Non/Semi-parametric learning from data with complex features

Xinyi Li
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Statistics and Probability Commons

Recommended Citation
Li, Xinyi, "Non/Semi-parametric learning from data with complex features" (2018). Graduate Theses and Dissertations. 17250.
https://lib.dr.iastate.edu/etd/17250

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Non/Semi-parametric learning from data with complex features

by

Xinyi Li

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Li Wang, Major Professor
Yehua Li
William Q. Meeker
Dan Nettleton
Zhengyuan Zhu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University
Ames, Iowa
2018

Copyright © Xinyi Li, 2018. All rights reserved.
# TABLE OF CONTENTS

| LIST OF TABLES | iv |
| LIST OF FIGURES | vi |
| ACKNOWLEDGEMENTS | viii |
| ABSTRACT | ix |
| CHAPTER 1. INTRODUCTION | 1 |
| CHAPTER 2. ADDITIVE PARTIALLY LINEAR MODELS FOR ULTRA-HIGH-DIMENSIONAL GENETIC DATA IN MAIZE SHOOT APICAL MERistem STUDY | 6 |
| 2.1 Introduction | 7 |
| 2.2 Variable Selection and Spline Pilot Estimators | 11 |
| 2.3 Estimation and Selection Consistency | 14 |
| 2.4 SBLL Estimators and Inference of the Nonparametric Components | 16 |
| 2.5 Simulation | 18 |
| 2.6 Application to SAM Data | 24 |
| 2.7 Discussion | 26 |
| Bibliography | 30 |
| CHAPTER 3. SPARSE MODEL IDENTIFICATION AND LEARNING FOR ULTRA-HIGH-DIMENSIONAL ADDITIVE PARTIALLY LINEAR MODELS | 34 |
| 3.1 Introduction | 35 |
| 3.2 Methodology | 40 |
| 3.2.1 Model Setup | 40 |
| 3.2.2 Spline Basis Approximation | 42 |
| 3.2.3 Adaptive Group LASSO Regularization | 43 |
| 3.3 Two-stage SBLL Estimator and Inference | 45 |
| 3.4 Implementation and Simulation | 50 |
| 3.4.1 Algorithm | 52 |
| 3.4.2 Simulation Studies | 54 |
| 3.5 Application | 60 |
| 3.6 Discussion | 64 |
| Bibliography | 67 |
CHAPTER 4. SPARSE LEARNING FOR IMAGE-ON-SCALAR REGRESSION WITH
APPLICATION TO IMAGING GENETICS STUDIES .......................... 70
  4.1 Introduction ........................................................................... 71
  4.2 Method .................................................................................. 75
    4.2.1 Model Setup .................................................................. 75
    4.2.2 Triangulations and Bivariate Spline Approximation ............. 76
  4.3 Adaptive group lasso estimator .............................................. 79
    4.3.1 Estimation and Selection Consistency ............................... 81
  4.4 Simulation Studies ................................................................. 84
  4.5 ADNI Data Analysis .............................................................. 91
Bibliography ............................................................................... 93

CHAPTER 5. GENERAL CONCLUSIONS ........................................ 98

APPENDIX A. SUPPLEMENTARY FOR “ADDITIVE PARTIALLY LINEAR MOD-
ELS FOR ULTRA-HIGH-DIMENSIONAL GENETIC DATA IN MAIZE SHOOT
APICAL MERISTEM STUDY” ......................................................... 102

APPENDIX B. SUPPLEMENTARY FOR “SPARSE MODEL IDENTIFICATION AND
LEARNING FOR ULTRA-HIGH-DIMENSIONAL ADDITIVE PARTIALLY LIN-
EAR MODELS” ........................................................................ 134

APPENDIX C. SUPPLEMENTARY FOR “SPARSE LEARNING FOR IMAGE-ON-
SCALAR REGRESSION WITH APPLICATION TO IMAGING GENETICS STUD-
IES” ....................................................................................... 165
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Selection results comparing the SBLL-AGLASSO and OLS-ALASSO.</td>
<td>22</td>
</tr>
<tr>
<td>Table 2.2</td>
<td>Estimation and cross-validation prediction results comparing the SBLL-AGLASSO, OLS-ALASSO and ORACLE*.</td>
<td>23</td>
</tr>
<tr>
<td>Table 2.3</td>
<td>Empirical coverage rates of 95% SCBs based on the SBLL-AGLASSO and ORACLE**.</td>
<td>24</td>
</tr>
<tr>
<td>Table 2.4</td>
<td>Selected SNPs and RNA transcripts in the SAM data example, and the prediction results.</td>
<td>28</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Statistics (B-i)–(B-vi) comparing the SMILE, SAPLM and SLM.</td>
<td>58</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Statistics (C-i)–(C-iv) comparing the SMILE, SAPLM and SLM.</td>
<td>59</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Estimation results comparing the ORACLE, SMILE, SAPLM and SLM.</td>
<td>61</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Coverage rates comparing the ORACLE, SMILE and SAPLM.</td>
<td>62</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>Selected SNPs and Transcripts by SMILE, SAPLM and SLM.</td>
<td>64</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Statistics for Correct Selection.</td>
<td>88</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>Statistics for False Selection.</td>
<td>89</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>Statistics for Estimation.</td>
<td>90</td>
</tr>
<tr>
<td>Table 4.4</td>
<td>Distribution of patients by diagnosis status and gender.</td>
<td>91</td>
</tr>
<tr>
<td>Table 4.5</td>
<td>Selected features for the ADNI data.</td>
<td>93</td>
</tr>
</tbody>
</table>
Table A.1  Cross validation results of the SBLL-AGLASSO and OLS-ALASSO for the SAM data example .................................................. 126

Table A.2  Selection and estimation results based on SNPs and RNA transcripts selected from real data analysis comparing the SBLL-AGLASSO, OLS-ALASSO and ORACLE* methods ........................................... 127

