boldsymbol{Y}$ was generated according to model (3.1). We evaluated the performance of our proposed model with 1000 simulations under $p = 50, 100$, $n = 1000, 2000$ and $q = 5$.

Table 9 and presents the selection and estimation accuracy of non-parametric function $m_k(\cdot)$ for discrete $\boldsymbol{G}$. We observed sample size n and minor allele frequency (P_a) of G_k were the determining factors in the performance of the proposed model and we present figure 3 to better visualize their impact. With the increase of sample size (from 1000 to 2000), the performance of the model increased. For example, the oracle percentage for $m_1(\cdot), \ldots, m_4(\cdot)$ increased from around 80% to 100%. The corresponding IMSE decreased significantly. These are consistent with the asymptotic theory of the proposed model. With the decrease of minor allele frequency for $\boldsymbol{G}_k$ (from 0.5 to 0.1), we observed a decrease in performance, both in terms of oracle percentage and model IMSE. For instance, in the case where $n =$

1000, the oracle percentages of $\{m_1(\cdot), m_2(\cdot)\}(P_a = 0.5)$, $\{m_3(\cdot), m_4(\cdot)\}(P_a = 0.3)$, and $\{m_5(\cdot), m_6(\cdot)\}(P_a = 0.1)$ were around 85%, 80%, and 23% respectively. Their IMSE also increased from 0.4 ($P_a = 0.5$) to 0.5 ($P_a = 0.3$), then to 1.3 ($P_a = 0.1$). We believed this is due to the fact that SNP with lower minor allele frequency provides less information. Overall, the proposed variable selection approach works better with larger sample and with common variant SNPs.

Table 9: Selection and estimation accuracy of $m_k(\cdot)$ for discrete $\boldsymbol{G}$

		$p = 50$			$p = 100$		
			IMSE			IMSE	
		Oracle %	Model	Oracle	Oracle %	Model	Oracle
n = 1000	$m_0(.)$	100.0%	1.63E-01	2.28E-01	100.0%	1.75E-01	2.29E-01
	$m_1(.)$	83.4%	4.29E-01	5.61E-01	83.1%	4.44E-01	6.08E-01
	$m_2(.)$	87.7%	3.82E-01	4.38E-01	87.9%	3.69E-01	4.50E-01
	$m_3(.)$	76.0%	5.43E-01	5.87E-01	75.5%	5.50E-01	6.17E-01
	$m_4(.)$	83.3%	4.37E-01	4.19E-01	79.6%	5.11E-01	4.50E-01
	$m_5(.)$	23.2%	1.35E+00	1.05E+00	22.1%	1.45E+00	1.18E+00
	$m_6(.)$	25.7%	1.27E+00	8.93E-01	23.7%	1.31E+00	9.22E-01
	$m_7(.)$	100.0%	5.09E-02	6.22E-02	99.9%	5.05E-02	5.86E-02
	$m_8(.)$	99.9%	6.08E-02	7.50E-02	99.9%	5.83E-02	6.75E-02
	$m_9(.)$	100.0%	1.05E-01	1.24E-01	99.9%	1.13E-01	1.21E-01
	Zero	99.0%	2.74E-03	0	99.2%	2.15E-03	0
n = 2000	$m_0(.)$	100.0%	7.47E-02	8.03E-02	100.0%	7.42E-02	8.35E-02
	$m_1(.)$	100.0%	9.46E-02	1.38E-01	99.9%	1.02E-01	1.49E-01
	$m_2(.)$	100.0%	9.52E-02	1.21E-01	99.9%	9.47E-02	1.21E-01
	$m_3(.)$	100.0%	1.07E-01	1.55E-01	99.8%	1.08E-01	1.50E-01
	$m_4(.)$	99.9%	1.05E-01	1.37E-01	99.8%	1.05E-01	1.30E-01
	$m_5(.)$	77.5%	4.84E-01	2.99E-01	75.1%	5.15E-01	3.08E-01
	$m_6(.)$	77.5%	4.93E-01	2.63E-01	73.5%	5.32E-01	2.53E-01
	$m_7(.)$	100.0%	1.89E-02	2.11E-02	100.0%	1.75E-02	1.97E-02
	$m_8(.)$	100.0%	2.17E-02	2.39E-02	100.0%	2.08E-02	2.33E-02
	$m_9(.)$	100.0%	4.47E-02	4.95E-02	100.0%	4.24E-02	4.56E-02
	Zero	99.3%	9.15E-04	0	99.5%	6.81E-04	0

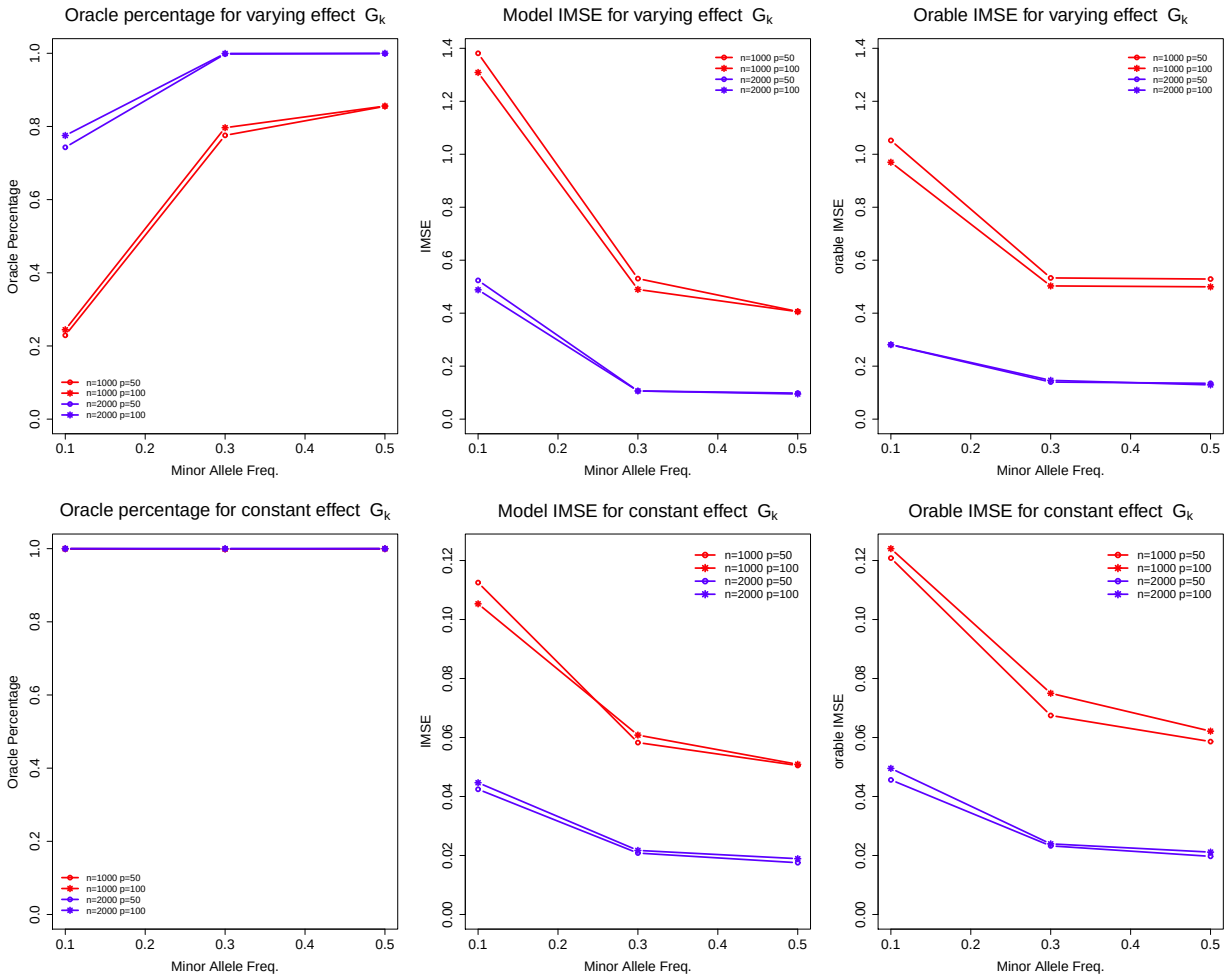

Figure 3: Selection and estimation accuracy of $m_k(\cdot)$ for discrete $\boldsymbol{G}$

Table 10 demonstrates the selection and estimation result of the loading parameter $\boldsymbol{\beta}$. The left and right panel correspond to the case where $p = 50$ and $p = 100$ case respectively and the upper and lower panel correspond to the case where $n = 1000$ and $n = 2000$ respectively. We observed sample size n was the determining factor in model performance. When sample size is large ($n = 2000$), the oracle percentage for non-zero loading covariates (β_1, β_2) was 100% and the oracle percentage for zero loading covariates $(\beta_3, \beta_4, \beta_5)$ was around 99%. The MSE for $\boldsymbol{\beta}$ was in the order of -3 to -5. When sample size is relatively small ($n = 1000$), although the oracle percentage was 100% for non-zero loading covariates,

the oracle percentage for zero loading parameters decreased to around 95%. Comparing between the case $p = 50$ and $p = 100$, we detected a small deterioration in selection accuracy for zero loading parameters with $n = 1000$. This is expected since model performance usually decreases with the increase of complexity. However, we did not detect such difference in performance with larger sample size ($n = 2000$).

Table 10: Estimation accuracy of $\boldsymbol{\beta}$ for discrete $\boldsymbol{G}$ ($\beta_1 = \beta_2 = \frac{1}{\sqrt{2}}$, $\beta_3 = \beta_4 = \beta_5 = 0$)

		$p = 50$			$p = 100$		
			MSE			MSE	
		Oracle %	Model	Oracle	Oracle %	Model	Oracle
n = 1000	β_1	100.0%	6.64E-04	7.28E-04	100.0%	5.84E-04	5.13E-04
	β_2	100.0%	7.37E-03	7.32E-03	100.0%	2.76E-03	3.27E-03
	β_3	95.2%	3.64E-04	0	95.2%	3.73E-04	0
	β_4	97.0%	1.21E-04	0	96.1%	4.18E-04	0
	β_5	96.1%	2.04E-04	0	94.7%	5.46E-04	0
n = 2000	β_1	100.0%	2.34E-04	2.22E-04	100.0%	2.20E-04	2.12E-04
	β_2	100.0%	2.31E-04	2.20E-04	100.0%	2.30E-03	2.37E-03
	β_3	98.9%	5.00E-05	0	98.9%	3.24E-05	0
	β_4	98.7%	5.44E-05	0	98.9%	4.49E-05	0
	β_5	98.9%	4.37E-05	0	99.0%	5.07E-05	0

Based on the simulation results with both continuous and discrete $\boldsymbol{G}$ variables, we inferred the following characteristics of the proposed model. (1) The proposed model performs reasonably well with large sample ($n = 1000$ or 2000). (2) The false positive rate for loading parameter $\boldsymbol{\beta}$ was around 5% when $n = 1000$. This is due to the implementation of LQA algorithm since it cannot shrink zero parameters to zero in some cases. (3) Compared to rare variant SNP ($P_a = 0.1$), the model performs better with common variant SNP ($P_a = 0.3$ or 0.5). We believe this is due to the fact that SNP with lower minor allele frequency provides less information.

3.5 Real Data Application

We demonstrated the utility of our model with a type 2 diabetes data set. The data set contains genotypes (SNPs), environments and phenotypic trait of interest for type 2 diabetes. The data set is consisted of two nested case-control cohort studies: the Nurses Health Study (NHS) and the Health Professional Fellow-up Study (HPFS) from the Gene Environmental Association Studies Consortium (GENVEA). Details of these two cohorts can be found from Coldditz and Hankinson (2005) and Rimm et al. (1991). Originally, the data set contained 3,391 females (NHS) and 2,599 males (HPFS). After data cleaning by removing subjects with unmatch genotypes and phenotypes, SNPs with more than 10% missing rate, MAF < 0.05, and deviation from Hardy-Weinberg equilibrium (p-value < 0.001), the data set contains 655,002 SNPs and a total of 5,865 subjects (2,494 males and 3,371 females), of which there were 2,733 cases and 3,132 controls.

There are 12 continuous covariates including: height, weight, age, alcohol consumption etc. We fit a marginal logistic regression model for all 12 factors. Based on their marginal p-values and the correlation between those factors, we decided to select 5 covariates as the environmental variables in the model. They were total physical activity (X_1 denoted as act), BMI (X_2), alcohol intake (X_3 denoted as alcohol), heme iron intake (X_4 denoted as heme), and glycemic load (X_5 denoted as gl). Based on the location of the SNPs, we mapped all SNPs to all known genes. Then we selected the genes with more than 30 SNPs. As a result, we obtained 2,178 genes. We applied the proposed variable selection approach by fitting one gene at a time to select significant SNPs and identify their effects. Since the first element of the loading parameter $\boldsymbol{\beta}$, β_1 has to be a non-zero positive number for identifiability purpose, we fit the proposed model five times, each time with a different variable as the first

component in $\boldsymbol{X}$. If a SNP is selected as varying or constant effect in all five models, this would suggest a convincing signal. We only considered results with a convincing signal, that is, SNPs showing non-zero effect in all the five fitted models by varying the order of the five environmental variables in $\boldsymbol{X}$.

In total, our model identified 13 varying effect SNPs and 26 constant effect SNPs. Here we presented one of the selected varying SNP as an example. Please refer to the supplemental material for the complete list of selected varying and constant SNPs. Fig 4 presents plots of marginal environmental effect and the interaction effect of a SNP rs6537663 with heme iron intake being the first loading covariate. With the increase of the index, we observed that the marginal effect first decreases then increases, followed by a rapid decrease as the total effect of the five environmental factors increases. For the interaction effect, it fluctuated around 0 as the total effect of the five environmental variables increases, indicating that the SNP is not responding (or insensitive) to the changes of the five environmental variables. As the index $\boldsymbol{X}'\boldsymbol{\beta}$ increases, the SNP reacts to the environmental changes, with a dramatic increased risk on type 2 diabetes after a certain threshold. This estimated effect implies that the genetic sensitivity of the SNP to the total effect of the five environmental variables follows a threshold model. This has practical applications as people are mostly OK with the daily environmental changes including dietary changes. However, such changes cannot pass a certain limit. Otherwise, disease may occur. Table 11 presents the selection and estimation results of the loading parameters $\boldsymbol{\beta}$ with heme iron intake being the first loading covariate. The model selects all loading parameters except the alcohol consumption (alcohol). We also observed that body mass index (bmi) has the largest effect which makes practical sense since BMI is positively associated with type 2 diabetes and is a risk factor for diabetes.

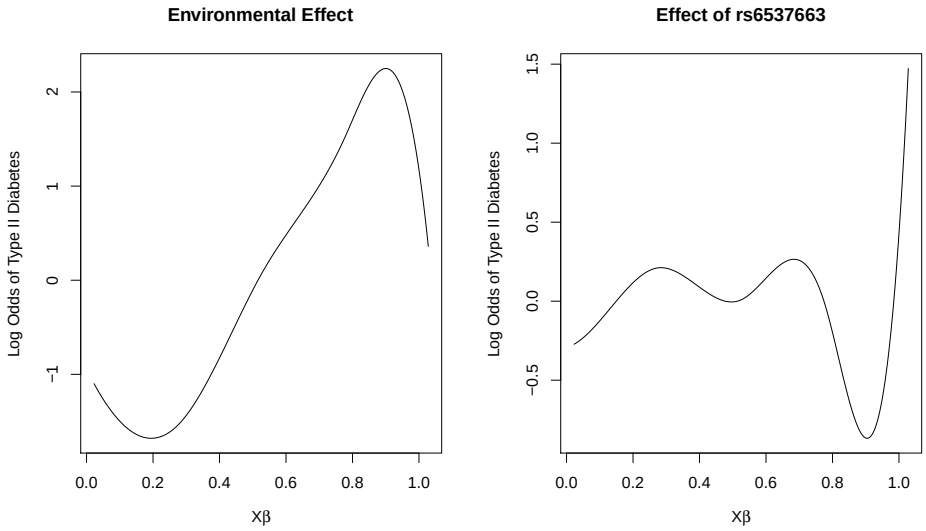

Figure 4: Plot of effects on a log odds scale for SNP rs6537663

Table 11: The estimated effect of $\boldsymbol{\beta}$ for SNP rs6537663.

act	bmi	alcohol	heme	gl
-0.1832	0.954445	0	0.215668	0.094647

Previous study (Sale et al. (2007) and Grant et al. (2006)) suggested that gene *TCF7L2* is associated with type 2 diabetes across multiple populations. Sale et al. (2007) reported strong association between type 2 diabetes and SNPs rs7903146 and rs7901695. Our model also selected SNP rs7901695 showing a constant effect (rs7903146 was not in our data set).