Table A.3  Comparison with existing works in the literature. ........................................ 131
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1</td>
<td>Plots of SBLL-AGLASSO estimator (dotted curve) and 95% SCB (dashed curves) of function components $\alpha_l(x_l)$, $l = 1, 2, 3$ (solid curve).</td>
<td>25</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Plots of SBLL-AGLASSO estimator (solid curve) and 95% SCB (dashed curves) for selected transcripts.</td>
<td>27</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Plots of the SMILE (dotted curve) and the 95% SCB (upper and lower dashed curves) of the nonparametric component $\phi_l(x_l)$, $l = 2, 3$ (solid curve).</td>
<td>62</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Plot of the SMILE (solid curve) and the 95% confidence band (upper and lower dashed curves) for the selected RNA transcript.</td>
<td>65</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>True and estimated varying coefficient functions.</td>
<td>87</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>Estimates of the coefficient functions of features for the ADNI data.</td>
<td>94</td>
</tr>
<tr>
<td>Figure A.1</td>
<td>Plots of the true function (black solid curve), ORACLE** estimator (blue solid curve), SBLL-AGLASSO estimator (red solid curve), ORACLE** 95% SCB (upper and lower blue dashed curves) and SBLL-AGLASSO 95% SCB (upper and lower red dashed curves).</td>
<td>129</td>
</tr>
</tbody>
</table>
Figure A.2  Plots of the true function (black solid curve), ORACLE** estimator (blue solid curve), SBLL-AGLASSO estimator (red solid curve), ORACLE** 95% SCB (upper and lower blue dashed curves) and SBLL-AGLASSO 95% SCB (upper and lower red dashed curves). ........................................ 130

Figure B.1  First stage selection results using different number of knots. .............. 136

Figure B.2  Average mean squared errors (AMSEs) of the SBLL estimators of $\phi_2$ and $\phi_3$.  138

Figure B.3  Coverage rates of the SCBs for $\phi_2$ and $\phi_3$. ............................. 139
ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all those who have kindly offered their hands to me in pursuing my doctoral degree.

First and foremost, I am genuinely grateful to my major professor, Dr. Li Wang, for generously guiding me and supporting me, with patience, with understanding, in various aspects of conducting scientific research. I am greatly inspired by her dedication, intelligence and diligence to research, and I have learned a lot from her about being a researcher and a data scientist.

I would like to thank Dr. Nettleton, who provided me the opportunity to work as a research assistant for Plant Science Institute Research Group, leading me to the new world of genetic studies. The collaboration with Dr. Nettleton benefited me in both conducting the research projects and writing scientific papers.

I am thankful to my committee members for their time and feedback on this work: Dr. Yehua Li, Dr. William Meeker and Dr. Zhengyuan Zhu.

I would also like to thank the faculty in the Statistics Department at Iowa State University. The instructions I received during my graduate study benefited me significantly, and I enjoyed the magical learning journey.

Last but not least, I would like to acknowledge my parents, for giving me spiritual light to go through the darkness. I would also like to thank my friends for their loving encouragement during my graduate life.
This thesis is mainly focused on developing novel and flexible non/semi-parametric statistical methods dealing with data with complex features. In recent years, advancement of high throughput technologies has made it possible to collect sophisticated high-dimensional datasets, such as microarray data, genome-wide single nucleotide polymorphism (SNP) data, and RNA sequencing (RNA-seq) data. These advances have caused an escalating demand for innovative dimension reduction tools to extract useful information from the huge amount of data, to visualize the underlying structure, and to facilitate the understanding and analysis of the data. The research undertaken in my thesis are described below.

In Chapter 2, we consider a semiparametric additive partially linear regression model (APLM) for analyzing ultra-high-dimensional data where both the number of linear components and the number of nonlinear components can be much larger than the sample size. We propose a two-step approach for estimation, selection and simultaneous inference of the components in the APLM. In the first step, the nonlinear additive components are approximated using polynomial spline basis functions, and a doubly penalized procedure is proposed to select nonzero linear and nonlinear components based on adaptive LASSO. In the second step, local linear smoothing is then applied to the data with the selected variables to obtain the asymptotic distribution of the estimators of the nonparametric functions of interest. The proposed method selects the correct model with probability approaching one under regularity conditions. The estimators of both the linear part and nonlinear part are consistent and asymptotically normal, which enables us to construct confidence
intervals and make inferences about the regression coefficients and the component functions. The performance of the method is evaluated by simulation studies. The proposed method is also applied to a dataset on the Shoot Apical Meristem (SAM) of maize genotypes.

In Chapter 3, we further consider the model identification problem, as long with variable selection, estimation and inference simultaneously for the additive partially linear model (APLM). APLM combines the flexibility of nonparametric regression with the parsimony of regression models, and has been widely used as a popular tool in multivariate nonparametric regression to alleviate the “curse of dimensionality”. A natural question raised in practice is the choice of structure in the nonparametric part, that is, whether the continuous covariates enter into the model in linear or nonparametric form. In this paper we present a comprehensive framework for simultaneous sparse model identification and learning for ultra-high-dimensional APLMs where both the linear and nonparametric components are possibly larger than the sample size. We propose a fast and efficient two-stage procedure. In the first stage, we decompose the nonparametric functions into a linear part and a nonlinear part. The nonlinear functions are approximated by constant spline bases, and a triple penalization procedure is proposed to select nonzero components using adaptive group LASSO. In the second stage, we refit data with selected covariates using higher order polynomial splines, and apply spline backfitted local linear smoothing to obtain asymptotic normality for the estimators. The procedure is shown to be consistent for model structure identification. It can identify zero, linear, and nonlinear components correctly and efficiently. Inference can be made on both linear coefficients and nonparametric functions. We conduct simulation studies to evaluate the performance of the method, and apply the proposed method to a dataset on the Shoot Apical Meristem (SAM) of maize genotypes for illustration.
In Chapter 4, motivated by recent advances in technology for brain imaging and high-throughput genotyping, we consider an imaging genetics approach to discover relationships between the interplay of genetic variation and environmental factors and measurements from imaging phenotypes. We propose an image-on-scalar regression method, in which the spatial heterogeneity of gene-environment interactions on imaging responses is investigated via an ultra-high-dimensional spatially varying coefficient model (SVCM). Bivariate splines on triangulations are used to represent the coefficient functions over an irregular two-dimensional domain of interest. When using the image-on-scalar regression method, a natural question raised in practice is if the coefficient function is really varying over space. In this paper, we present a unified approach for simultaneous sparse learning and model structure identification (i.e., varying and constant coefficients separation). Our method can identify zero, nonzero constant and spatially varying components correctly and efficiently. The estimators of constant coefficients and varying coefficient functions are consistent. The performance of the method is evaluated by a few simulation examples and a brain mapping study based on the Alzheimer’s Disease Neuroimaging Initiative data.
CHAPTER 1. INTRODUCTION

This dissertation is mainly focused on developing novel and flexible non/semi-parametric statistical methods, as well as designing fast and efficient computational algorithms for solving problems arising from data with complex features. These features include high-dimensional data, data distributed over irregular domains, functional or repeatedly measured data. In recent years, advancement of high throughput technologies has made it possible to collect sophisticated high-dimensional datasets, such as microarray data, genome-wide single nucleotide polymorphism (SNP) data, and RNA sequencing (RNA-seq) data. These advances have caused an escalating demand for innovative dimension reduction tools to extract useful information from the huge amount of data, to visualize the underlying structure, and to facilitate the understanding and analysis of the data. Compared with conventional datasets where there are many fewer variables than observations, large-dimensional datasets involve a number of variables equal to or even much larger than the number of observations. These scenarios correspond to $p \ll n$, $p \approx n$ and $p \gg n$, respectively, with $p$ the number of variables and $n$ the number of observations. The last two scenarios challenge a key assumption in classical statistics: the number of observations grows large relative to the number of parameters. The research undertaken in my dissertation are described below.

Motivated by the maize shoot apical meristem (SAM) data, which consists phenotypes, SNP data and RNA-seq data, we have completed two papers to conduct the genome wide association study (GWAS). The first and second parts of my dissertation consider a class of semiparametric additive partially linear models (APLMs) to study variable selection, estimation and inference
simultaneously for ultra-high-dimensional data; and in the second paper, we also consider model identification based on the work in the first paper.

Specifically, in Chapter 2, we consider a semiparametric additive partially linear regression model (APLM) for analyzing ultra-high-dimensional data where both the number of linear components and the number of nonlinear components can be much larger than the sample size. We propose a two-step approach for estimation, selection and simultaneous inference of the components in the APLM. In the first step, the nonlinear additive components are approximated using polynomial spline basis functions, and a doubly penalized procedure is proposed to select nonzero linear and nonlinear components based on adaptive LASSO. A difficulty here is that the covariates in the parametric components and those in the nonparametric components could be dependent. To resolve the dependence between the covariates in the nonparametric and parametric parts, we consider a projection of the covariates in the parametric part to the space generated by the spline basis for the covariates in the nonparametric part and study the properties of the projection. The ultra-high dimensionality raises challenging issues for the investigation of the projection properties.

After variable selection, we would like to provide an inferential tool for the linear and nonparametric components. It is well known that kernel estimation in high dimension would be extremely computationally intensive. The spline method, on the other hand, is very fast, but the rate of convergence is only established in mean squares sense, and there is no asymptotic distribution or uniform convergence, so no measures of confidence can be assigned to the estimators. In this paper, we propose a two-step “spline-backfitted local linear smoothing” (SBLL) procedure to the APLM for estimation, selection and simultaneous inference of the components. In the first stage, we approximate the nonparametric functions, by undersmoothed polynomial spline functions. We perform variable selection for the APLM using a double penalized procedure to identify impor-
tant variables, which is crucial to obtain efficient estimators for the non-zero components. We show that the proposed model selection for both parametric and nonparametric terms is consistent, the estimators of the nonzero linear coefficients are asymptotically normal, and the estimators of the nonzero nonparametric functions are $L_2$-norm consistent. In the second stage, we apply the one-step backfitting method using pilot coefficient estimators and spline estimators for the components selected in the first stage followed by local linear estimators. A simultaneous confidence band (SCB) is provided for each nonparametric component. The success of our method lies in the well-known “reducing bias by undersmoothing” and “averaging out the variance” principles. Figuratively speaking, our two-step method can be viewed as a hammer-knife tool that first slams a huge clump into smaller pieces in one hit (the spline step), selects the important pieces (adaptive lasso), and then cuts each important piece into an exact shape (univariate local linear smoothing). The SBLL estimator we propose combines the best features of both local linear and spline methods.

Chapter 3 is an extension of the ultra-high-dimensional APLM studies in Chapter 2. When applying APLMs to the SAM data, a natural question raised from the biologists is the choice of structure in the nonparametric part, that is, whether the quantitative genetic variables enter into model in linear or nonparametric form. We present a comprehensive framework for simultaneous sparse model identification and learning for APLMs. For this purpose, we propose an efficient two-stage procedure for ultra-high-dimensional APLMs where both the linear and nonparametric components are possibly larger than the sample size. In the first stage, we decompose the nonparametric functions into linear and purely nonlinear functions. The purely nonlinear functions are approximated by constant spline basis, and a triple penalization procedure is proposed to select nonzero components using adaptive group LASSO. In the second stage, we refit data with selected covariates, and apply local linear smoothing to obtain the asymptotic normality for the estimators.
The procedure is shown to be consistent for model structure identification. It can identify zero, linear, and nonlinear components correctly and efficiently. Inference can be made on both linear coefficients and nonparametric functions. We conduct simulation studies to evaluate the performance of the method, and apply the proposed method to a dataset on the Shoot Apical Meristem (SAM) of maize genotypes for illustration.