3.6 Discussion

Gene-environment interaction has been one of the major components in genetic association studies. In this paper, we developed a 3 stage iterative variable selection approach for generalized varying multi-index coefficient model with binary responses. Our goal was to

identify varying, constant and zero effects as well as to select non-zero loading parameters. Biologically speaking, our approach is attractive since it offered a novel way to look at G×E interaction from a systems genetics perspective. Our model is flexible to detect non-linear interactions. It should be preferred when the gene effect is non-linearly modified by simultaneous exposure to multiple environmental factors. Statistically speaking, gVMICM treats the effect of multiple variables $\boldsymbol{X}$ as a single index, thus reducing model dimension and alleviating the curse of dimensionality.

In a typical G×E study, there are usually hundreds of thousands of genotype (SNPs) and a couple dozens environmental variables. It is important to reduce the dimension of gene predictors first before apply our method. For example, one could fit a marginal model between the response and every genotype, then select a reasonable number of gene predictors to fit the variable selection model. In human genome, a pathway usually contains a wide range of genes and each gene could contain ten to hundred of SNPs. In this chapter, we implemented the proposed method focusing a gene as a unit to select varying and constant effect SNPs. We could implement such method in every pathway to select significant genes or SNPs of importance. Alternatively, we could apply principle component analysis (PCA) (Jolliffe (2002)) or sparse principle component analysis (sPCA) (Zou et al. (2006)) to summarize the SNP information in a gene or pathway to several principle components (PCs), apply the proposed method to select significant PCs.

In chapter 2, we proposed a 3 stage variable selection approach for VMICM with continuous responses. Then we generalized such approach to binary responses in chapter 3. The regression model applied in chapter 2 is essentially a mean regression model in which one is interested in the conditional mean. When there are outliers or the nature of interest is not on the mean rather than on different quantiles, a quantile regression model might be

a natural choice. For example, when studying the effect of genes on birth weight, people are typically interested in the effect of SNPs on the lower or upper quantile of the birth weight because extremely low or high birth weight may pose potential risk later in life. In such cases, it is essential to extend the proposed method to a quantile regression setup to select important genes modulated by multiple environmental exposures to affect a trait of interest such as birth weight. This will be addressed in chapter 4.

Chapter 4

Variable Selection for Quantile VMICM

4.1 Introduction

Over the past decades, there has been a growing interest in identifying gene-environment (G×E) interaction in scientific research. Gene-environment interaction was defined as a different effect of a genotype on disease risk under different environmental exposures (Ottman (1996)). Traditionally, G×E interactions has been investigated based on a single environmental exposure model. However, more and more epidemiological studies reveal that disease risk can be modified by simultaneously exposure to multiple environmental factors (Carpenter et al. (2002) and Sexton and Hattis (2007)). When multiple environmental factors are analyzed, the model dimension can increase dramatically with the inclusion of the interaction terms, which lead to estimation instability and large standard errors (curse of dimensionality). To ease such burden, a varying multi-index coefficient model(VMICM) (Liu et al. (2016)) can be applied to model the interaction between genetic factors $\boldsymbol{G}_k$ and a mixture of environmental factors $\boldsymbol{X}$ by

$$\boldsymbol{Y} = \sum_{k=0}^{p} m_k(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{G}_k + \boldsymbol{\epsilon}$$

where $m_k(u), k = 0, 1, \ldots, p$ are continuous smooth functions; $\boldsymbol{\beta}$ is q-dimensional loading parameter; and $\boldsymbol{\epsilon}$ is the random error. We have proposed an iterative 3 steps variable selection method for VMICM in chapter 2, which classifies the non-parametric smooth function $m_k(u)$ into three categories: varying, constant and zero. The goal of this paper is to generalize such approach to a quantile regression setting.

The quantile VMICM is a important alternative to the conditional mean models for analyzing G×E interaction. First, comparing conditional mean regression, modelling conditional quantiles offers a far more comprehensive understanding of the distribution of the response variable. In many applications, the impact of the genetic factor G on the response varies at different quantiles of the distribution. People are more interested in the quantiles rather than the mean. For example, in a study to find genes associated with birth weight, extremely low or high birth weight could be problematic since it could cause complications later in life. In this case, one would be interested in identifying genes or SNPs affecting low or high birth weight with their effect modified by environmental exposures. Second, even when we are interested in the center of the conditional distribution of the response variable, median regression (quantile regression with $\tau = 0.5$) can provides more robust estimators, especially when there are outliers in the trait distribution.

With the recent advancement in biotechnoloty, we are now able to collect hundreds of thousands of single nucleotide polymorphisms (SNPs) data, at the same time with relatively low cost. Such advancement renders traditional model selection methods such as forward/backword selection or methods based on AIC/BIC information obsolete. Recently, variable selection via penalized regression has been gaining popularity. The idea is to add a penalty term to the loss(likelihood) function. With different choices of penalty functions, the estimator could possess different properties. Fan and Li (2001) proposed three

important criteria for penalized estimator: sparsity, unbiasedness and continuity. They also characterized oracle property, meaning that the model performs as well as if the true model is known in advance. For instance, adaptive LASSO (Zou (2006)), smoothly clipped absolute deviation (SCAD) (Fan and Li (2001)) and minimax concave penalty (MCP) (Zhang (2010)) possess oracle property.

In this work, we proposed a quantile regression based variable selection method built upon the VMICM model to identify how genetic effects are modified by simultaneous exposure to multiple environmental factors to affect a disease trait. We adopted the MCP penalty in our modeling framework. We evaluated the method through both simulation and real data applications. The proposed variable selection method for quantile regression enriches the literature about variable selection for quantile regression.

The rest of the paper is organized as follows. Section 2 introduces the proposed variable selection method, how to formulate the penalized quantile loss function, how to iteratively optimize the penalized quantile loss function as well as how to select various tuning parameters and initial value for $\boldsymbol{\beta}$. In section 3, we evaluated the finite sample performance of our model via monte carlo simulation. We applied our approach to a birth weight data set in section 4, followed by a discussion.

4.2 Variable selection for quantile regression with VMICM

Throughout the paper, superscript ^{T} is used to denote matrix transpose; $||.||_p$ is used to denote L_p norm; $log(a)$ is used to denote natural logarithm of a. For the sake of simplicity, we use constant and non-zero constant interchangeably.

4.2.1 Model setup

For a random sample of the data $\{\boldsymbol{Y}_{n\times 1}, \boldsymbol{X}_{n\times q}, \boldsymbol{G}_{n\times(p+1)}\}$ with size n, we assume the following model:

$$\boldsymbol{Y} = \sum_{k=0}^{p} m_k(\boldsymbol{X}\boldsymbol{\beta}(\tau), \tau)\boldsymbol{G}_k + \boldsymbol{\epsilon}(\tau) \tag{4.1}$$

where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ is a continuous response variable; $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_q)$ is a continuous q-dim environmental variable; $\boldsymbol{G}_{n\times(p+1)} = (\boldsymbol{G}_0, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_p)$ where $\boldsymbol{G}_0 = (1, \ldots, 1)^T$ and $\boldsymbol{G}_k$ is a continuous or discrete genetic vector of length n for $k = 1, 2, \ldots p$. Our parameters of interest $\{m_k(u, \tau)\}_{k=0,1,\ldots,p}$ are $p+1$ unknown non-parametric functions conditional on quantile τ; $\boldsymbol{\beta}(\tau) = (\beta_1(\tau), \ldots, \beta_q(\tau))^T$ is the loading parameters conditional on quantile τ. $\boldsymbol{\epsilon}(\tau)$ is an unknown random error satisfies $P(\boldsymbol{\epsilon}(\tau) < 0|\boldsymbol{X}, \boldsymbol{G}) = \tau$ for some specific quantile $0 < \tau < 1$. The case with $\tau = 0.5$ corresponds to median regression. For ease of presentation, all parameters of interest are τ-specific. For instance, we use $\boldsymbol{\beta}$ and $m_k(u)$ to represents $\boldsymbol{\beta}(\tau)$ and $m_k(u, \tau)$, respectively. We denote the quantile VMICM model as τVMICM.

4.2.2 Estimation method

Our goal is to select and estimate unknown functions $\{m_k(\cdot)\}_{k=0,1,\ldots,p}$ and unknown loading parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$. For the sake of identifiability, we assumed $||\boldsymbol{\beta}||_2 = 1$ and $\beta_1 > 0$, and $m_k(\cdot)$ cannot has the form of $m_k(\boldsymbol{u}) = \boldsymbol{\alpha}^T\boldsymbol{u}\boldsymbol{\beta}^T\boldsymbol{u} + \boldsymbol{\gamma}^T\boldsymbol{u} + c$. Please refer to Schumaker (2007) for details of the construction of B-splines basis function. Given the number of interior knots K and the degree of the B-spline basis function h, we can approximate $m_k(u)$ by

$$m_k(u) \approx \gamma_{k1} + \bar{\boldsymbol{B}}(u)\boldsymbol{\gamma}_{k*} \tag{4.2}$$

where $\boldsymbol{\gamma}_{k*} = (\gamma_{k2}, \gamma_{k3}, \ldots, \gamma_{kL})^T$, $\boldsymbol{\gamma}_k = (\gamma_{k1}, \boldsymbol{\gamma}_{k*}^T)^T$ and $L = K + h + 1$. Hence, (4.1) can be rewritten as :

$$\boldsymbol{Y} = \sum_{k=0}^{p} [\gamma_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{\gamma}_{k*}]\boldsymbol{G}_k + \boldsymbol{\epsilon}. \tag{4.3}$$

Here, the estimation of non-parametric function $m_k(u)_{k=0,1,\ldots,p}$ and its loading parameter $\boldsymbol{\beta}$ is transformed to the estimation of $\{\gamma_{k1}, \boldsymbol{\gamma}_{k*}\}_{k=0,1,\ldots,p}$ and $\boldsymbol{\beta}$. This B-spline approximation (4.2) also enable us to separate the constant effect of $\boldsymbol{G}_k$ on $\boldsymbol{Y}$ from its joint effect with $\boldsymbol{X}$ on $\boldsymbol{Y}$. (1) If $||\boldsymbol{\gamma}_{k*}||_2 \neq 0$, then $\boldsymbol{G}_k$ and $\boldsymbol{X}$ jointly affect $\boldsymbol{Y}$; (2) If $||\boldsymbol{\gamma}_{k*}||_2 = 0$ and $|\gamma_{k1}| \neq 0$, then $\boldsymbol{G}_k$ has a constant effect on $\boldsymbol{Y}$; (3) If $||\boldsymbol{\gamma}_{k*}||_2 = 0$ and $|\gamma_{k1}| = 0$ then $\boldsymbol{G}_k$ has no effect on $\boldsymbol{Y}$ at all.

Following the idea proposed in previous chapters, we adopted the following penalized regression approach and defined the objective function as $Q_\tau(\boldsymbol{\beta}, \boldsymbol{\gamma})$:

$$\begin{aligned} Q_\tau(\boldsymbol{\beta}, \boldsymbol{\gamma}) = &\sum_{i=1}^{n} \rho_\tau\big(Y_i - \sum_{k=0}^{p} [\gamma_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{\gamma}_{k*}]G_{ik}\big) + n\sum_{k=1}^{p} p_{\lambda_1}(||\boldsymbol{\gamma}_{k*}||_2) \\ &+ n\sum_{k=1}^{p} p_{\lambda_2}(|\gamma_{k1}|)I(||\boldsymbol{\gamma}_{k*}||_2 = 0) + n\sum_{d=2}^{q} p_{\lambda_3}(|\beta_d|) \end{aligned} \tag{4.4}$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ is the quantile loss function; $p_{\lambda_1}(\cdot), p_{\lambda_2}(\cdot), p_{\lambda_3}(\cdot)$ are penalty functions of the corresponding parameters; and $I(\cdot)$ is an indicator function which equals 1 if the condition in the parentheses is satisfied, and 0 otherwise. From the construction of the penalty function, we only penalize γ_{k1} if $||\boldsymbol{\gamma}_{k*}||_2 = 0$, meaning that we are interested in whether the non-parametric function $m_k(\cdot)$ is zero or a non-zero constant only when it is not varying. No penalty is applied to the intercept function $m_0(\cdot)$ as both $\boldsymbol{\gamma}_{0*}$ and γ_{01} are not involved in the penalty term. No penalty is applied to the coefficient β_1 due to the constrain:

$\beta_1 > 0$. For the penalty function, we use MCP penalty proposed by Zhang (Zhang (2010)) which is defined as $p(x, \lambda) = \lambda \int_0^x (1 - \frac{s}{\tau\lambda})_+ ds$ with the regularization parameters $\tau > 0$ and $\lambda > 0$.

4.2.3 Estimation algorithm

In the previous chapters, we proposed a 3 step iterative approach based on the VMICM model for both continuous and binary responses. The methods classify the non-parametric functions $m_k(\cdot)$ into 3 categories: varying, constat or zero, denoted by V, C and Z respectively. Here, we generalized the estimation algorithm to a quantile regression setting.

Step 1: For given $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}^{(0)}$, the step 1 estimator of $\boldsymbol{\gamma}$, $\hat{\boldsymbol{\gamma}}^{(1)} = \{\hat{\gamma}_{k1}^{(1)}, \hat{\boldsymbol{\gamma}}_{k*}^{(1)T}\}_{k=0,1,\ldots,p}^{T}$ can be obtained via optimizing the following grouped penalized regression

$$\hat{\boldsymbol{\gamma}}^{(1)} = \min_{\boldsymbol{\gamma}} Q_1(\boldsymbol{\gamma}|\lambda_1, \hat{\boldsymbol{\beta}}^{(0)})$$

where

$$Q_1(\boldsymbol{\gamma}|\lambda_1, \hat{\boldsymbol{\beta}}^{(0)}) = \sum_{i=1}^{n} \rho_\tau\big(Y_i - \sum_{k=0}^{p}[\gamma_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta}^{(0)})\boldsymbol{\gamma}_{k*}]G_{ik}\big) + n\sum_{k=1}^{p} p_{\lambda_1}(||\boldsymbol{\gamma}_{k*}||_2).$$

Instead of penalizing each coordinate of $\boldsymbol{\gamma}_{k*} = (\gamma_{k2}, \ldots, \gamma_{kL})^T$ separately, we penalized the L_2 norm of $\boldsymbol{\gamma}_{k*}$ since we would like to assess the presence of the varying effect of $\boldsymbol{G}_k$ on the response variable. Step 1 classifies $m_k(\cdot), k = 1, \ldots, p$ into two categories: varying(V) or non-varying(NV) where $m_k(\cdot) \in V$ if $||\hat{\boldsymbol{\gamma}}_{k*}^{(1)}||_2 > 0$ and $m_k(\cdot) \in NV$ if $||\hat{\boldsymbol{\gamma}}_{k*}^{(1)}||_2 = 0$.

Step 2: Based on the step 1 estimators of the B-spline coefficients $\boldsymbol{\gamma}$, the step 2 estimators $\hat{\boldsymbol{\gamma}}^{(2)} = \{(\hat{\gamma}_{k1}^{(2)}, \hat{\boldsymbol{\gamma}}_{k*}^{(2)})_{k\in V}, (\hat{\gamma}_{k1}^{(2)})_{k\in C}\}$ can be obtained via the penalized regression. Note that

$\hat{\boldsymbol{\gamma}}_{k*}^{(2)} = 0$ automatically if $\hat{\boldsymbol{\gamma}}_{k*}^{(1)} = 0$. We obtained the estimator by,

$$\hat{\boldsymbol{\gamma}}^{(2)} = \min_{\boldsymbol{\gamma}} Q_2(\boldsymbol{\gamma}|\lambda_2, \boldsymbol{\beta}^{(0)}, \hat{\boldsymbol{\gamma}}^{(1)})$$

where

$$\begin{aligned} Q_2(\boldsymbol{\gamma}|\lambda_2, \boldsymbol{\beta}^{(0)}, \hat{\boldsymbol{\gamma}}^{(1)}) &= \sum_{i=1}^{n} \rho_\tau\big(Y_i - \sum_{k\in V}[\gamma_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta}^{(0)})\boldsymbol{\gamma}_{k*}]G_{ik} - \sum_{k\in C}\gamma_{k1}^{(2)}G_{ik}\big) \\ &\quad + n\sum_{k=1}^{p} p_{\lambda_2}(|\gamma_{k1}^{(2)}|)I(||\hat{\boldsymbol{\gamma}}_{k*}^{(1)}||_2 = 0). \end{aligned} \tag{4.5}$$

Based on the initial estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}^{(0)}$, we can obtain the estimators of the B-splines coefficients $\boldsymbol{\gamma}$, $\hat{\boldsymbol{\gamma}}^{(2)}$ and classify $m_k(\cdot)$ $k = 1, \ldots, p$ into V, C or 0.

Step 3: We obtained $\hat{\boldsymbol{\beta}}$ via the penalized regression by

$$\hat{\boldsymbol{\beta}} = \min_{||\boldsymbol{\beta}||_2=1} Q_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}^{(2)})$$

where

$$Q_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}^{(2)}) = \sum_{i=1}^{n} \rho_\tau\big(Y_i - \sum_{k=0}^{p}[\hat{\gamma}_{k1}^{(2)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}_{k*}^{(2)}]G_{ik}\big) + n\sum_{d=2}^{q} p_{\lambda_3}(|\beta_d|).$$

Step 4: Set $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}$, then iterate step 1 to 3 until convergence. Denote $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$ as the converged estimators.