Our study in Chapter 4 is motivated by some recent successful findings in imaging genetics studies facilitated by the Alzheimer’s Disease Neuroimaging Initiative (ADNI). Recent advances in both high throughput genotyping and brain imaging techniques have capacitated imaging genetics becoming an emerging discipline, which focuses on exploring the genetic influence on structural and functional imaging variations. Compared to traditional case-control designs, imaging genetics outshines in identifying underlying genes by employing imaging measures as phenotypes, such as the fludeoxyglucose positron emission tomography (FDG-PET) images. Single nucleotide polymorphisms (SNPs) and other polymorphisms in several genes, including Apolipoprotein E (APOE), have been demonstrated to be related to neuroimaging measures in brain disorders, such as mild cognitive impairment (MCI) and Alzheimer’s Disease (AD) (e.g., Kim et al. (2009); Ashford and Mortimer (2002)). However, it is of great interest to identify other genes that play a role in the development and progression of MCI and AD. In this paper, we handle all the predictors (e.g., environmental factors, genetic factors and their interactions) jointly when investigating the association between imaging responses and scalar predictors. We develop an efficient method for G×E interaction identification to address the high dimensionality of both the imaging and genomic data (Shen et al., 2010; Stein et al., 2010a,b). Although interaction selection has drawn much attention in the literature (Hao and Zhang, 2014; Kong et al., 2017; Li and Liu, 2018), effectively relating hundreds of thousands of predictors to large-scale imaging data remains a challenging task. In this
paper, we develop a novel statistical methodology that integrates the imaging data, environmental data and ultra-high dimensional genetic data in a principled functional regression and variable selection framework. To model the effect of G×E interaction, the varying coefficient model (VCM) introduced by Hastie and Tibshirani (1993) provides a flexible modeling approach. We are particularly interested in the spatially varying-coefficient model (SVCM) as it is powerful for modeling the nonstationarity of regression coefficients over space; see Zhu et al. (2014), Mu et al. (2018) and among others. Here the imaging response variables are associated with the scalar covariates through the functional linear regression, but the regression coefficients can vary from location to location and are modeled as a nonparametric function of spatial coordinates. We propose a unified framework to address the following questions: (i) how to identify those important main genetic and environmental factors and interactions; (ii) how to estimate the coefficient functions for these important variables; and (iii) how to separate nonzero constant and spatially varying components.

We propose a bivariate spline smoothing method to preserve important features (shape and/or smoothness) of imaging data. Two regularization operations are imposed to simultaneously identify the constant and spatially varying coefficient functions, removing insignificant predictors, and estimating the remaining coefficients and/or coefficient functions. We derive model selection consistency for the proposed method and show that it possesses the oracle property when the dimension of covariates exceeds the sample size.
CHAPTER 2. ADDITIVE PARTIALLY LINEAR MODELS FOR ULTRA-HIGH-DIMENSIONAL GENETIC DATA IN MAIZE SHOOT APICAL MERISTEM STUDY

A paper to be submitted to the Journal of Agricultural, Biological, and Environmental Statistics

Xinyi LI, Li WANG, and Dan NETTLETON

Department of Statistics, Iowa State University, Ames, IA 50011

Abstract

We consider a semiparametric additive partially linear regression model (APLM) for analyzing ultra-high-dimensional data where both the number of linear components and the number of nonlinear components can be much larger than the sample size. We propose a two-step approach for estimation, selection and simultaneous inference of the components in the APLM. In the first step, the nonlinear additive components are approximated using polynomial spline basis functions, and a doubly penalized procedure is proposed to select nonzero linear and nonlinear components based on adaptive LASSO. In the second step, local linear smoothing is then applied to the data with the selected variables to obtain the asymptotic distribution of the estimators of the nonparametric functions of interest. The proposed method selects the correct model with probability approaching one under regularity conditions. The estimators of both the linear part and nonlinear part are consistent and asymptotically normal, which enables us to construct confidence intervals and make
inferences about the regression coefficients and the component functions. The performance of the method is evaluated by simulation studies. The proposed method is also applied to a dataset on the Shoot Apical Meristem (SAM) of maize genotypes.

**Keywords:** Dimension reduction; inference for ultra-high-dimensional data; semiparametric; spline-backfitted local polynomial; variable selection.

### 2.1 Introduction

Advancement of high throughput technologies has made it possible to collect sophisticated high-dimensional datasets, such as microarray data, genome-wide single nucleotide polymorphisms (SNP) data, and RNA sequencing (RNA-seq) data. These advances have caused an escalating demand for innovative dimension reduction tools to extract useful information from the huge amount of data, to visualize the underlying structure, and to facilitate the understanding and analysis of the data. In a maize Shoot Apical Meristem (SAM) study (Leiboff et al., 2015), utilizing a 1.3 million SNP dataset and RNA-seq read counts for 39,656 genes, researchers are interested in identifying candidate SNPs and genes whose transcript abundance levels are associated with SAM morphological variation. Subsequent analyses of candidate genes can be used to improve understanding of the genetic control of maize SAM morphological diversity.

Compared with conventional datasets where there are many fewer variables than observations, large-dimensional datasets involve a number of variables equal to or even much larger than the number of observations. These scenarios correspond to \( p \ll n, p \approx n \) and \( p \gg n \), respectively, with \( p \) the number of variables and \( n \) the number of observations. The last two scenarios challenge a key assumption in classical statistics: the number of observations grows large relative to the number of parameters.
In terms of statistical accuracy, dimension reduction and variable selection play pivotal roles in analyzing high-dimensional data. Increasing model sparsity enforces a lower dimensional model structure. In turn, it makes inference more tractable, models easier to interpret, and leads to more robustness against noise. Regularization techniques that enforce sparsity have been widely studied in the literature for more than a decade, including the lasso Tibshirani (1996), adaptive lasso Zou (2006), and the smoothly clipped absolute deviation penalty Fan and Li (2001). The goal of the proposed research is to develop novel dimension reduction methodology that is both computationally efficient and statistically optimal for high-dimensional data that have become more and more prevalent in many applications.

The “curse of dimensionality” is a significant obstacle in high-dimensional data analysis. To deal with this problem, we will investigate both approaches for modeling massive heterogeneous data using a semiparametric additive partially linear model that combines the flexibility of non-parametric regression with the parsimony of linear regression. To be specific, suppose for the $i$th subject, we observe $(Y_i, Z_{(i)}, X_{(i)})$, where $Z_{(i)} = (Z_{i1}, \ldots, Z_{ip_1})^\top$ is a $p_1$-dimensional vector of covariates and $X_{(i)} = (X_{i1}, \ldots, X_{ip_2})^\top$ is a $p_2$-dimensional vector of covariates, $i = 1, \ldots, n$. We assume $\{(Y_i, Z_{(i)}, X_{(i)})\}_{i=1}^n$ is an independent and identically distributed sample of size $n$ from the distribution of $(Y, Z, X)$, satisfying the following model:

$$Y_i = \mu + \sum_{k=1}^{p_1} Z_{ik} \beta_k + \sum_{l=1}^{p_2} \alpha_l(X_{il}) + \epsilon_i, \quad (2.1)$$

where $\mu$ is the intercept, $\beta_k$’s are some unknown coefficients, $\{\alpha_l(\cdot)\}_{l=1}^{p_2}$ are unknown smooth functions, and $\epsilon_i$’s are independent and identically distributed random errors with mean zero and variance $\sigma^2$. Without loss of generality, we assume that $\{X_{il}\}_{i=1}^n$ can be rescaled into $[0, 1]$ for each $l = 1, \ldots, p_2$, and each $\alpha_l(\cdot)$ is centered with $\mathbb{E} \alpha_l(X_{il}) = 0$ to make model (2.1) identifiable.
The APLM is particularly convenient when $Z$ is a vector of discrete, i.e., categorical variables. Another interesting aspect of the APLM is that it allows $X$ and $Z$ to be dependent. Specifically, the APLM allows $Z$ to be a deterministic function of $X$, thus allowing high-dimensional variables to enter the model parametrically. In the SAM data analysis, the vector $X$ contains thousands of RNA-seq expression measurements, while the vector $Z$ contains SNP genotypes. Our model allows for $Z$ and $X$ to be correlated.

Estimation and inference for APLMs have been well studied in literature (Opsomer and Ruppert, 1997), and several approaches have been proposed to estimate the linear and nonlinear components in APLMs. The first is the classic backfitting approach; see the details in Hastie and Tibshirani (1990). The second is the marginal integration approach proposed by Linton and Nielsen (1995); see Linton (1997) and Sperlich et al. (2002). Mammen et al. (1999) introduced a new smooth backfitting estimator. Marx and Eilers (1998) studied penalized splines, which share most of the practical benefits of smoothing spline methods, combined with ease of use and reduction of the computational cost of backfitting Generalized Additive Models (GAMs). For regression spline estimators; see Xie and Huang (2009), Liu et al. (2011) and Wang et al. (2011).

Motivated by a dataset from the SAM project (see the details in Section 2.6), we first study variable selection for APLMs in the setting that both the dimension of the linear components and the dimension of nonlinear components are ultra-high. Existing work on penalized semiparametric regression has been largely limited to cases where either dimension is fixed; see, for example, Liang and Li (2009), Ding et al. (2011), Liu et al. (2011), Kai et al. (2011), Wang et al. (2011) and Sherwood and Wang (2016). Existing semiparametric works considering selection and estimation procedures for both linear and nonlinear components, to our best knowledge, do not allow the number of nonlinear components to be much larger than the number of observations $n$. 
Important progress in the large $p$ setting has been recently made by Huang et al. (2010) for ultra-high-dimensional additive models. Compared with Huang et al. (2010), which involves only nonparametric components, the establishment of the asymptotic distribution of the estimators for the parametric terms is quite challenging given that the number of covariates for both parametric and nonparametric terms are much larger than $n$, and the convergence rate for the nonparametric component estimators is slower than root-$n$. A difficulty here is that the covariates in the parametric components and those in the nonparametric components could be dependent. To resolve the dependence between the covariates in the nonparametric and parametric parts, we consider a projection of the covariates in the parametric part to the space generated by the spline basis for the covariates in the nonparametric part and study the properties of the projection. The ultra-high dimensionality raises challenging issues for the investigation of the projection properties.

After variable selection, we would like to provide an inferential tool for the linear and nonparametric components. It is well known that kernel estimation in high dimension would be extremely computationally intensive. The spline method, on the other hand, is very fast, but the rate of convergence is only established in mean squares sense, and there is no asymptotic distribution or uniform convergence, so no measures of confidence can be assigned to the estimators.

In this paper, we propose a two-step “spline-backfitted local linear smoothing” (SBLL) procedure to the APLM for estimation, selection and simultaneous inference of the components. In the first stage, we approximate the nonparametric functions $\alpha_l(\cdot)$ ($l = 1, \ldots, p_2$), by undersmoothed polynomial spline functions. We perform variable selection for the APLM using a double penalized procedure to identify important variables, which is crucial to obtain efficient estimators for the non-zero components. We show that the proposed model selection for both parametric and nonparametric terms is consistent, the estimators of the nonzero linear coefficients are asymptotically
normal, and the estimators of the nonzero nonparametric functions are \( L_2 \)-norm consistent. In the second stage, we apply the one-step backfitting method (Wang and Yang, 2007) using pilot coefficient estimators and spline estimators for the components selected in the first stage followed by local linear estimators. A simultaneous confidence band (SCB) is provided for each nonparametric component. The success of our method lies in the well-known “reducing bias by undersmoothing” and “averaging out the variance” principles. Figuratively speaking, our two-step method can be viewed as a hammer-knife tool that first slams a huge clump into smaller pieces in one hit (the spline step), selects the important pieces (adaptive lasso), and then cuts each important piece into an exact shape (univariate local linear smoothing). The SBLL estimator we propose combines the best features of both local linear and spline methods.

The rest of the paper is organized as follows. In Section 2.2, we describe the first stage spline smoothing and propose a doubly penalized regularization method for simultaneous estimation and variable selection. The theoretical properties on selection consistency and rates of convergence for the estimators are developed in Section 2.3. Section 2.4 introduces the spline-backfitted local linear estimators and simultaneous confidence bands (SCBs) for the nonparametric components. The performance of the estimators is assessed by simulations in Section 2.5 and illustrated by application to the SAM data in Section 2.6. Some concluding remarks are given in Section 2.7. Technical details are provided in Supplemental Materials.

### 2.2 Variable Selection and Spline Pilot Estimators

We approximate the smooth functions \( \{\alpha_l(\cdot)\}_{l=1}^{p_2} \) in (2.1) by polynomial splines for their simplicity in computation. For example, for each \( l = 1, \ldots, p_2 \), let \( \nu_l \) be a partition of \([0, 1]\), with \( N_n \) interior knots \( \nu_l = \{0 = \nu_{l,0} < \nu_{l,1} < \cdots < \nu_{l,N_n} < \nu_{l,N_n+1} = 1\} \). The polynomial splines of order
\( \varrho + 1 \) are polynomial functions with \( \varrho \)-degree (or less) on intervals \([v_{l,j}, v_{l,j+1})\), \( j = 0, \ldots, N_n - 1 \), and \([v_{l,N_n}, v_{l,N_n+1})\), and have \( \varrho - 1 \) continuous derivatives globally. Let \( B_l = B^\varphi_l([0,1], v_l) \) be the space of such polynomial splines, and \( B^0_l = \{ b \in B_l : E_b(X_l) = 0, E_b^2(X_l) < \infty \} \). This ensures that the spline functions are centered; see for example Xue and Yang (2006), Wang and Yang (2007) and Wang et al. (2014).