With this iteration approach, we still need to select the tuning parameters $\lambda_1, \lambda_2, \lambda_3$, order h and number of interior knots K for the B-spline approximation, as well as a reasonable initial value for $\boldsymbol{\beta}$.

4.2.4 Selection of parameters

For a traditional linear mean regression model, Bayesian Information Criterion (BIC) (Schwarz (1978)) has been a popular choice in the selection of shrinkage parameters. Lee et al. (2014) provided theoretical justification in the use of BIC in quantile regression models. Hence, we use BIC as our selection criterion for shrinkage parameters, and the order h and the number of interior knots K of the B-spline basis functions.

4.2.4.1 Selection of the tuning parameters $\lambda_1, \lambda_2, \lambda_3$

Step 1: We took λ_1 as the minimizer of

$$BIC(\lambda_1) = log \sum_{i=1}^{n} \rho_\tau \big(Y_i - \sum_{k=0}^{p} [\hat{\gamma}_{k1}^{(\lambda_1)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})\hat{\boldsymbol{\gamma}}_{k*}^{(\lambda_1)}] G_{ik}\big) + \frac{log(n)}{2n} * df_{\lambda_1}$$

where $\{\hat{\gamma}_{k1}^{(\lambda_1)}, \hat{\boldsymbol{\gamma}}_{k*}^{(\lambda_1)}\}_{k=0,1,...,p}$ are the minimizers of $Q_1(\boldsymbol{\gamma}|\lambda_1, \hat{\boldsymbol{\beta}}^{(0)})$ defined above; $\hat{\boldsymbol{\beta}}^{(0)}$ is chosen as the estimator from previous iteration; and df_{λ_1} is defined as the total number of non-zero coefficients if λ_1 is the penalized parameter.

Step 2: We took λ_2 as the minimizer of

$$BIC(\lambda_2) = log \sum_{i=1}^{n} \rho_\tau \big(Y_i - \sum_{k=0}^{p} [\hat{\gamma}_{k1}^{(\lambda_2)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})\hat{\boldsymbol{\gamma}}_{k*}^{(\lambda_2)}] G_{ik}\big) + \frac{log(n)}{2n} * df_{\lambda_2}$$

where $\{\hat{\gamma}_{k1}^{(\lambda_2)}, \hat{\boldsymbol{\gamma}}_{k*}^{(\lambda_2)}\}_{k=0,1,...,p}$ are the minimizers of $Q_2(\boldsymbol{\gamma}|\lambda_2, \hat{\boldsymbol{\beta}}^{(0)})$ defined above and df_{λ_2} is defined as the total number of non-zero coefficients if λ_2 is the penalized parameter.

Step 3: We took λ_3 as the minimizer of

$$BIC(\lambda_3) = log \sum_{i=1}^{n} \rho_\tau \big(Y_i - \sum_{k=0}^{p} [\hat{\gamma}_{k1}^{(2)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda_3)})\hat{\boldsymbol{\gamma}}_{k*}^{(2)}] G_{ik}\big) + \frac{log(n)}{2n} * df_{\lambda_3}$$

where $\hat{\boldsymbol{\beta}}^{(\lambda_3)}$ are the minimizers of $Q_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}^{(2)})$ defined above and df_{λ_3} is defined as the total number of non-zero $\boldsymbol{\beta}$ if λ_3 is the penalized parameter.

To find the optimal tuning parameters, $\lambda_1, \lambda_2, \lambda_3$ are searched over a grid of exponentially decreasing values with the minimum to be 1E-3, and the maximum of $\lambda_1, \lambda_2, \lambda_3$ are set to be the minimum value such that all of the penalized estimators are 0. For ease of computation, the number of grid points is set to be 100.

4.2.4.2 Selection of the order h and the number of interior knots K

As we discussed in previous chapters, higher order of the B-spline basis function implies more complex functions and leads to less interpretable effect. From a practical point of view, the interaction effect is less likely to be highly nonlinear. Thus, we searched optimal h over the set $h \in \{2, 3, 4\}$ to avoid any complications in interpretation. As for the number of interior knots K, we searched the optimal K in $\mathscr{K} = \{2, 3, 4, 5\}$. For every combination of K and h, we fit the following intercept only model,

$$Y = m_0(X\boldsymbol{\beta}) + \boldsymbol{\epsilon} \tag{4.6}$$

Again, as we discussed in previous chapters, this is to avoid computational burden. The optimal K and h are those that minimize $log \sum_{i=1}^{n} \rho_\tau(\hat{Y}_i - Y_i) + \frac{log(n)}{2n}(K + h + 1)$, where $\hat{Y}_i$ is the estimation of the i-th subject under model (4.6).

4.2.5 Selection of the initial values

The initial value of $\boldsymbol{\beta}$ for a single index model is usually set to be $(1, 0, \ldots, 0)^T$ or $(\frac{1}{\sqrt{q}}, \ldots, \frac{1}{\sqrt{q}})^T$. However, neither works well in our simulations. Hence, we get the initial

value by fitting the intercept only model (4.6).

4.3 Simulation

We conducted extensive simulation to investigate the finite sample performance of the proposed selection approach under the τVMICM model. The performance is evaluated in several ways: (1) the selection accuracy (oracle percentage) of the non-parametric function $m_k(u)$; (2) IMSE of $\hat{m}(u)$; (3) the selection accuracy (oracle percentage) of $\boldsymbol{\beta}$ and (4) MSE of $\boldsymbol{\beta}$. The performance is evaluated over 1000 simulation runs.

The oracle percentage of $m_k(u)$ is defined as the percentage of correct classification of $m_k(u)$. For instance, if $m_k(u) \in V$, and $m_k(\cdot)$ is classified as varying for g times, then the oracle percentage of $m_k(\cdot)$ is calcualted as $\frac{g}{1000} \times 100\%$. IMSE of $m_k(\cdot)$ is defined as

$$\frac{1}{1000}\sum_{r=1}^{1000}[\frac{1}{100}\sum_{j=1}^{100}(\hat{m}_k(u_j)^r - m_k(u_j))^2]$$

where $\hat{m}_k(u_j)^r$ is the fitted value of $m_k(u_j)$ for the r-th simulation and u_j is taken at the $j-\%$ quantile among the range of $\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(r)}$. The oracle percentage of $\boldsymbol{\beta}$ is defined as the percentage of correct selection of $\boldsymbol{\beta}$. For example, if $\beta_d \neq 0$ and β_d is selected to be non-zero for g times, then the oracle percentage of β_d is calculated as $\frac{g}{1000} \times 100\%$. MSE of β_d is calculated as $\frac{1}{1000}\sum_{r=1}^{1000}(\hat{\beta}_d^{(r)} - \beta_d)^2$ where $\hat{\beta}_d^{(r)}$ is the estimator for β_d in the $r-th$ simulation.

4.3.1 Simulation Setting

Our data was generated according to the model,

$$\boldsymbol{Y} = m_0(\boldsymbol{X}\boldsymbol{\beta}) + \sum_{k=1}^{p} m_k(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{G}_k + \boldsymbol{\epsilon}(\tau).$$

Five ($q = 5$) independent environmental factors were generated with each one being generated from a $Unif(0,1)$ distribution. For the loading parameter $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)^T$, we set $\beta_1 = \beta_2 = \frac{1}{\sqrt{2}}$ and the rest $\beta_j's$ were set as zeros. $\boldsymbol{\epsilon}$ was generated from a $N(0,1)$ distribution and $\boldsymbol{\epsilon}(\tau) = \boldsymbol{\epsilon} - F^{-1}(\tau)$ where F denotes the CDF of $\boldsymbol{\epsilon}$. $F^{-1}(\tau)$ was subtracted from $\boldsymbol{\epsilon}$ to make sure the τth-quantile of $\boldsymbol{\epsilon}(\tau)$ is zero. For the genetic factor $\boldsymbol{G}$, we evaluated the performance of our model with both continuous and discrete variables.

4.3.2 The continuous case

We first evaluated the performance of our approach with continuous genetic predictors $\boldsymbol{G}$ which marginally followed a $N(0,1)$ distribution. The non-parametric functions $m_k(u)$ were set to be: $m_0(u) = 2sin(2\pi u)$, $m_1(u) = 2cos(\pi u) + 2$ and $m_2(u) = sin(2\pi u) + cos(\pi u) + 1$; $m_3(u) = 2$ and $m_4(u) = 2.5$ which were non-zero constants; $m_k(u) = 0$ for $k = 5, \ldots, p$ which were zero effects. We simulated the data under $p = 50, 100$, $\tau = 0.25, 0.5, 0.75$, and $n = 2000$.

Table 12 presents the selection and estimation result for non-parametric functions $m_k(u)$ with continuous predictor $\boldsymbol{G}$. For varying and constant coefficient function ($m_1(\cdot)$ to $m_4(\cdot)$), the oracle percentage was at 100% and model IMSE was in the order of -2 for all cases. These suggest that our model could correctly select and estimate non-zero effect predictors at all quantiles. Between the median regression case ($\tau = 0.5$) and the cases where $\tau = 0.25, 0.75$,

the false positive rate was much lower for median regression case (around 1%) than that of the case $\tau = 0.25, 0.75$ (around 10%). Also the model IMSE of the case $\tau = 0.5$ was smaller than the model IMSE of the case $\tau = 0.25, 0.75$. These are expected since median regression usually provides the most accurate estimator among different quantiles. Between the case $p = 50$ and the case $p = 100$, we did not observe an significant difference in model performance. Overall, the proposed variable selection approach can correctly select non-zero effect predictors at all quantiles. Compared with median regression, although the false positive rate and model IMSE were higher in the case $\tau = 0.25, 0.75$, they were still decent.

Table 12: Selection and estimation accuracy of $m_k(\cdot)$ for continuous $\boldsymbol{G}$

		p = 50			p = 100		
0.25	$m_0(.)$	100.0%	2.78E+00	2.96E-02	100.0%	2.67E+00	2.97E-02
	$m_1(.)$	100.0%	8.16E-02	6.16E-03	100.0%	7.07E-02	6.16E-03
	$m_2(.)$	100.0%	9.59E-02	1.30E-02	100.0%	7.80E-02	1.31E-02
	$m_3(.)$	100.0%	2.58E-02	8.77E-04	100.0%	2.19E-02	9.71E-04
	$m_4(.)$	100.0%	2.56E-02	1.02E-03	99.8%	2.12E-02	9.80E-04
	Zero	88.8%	1.03E-02	0	90.6%	7.27E-03	0
0.5	$m_0(.)$	100.0%	7.49E-02	2.88E-02	100.0%	7.70E-02	2.91E-02
	$m_1(.)$	100.0%	4.98E-02	5.07E-03	100.0%	4.89E-02	5.31E-03
	$m_2(.)$	100.0%	5.72E-02	1.22E-02	100.0%	5.77E-02	1.23E-02
	$m_3(.)$	99.7%	1.25E-03	7.90E-04	99.9%	1.30E-03	7.90E-04
	$m_4(.)$	99.9%	1.50E-03	7.91E-04	99.8%	1.48E-03	8.29E-04
	Zero	98.7%	1.00E-04	0	99.1%	6.26E-05	0
0.75	$m_0(.)$	100.0%	2.60E+00	3.01E-02	100.0%	2.60E+00	3.13E-02
	$m_1(.)$	100.0%	6.25E-02	6.08E-03	100.0%	6.17E-02	6.18E-03
	$m_2(.)$	100.0%	7.12E-02	1.35E-02	100.0%	7.14E-02	1.37E-02
	$m_3(.)$	99.9%	2.42E-02	8.58E-04	100.0%	2.18E-02	9.03E-04
	$m_4(.)$	99.9%	2.62E-02	9.93E-04	100.0%	2.20E-02	9.63E-04
	Zero	88.5%	1.05E-02	0	90.7%	7.19E-03	0

Table 13 presents the selection and estimation results for the loading parameters $\boldsymbol{\beta}$. For non-zero loading β's (β_1 and β_2), the oracle percentage was close to 100% in all cases. This

suggests our model could correctly select non-zero loading parameters. For zero loading β's (β_3, β_4, and β_5), the oracle percentage for the median regression case was around 98.5%. Compared between the case $\tau = 0.25$ and $\tau = 0.75$, oracle percentage for the case $\tau = 0.25$ (99%) was slightly higher than that of the case $\tau = 0.75$ (96%), while the MSE for the case $\tau = 0.25$ is higher than that of the case $\tau = 0.75$. Between the case $p = 50$ and $p = 100$, we did not observed a difference in model performance. Overall, the proposed model can correctively select and estimate loading covariates with reasonably high accuracy at all quantiles.

Table 13: Selection and estimation accuracy of $\boldsymbol{\beta}$

		p = 50			p = 100		
τ		Oracle %	Model	Oracle	Oracle %	Model	Oracle
0.25	β_1	100.0%	1.20E-02	4.76E-05	100.0%	6.71E-03	4.36E-05
	β_2	99.8%	1.47E-02	4.75E-05	100.0%	7.69E-03	4.37E-05
	β_3	99.9%	7.15E-05	0	99.9%	5.81E-05	0
	β_4	99.9%	2.09E-04	0	99.9%	2.93E-04	0
	β_5	99.9%	9.60E-05	0	99.9%	9.53E-05	0
0.5	β_1	100.0%	5.29E-05	3.78E-05	100.0%	5.33E-05	3.54E-05
	β_2	100.0%	5.29E-05	3.78E-05	100.0%	5.34E-05	3.54E-05
	β_3	98.6%	1.52E-06	0	98.3%	1.23E-06	0
	β_4	98.3%	3.33E-06	0	98.8%	3.24E-06	0
	β_5	98.7%	1.85E-06	0	98.9%	1.45E-06	0
0.75	β_1	100.0%	4.40E-03	4.81E-05	100.0%	3.60E-03	4.89E-05
	β_2	100.0%	4.51E-03	4.80E-05	100.0%	3.49E-03	4.88E-05
	β_3	95.3%	4.14E-05	0	96.3%	1.87E-05	0
	β_4	95.9%	5.31E-05	0	96.2%	4.23E-05	0
	β_5	95.8%	3.85E-05	0	95.8%	2.32E-05	0

4.3.3 The discrete case

We continued to evaluate the performance of our model with discrete genetic predictors $\boldsymbol{G}$. One of many applications of our model is to select significant single nucleotide polymorphism

(SNP) in a gene or pathway. All SNPs take values of 0, 1 and 2 to represent aa, Aa, and AA genotype under an additive genetic model. We simulated $\boldsymbol{G}$ according to the following probability distribution function,

$$P(G_{ij}=0)=MAF^2, P(G_{ij}=1)=2*MAF(1-MAF), P(G_{ij}=2)=(1-MAF)^2$$

where G_{ij} is the j-th predictor of the i-th subject, $i=1,\ldots,n$ and $j=1,\ldots,p$. The data were simulated under $p=50,100$, $\tau=0.25,0.5,0.75$, and $n=2000$. We set the non-parametric function $m_k(u)$ and the corresponding MAF for $\boldsymbol{G}_k$ as follows:

Table 14: Setup for $m_k(u)$

Function	MAF of $\boldsymbol{G}_k$
$m_0(u)=2sin(2\pi u)$	NA
$m_1(u)=2cos(\pi u)+2$	0.5
$m_2(u)=sin(2\pi u)+cos(\pi u)+1$	0.5
$m_3(u)=2cos(\pi u)+2$	0.3
$m_4(u)=sin(2\pi u)+cos(\pi u)+1$	0.3
$m_5(u)=2cos(\pi u)+2$	0.1
$m_6(u)=sin(2\pi u)+cos(\pi u)+1$	0.1
$m_7(u)=2$	0.5
$m_8(u)=2$	0.3
$m_9(u)=2$	0.1
$m_k(u)=0, k>9$	Unif(0.05,0.5)

From the simulation setup, we would be able to evaluate how our proposed variable selection method perform with SNPs of a wide range of MAF. Table 15 presents the selection and estimation result for the non-parametric functions with discrete genetic predictors. Compared to the case where $\tau=0.25,0.75$, the median regression case ($\tau=0.5$) performed better. For instance, the oracle percentage for varying effect G_k was around 99% for median regression, while that of the case $\tau=0.25,0.75$ ranged from 90% to 97% ($p=50$) and 84%

to 91% ($p = 100$). Further, the model IMSE for the median regression case was smaller than that of the case $\tau = 0.25, 0.75$. Between the case where $p = 50$ and $p = 100$, the case $p = 50$ performed slightly better both in terms of oracle percentage and model IMSE with varying effect predictors. Overall, the model performed reasonably well across different quantiles. In addition, we observed the model performance was associated with minor allele frequency of G_k. We prepared figure 5 to better visualize such trend. With an increase of minor allele frequency of G_k (from 0.1 to 0.5), we observed an increase in oracle percentage as well as a decrease in model IMSE. These suggest that the model performs better with common variant SNPs. It is expected since rare variant SNP has less information.