For any \( L^2 \)-integrable function \( \alpha(\cdot) \), let \( \| \alpha \|_2 = \{ \int_0^1 \alpha^2(x) f(x) dx \}^{1/2} \) be its \( L^2 \) norm, where \( f(x) \) is the density function of \( X \). Let \( \{ B_{lj}(\cdot) \}_{j=1}^{J_n} \) be a set of standardized spline basis functions for \( B^0_l \) with dimension \( J_n = N_n + \varrho \), where \( B_{lj}(x_l) = b_{lj}(x_l)/\|b_{lj}\|_2 \) \( (j = 1, \ldots, J_n) \), so that \( EB_{lj}(X_l) \equiv 0 \), \( EB_{lj}^2(X_l) \equiv 1 \). Suppose the nonlinear component can be well approximated by a spline function so that, for all \( x_l \in [0,1] \),

\[
\alpha_l(x_l) \approx \sum_{j=1}^{J_n} \gamma_{lj} B_{lj}(x_l) = \mathbf{B}_l^\top(x_l) \mathbf{\gamma}_l \quad (l = 1, \ldots, p_2),
\]

(2.2)

where \( \mathbf{B}_l(x_l) = (B_{l1}(x_l), \ldots, B_{lJ_n}(x_l))^\top \) and \( \mathbf{\gamma}_l = (\gamma_{l1}, \ldots, \gamma_{lJ_n})^\top \) is a vector of coefficients.

To perform simultaneous variable selection and model estimation, we propose minimizing the penalized sum of squares

\[
\sum_{i=1}^n \left[ Y_i - \mu - \left( \sum_{k=1}^{p_1} Z_{ik} \beta_k + \sum_{l=1}^{p_2} \alpha_l(X_{il}) \right) \right]^2 + \sum_{k=1}^{p_1} p_{\lambda_1}(|\beta_k|) + \sum_{l=1}^{p_2} p_{\lambda_2}(|\alpha_l|_2).
\]

(2.3)

The penalty function \( p_{\lambda_n}(\cdot) \) can be the \( L_1 \)-penalty with \( p_{\lambda_n}(|\cdot|) = \lambda_n |\cdot| \) which provides a lasso estimator, or the \( L_2 \) penalty \( p_{\lambda_n}(|\cdot|) = \lambda_n |\cdot|^2 \) which produces a ridge-type estimator. However, we don’t apply the \( L_0 \) penalty here as it is highly computationally intensive and unstable. The advantage of choosing the penalization using \( |\alpha_l|_2 \) is that it no longer relies on a particular choice of spline basis. This type of penalization ensures that the coefficients within the same nonparametric component are treated as an entire group in model selection, and therefore it achieves the same effect as the group-wise model selection approach (Yuan and Lin, 2006).
For any vector \( a \in \mathbb{R}^p \), denote \( \|a\|_2 = (\sum_{i=1}^{p} |a_i|^2)^{1/2} \) as the \( L_2 \) norm of \( a \). Following from (2.2) and (2.3), it is approximately equivalent to consider the problem of minimizing

\[
\sum_{i=1}^{n} \left[ Y_i - \mu - \left( \sum_{k=1}^{p_1} Z_{ik} \beta_k + \sum_{l=1}^{p_2} B_l^T(X_{il}) \gamma_l \right) \right]^2 + \sum_{k=1}^{p_1} p_{\lambda_1}(|\beta_k|) + \sum_{l=1}^{p_2} p_{\lambda_2}(\|\gamma_l\|_2). \tag{2.4}
\]

In the following, denote \( \beta = (\beta_1, \ldots, \beta_{p_1})^T \) a \( p_1 \)-dimensional vector, and \( \gamma = (\gamma_1^T, \ldots, \gamma_{p_2}^T)^T \) a length \( (p_2J_n) \) vector. Let \( \theta^T = (\beta^T, \gamma^T) = (\beta_1, \ldots, \beta_{p_1}, \gamma_1, \ldots, \gamma_{p_2}) = (\theta_1^T, \ldots, \theta_{m_1}^T, \ldots, \theta_{p_1+p_2}^T) \), where \( \theta_m = \beta_m I\{1 \leq m \leq p_1\} + \gamma_{m-p_1} I\{p_1 + 1 \leq m \leq p_1 + p_2\}, \) with \( I(\cdot) \) an indicator function. Furthermore, let \( \bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i \), \( Y = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})^T \) be an \( n \)-length vector, and \( Z = (Z_1, \ldots, Z_{p_1}) \) be an \( n \times p_1 \) matrix, where \( Z_k = (Z_{ik}, \ldots, Z_{nk})^T (k = 1, \ldots, p_1) \). Let \( B = (B_1, \ldots, B_{p_2}) \) be a dimension \( n \times (p_2J_n) \) matrix, where \( B_l = (B_l(X_{1l}), \ldots, B_l(X_{nl}))^T (l = 1, \ldots, p_2) \) is a dimension \( n \times J_n \) matrix. Let \( D = (Z_1, \ldots, Z_{p_1}, B_1, \ldots, B_{p_2}) \equiv (D_1, \ldots, D_m, \ldots, D_{p_1+p_2}) \) be an \( n \times (p_1 + p_2J_n) \) matrix, where \( D_m = Z_m I\{1 \leq m \leq p_1\} + B_m - p_1 I\{p_1 + 1 \leq m \leq p_1 + p_2\} \), an \( n \times d_m \) submatrix of \( D \) with \( d_m = I(1 \leq m \leq p_1) + J_n I\{p_1 + 1 \leq m \leq p_1 + p_2\} \). Then the group lasso estimator for the objective function in (2.4) with the lasso penalty is

\[
\tilde{\theta} \equiv (\beta^T, \gamma^T)^T = \arg \min_{\theta} \left\{ \|Y - D\theta\|_2^2 + \lambda_1 \sum_{k=1}^{p_1} |\beta_k| + \lambda_2 \sum_{l=1}^{p_2} \|\gamma_l\|_2 \right\}. \tag{2.5}
\]

Consequently, the group lasso estimator of \( \alpha_l \) is \( \tilde{\alpha}_l(x) = \sum_{j=1}^{J_n} \tilde{\gamma}_{lj} B_{lj}(x) \) \( (l = 1, \ldots, p_2) \).