Table 15: Selection and estimation accuracy for $m_k(u)$ with discrete $\boldsymbol{G}$

		$p = 50$			$p = 100$		
			IMSE			IMSE	
τ		Oracle %	Model	Oracle	Oracle %	Model	Oracle
0.25	Intercept	100.0%	7.54E-01	2.47E-02	100.0%	7.98E-01	2.44E-02
	Varying	97.5%	1.44E-01	2.30E-02	91.4%	2.23E-01	2.32E-02
	Constant	94.2%	1.66E-02	3.19E-03	95.5%	1.66E-02	3.22E-03
	Zero	93.9%	4.71E-03	0	96.0%	3.05E-03	0
0.5	Intercept	100.0%	4.84E-02	2.35E-02	100.0%	4.97E-02	2.30E-02
	Varying	99.9%	9.14E-02	2.00E-02	99.3%	9.84E-02	1.98E-02
	Constant	95.4%	7.52E-03	2.68E-03	95.5%	9.30E-03	2.76E-03
	Zero	93.6%	2.87E-03	0	93.8%	2.87E-03	0
0.75	Intercept	100.0%	1.02E+00	2.53E-02	100.0%	1.09E+00	2.48E-02
	Varying	90.8%	2.29E-01	2.33E-02	84.1%	3.35E-01	2.30E-02
	Constant	96.7%	1.35E-02	3.19E-03	98.6%	1.26E-02	3.14E-03
	Zero	96.8%	2.09E-03	0	98.5%	9.24E-04	0

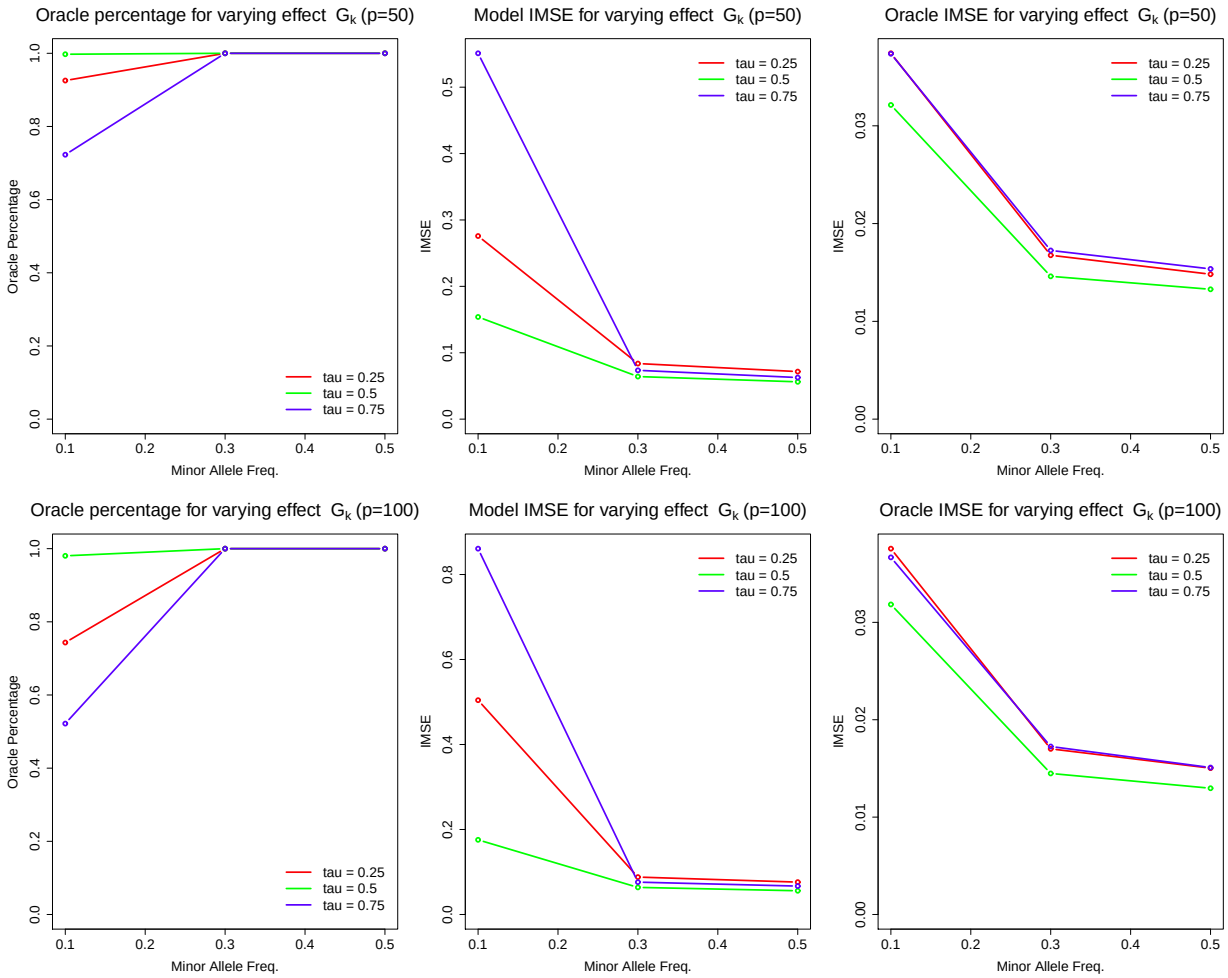

Figure 5: Selection and estimation accuracy of $m_k(\cdot)$ for discrete $\boldsymbol{G}$

Table 16 shows the selection and estimation of loading parameters $\boldsymbol{\beta}$ with discrete $\boldsymbol{G}$. The proposed model select and estimate non-zero loading β with high precision (100% in all cases). For zero loading parameters (β_3, β_4, and β_5), the false positive rate was around 0%, 1.5%, and 6.5% for the case $\tau = 0.25$, $\tau = 0.5$, and $\tau = 0.75$ respectively. The MSE for zero loading β's was in the order of -5 to -7, suggesting the proposed model shrunk it very close to 0.

Table 16: Selection and estimation accuracy for $\boldsymbol{\beta}$ with discrete $\boldsymbol{G}$

		$p = 50$			$p = 100$		
			MSE			MSE	
τ		Oracle %	Model	Oracle	Oracle %	Model	Oracle
0.25	β_1	100.0%	3.98E-04	4.42E-05	100.0%	4.63E-04	4.75E-05
	β_2	100.0%	4.04E-04	4.43E-05	100.0%	4.69E-04	4.74E-05
	β_3	100.0%	0	0	100.0%	0	0
	β_4	100.0%	0	0	100.0%	0	0
	β_5	100.0%	0	0	100.0%	0	0
0.5	β_1	100.0%	5.20E-05	3.96E-05	100.0%	5.30E-05	3.97E-05
	β_2	100.0%	5.21E-05	3.97E-05	100.0%	5.30E-05	3.97E-05
	β_3	99.1%	2.27E-07	0	99.1%	1.00E-06	0
	β_4	98.9%	3.43E-07	0	98.4%	8.12E-07	0
	β_5	98.9%	5.45E-07	0	98.7%	5.87E-07	0
0.75	β_1	100.0%	2.75E-04	5.13E-05	100.0%	3.70E-04	4.69E-05
	β_2	100.0%	2.84E-04	5.15E-05	100.0%	3.59E-04	4.69E-05
	β_3	93.6%	2.52E-05	0	93.0%	2.26E-05	0
	β_4	94.3%	2.23E-05	0	93.9%	1.80E-05	0
	β_5	93.5%	3.46E-05	0	93.3%	3.08E-05	0

Based on the simulation results described above, we were able to conclude the followings. (1) The proposed model selects and estimates G_k with reasonably high precision. (2) Compared to rare variant SNPs ($P_a = 0.1$), the model performed better with common variant SNPs ($P_a = 0.3, 0.5$). (3) The model could correctly select and estimate non-zero loading β. At last, (4) the false positive rate for zero loading β was low for $\tau = 0.25, 0.5$, and it was slightly higher for $\tau = 0.75$.

4.4 Real Data Application

We applied the proposed variable selection approach to a birth weight data set, which is obtained from Gene Environment Association Studies initiative (GENEVA) funded by the

Genes, Environment and Health Initiative (GEI). Epidemiological studies often suggested that birth weight is strongly associated with morbidity and mortality risk during the first year, and risk of many diseases in adulthood. Birth weight is affected by fetal genes and maternal environment. We first performed data cleaning, removing SNPs with more that 5% missing, SNPs with minor allele frequency < 0.05 and SNPs deviates from Hardy-Weinberg equilibrium (p-value < 0.001). After this cleaning step, the data set contains 1,126 subjects and 590,913 SNPs. For the environmental factors $\boldsymbol{X}$, based on the marginal p-value (< 0.05) when regressing birth weight with each X variable, we select 3 environmental factors: mother's mean OGTT(oral glucose tolerance test) diastolic blood pressure (X_1), mother's one hour OGTT glucose level (X_2) and mother's mean OGTT systolic blood pressure (X_3).

We first map all SNPs to all known genes based on its location. Then we select the genes which contain ≥ 30 SNPs. As a result, we get 2,076 genes. Then we fit the proposed model to all genes at τ=0.25, 0.5, and 0.75. Since the first element of the loading parameter $\boldsymbol{\beta}$ has to be a non-zero positive number ($\beta > 0$), we fit the proposed model three times by varying the order of the X variable inside the index function. If the SNP is selected as varying or constant in all the three cases, this would suggest a convincing signal. Therefore, we only consider the cases where all three models return the same effect classification result.

Our model identified 122 genes with constat effect and no gene with varying effect, indicating that these genes are not sensitive to the changes of those three environmental variables, i.e., no significant G×E effect. Consider gene *ST3GAL1* located on chromosome 8 as an example. It contains 39 SNPs in our data set. Table 17 presents the estimated SNP effect in gene *ST3GAL1*. The left, middle, and right column correspond to the case where $\tau = 0.25$, $\tau = 0.50$, and $\tau = 0.75$, respectively. Among those, SNPs rs13267049, rs6986303, and rs6990329 have effect on response in lower quantile ($\tau = 0.25$). SNP rs2142306 only has

effect for the $\tau = 0.5$ quantile, and SNPs rs2736860, rs9643299, and rs7460764 have effect when $\tau = 0.75$. Interestingly, for SNP rs7831227, we observed an increase in negative effect with the increase of the quantile of the response variable. This suggests that the SNP has different effect at different quantile of the birth weight. A genome-wide associations study in 2016 (Horikoshi et al. (2016)) suggested that only 15% of the variance in birth weight was captured by genetic variations. Thus, it is not surprise to see the limited genetic variants with relatively small effect being selected.

Table 17: Effect of SNPs in gene *ST3GAL1*

	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
rs13267049	0.0260	0	0
rs2736860	0	0	-0.0290
rs2142306	0	0.0720	0
rs6986303	0.1129	0	0
rs6990329	0.1411	0	0
rs9643299	0	0	-0.0489
rs7460764	0	0	-0.1077
rs7831227	-0.0294	-0.0801	-0.1007

Table 18 presents the parameter estimates for the marginal effect of environmental factors on birth weight. The first, second, and third row correspond to the case where $\tau = 0.25$, $\tau = 0.50$, and $\tau = 0.75$ respectively. We observed different results at different quantiles, suggesting that the environmental effects are different at different quantiles for the birth weight. For example, at lower quantile (0.25), X_2 (mother's one hour OGTT glucose level) and X_3 (mother's systolic blood pressure) showed strongest effect. At higher quantile (0.75), X_1 (mother's diastolic blood pressure) and X_3 showed strongest effect. For median regression case ($\tau = 0.5$), we observed a strong effect for X_1, while X_2 was not selected.

Figure 6 represents the effect of environmental factors on different quantiles of birth

Table 18: Estimated Loading Parameter for Gene $ST3GAL1$

	β_1	β_2	β_3
$\tau = 0.25$	0.272	0.707	0.653
$\tau = 0.5$	0.895	0.000	-0.445
$\tau = 0.75$	0.637	-0.288	-0.715

weight. The red, blue, and black curve correspond to the effect at quantile 0.25, 0.5, and 0.75, respectively. Since the estimated $\boldsymbol{\beta}$ is different at different quantile, the span differs. We observed a higher fitted birth weight at a higher quantile. For $\tau = 0.75$, with the increased index $\boldsymbol{X\beta}$, we first observed a quick decrease in fitted birth weight, then it fluctuates around 3.3. Since the loading coefficient estimates for β_2 and β_3 are negative, this implies that higher values for mother's one hour OGTT glucose level and mother's systolic blood pressure potentially contribute to higher birth weight at the upper quantile. For the median regression, the fitted birth weight first decreases from 3.3 to 3.0, then slowly increases to 3.2. Both larger values in Mother's diastolic blood pressure and Mother's systolic blood pressure contribute to relatively higher birth weight. For $\tau = 0.25$, we saw a positive trend in the total effect as the index increases.

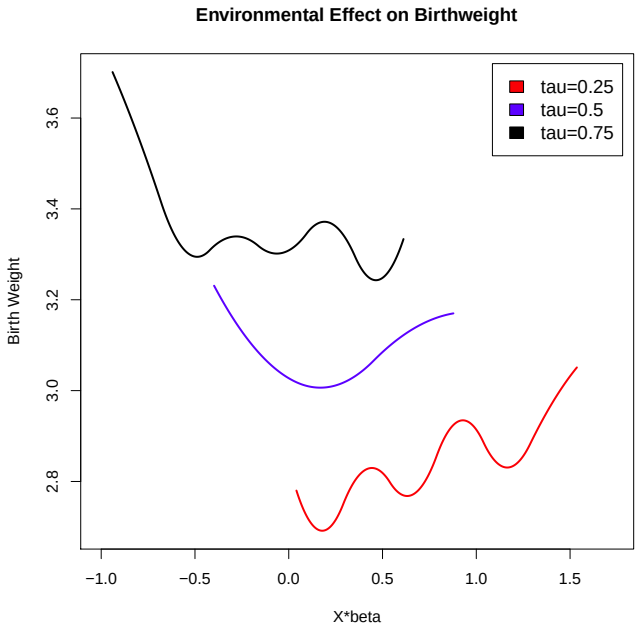

Figure 6: Plot of interaction effect effects

4.5 Discussion

Varying multi-index coefficient model is a novel way to model non-linear interaction between genetic variants $\boldsymbol{G}$ and a mixture of environmental factors $\boldsymbol{X}$. In previous chapters, we proposed a 3 step iterative variable selection approach with the goal of selecting varying and constat effect genetic variants as well as non-zero loading parameters. In this paper, we generalized such approach to a conditional quantile regression setting. Compared to condition mean regression, condition quantile regression possesses several advantages. First, modelling the data at several quantiles offers a more comprehensive way to understand the distribution of the response variable. In many applications, the effect of $\boldsymbol{G}_k$ on the response differs at different quantiles. Second, quantile regression is robust to extreme observations. By looking at different quantiles, novel insights about the underlying genetic mechanism

could be revealed.

From the setup of τVMICM, the environmental factors $\boldsymbol{X}$ have to be continuous due to the model constrain. Nevertheless, discrete factors such as gender and smoking status might possess significant effect. In this case, we could easily generalize τVMICM to a partial linear τVMICM model as $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\alpha} + (\boldsymbol{Z}\boldsymbol{G})\boldsymbol{\delta} + \sum_{k=0}^{p} m_k(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{G}_k + \boldsymbol{\epsilon}(\tau)$ where $\boldsymbol{Z}$ represent the discrete environmental factors and $\boldsymbol{\delta}$ represent the interaction of gene with discrete environmental variables. The proposed variable selection approach could be modified to accommodate these changes.

Although varying multi-index coefficient model enjoys many advantages over traditional G×E model, it is not without its limitations. One of which is the constrains on its loading parameters, $||\boldsymbol{\beta}||_2 = 1$ and $\beta_1 > 0$. Potentially, it would lead to different selection results if we vary the order of loading covariates. In the application to real data, we can fit several different models by varying the first loading covariate. If all the models return the same selection results, then we are convinced such finding is valid. The constrains also limit the interpretation of fitted environmental parameters, it is difficult to characterize the effect of a single environmental factor. If environmental effect is our primary interest, we should consider other modelling technique.

In the case study, the estimated interaction coefficients are all constants, indicating no G×E interaction for birth weight. As the study is focused on the Thai population, this cannot rule out the possibility that the mother's conditions may act on fetus' genome to affect fetal growth in other populations. Further investigation is needed to confirm this.

Chapter 5

Conclusion and future work

5.1 Conclusion

The main goal of this dissertation is to develop variable selection methods to identify non-linear G×E interactions. We first proposed to use varying multi-index coefficient models since it allows non-linear interaction between genetic factors and multiple environmental factors. In Chapter 2, we proposed a 3 step iterative variable selection approach for VMICM via a penalized regression. It separates gene effects into three categories: varying, constant and zero. It could also select non-zero loading parameters for environmental factors. In Chapter 3, we generalized such approach to a generalized regression setting with binary responses. Following the work of Chapter 2, we extended the proposed variable selection approach to a quantile regression setting in Chapter 4, since it provided a more comprehensive understanding of the data and offered more robust estimators.