We use the group lasso (GLASSO) estimator \( \tilde{\theta} \) to obtain weights for adaptive group lasso by setting \( w_m = \|\tilde{\theta}_m\|_2^{-1} I\{\|\tilde{\theta}_m\|_2 > 0\} + \infty \times I\{\|\tilde{\theta}_m\|_2 = 0\} \), where by convention, \( \infty \times 0 = 0 \). Then the adaptive group lasso objective function is

\[
L_{n2}(\theta; \lambda_3, \lambda_4) = \|Y - D\theta\|_2^2 + \lambda_3 \sum_{k=1}^{p_1} w_k |\beta_k| + \lambda_4 \sum_{l=1}^{p_2} w_{l+p_1} \|\gamma_l\|_2.
\]

The adaptive group lasso (AGLASSO) estimator is \( \hat{\theta} \equiv (\hat{\beta}^T, \hat{\gamma}^T)^T = \arg \min_{\theta} L_{n2}(\theta; \lambda_3, \lambda_4) \). Then the AGLASSO estimator of \( \alpha_l \) is \( \hat{\alpha}_l(x) = B_l^T(x) \hat{\gamma} = \sum_{j=1}^{J_n} \hat{\gamma}_{lj} B_{lj}(x) \) \( (l = 1, \ldots, p_2) \).
2.3 Estimation and Selection Consistency

In this section, we define predictors $X_l$ and $Z_k$ as redundant in model (2.1), if and only if $\alpha_l(X_l) = 0$ almost surely and $\beta_k = 0$. Suppose there is only an unknown subset of predictors that are relevant in (2.1) with non-zero components. We are interested in identifying such subsets of relevant predictors consistently while estimating the nonzero components in (2.1) simultaneously.

We establish the asymptotic properties of the penalized parametric and nonparametric components estimators in the following theorems. We assume that in the true model only the first $s_1$ $(0 \leq s_1 \leq p_1)$ linear components and the first $s_2$ $(0 \leq s_2 \leq p_2)$ nonlinear components are nonzero, and the remaining components are all zeros. More specifically, we define active linear index set as $S_z = \{1, \ldots, s_1\}$ or the empty set if $s_1 = 0$, active nonlinear index set as $S_x = \{1, \ldots, s_2\}$ or the empty set if $s_2 = 0$, inactive linear index set as $N_z = \{s_1 + 1, \ldots, p_1\}$ and inactive nonlinear index set as $N_x = \{s_2 + 1, \ldots, p_2\}$. Further, let

$$S = S_z \cup \{l + p_1 : l \in S_x\}, \quad N = N_z \cup \{l + p_1 : l \in N_x\}$$

(2.6)

be the active and inactive index set, respectively.

In the following, to avoid confusion, we use $\beta_0 = (\beta_{01}, \ldots, \beta_{0p_1})^\top$ and $\alpha_0 = (\alpha_{01}, \ldots, \alpha_{0p_2})^\top$ to denote the true parameters and functions in model (2.1). Let $\beta_0 = (\beta_{0,S_z}^\top, \beta_{0,N_z}^\top)^\top$, where $\beta_{0,S_z}$ consists of all $s_1$ nonzero components of $\beta_0$, and $\beta_{0,N_z} = \mathbf{0}$ without loss of generality. In a similar fashion to $S$, denote

$$\hat{S} = \{1 \leq m \leq p_1 + p_2 : \|\hat{\theta}_m\|_2 > 0\}.$$

(2.7)

The following theorems establish the asymptotic properties of the adaptive group lasso estimators. We only state the main results here. The proofs are provided in Supplemental Materials.
Theorem 2.1. Suppose that Assumptions (A1) – (A6) described in Supplementary A hold. For \( \hat{S} \) in (2.7), as \( n \to \infty \), \( \Pr(\hat{S} = S) \to 1 \).

Theorem 2.2. Suppose that Assumptions (A1) – (A6) described in Supplementary A hold. Then

\[
\begin{align*}
\sum_{k \in S_z} |\tilde{\beta}_k - \beta_{0k}|^2 &= O_p(J_n/n) + O(J_n^{-2d}) + O_p \left\{ \left( \lambda_{n3}^2 + \lambda_{n4}^2 \right)/n^2 \right\}, \\
\sum_{l \in S_z} \|\tilde{\alpha}_l - \alpha_{0l}\|_2^2 &= O_p(J_n/n) + O(J_n^{-2d}) + O_p \left\{ \left( \lambda_{n3}^2 + \lambda_{n4}^2 \right)/n^2 \right\}.
\end{align*}
\]

Theorem 2.1 shows the proposed method can consistently distinguish nonzero components from zero components. Theorem 2.2 gives the convergence rates of the estimators.

Next we establish the asymptotic normal distribution for the parametric estimators. To make \( \beta_{0,S_z} \) estimable at the root-\( n \) rate, we need a condition to ensure \( X \) and \( Z \) are not functionally related. Define \( F_+ = \{ \phi(x) = \sum_{i=1}^{s_z} \phi_i(x_i), \ E\phi_i(x_i) = 0, \ \|\phi_i\|_2 < \infty \} \) the Hilbert space of theoretically centered \( L_2 \) additive functions on \([0,1]^{s_z}\). For any \( k = 1, \ldots, s_1 \), let \( z_k \) be the coordinate mapping that maps \( Z \) to its \( k \)th component so that \( z_k(Z) = Z_k \), and let \( \psi_k = \arg\min_{\psi \in F_+} \|z_k - \psi\|_2^2 = \arg\min_{\psi \in F_+} E\{Z_k - \psi(X)\}^2 \) be the orthogonal projection of \( z_k \) onto \( F_+ \). Let \( \tilde{Z}_{S_z} = \{\psi_1(X), \ldots, \psi_{s_1}(X)\}^\top \) and \( Z_{S_z} = (Z_1, \ldots, Z_{s_1})^\top \).