In conclusion, this dissertation contributes to the literature in two ways. From the methodological perspective, it contributes to the methods development in variable selection under a nonparametric varying index coefficients model framework. For the selection of the nonparametric coefficients, we separated three types of effects rather than just selecting zero vs non-zero functions. This complicates our selection procedure and distinguishes our methods to existing ones in the literature (e.g., Feng and Xue (2013)). Theoretical properties

of the selection methods were evaluated. Our methods have practical meanings and enrich the literature of variable selection.

From the application perspective, our method development is well motivated by empirical studies to evaluate the joint effect of multiple environmental exposures and how they interact with genes to affect a disease trait. By taking gene or pathway information into account, we were able to select important players in a gene set. The method developed under the quantile regression framework makes much biological sense in certain trait such as birth weight. Novel insights are expected under the proposed models.

5.2 Future work

In the simulation studies, we assumed any two genetic variants G_k and G_l are independent. However, it would be more desirable if we consider linkage disequilibrium structure among SNPs. More specifically, we could set the correlation between G_k and G_l to be $\rho^{|k-l|}$ where $\rho = 0.3$ for low correlation case, $\rho = 0.5$ for median correlation case, and $\rho = 0.9$ for high correlation case. To demonstrate the robustness of median regression compared to mean regression, we could consider the case where $\boldsymbol{\epsilon}$ follows a t distribution with 3 degree of freedom.

In Chapter 2 and 3, we theoretically proved that our penalized estimators are consistent in both estimation and selection under a fixed number of parameters. It could be more desirable if we could prove the selection consistency when the number of parameters increases as the sample size increases. Also it could be beneficial if we could demonstrate the consistency of the penalized estimators in the quantile regression setting in Chapter 4. Further, generalizations of the proposed selection approach to other generalized regression setting

such as poisson or categorical variable could be done with modification to the likelihood function. At last, extension of the proposed model to longitudinal data could be considered.

We will consider the aforementioned mentioned future work and continue to investigate along those line.

APPENDICES

Appendix A

Real Data Results

A.1 Real data results of gVMICM

Table 19 presented all the varying effect SNPs selected. Similarly, Table 20 presented all the constant effect SNPs selected.

Table 19: List of SNPs with a varying effect.

GeneID	SNP
GeneID:440600	rs6537663
GeneID:2590	rs6666516
GeneID:729993	rs1015431
GeneID:54768	rs4788621
GeneID:117532	rs7509377
GeneID:758	rs5766384
GeneID:23395	rs4311249
GeneID:647107	rs2404825
GeneID:8633	rs3775049
GeneID:2185	rs6557991
GeneID:4915	rs6559870
GeneID:19	rs4742969
GeneID:286205	rs2416996

Table 20: List of SNPs with a constant effect.

GeneID	SNP
GeneID:114827	rs3815792
GeneID:2899	rs12118788
GeneID:260425	rs11102660
GeneID:9857	rs2293990
GeneID:2590	rs9308482
GeneID:6934	rs7901695
GeneID:55742	rs7101596
GeneID:867	rs4489755
GeneID:10867	rs740771
GeneID:57494	rs11047510
GeneID:196385	rs11058132
GeneID:64328	rs1961415
GeneID:23348	rs7326971
GeneID:23348	rs7991210
GeneID:57099	rs16962542
GeneID:11060	rs16970994
GeneID:25780	rs6708570
GeneID:100505498	rs6730602
GeneID:117532	rs11696526
GeneID:29780	rs5765571
GeneID:25814	rs713999
GeneID:9620	rs11090812
GeneID:23429	rs17009630
GeneID:80254	rs11710699
GeneID:8633	rs10516957
GeneID:157680	rs1788161

Appendix B

Algorithm

B.1 Algorithm for VMICM

Here we presents the algorithm used in each steps:

Step 1: To minimize the objective function $Q_1(\boldsymbol{\gamma}|\lambda_1, \hat{\boldsymbol{\beta}}^{(0)})$, we implemented the group coordinate descent algorithm. The design matrix has the form

$$\boldsymbol{D} = (\boldsymbol{G}_0, \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})\boldsymbol{G}_0, \boldsymbol{G}_1, \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})\boldsymbol{G}_1, \ldots, \boldsymbol{G}_p, \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})\boldsymbol{G}_p)_{n\times(L*(p+1))}$$

with the corresponding parameters $(\gamma_{01}, \boldsymbol{\gamma}_{0*}, \gamma_{11}, \boldsymbol{\gamma}_{1*}, \ldots, \gamma_{0p}, \boldsymbol{\gamma}_{p*})$, where $\boldsymbol{\gamma}_{k*}$ has a length of $L-1$. We assigned a grouping index for each of the parameters (from 0 to M, where $M+1$ is the number of groups). Parameters with the same grouping index were in the same group and they were penalized as a group, parameters with grouping index 0 were not penalized.

Denote $\boldsymbol{D}_m$ as the design matrix for group m, $m = 0, 1, \ldots M$. Given the penalty parameter λ_1 and the MCP tuning parameter γ^{MCP}, $\boldsymbol{\gamma}^{(1)}$ is obtained through the following iteration:

(0) Perform Q-R decomposition on all $\boldsymbol{D}_m$, i.e., $\boldsymbol{D}_m = \boldsymbol{Q}_m\boldsymbol{R}_m$, $m = 0, 1, 2 \ldots M$, where $\boldsymbol{Q}_m^T\boldsymbol{Q}_m = \boldsymbol{I}$ and $\boldsymbol{R}_m$ is an upper triangular matrix. Hence, $\boldsymbol{Q}_m$ is the normalized design matrix for group m.

(1) For given initial values for $\boldsymbol{\gamma}$, denoted as $\hat{\boldsymbol{\gamma}}^{initial} = \{\hat{\boldsymbol{\gamma}}_m^{initial}\}, m = 0, 1, \ldots, M$. We obtained OLS estimate $\hat{\boldsymbol{\gamma}}_m^{OLS}$ via $\hat{\boldsymbol{\gamma}}_m^{OLS} = \boldsymbol{Q}_m^T(Y - \boldsymbol{Q}_{-m}\tilde{\boldsymbol{\gamma}}_{-m}) = \boldsymbol{Q}_m^T\boldsymbol{Y} - \boldsymbol{Q}_m^T\boldsymbol{Q}_{-m}\tilde{\boldsymbol{\gamma}}_{-m}$ where subscript $\boldsymbol{Q}_{-m}$ represents the normalized design matrix without group m and $\tilde{\boldsymbol{\gamma}}_{-m}$ represents the most updated values for $\boldsymbol{\gamma}$ without group m.

(2) For group 0, set $\hat{\boldsymbol{\gamma}}_0 = \hat{\boldsymbol{\gamma}}_0^{OLS}$.

(3) For all other groups $m = 1, \ldots, M$, obtain the MCP estimate $\hat{\boldsymbol{\gamma}}_m$ via

$\hat{\boldsymbol{\gamma}}_m = \hat{\boldsymbol{\gamma}}_m^{OLS}$ if $||\hat{\boldsymbol{\gamma}}_m^{OLS}||_2 > \lambda * \tau$

$\hat{\boldsymbol{\gamma}}_m = \frac{\tau}{\tau - 1} S(\hat{\boldsymbol{\gamma}}_m^{OLS}, \lambda)$ if $\hat{\boldsymbol{\gamma}}_m^{OLS} \leq \lambda * \tau$ where $S(\hat{\boldsymbol{\gamma}}_m^{OLS}, \lambda) = (1 - \frac{\lambda}{||\hat{\boldsymbol{\gamma}}_m^{OLS}||_2})_+ * \hat{\boldsymbol{\gamma}}_m^{OLS}$

(4) Updated $\hat{\boldsymbol{\gamma}}_m^{initial}$ in step (1) by $\hat{\boldsymbol{\gamma}}_m$. Iterate (1) through (4) until convergence

(5) We adjusted the converged estimator $\hat{\boldsymbol{\gamma}}_m$ by $\hat{\boldsymbol{\gamma}}_m^{final} = \boldsymbol{R}_m^{-1}\hat{\boldsymbol{\gamma}}_m$, for $m = 0, 1, \ldots, M$.

Step 2: To obtained $\hat{\boldsymbol{\gamma}}^{(2)} = \min_{\boldsymbol{\gamma}} Q_2(\boldsymbol{\gamma}|\lambda_2, \boldsymbol{\beta}^{(0)}, \hat{\boldsymbol{\gamma}}^{(1)})$ where

$$Q_2(\boldsymbol{\gamma}|\lambda_2, \boldsymbol{\beta}^{(0)}, \hat{\boldsymbol{\gamma}}^{(1)}) = \sum_{i=1}^{n}\{Y_i - \sum_{k\in V}[\gamma_{k1}^{(2)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta}^{(0)})\boldsymbol{\gamma}_{k*}^{(2)}]G_k - \sum_{k\in C}\gamma_{k1}^{(2)}G_k\}^2 + n\sum_{k=1}^{p} p_{\lambda_2}(|\gamma_{k1}^{(2)}|)I(||\hat{\boldsymbol{\gamma}}_{k*}^{(1)}||_2 = 0)$$

We implement standard coordinate descent algorithm and the design matrix is:

$$\boldsymbol{D}^{(2)} = \left(\{\boldsymbol{G}_j, \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})\boldsymbol{G}_j\}_{j\in V}, \{\boldsymbol{G}_k, \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})\boldsymbol{G}_k\}_{k\in C}\right)$$

Step 3 : To obtain $\hat{\boldsymbol{\beta}}$, we minimized

$$Q_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}^{(2)}) = ||\boldsymbol{Y} - \sum_{k=0}^{p}[\hat{\gamma}_{k1} + \bar{B}(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}_{k*}]\boldsymbol{G}_k||_2 + n\sum_{d=2}^{q} p_{\lambda_3}(|\beta_d|)$$

For an given initial value $\tilde{\boldsymbol{\beta}}$ and doing a local linear approximation of $\bar{B}(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}_{k*}\boldsymbol{G}_k$ at $\tilde{\boldsymbol{\beta}}$, we have

$$\bar{B}(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}_{k*}\boldsymbol{G}_k \approx \bar{B}(\boldsymbol{X}\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_{k*}\boldsymbol{G}_k + \bar{B}'(\boldsymbol{X}\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_{k*}\boldsymbol{X}\boldsymbol{G}_k(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})$$

For the $d-th$ coordinate of $\boldsymbol{\beta}$, β_d, we have

$$\bar{B}(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}^*_k\boldsymbol{G}_k \approx \bar{B}(\boldsymbol{X}\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}^*_k\boldsymbol{G}_k + \bar{B}'(\boldsymbol{X}\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}^*_k\boldsymbol{X}_d\boldsymbol{G}_k(\beta_d - \tilde{\beta}_d)$$

Thus we could obtain $\hat{\beta}_d$ by minimizing the following:

$$Q_d = ||\boldsymbol{Y}^*_d - \boldsymbol{X}^*_d\beta_d||_2^2 + np_{\lambda_3}(|\beta_d|)$$

where

$$\boldsymbol{Y}^*_d = \boldsymbol{Y} - \textstyle\sum_{k=0}^{p}[\hat{\gamma}_{k1}\boldsymbol{G}_k + \bar{B}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}^*_k\boldsymbol{G}_k - \bar{B}'(\boldsymbol{X}\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}^*_k\boldsymbol{G}_k\boldsymbol{X}_d\tilde{\beta}_d]$$

$$\boldsymbol{X}^*_d = \textstyle\sum_{k=0}^{p}\bar{B}'(\boldsymbol{X}\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}^*_k\boldsymbol{G}_k\boldsymbol{X}_d$$

Hence, we obtained the MCP penalized estimator $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}^*_1, \ldots, \hat{\beta}^*_q)^T$ via the following

(1) Given an initial estimate of $\boldsymbol{\beta}$, denoted as $\tilde{\boldsymbol{\beta}}$, calculate $\boldsymbol{Y}^*_d and \boldsymbol{X}^*_d$ according to the above formula.

(2) Normalize $\boldsymbol{X}^*_d$ by $\boldsymbol{X}^{*'}_d = \boldsymbol{X}^*_d/||\boldsymbol{X}^*_d||_2$.

(3) Calculate $\hat{\beta}^{OLS}_d = \boldsymbol{X}^{*'T}_d\boldsymbol{Y}^*_d$.

(4) Set $\hat{\beta}^*_1 = \hat{\beta}^{OLS}_1$ and for $d \neq 1, \hat{\beta}^*_d = \frac{(\hat{\beta}^{OLS}_d - \lambda)_+}{1-1/\gamma^{MCP}}$ if $|\hat{\beta}^{OLS}_d| \leq \lambda\gamma^{MCP}$ and $\hat{\beta}^*_d = \hat{\beta}^{OLS}_d$ if $|\hat{\beta}^{OLS}_d| > \lambda\gamma^{MCP}$.

(5) Adjust $\hat{\beta}^*_d$ by $\hat{\beta}^{adjusted}_d = \hat{\beta}^*_d * ||\boldsymbol{X}^*_d||_2$.

(6) Repeat steps (1) through (5) for all β_d, $d = 1, \ldots, q$.

(7) Normalize $\hat{\boldsymbol{\beta}}^{adjusted}$, i.e. $\hat{\beta}_d = \dfrac{\hat{\beta}_d^{adjusted}}{||\hat{\boldsymbol{\beta}}^{adjusted}||_2} * sign(\beta_1^{adjusted})$.

(8) Update $\tilde{\boldsymbol{\beta}}$ in step (1) with $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_q)^T$ and iterate step (1) through (6) until convergence.

B.2 Algorithm for model (2.6)

For the intercept only model, $Y = m_0(\boldsymbol{X\beta}) + \boldsymbol{\epsilon}$, the following contain the steps in estimation.

(0) We approximated $m_0(\boldsymbol{X\beta})$ with B-spline basis function: $m_0(X\boldsymbol{\beta}) \approx \gamma_{01} + \bar{B}(\boldsymbol{X\beta})\boldsymbol{\gamma}_{0*}$.

(1) Given initial value for $\boldsymbol{\beta}$, denoted as $\hat{\boldsymbol{\beta}}$, let the design matrix $\boldsymbol{D} = (\mathbf{1}, \bar{\mathbf{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}))$, denote $\boldsymbol{\gamma} = (\gamma_{01}, \boldsymbol{\gamma}_{0*})$, we estimated $\boldsymbol{\gamma}$ with $\hat{\boldsymbol{\gamma}} = (\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{Y}$.

(2) Obtain $\boldsymbol{\beta}^{updated}$ via minimizing $||\boldsymbol{Y} - \hat{\gamma}_{01} - \bar{B}(\boldsymbol{X\beta})\hat{\boldsymbol{\gamma}}_{0*}||_2^2$ with Newton-Raphson algorithm.

(3) Replace $\hat{\boldsymbol{\beta}}$ by $\boldsymbol{\beta}^{updated}$ in step (1), and iterate until convergence.

B.3 Algorithm for gVMICM

Here we showed the algorithm used in gVMICM:

Step 1 & *Step 2* followed directly from the group coordinate descent algorithm described before, hence the details are omitted.

Step 3:

$$\hat{\boldsymbol{\beta}} = \max_{||\boldsymbol{\beta}||_2=1} M_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}^{(2)}) = \max_{||\boldsymbol{\beta}||_2=1} \left(l(\hat{\boldsymbol{\gamma}}^{(2)}, \boldsymbol{\beta}) - n\sum_{d=2}^{q} p_{\lambda_3}(|\beta_d|) \right)$$

We implemented the local quadratic approximation technique proposed by Fan and Li(2001)(Fan and Li (2001)). Denote $\tilde{\boldsymbol{\beta}}$ to be the most updated value of $\boldsymbol{\beta}$, by Taylor expansion at $\tilde{\boldsymbol{\beta}}$, we have

$$l(\hat{\boldsymbol{\gamma}}^{(2)}, \boldsymbol{\beta}) \approx l(\hat{\boldsymbol{\gamma}}^{(2)}, \tilde{\boldsymbol{\beta}}) + \bigtriangledown l(\hat{\boldsymbol{\gamma}}^{(2)}, \tilde{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T \bigtriangledown^2 l(\hat{\boldsymbol{\gamma}}^{(2)}, \tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})$$

where

$$\bigtriangledown l(\hat{\boldsymbol{\gamma}}^{(2)}, \tilde{\boldsymbol{\beta}}) = (\frac{\partial l(\hat{\boldsymbol{\gamma}}^{(2)}, \boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial l(\hat{\boldsymbol{\gamma}}^{(2)}, \boldsymbol{\beta})}{\partial \beta_q})^T \Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} \text{ is the gradient}$$

$$\bigtriangledown^2 l(\hat{\boldsymbol{\gamma}}^{(2)}, \tilde{\boldsymbol{\beta}}) = \left[\frac{\partial^2 l(\hat{\boldsymbol{\gamma}}^{(2)}, \boldsymbol{\beta})}{\partial \beta_j \partial \beta_l}\right]_{1 \le j,l \le q} \Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} \text{ is the hessian matrix}$$

and

$$\frac{\partial l(\hat{\boldsymbol{\gamma}}^{(2)}, \boldsymbol{\beta})}{\partial \beta_j} = (\boldsymbol{Y} - \frac{1}{1 + e^{-\boldsymbol{\mu}^{B^{(3)}}}})^T \frac{\partial \boldsymbol{\mu}^{B^{(3)}}}{\partial \beta_j}$$

$$\frac{\partial^2 l(\hat{\boldsymbol{\gamma}}^{(2)}, \boldsymbol{\beta})}{\partial \beta_j \partial \beta_l} = (\boldsymbol{Y} - \frac{1}{1 + e^{-\boldsymbol{\mu}^{B^{(3)}}}})^T \frac{\partial \boldsymbol{\mu}^{B^{(3)}}}{\partial \beta_j \partial \beta_l} - \left(\frac{e^{-\boldsymbol{\mu}^{B^{(3)}}}}{(1 + e^{-\boldsymbol{\mu}^{B^{(3)}}})^2}\right)^T \frac{\partial \boldsymbol{\mu}^{B^{(3)}}}{\partial_{\beta_j}} \cdot \frac{\partial \boldsymbol{\mu}^{B^{(3)}}}{\partial_{\beta_l}}$$

with

$$\boldsymbol{\mu}^{B^{(3)}} = \sum_{k=0}^{p} [\hat{\gamma}_{k1}^{(2)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}_{k*}^{(2)}] \cdot \boldsymbol{G}_k$$

$$\frac{\partial \boldsymbol{\mu}^{B^{(3)}}}{\partial \beta_j} = \sum_{k=0}^{p} \left(\bar{\boldsymbol{B}}'(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}_{k*}^{(2)} \cdot \boldsymbol{X}_j \cdot \boldsymbol{G}_k\right)$$

$$\frac{\partial \boldsymbol{\mu}^{B^{(3)}}}{\partial \beta_j \partial \beta_l} = \sum_{k=0}^{p} \left(\bar{\boldsymbol{B}}''(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}_{k*}^{(2)} \cdot \boldsymbol{X}_j \cdot \boldsymbol{X}_l \cdot \boldsymbol{G}_k\right)$$

Let $\boldsymbol{\Sigma}_{\lambda_3}(\tilde{\boldsymbol{\beta}}) = diag(0, \frac{p'_{\lambda_3}(\hat{\beta_2})}{|\hat{\beta_2}|}, \ldots, \frac{p'_{\lambda_3}(\hat{\beta_q})}{|\hat{\beta_q}|})$. Then, we updated $\boldsymbol{\beta}$ with

$$\boldsymbol{\beta}^* = \tilde{\boldsymbol{\beta}} - \Big[\bigtriangledown^2 l(\hat{\boldsymbol{\gamma}}^{(2)}, \tilde{\boldsymbol{\beta}}) + n\boldsymbol{\Sigma}_{\lambda_3}(\tilde{\boldsymbol{\beta}})\Big]^{-1} \Big[\bigtriangledown l(\hat{\boldsymbol{\gamma}}^{(2)}, \tilde{\boldsymbol{\beta}}) + n\boldsymbol{\Sigma}_{\lambda_3}(\tilde{\boldsymbol{\beta}})\tilde{\boldsymbol{\beta}}\Big]$$

Then we standardized $\boldsymbol{\beta}$ with $\boldsymbol{\beta}^{updated} = sign(\beta_1^*)\frac{\boldsymbol{\beta}^*}{||\boldsymbol{\beta}^*||_2}$. We iterate the above steps until convergence.

B.4 Algorithm for quantile VMICM

Here, we presented the algorithm used in the quantile VMICM model.

Step 1&2: The algorithm used in these steps followed directly from Peng and Wang(2015) and we implemented the R-package "rqPen" (Sherwood and Maidman (2016)).

Step 3: Obtain $\hat{\boldsymbol{\beta}} = \min_{||\boldsymbol{\beta}||_2=1} Q_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}^{(2)})$ where

$$Q_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \rho_\tau\big(Y_i - \sum_{k=0}^{p}[\hat{\gamma}_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}_{k*}]G_{ik}\big) + n\sum_{d=2}^{q} p_{\lambda_3}(|\beta_d|)$$

Since $\bar{B}(\boldsymbol{X}\boldsymbol{\beta})$ is not a linear function of $\boldsymbol{\beta}$, we adopted the idea of first order approximation for $\bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})$. Denote $\tilde{\boldsymbol{\beta}}$ as the most updated value of $\boldsymbol{\beta}$. We have

$$\bar{B}(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}_{k*}\boldsymbol{G}_k \approx \bar{B}(\boldsymbol{X}\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_{k*}\boldsymbol{G}_k + \bar{B}'(\boldsymbol{X}\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_{k*}\boldsymbol{X}\boldsymbol{G}_k(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})$$

For individual β_d, $d = 1, \ldots q$, we have

$$\bar{B}(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}_k^*\boldsymbol{G}_k \approx \bar{B}(\boldsymbol{X}\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_k^*\boldsymbol{G}_k + \bar{B}'(\boldsymbol{X}\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_k^*\boldsymbol{X}_d\boldsymbol{G}_k(\beta_d - \tilde{\beta}_d)$$

Then we could obtain $\hat{\beta}_d$ by minimizing the following:

$$Q_d = \rho_\tau(\boldsymbol{Y}_d^* - \boldsymbol{X}_d^*\beta_d) + np_{\lambda_3}(|\beta_d|)$$

where $\boldsymbol{Y}_d^* = \boldsymbol{Y} - \sum_{k=0}^{p}[\hat{\gamma}_{k1}\boldsymbol{G}_k + \bar{B}(\boldsymbol{X}^T\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_k^*\boldsymbol{G}_k - \bar{B}'(\boldsymbol{X}\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_k^*\boldsymbol{G}_k\boldsymbol{X}_d\tilde{\beta}_d]$ and $\boldsymbol{X}_d^* = \sum_{k=0}^{p}\bar{B}'(\boldsymbol{X}\tilde{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_k^*\boldsymbol{G}_k\boldsymbol{X}_d$. Hence, we obtained the MCP penalized estimator $\hat{\boldsymbol{\beta}}^* = (\beta_1^*, \ldots, \beta_q^*)^T$ via the following steps:

(1) For given initial value for $\boldsymbol{\beta}$, denoted as $\tilde{\boldsymbol{\beta}}$, calculate $\boldsymbol{Y}_d^*$ and $\boldsymbol{X}_d^*$ according to the above formula.

(2) Obtain $\hat{\boldsymbol{\beta}}^* = (\beta_1^*, \ldots, \beta_q^*)^T$ via implementing the iterative coordinate descent algorithm for quantile regression.

(3) Standardize $\hat{\boldsymbol{\beta}}^*$ via $\hat{\boldsymbol{\beta}} = sign(\hat{\beta}_1^*)\frac{\hat{\boldsymbol{\beta}}^*}{||\hat{\boldsymbol{\beta}}^*||_2}$.

(4) Iterate steps (1) to (3) until convergence.

B.5 Algorithm for model (4.6)

Consider the intercept only model $Y = m_0(\boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\epsilon}$, we proposed an iterative algorithm to estimate its parameters in the following.

(0) Approximate $m_0(\boldsymbol{X}\boldsymbol{\beta})$ with the B-spline basis function, i.e., $m_0(X\boldsymbol{\beta}) \approx \gamma_{01} + \bar{B}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{\gamma}_{0*}$.

(1) For given initial value for $\boldsymbol{\beta}$, denoted as $\hat{\boldsymbol{\beta}}$, obtain $\hat{\gamma}_{01}$ and $\hat{\boldsymbol{\gamma}}_{0*}$ via the "rq" function in R (in package "quantreg"), with $\bar{B}(\boldsymbol{X}\hat{\boldsymbol{\beta}})$ being the design matrix.

(2) Obtain $\boldsymbol{\beta}^{updated}$ via minimizing $\sum_{i=1}^{n}\rho_\tau(\boldsymbol{Y}_i - \hat{\gamma}_{01} - \bar{B}(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}_{0*})$ with the linear approximation method described in Appendix B.4.

(3) Update $\hat{\boldsymbol{\beta}}$ by $\boldsymbol{\beta}^{updated}$, and iterate until convergence.

Appendix C

Proof of Theorems

Some notations: Denote the space of Lipschitz continuous functions for any fixed constant c as $Lip([a,b],c) = \{f : |f(x_1) - f(x_2)| \leq c|x_1 - x_2|, \forall x_1, x_2 \in [a,b]\}$. Let $C^{(p)}[a,b] = \{f : f^{(p)} \in C[a,b]\}$ be the space of the pth order smooth functions.

C.1 Proof of Theorem 2.3.1

We assume the following regularity conditions:

(A1) The density function $f_{U(\boldsymbol{\beta})}(\cdot)$ of random variable $U(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$ is bounded away from 0 on $\boldsymbol{\Omega} = \{\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X} \in \mathscr{X}\}$, where $\mathscr{X}$ is the compact support of $\boldsymbol{X}$. There exists a constant c_1, such that $f_{U(\boldsymbol{\beta})}(\cdot) \in Lip([a,b],c_1)$.

(A2) For $k = 0, 1, \ldots, p$, the non-parametric function $m_k(\cdot) \in C^{(r)}$ and $r \geq 2$.

(A3) $E(||\boldsymbol{G}||^6) < \infty$ and $E(|\boldsymbol{\epsilon}|^6) < \infty$.

(A4) The matrix $\boldsymbol{M}(u) = E(\boldsymbol{G}\boldsymbol{G}^T|\boldsymbol{X}\boldsymbol{\beta} = u)$ is positive definite, each element of $\boldsymbol{M}(u) \in Lip([a,b],c_4)$.

(A5) Let $b_n = \max_{k,l}\{p''_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}||_2), p''_{\lambda_2}(|\boldsymbol{\gamma}^0_{k1}|), p''_{\lambda_3}(|\beta^0_d|), \boldsymbol{\gamma}^0_{k*} \neq 0, \boldsymbol{\gamma}^0_{k1} \neq 0, \beta^0_l \neq 0\}$ for $k = 1, \ldots, p, d = 2, \ldots, q$. Then $b_n \to 0$ as $n \to 0$

(A6)

$$\liminf_{n\to\infty} \liminf_{||\boldsymbol{\gamma}_{k*}||_2 \to 0^+} \frac{1}{\lambda_1}|p'_{\lambda_1}(||\boldsymbol{\gamma}_{k*}||_2)| > 0 \text{ for } k = v+1, \ldots, p$$

$$\liminf_{n\to\infty} \liminf_{|\boldsymbol{\gamma}_{k1}|\to 0^+} \frac{1}{\lambda_2}|p'_{\lambda_2}(|\boldsymbol{\gamma}_{k1}|)| > 0 \text{ for } k = c+1, \ldots, p$$

$$\liminf_{n\to\infty} \liminf_{|\beta_d|\to 0^+} \frac{1}{\lambda_3}|p'_{\lambda_3}(|\beta_d|)| > 0 \text{ for } d = s+1, \ldots, q$$

(A7) Let $c_1, \ldots, c_K$ be the interior knots of $[a, b]$, where $a = \inf\{u : u \in \boldsymbol{\Omega}\}$, $b = \sup\{u : u \in \boldsymbol{\Omega}\}$ and $c_0 = 1$, $c_{K+1} = b$, $h_i = c_i - c_{i-1}$, $h = \max\{h_i\}$. Then exists a constant C_7 such that $\frac{h}{\min\{h_i\}} < C_7$ and $\max\{h_{i+1}h_i\} = o(K^{-1})$.

Lemma C.1.1. *If $m_k(u), k = 0, 1, \ldots, p$ satisfy condition (A2), then there exists a constant $C > 0$ such that*

$$\sup_{u\in\boldsymbol{\Omega}} |m_k(u) - \gamma_{k1} - \bar{\boldsymbol{B}}(u)\boldsymbol{\gamma}_{k*}| \leq CK^{-r}.$$

Proof: This result follows directly from a standard B-spline theory.

Denote $\boldsymbol{\phi} = (\beta_2, \ldots \beta_q)^T$, hence $\boldsymbol{\beta} = (\sqrt{1 - ||\boldsymbol{\phi}||_2^2}, \boldsymbol{\phi}^T)^T$ and $\boldsymbol{\phi}^0$ is the true value of $\boldsymbol{\phi}$. To show the consistency of $\hat{\boldsymbol{\beta}}$, it is equivalent to show the consistency of $\hat{\boldsymbol{\phi}} = (\hat{\beta}_2, \cdots, \hat{\beta}_q)$. Let $\alpha_n = n^{-r/(2r+1)} + a_n$, $\boldsymbol{\gamma} = \boldsymbol{\gamma}^0 + \alpha_n\boldsymbol{\tau}_1$, $\boldsymbol{\phi} = \boldsymbol{\phi}^0 + \alpha_n\boldsymbol{\tau}_2$ and $\boldsymbol{\tau} = (\boldsymbol{\tau}_1, \boldsymbol{\tau}_2)$, where $\boldsymbol{\tau}_1 = (\tau_{01}, \boldsymbol{\tau}_{0*}, \ldots, \tau_{p1}, \boldsymbol{\tau}_{p*})$ and $\{\tau_{k1}, \boldsymbol{\tau}_{k*}\}$ corresponds to the B-spline coefficients $\gamma_{k1}, \boldsymbol{\gamma}_{k*}$; $\boldsymbol{\tau}_2 = (\tau_1^{\phi}, \ldots, \tau_{q-1}^{\phi})$ and τ_l^{ϕ} corresponds to ϕ_l.

To show the consistency of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\phi}}$, we need to show that $\forall\epsilon > 0$, $\exists$ a large enough C, so that:

$$P\big(\inf_{||\boldsymbol{\tau}||=C}\{Q_1(\boldsymbol{\gamma}, \boldsymbol{\phi})\} > Q_1(\boldsymbol{\gamma}^0, \boldsymbol{\phi}^0)\big) \geq 1 - \epsilon \tag{C.1}$$

where

$$Q_1(\boldsymbol{\gamma}, \boldsymbol{\phi}) = g(\boldsymbol{\gamma}, \boldsymbol{\phi}) + n\sum_{k=1}^{p} p_{\lambda_1}(||\boldsymbol{\gamma}_{k*}||_2) + n\sum_{k=1}^{p} p_{\lambda_2}(|\gamma_{k1}|)I(||\boldsymbol{\gamma}_{k*}||_2 = 0) + n\sum_{l=1}^{q-1} p_{\lambda_3}(|\phi_l|)$$

and $g(\boldsymbol{\gamma}, \boldsymbol{\phi}) = ||\boldsymbol{Y} - \sum_{k=0}^{p}(\gamma_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{\phi})\boldsymbol{\gamma}_{k*})\boldsymbol{G}_k||_2^2$ is the squared error loss. If (C.1) holds, we

can say with probability at least $1-\epsilon$, there exists a local minimum in the ball $\{(\boldsymbol{\gamma}^0, \boldsymbol{\phi}^0)+\alpha_n * \boldsymbol{\tau}; ||\boldsymbol{\tau}|| \leq C\}$. Hence, there exists a local minimizer such that $||(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}}) - (\boldsymbol{\gamma}^0, \boldsymbol{\phi}^0)|| = O_p(\alpha_n)$.

Let

$$
\begin{aligned}
D_n(\boldsymbol{\tau}) &= \frac{1}{K}\{Q_1(\boldsymbol{\gamma}, \boldsymbol{\phi}) - Q_1(\boldsymbol{\gamma}^0, \boldsymbol{\phi}^0)\} = \frac{1}{K}\{Q_1(\boldsymbol{\gamma}^0 + \alpha_n \boldsymbol{\tau}_1, \boldsymbol{\phi}^0 + \alpha_n \boldsymbol{\tau}_2) - Q_1(\boldsymbol{\gamma}^0, \boldsymbol{\phi}^0)\} \\
&= \frac{1}{K}\Big\{ g(\boldsymbol{\gamma}^0 + \alpha_n \boldsymbol{\tau}_1, \boldsymbol{\phi}^0 + \alpha_n \boldsymbol{\tau}_2) - g(\boldsymbol{\gamma}^0, \boldsymbol{\phi}^0) \\
&+ n \sum_{k=1}^{p} \left[p_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*} + \alpha_n \boldsymbol{\tau}_{k*}||_2) - p_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}||_2) \right] \\
&+ n \sum_{k=1}^{p} \left[p_{\lambda_2}(|\gamma^0_{k1} + \alpha_n \tau_{k1}|) I(||\boldsymbol{\gamma}^0_{k*} + \alpha_n \tau_{k*}||_2 = 0) - p_{\lambda_2}(|\gamma^0_{k1}|) I(||\boldsymbol{\gamma}^0_{k*}||_2 = 0) \right] \\
&+ n \sum_{j=1}^{q-1} \left[p_{\lambda_3}(|\phi^0_j + \alpha_n \tau^{\phi}_j|) - p_{\lambda_3}(|\phi^0_j|) \right] \Big\}
\end{aligned}
$$

Since $p_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}||_2)] = 0$ for $k = v+1, \ldots, p$ and $p_{\lambda_3}(|\phi^0_j|) = 0$ for $j = s+1, \ldots, q-1$ and $I(||\boldsymbol{\gamma}^0_{k*}||_2 = 0) = 0$ for $k = 1, \ldots, v$, we have

$$
\begin{aligned}
D_n(\boldsymbol{\tau}) \geq \frac{1}{K}\Big\{ & g(\boldsymbol{\gamma}^0 + \alpha_n \boldsymbol{\tau}_1, \boldsymbol{\phi}^0 + \alpha_n \boldsymbol{\tau}_2) - g(\boldsymbol{\gamma}^0, \boldsymbol{\phi}^0) \\
& + n \sum_{k=1}^{v} [p_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*} + \alpha_n \boldsymbol{\tau}_{k*}||_2) - p_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}||_2)] \\
& + n \sum_{k=v+1}^{p} \left[p_{\lambda_2}(|\gamma^0_{k1} + \alpha_n \tau_{k1}|) - p_{\lambda_2}(|\gamma^0_{k1}|) \right] \\
& + n \sum_{j=1}^{s-1} [p_{\lambda_3}(|\phi^0_j + \alpha_n \tau^{\phi}_j|) - p_{\lambda_3}(|\phi^0_j|)] \Big\}
\end{aligned}
$$

Then by Taylor series expansion at $(\boldsymbol{\gamma}^0, \boldsymbol{\phi}^0)$, we have

$$
\begin{aligned}
D_n(\boldsymbol{\tau}) \geq & \frac{\alpha_n}{K}(\frac{\partial g}{\partial \boldsymbol{\gamma}^0}, \frac{\partial g}{\partial \boldsymbol{\phi}^0})\boldsymbol{\tau}^T - \frac{1}{2K} n\alpha_n^2 \boldsymbol{\tau} \boldsymbol{I}(\boldsymbol{\gamma}^0, \boldsymbol{\phi}^0)\boldsymbol{\tau}^T(1+o_p(1)) \\
& + \frac{n}{K}\sum_{k=1}^{v}\left[\alpha_n p'_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}||_2)\frac{\boldsymbol{\gamma}^0_{k*}}{||\boldsymbol{\gamma}^0_{k*}||_2}\boldsymbol{\tau}^T_{k*} + \alpha_n^2 p''_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}||_2)\boldsymbol{\tau}_{k*}\boldsymbol{\tau}^T_{k*}(1+o_p(1))\right] \\
& + \frac{n}{K}\sum_{k=v+1}^{p}\left[\alpha_n p'_{\lambda_2}(|\gamma^0_{k1}|)sign(\gamma^0_{k1})\tau_{k1} + \alpha_n^2 p''_{\lambda_2}(|\gamma^0_{k1}|)(\tau_{k1})^2(1+o_p(1))\right] \\
& + \frac{n}{K}\sum_{j=1}^{s-1}\left[\alpha_n p'_{\lambda_3}(|\phi^0_j|)sign(\phi^0_j)\tau^\phi_j + \alpha_n^2 p''_{\lambda_3}(|\phi^0_j|)(\tau^\phi_j)^2(1+o_p(1))\right] \\
& := S_1 + S_2 + S_3 + S_4 + S_5
\end{aligned}
$$

where $\boldsymbol{I}(\boldsymbol{\gamma}^0, \boldsymbol{\phi}^0)$ is the Fisher's information matrix.

By standard arguments, S_1 is of the order $O_p(1+n^{r/(2r+1)}\alpha_n)||\boldsymbol{\tau}||$, S_2 is of the order $O_p(1+2n^{r/(2r+1)}\alpha_n)||\boldsymbol{\tau}||^2$ and for large enought C, S_2 dominates S_1 uniformly in $||\boldsymbol{\tau}|| = C$. Further, by Taylor expansion at $\boldsymbol{\gamma}^0$, we have

$$
\begin{aligned}
S_3 &\leq \frac{n}{K}\alpha_n a_n \sum_{k=1}^{v}\frac{\boldsymbol{\gamma}^0_{k*}}{||\boldsymbol{\gamma}^0_{k*}||_2}\boldsymbol{\tau}^T_{k*} + \frac{n}{K}\alpha_n^2 \max_k\{p''_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}||_2)\}\sum_{k=1}^{v}\boldsymbol{\tau}_{k*}\boldsymbol{\tau}^T_{k*} \\
&\leq \frac{n}{K}\alpha_n^2\sqrt{v} + \frac{n}{K}\alpha_n^2 C\max_k\{p''_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}||_2)\}
\end{aligned}
$$

Since $\max_k\{p''_{\lambda_1}\} \to 0$, S_3 is dominated by S_2.

For S_4 and S_5, we have

$$
\begin{aligned}
S_4 &\leq \alpha_n a_n \frac{n}{K}\sum_{k=v+1}^{p}\tau_{k1} + \frac{n}{K}\alpha_n^2\max_k\{p''_{\lambda_2}(|\gamma^0_{k1}|)\}\sum_{k=v+1}^{p}(\tau_{k1})^2 \\
&\leq \frac{n}{K}\alpha_n^2 C + \frac{n}{K}\alpha_n^2 C^2\max_k\{p''_{\lambda_2}(|\gamma^0_{k1}|)\}
\end{aligned}
$$

and

$$\begin{aligned} S_5 &\leq \alpha_n a_n \frac{n}{K} \sum_{j=1}^{s} \tau_j^{\phi} + \frac{n}{K} \alpha_n^2 \max_j \{p''_{\lambda_3}(|\phi_j^0|)\} \sum_{j=1}^{s} (\tau_j^{\phi})^2 \\ &\leq \frac{n}{K} \alpha_n^2 C + \frac{n}{K} \alpha_n^2 C^2 \max_j \{p''_{\lambda_3}(|\phi_j^0|)\} \end{aligned}$$

Similarly, we have S_4 and S_5 being dominated by S_2. Hence, by choosing a large enough C, we have $||(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}}) - (\boldsymbol{\gamma}^0, \boldsymbol{\phi}^0)|| = O_p(\alpha_n)$. This proves the consistency of the penalized least square estimator $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}})$.

C.2 Proof of Theorem 2.3.2

Without loss of generality, we denote $\boldsymbol{\phi} = (\boldsymbol{\phi}^{nz}, \boldsymbol{\phi}^{z})$, where $\boldsymbol{\phi}^{nz} = (\phi_1, \ldots, \phi_{s-1})$ and $\boldsymbol{\phi}^{z} = (\phi_s, \ldots, \phi_{q-1})$. From $\lambda_{max} \to 0$, it is very easy to see that $a_n = 0$ for large n. Then by Theorem 2.3.1, it is sufficient to show for $\boldsymbol{\phi}^{nz}$,

$$||\phi_j - \phi_j^0||_2 = O_p(n^{-r/(2r+1)}), j = 1, \ldots, s-1$$

and for $\boldsymbol{\phi}^{z}$, for some given small $\epsilon = Cn^{-r/(2r+1)}$, when $n \to \infty$, with probability approaching 1, for $j = s, \ldots, q-1$, we have

$$\frac{\partial Q_1(\boldsymbol{\gamma}, \boldsymbol{\phi})}{\partial \phi_j} > 0 \text{ when } 0 < \phi_j < \epsilon \text{ and } \frac{\partial Q_1(\boldsymbol{\gamma}, \boldsymbol{\phi})}{\partial \phi_j} < 0 \text{ when } -\epsilon < \phi_j < 0$$

Since we have

$$\frac{\partial Q_1(\boldsymbol{\gamma}, \boldsymbol{\phi})}{\partial \phi_j} = \frac{\partial g(\boldsymbol{\gamma}, \boldsymbol{\phi})}{\partial \phi_j} + n p_{\lambda_3}(|\phi_j|) sign(\phi_j).$$

Do Taylor expansion at $\boldsymbol{\phi}^0$ for $\frac{\partial g(\boldsymbol{\gamma},\boldsymbol{\phi})}{\partial \phi_j}$ only, we have

$$\frac{\partial Q_1(\boldsymbol{\gamma},\boldsymbol{\phi})}{\partial \phi_j} = \frac{\partial g(\boldsymbol{\gamma},\boldsymbol{\phi}^0)}{\partial \phi_j} + \sum_{l=1}^{q-1} \frac{\partial^2 g(\boldsymbol{\gamma},\boldsymbol{\phi}^0)}{\partial \phi_j \partial \phi_l}(\phi_l - \phi_l^0) + \sum_{l=1}^{q-1}\sum_{k=1}^{q-1} \frac{\partial^3 g(\boldsymbol{\gamma},\boldsymbol{\phi}^*)}{\partial \phi_j \partial \phi_l \partial \phi_k}(\phi_l - \phi_l^0)(\phi_k - \phi_k^0) + n p_{\lambda_3}(|\phi_j|) sign(\phi_j)$$

where $\boldsymbol{\phi}^*$ lies between $\boldsymbol{\phi}^0$ and $\boldsymbol{\phi}$. After some calculation, we have

$$\frac{\partial Q_1(\boldsymbol{\gamma},\boldsymbol{\phi})}{\partial \phi_j} = n\lambda_3 \{ \frac{1}{\lambda_3} p'_{\lambda_3}(|\phi_j|) sign(\phi_j) + O_p(\frac{1}{\lambda_3} n^{-r/(2r+1)}) \}$$

Since $\lim_{n\to\infty} \liminf_{\phi_j \to 0} \frac{1}{\lambda_3} p'_{\lambda_3}(|\phi_j|) > 0$ and $\frac{1}{\lambda_3} n^{-r/(2r+1)} \to 0$, we can conclude that the sign of $\frac{\partial Q_1(\boldsymbol{\gamma},\boldsymbol{\phi})}{\partial \phi_j}$ is completely determined by sign of ϕ_j. Hence, we have proven $\hat{\beta}_j = 0$ for $j = s+1, \ldots, q$

For (ii) & (iii), applying similar arguments as in (i), we immediately have, with probability approaching 1, $\hat{\boldsymbol{\gamma}}_{k*} = 0$ for $k = v+1, \ldots, p$ and $\hat{\gamma}_{k1} = 0$ for $k = c+1, \ldots, p$. Then by $\sup_u \boldsymbol{B}(u) = O(1)$ and $\hat{m}_k(\cdot) = \hat{\gamma}_{k0} + \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_{k*}$, we have proven $\hat{m}_k(\cdot) = c_k$ for $k = v+1, \ldots, c$ where c_k is some constant and $\hat{m}_k(\cdot) = 0$ for $k = c+1, \ldots, p$.

C.3 Proof of Theorem 3.3.1

We assume the following regularity conditions:

(A1) The density function $f_{U(\boldsymbol{\beta})}(\cdot)$ of a random variable $U(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$ is bounded away from 0 on $\boldsymbol{\Omega} = \{\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X} \in \mathscr{X}\}$, where $\mathscr{X}$ is the compact support of $\boldsymbol{X}$. And there exists a constant c_1, such that $f_{U(\boldsymbol{\beta})}(\cdot) \in Lip([a,b], c_1)$.

(A2) For $k = 0, 1, \ldots, p$, the non-parametric function $m_k(\cdot) \in C^{(r)}$ and $r \geq 2$.

(A3) $E(||\boldsymbol{G}||^6) < \infty$.

(A4) The matrix $\boldsymbol{M}(u) = E(\boldsymbol{G}\boldsymbol{G}^T|\boldsymbol{X}\boldsymbol{\beta} = u)$ is positive definite, and each element of $\boldsymbol{M}(u) \in Lip([a,b], c_4)$

(A5) Let $b_n = \max_{k,l}\{p''_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}||_2), p''_{\lambda_2}(|\boldsymbol{\gamma}^0_{k1}|), p''_{\lambda_3}(|\beta^0_d|), \boldsymbol{\gamma}^0_{k*} \neq 0, \boldsymbol{\gamma}^0_{k1} \neq 0, \beta^0_l \neq 0\}$ for $k = 1, \ldots, p, d = 2, \ldots, q$, then $b_n \to 0$ as $n \to 0$.

(A6)

$$\liminf_{n\to\infty} \liminf_{||\boldsymbol{\gamma}_{k*}||_2 \to 0^+} \frac{1}{\lambda_1}|p'_{\lambda_1}(||\boldsymbol{\gamma}_{k*}||_2)| > 0 \text{ for } k = v+1, \ldots, p$$

$$\liminf_{n\to\infty} \liminf_{|\boldsymbol{\gamma}_{k1}| \to 0^+} \frac{1}{\lambda_2}|p'_{\lambda_2}(|\boldsymbol{\gamma}_{k1}|)| > 0 \text{ for } k = c+1, \ldots, p$$

$$\liminf_{n\to\infty} \liminf_{|\beta_d| \to 0^+} \frac{1}{\lambda_3}|p'_{\lambda_3}(|\beta_d|)| > 0 \text{ for } d = s+1, \ldots, q$$

(A7) Let $c_1, \ldots, c_K$ be the interior knots of $[a,b]$, $a = \inf\{u : u \in \boldsymbol{\Omega}\}$, $b = \sup\{u : u \in \boldsymbol{\Omega}\}$ and $c_0 = 1$, $c_{K+1} = b$, $h_i = c_i - c_{i-1}$, $h = \max\{h_i\}$. Then there exists a constant C_7 such that $\frac{h}{\min\{h_i\}} < C_7$ and $\max\{h_{i+1}h_i\} = o(K^{-1})$.

Let $\boldsymbol{\phi} = (\beta_2, \ldots, \beta_q)^T$, and we have $\boldsymbol{\beta} = (\sqrt{1 - ||\boldsymbol{\phi}||_2^2}, \boldsymbol{\phi}^T)^T$, hence the restriction $||\boldsymbol{\beta}|| = 1$ and $\beta_1 > 0$ is equivalent to $||\boldsymbol{\phi}||_2 < 1$. To show the consistency of $\hat{\boldsymbol{\beta}}$, it is enough to show the consistency of $\hat{\boldsymbol{\phi}}$. Let $\alpha_n = n^{-r/(2r+1)} + a_n$, $\boldsymbol{\gamma} = \boldsymbol{\gamma}^0 + \alpha_n\boldsymbol{\tau}_1$, $\boldsymbol{\phi} = \boldsymbol{\phi}^0 + \alpha_n\boldsymbol{\tau}_2$ and $\boldsymbol{\tau} = (\boldsymbol{\tau}_1, \boldsymbol{\tau}_2)$, where $\tau_1 = (\tau_{01}, \boldsymbol{\tau}_{0*}, \ldots, \tau_{p1}, \boldsymbol{\tau}_{p*})$ and $\{\tau_{k1}, \boldsymbol{\tau}_{k*}\}$ corresponds to the B-spline coefficients $\gamma_{k1}, \boldsymbol{\gamma}_{k*}$; $\boldsymbol{\tau}_2 = (\tau^\phi_1, \ldots, \tau^\phi_{q-1})$ and τ^ϕ_l corresponds to ϕ_l.

To show consistency of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\phi}}$, we need to show $\forall \epsilon > 0$, $\exists$ a large enough C, so that:

$$P\Big(\sup_{||\boldsymbol{\tau}||=C}\{M(\boldsymbol{\gamma}, \boldsymbol{\phi})\} < M(\boldsymbol{\gamma}^0, \boldsymbol{\phi}^0)\Big) \geq 1 - \epsilon \tag{C.2}$$

where

$$M(\boldsymbol{\gamma},\boldsymbol{\phi}) = l(\boldsymbol{\gamma},\boldsymbol{\phi}) - n\sum_{k=1}^{p} p_{\lambda_1}(||\boldsymbol{\gamma}_{k*}||_2) - n\sum_{k=1}^{p} p_{\lambda_2}(|\gamma_{k1}|)I(||\boldsymbol{\gamma}_{k*}||_2 = 0) - n\sum_{l=1}^{q-1} p_{\lambda_3}(|\phi_l|)$$

and $l(\boldsymbol{\gamma},\boldsymbol{\phi})$ is the log-likelihood function defined above. If (C.2) holds, we can see, with probability at least $1-\epsilon$, there exists a local maximum in the ball $\{(\boldsymbol{\gamma}^0,\boldsymbol{\phi}^0) + \alpha_n * \boldsymbol{\tau}; ||\boldsymbol{\tau}|| \leq C\}$. Hence, there exists a local maximizer such that $||(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\phi}}) - (\boldsymbol{\gamma}^0,\boldsymbol{\phi}^0)|| = O_p(\alpha_n)$.

Let

$$\begin{aligned}
D_n(\boldsymbol{\tau}) &= \frac{1}{K}\{M(\boldsymbol{\gamma},\boldsymbol{\phi}) - M(\boldsymbol{\gamma}^0,\boldsymbol{\phi}^0)\} = \frac{1}{K}\{M(\boldsymbol{\gamma}^0 + \alpha_n\boldsymbol{\tau}_1, \boldsymbol{\phi}^0 + \alpha_n\boldsymbol{\tau}_2) - M(\boldsymbol{\gamma}^0,\boldsymbol{\phi}^0)\} \\
&= \frac{1}{K}\Big\{\big[l(\boldsymbol{\gamma}^0 + \alpha_n\boldsymbol{\tau}_1, \boldsymbol{\phi}^0 + \alpha_n\boldsymbol{\tau}_2) - l(\boldsymbol{\gamma}^0,\boldsymbol{\phi}^0)\big] \\
&- n\sum_{k=1}^{p}\big[p_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*} + \alpha_n\boldsymbol{\tau}_{k*}||_2) - p_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}||_2)\big] \\
&- n\sum_{k=1}^{p}\big[p_{\lambda_2}(|\gamma^0_{k1} + \alpha_n\tau_{k1}|)I(||\boldsymbol{\gamma}^0_{k*} + \alpha_n\tau_{k*}||_2 = 0) - p_{\lambda_2}(|\gamma^0_{k1}|)I(||\boldsymbol{\gamma}^0_{k*}||_2 = 0)\big] \\
&- n\sum_{j=1}^{q-1}\big[p_{\lambda_3}(|\phi^0_j + \alpha_n\tau^{\phi}_j|) - p_{\lambda_3}(|\phi^0_j|)\big]\Big\}
\end{aligned}$$

Since $p_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}||_2)] = 0$ for $k = v+1,\ldots,p$ and $p_{\lambda_3}(|\phi^0_j|) = 0$ for $j = s+1,\ldots,q-1$ and $I(||\boldsymbol{\gamma}^0_{k*}||_2 = 0) = 0$ for $k = 1,\ldots,v$, we have

$$
\begin{aligned}
D_n(\boldsymbol{\tau}) \;\leq\; & \frac{1}{K}\Big\{ l(\boldsymbol{\gamma}^0+\alpha_n\boldsymbol{\tau}_1, \boldsymbol{\phi}^0+\alpha_n\boldsymbol{\tau}_2) - l(\boldsymbol{\gamma}^0,\boldsymbol{\phi}^0) \\
& -n\sum_{k=1}^{v}[p_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}+\alpha_n\boldsymbol{\tau}_{k*}||_2) - p_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}||_2)] \\
& -n\sum_{k=v+1}^{p}\big[p_{\lambda_2}(|\gamma^0_{k1}+\alpha_n\tau_{k1}|) - p_{\lambda_2}(|\gamma^0_{k1}|)\big] \\
& -n\sum_{j=1}^{s-1}[p_{\lambda_3}(|\phi^0_j+\alpha_n\tau^{\phi}_j|) - p_{\lambda_3}(|\phi^0_j|)]\Big\}
\end{aligned}
$$

Then by Taylor expansion at $(\boldsymbol{\gamma}^0,\boldsymbol{\phi}^0)$, we have

$$
\begin{aligned}
D_n(\boldsymbol{\tau}) \;\leq\; & \frac{\alpha_n}{K}(\frac{\partial l}{\partial\boldsymbol{\gamma}^0},\frac{\partial l}{\partial\boldsymbol{\phi}^0})\boldsymbol{\tau}^T - \frac{1}{2K}n\alpha_n^2\boldsymbol{\tau}\boldsymbol{I}(\boldsymbol{\gamma}^0,\boldsymbol{\phi}^0)\boldsymbol{\tau}^T(1+o_p(1)) \\
& -\frac{n}{K}\sum_{k=1}^{v}\big[\alpha_n p'_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}||_2)\frac{\boldsymbol{\gamma}^0_{k*}}{||\boldsymbol{\gamma}^0_{k*}||_2}\boldsymbol{\tau}^T_{k*} + \alpha_n^2 p''_{\lambda_1}(||\boldsymbol{\gamma}^0_{k*}||_2)\boldsymbol{\tau}_{k*}\boldsymbol{\tau}^T_{k*}(1+o_p(1))\big] \\
& -\frac{n}{K}\sum_{k=v+1}^{p}\big[\alpha_n p'_{\lambda_2}(|\gamma^0_{k1}|)sign(\gamma^0_{k1})\tau_{k1} + \alpha_n^2 p''_{\lambda_2}(|\gamma^0_{k1}|)(\tau_{k1})^2(1+o_p(1))\big] \\
& -\frac{n}{K}\sum_{j=1}^{s-1}\big[\alpha_n p'_{\lambda_3}(|\phi^0_j|)sign(\phi^0_j)\tau^{\phi}_j + \alpha_n^2 p''_{\lambda_3}(|\phi^0_j|)(\tau^{\phi}_j)^2(1+o_p(1))\big] \\
& := S_1 - S_2 - S_3 - S_4 - S_5
\end{aligned}
$$

where $\boldsymbol{I}(\boldsymbol{\gamma}^0,\boldsymbol{\phi}^0)$ is the Fisher's information matrix. By standard arguments of likelihood theory, S_1 is of the order $O_p(1+n^{r/(2r+1)}\alpha_n)||\boldsymbol{\tau}||$, S_2 is of the order $O_p(1+2n^{r/(2r+1)}\alpha_n)||\boldsymbol{\tau}||^2$ and for large enough C, S_2 dominates S_1 uniformly in $||\boldsymbol{\tau}|| = C$.

Further, we have

$$S_3 \leq \frac{n}{K}\alpha_n a_n \sum_{k=1}^{v} \frac{\boldsymbol{\gamma}_{k*}^0}{||\boldsymbol{\gamma}_{k*}^0||_2}\boldsymbol{\tau}_{k*}^T + \frac{n}{K}\alpha_n^2 \max_k\{p''_{\lambda_1}(||\boldsymbol{\gamma}_{k*}^0||_2)\}\sum_{k=1}^{v}\boldsymbol{\tau}_{k*}\boldsymbol{\tau}_{k*}^T$$
$$\leq \frac{n}{K}\alpha_n^2\sqrt{v}||\boldsymbol{\tau}|| + \frac{n}{K}\alpha_n^2||\boldsymbol{\tau}||^2 \max_k\{p''_{\lambda_1}(||\boldsymbol{\gamma}_{k*}^0||_2)\}$$

Since $\max_k\{p''_{\lambda_1}\} \to 0$, we have S_3 dominated by S_2.

For S_4 and S_5, we have

$$\begin{aligned} S_4 &\leq \alpha_n a_n \frac{n}{K}\sum_{k=v+1}^{p}\tau_{k1} + \frac{n}{K}\alpha_n^2 \max_k\{p''_{\lambda_2}(|\gamma_{k1}^0|)\}\sum_{k=v+1}^{p}(\tau_{k1})^2 \\ &\leq \frac{n}{K}\alpha_n^2||\boldsymbol{\tau}|| + \frac{n}{K}\alpha_n^2||\boldsymbol{\tau}||^2 \max_k\{p''_{\lambda_2}(|\gamma_{k1}^0|)\} \end{aligned}$$

and

$$\begin{aligned} S_5 &\leq \alpha_n a_n \frac{n}{K}\sum_{l=1}^{s-1}\tau_l^{\phi} + \frac{n}{K}\alpha_n^2 \max_l\{p''_{\lambda_3}(|\phi_l^0|)\}\sum_{l=1}^{s-1}(\tau_l^{\phi})^2 \\ &\leq \frac{n}{K}\alpha_n^2||\boldsymbol{\tau}|| + \frac{n}{K}\alpha_n^2||\boldsymbol{\tau}||^2 \max_l\{p''_{\lambda_3}(|\phi_l^0|)\} \end{aligned}$$

Similarly, we have S_4 and S_5 dominated by S_2. Hence, by choosing a large enough C, we have $||(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}}) - (\boldsymbol{\gamma}^0, \boldsymbol{\phi}^0)|| = O_p(\alpha_n)$. Hence the consistency of penalized least squares estimator $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\phi}})$ is proven.

C.4 Proof of Theorem 3.3.2

For ease of notation, we denote $\boldsymbol{\phi} = (\boldsymbol{\phi}^{nz}, \boldsymbol{\phi}^{z})$, where $\boldsymbol{\phi}^{nz} = (\phi_1, \ldots, \phi_{s-1})$ and $\boldsymbol{\phi}^{z} = (\phi_s, \ldots, \phi_{q-1})$. From $\lambda_{max} \to 0$, it is very easy to see $a_n = 0$ for large n. Then by Theorem

3.3.1, it is sufficient to show for $\boldsymbol{\phi}^{nz}$, it satisfies,

$$||\phi_l - \phi_l^0||_2 = O_p(n^{-r/(2r+1)}), l = 1, \ldots, s-1$$

and for $\boldsymbol{\phi}^z$ and for some given small $\epsilon = Cn^{-r/(2r+1)}$, when $n \to \infty$, with probability approaching 1, for $l = s, \ldots, q-1$, it satisfies,

$$\frac{\partial M(\boldsymbol{\gamma}, \boldsymbol{\phi})}{\partial \phi_l} < 0 \text{ when } 0 < \phi_l < \epsilon \text{ and } \frac{\partial M(\boldsymbol{\gamma}, \boldsymbol{\phi})}{\partial \phi_l} > 0 \text{ when } -\epsilon < \phi_l < 0$$

Since we have

$$\frac{\partial M(\boldsymbol{\gamma}, \boldsymbol{\phi})}{\partial \phi_l} = \frac{\partial l(\boldsymbol{\gamma}, \boldsymbol{\phi})}{\partial \phi_l} - np'_{\lambda_3}(|\phi_l|)sign(\phi_l)$$

Do Taylor expansion at $\boldsymbol{\phi}^0$ for $\frac{\partial l(\boldsymbol{\gamma}, \boldsymbol{\phi})}{\partial \phi_l}$ only, we have

$$\begin{aligned}\frac{\partial M(\boldsymbol{\gamma}, \boldsymbol{\phi})}{\partial \phi_l} = &\frac{\partial l(\boldsymbol{\gamma}, \boldsymbol{\phi}^0)}{\partial \phi_l} + \sum_{k=1}^{q-1} \frac{\partial^2 l(\boldsymbol{\gamma}, \boldsymbol{\phi}^0)}{\partial \phi_l \partial \phi_k}(\phi_k - \phi_k^0) \\ &+ \sum_{k=1}^{q-1}\sum_{j=1}^{q-1} \frac{\partial^3 l(\boldsymbol{\gamma}, \boldsymbol{\phi}^*)}{\partial \phi_l \partial \phi_k \partial \phi_j}(\phi_k - \phi_k^0)(\phi_j - \phi_j^0) - np'_{\lambda_3}(|\phi_j|)sign(\phi_j)\end{aligned}$$

where $\boldsymbol{\phi}^*$ lies between $\boldsymbol{\phi}^0$ and $\boldsymbol{\phi}$. After some calculation, we have

$$\frac{\partial M(\boldsymbol{\gamma}, \boldsymbol{\phi})}{\partial \phi_l} = n\lambda_3 \Big\{ -\frac{1}{\lambda_3} p'_{\lambda_3}(|\phi_l|)sign(\phi_l) + O_p(\frac{1}{\lambda_3} n^{-r/(2r+1)}) \Big\}$$

Since $\lim_{n\to\infty} \liminf_{\phi_j \to 0} \frac{1}{\lambda_3} p'_{\lambda_3}(|\phi_j|) > 0$ and $\frac{1}{\lambda_3} n^{-r/(2r+1)} \to 0$, we can conclude that the sign of $\frac{\partial M(\boldsymbol{\gamma}, \boldsymbol{\phi})}{\partial \phi_j}$ is completely determined by the sign of ϕ_j. Hence, we prove $\hat{\beta}_j = 0$ for $j = s+1, \ldots, q$

For (ii) & (iii), applying similar arguments as in (i), we immediately have, with probability

approaching 1, $\hat{\boldsymbol{\gamma}}_{k*} = 0$ for $k = v+1, \ldots, p$ and $\hat{\gamma}_{k1} = 0$ for $k = c+1, \ldots, p$. Then by $\sup_u \boldsymbol{B}(u) = O(1)$ and $\hat{m}_k(\cdot) = \hat{\gamma}_{k0} + \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\gamma}}_{k*}$, we prove $\hat{m}_k(\cdot) = c_k$ for $k = v+1, \ldots, c$, where c_k is some constant and $\hat{m}_k(\cdot) = 0$ for $k = c+1, \ldots, p$.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Antoniadis, A., & Fan, J. (2001). Regularization of wavelet approximations. Journal of the American Statistical Association, 96(455), 939-967.

[2] Ben Sherwood and Adam Maidman (2016). rqPen: Penalized Quantile Regression. R package version 1.5.1. https://CRAN.R-project.org/package=rqPen

[3] Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. The annals of applied statistics, 5(1), 232.

[4] Cai, Z., Fan, J., & Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. Journal of the American Statistical Association, 95(451), 888-902.

[5] Carpenter, D. O., Arcaro, K., & Spink, D. C. (2002). Understanding the human health effects of chemical mixtures. Environmental Health Perspectives, 110(Suppl 1), 25.

[6] Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. Heredity, 87(1), 52-58.

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

[8] Colditz, G. A., & Hankinson, S. E. (2005). The Nurses' Health Study: lifestyle and health among women. Nature Reviews Cancer, 5(5), 388-396.

[9] Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. The Annals of statistics, 32(2), 407-499.

[10] Fan, J., & Zhang, W. (1999). Statistical estimation in varying coefficient models. Annals of Statistics, 1491-1518.

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

[12] Feng, S., & Xue, L. (2013). Variable selection for single-index varying-coefficient model. Frontiers of Mathematics in China, 8(3), 541-565.

[13] Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. Technometrics, 35(2), 109-135.

[14] Friedman, J., Hastie, T., Hfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. The Annals of Applied Statistics, 1(2), 302-332.

[15] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1), 1.

[16] Grant, S. F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J., ... & Styrkarsdottir, U. (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nature genetics, 38(3), 320-323.

[17] Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. Journal of the Royal Statistical Society. Series B (Methodological), 757-796.

[18] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

[19] Horikoshi, M., Beaumont, R. N., Day, F. R., Warrington, N. M., Kooijman, M. N., Fernandez-Tajes, J., ... & Bradfield, J. P. (2016). Genome-wide associations for birth weight and correlations with adult disease. Nature, 538(7624), 248-252.

[20] Hoover, D. R., Rice, J. A., Wu, C. O., & Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika, 809-822.

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

[22] Jolliffe, I. (2002). Principal component analysis. John Wiley & Sons, Ltd.

[23] Lee, E. R., Noh, H., & Park, B. U. (2014). Model selection via Bayesian information criterion for quantile regression models. Journal of the American Statistical Association, 109(505), 216-229.

[24] Liu, X., Cui, Y., & Li, R. (2016). Partial linear varying multi-index coefficient model for integrative gene-environment interactions. Statistica Sinica, 26, 1037-1060.

[25] Ma, S., Yang, L., Romero, R., & Cui, Y. (2011). Varying coefficient model for geneenvironment interaction: a non-linear look. Bioinformatics, 27(15), 2119-2126.

[26] Ma, S., & Song, P. X. K. (2015). Varying index coefficient models. Journal of the American Statistical Association, 110(509), 341-356.

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

[28] Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. The American Journal of Human Genetics, 74(4), 765-769.

[29] Ottman, R. (1996). Geneenvironment interaction: definitions and study designs. Preventive medicine, 25(6), 764.

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

[31] Rimm, E. B., Giovannucci, E. L., Willett, W. C., Colditz, G. A., Ascherio, A., Rosner, B., & Stampfer, M. J. (1991). Prospective study of alcohol consumption and risk of coronary disease in men. The Lancet, 338(8765), 464-468.

[32] Sale, M. M., Smith, S. G., Mychaleckyj, J. C., Keene, K. L., Langefeld, C. D., Leak, T. S., ... & Freedman, B. I. (2007). Variants of the transcription factor 7-like 2 (TCF7L2) gene are associated with type 2 diabetes in an African-American population enriched for nephropathy. Diabetes, 56(10), 2638-2642.

[33] Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., ... & Zhu, J. (2008). Mapping the genetic architecture of gene expression in human liver. PLoS Biol, 6(5), e107. Chicago

[34] Schumaker, L. (2007). Spline functions: basic theory. Cambridge University Press.

[35] Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 6(2), 461-464.

[36] Sexton, K., & Hattis, D. (2007). Assessing cumulative health risks from exposure to environmental mixturesthree fundamental questions. Environmental Health Perspectives, 825-832.

[37] Tang, Y., Wang, H. J., Zhu, Z., & Song, X. (2012). A unified variable selection approach for varying coefficient models. Statistica Sinica, 601-628.

[38] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

[39] Yang, X., Zhang, B., Molony, C., Chudin, E., Hao, K., Zhu, J., ... & Guengerich, F. P. (2010). Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. Genome research, 20(8), 1020-1036.

[40] Yu, Y., & Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. Journal of the American Statistical Association, 97(460), 1042-1054.

[41] Xia, Y., & Li, W. K. (1999). On the estimation and testing of functional-coefficient linear models. Statistica Sinica, 735-757.

[42] Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1), 49-67.

[43] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. The Annals of statistics, 38(2), 894-942.

[44] Zhou, X., & You, J. (2004). Wavelet estimation in varying-coefficient partially linear regression models. Statistics & probability letters, 68(1), 91-104.

[45] Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), 1418-1429.

[46] Zou, H., & Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. Annals of statistics, 36(4), 1509.

[47] Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. Journal of computational and graphical statistics, 15(2), 265-286.