Theorem 2.3. Suppose Assumptions (A1) – (A6) hold described in Supplementary A hold. If the functions \( \psi_k(\cdot) \in \mathcal{H}^{(d)}, \ k = 1, \ldots, s_2, \ J_n^{^{-2d}} \to 0, \ n^{-1/2}(\lambda_{n3} + \lambda_{n4}) \to 0, \) as \( n \to \infty \), then the regression coefficient estimators in active set \( \hat{\beta}_{S_z} \) are asymptotically normal, that is, \( (n\Sigma)^{1/2} (\hat{\beta}_{S_z} - \beta_{0,S_z}) \to N(0, I_{s_1}) \) in distribution, where \( I_{s_1} \) is an \( s_1 \times s_1 \) identity matrix and \( \Sigma = \sigma^{-2}E[(Z_{S_z} - \tilde{Z}_{S_z})(Z_{S_z} - \tilde{Z}_{S_z})^\top] \).
2.4 SBLL Estimators and Inference of the Nonparametric Components

After model selection, our next step is to conduct statistical inference for the nonparametric component functions of those important variables. Although the one step penalized estimation in Section 2.2 can quickly identify the nonzero nonlinear components, the asymptotic distribution is not available for the resulting estimators. We extend the one-step backfitting method in Wang and Yang (2007) and Wang and Yang (2009) to our second stage study using pilot spline estimators in the first stage followed by local linear estimators.

The basic idea is that for every \( l = 1, \ldots, p \), we estimate the \( l \)th additive function \( \alpha_0^l(\cdot) \) in model (2.1) nonparametrically by assuming that the parameter vector \( \beta_0 \) and other nonparametric components \( \alpha_0^l = \{ \alpha_0^{l'}(\cdot) : l' = 1, \ldots, p, l' \neq l \} \) are known. The problem turns into a univariate function estimation problem. One can obtain an oracle smoother \( \hat{\alpha}_l^o \) by smoothing \( \{(Y_{il}, X_{il})\}_{i=1}^n \), where \( Y_{il} = Y_i - \sum_{k \in S_z} Z_{ik}\hat{\beta}_k + \sum_{l' \in S_x \backslash \{l\}} \alpha_0^{l'}(X_{il'}) \).

Define \( \hat{S}_z = \{ k = 1, \ldots, p_1 : |\hat{\beta}_k| > 0 \} \) and \( \hat{S}_x = \{ l = 1, \ldots, p_2 : \|\hat{\gamma}_l\|_2 > 0 \} \). Since the true parameter vector \( \beta_0 \) and functions \( \{ \alpha_0^{l'} : l' \in S_x \backslash \{l\} \} \) are unknown, we replace those with the pilot estimators \( \{ \hat{\beta}_k : k \in \hat{S}_z \} \) and functions \( \{ \hat{\alpha}_l : l' \in \hat{S}_x \backslash \{l\} \} \), and obtain the pseudo-responses \( \hat{Y}_l = Y_i - \sum_{k \in \hat{S}_z} Z_{ik}\hat{\beta}_k + \sum_{l' \in \hat{S}_x \backslash \{l\}} \hat{\alpha}_l(X_{il'}) \).

Denote \( K \) a continuous kernel function, and let \( K_h(t) = K(t/h)/h \) be a rescaling of \( K \), where \( h \) is usually called the bandwidth. Thus, for any \( l \in \hat{S}_z \), we obtain the oracle local linear smoother \( \hat{\alpha}_l^o \) and the spline-backfitted local linear (SBLL) estimator \( \hat{\alpha}_l^{SBLL}(x_l) \) by applying the following local linear smoothing on \( \{(Y_{il}, X_{il})\}_{i=1}^n \) and \( \{ (\hat{Y}_{il}, X_{il}) \}_{i=1}^n \), respectively:

\[
\left( \hat{\alpha}_l^o(x_l), \hat{\alpha}_l^{SBLL}(x_l) \right) = (1 \ 0) (X_l^T W_l X_l)^{-1} X_l^T W_l (Y_l, \hat{Y}_l),
\]  
(2.8)
where the oracle and pseudo-response vectors are $Y_t = (Y_{1t}, \ldots, Y_{nt})^\top$ and $\hat{Y}_t = (\hat{Y}_{1t}, \ldots, \hat{Y}_{nt})^\top$, respectively, and the weight and “design” matrices are

$$W_t = n^{-1} \text{diag}\{K_h(X_{it} - x_t)\}_{i=1}^n, \quad X_t^\top = \begin{pmatrix} 1 & \ldots & 1 \\ X_{1t} - x_t & \ldots & X_{nt} - x_t \end{pmatrix}.$$ 

Asymptotic properties of smoothers of $\hat{\alpha}_t^\text{SBLL}(x_l)$ $(l \in S_x)$ can be easily established based on these assumptions. Specifically, let $\mu_2(K) = \int u^2 K(u) du$, and let $f_t$ be the probability density function of $X_t$, then under Assumptions (B1) and (B2) given in Supplementary A,

$$(nh)^{1/2} \left\{ \hat{\alpha}^\text{SBLL}_t(x_l) - \alpha_0(x_l) - b_t(x_l)h^2 \right\} \rightarrow N \left\{ 0, v_t^2(x_l) \right\} (l \in S_x), \quad (2.9)$$

in distribution as $n \rightarrow \infty$, where $b_t(x_l) = \mu_2(K)\alpha''_0(x_l)/2$, $v_t^2(x_l) = \|K\|^2 F_t^{-1}(x_l)s^2$.

The following theorem states that the uniform magnitude of difference between $\hat{\alpha}^\text{SBLL}_t(x_l)$ and $\hat{\alpha}^\text{o}_t(x_l)$ is of order $o_p\{(nh)^{-1/2}\}$, which is dominated by the asymptotic uniform size of $\hat{\alpha}^\text{o}_t(x_l) - \alpha_0(x_l)$. As a result, $\hat{\alpha}^\text{SBLL}_t(x_l)$ will have the same asymptotic distribution as $\hat{\alpha}^\text{o}_t(x_l)$.

**Theorem 2.4.** Under Assumptions (A1) – (A6) and (B1) – (B3) described in Supplementary A hold, the SBLL estimator $\hat{\alpha}^\text{SBLL}_t(x_l)$ given in (2.8) satisfies

$$\sup_{x_l \in [0,1]} \left| \hat{\alpha}^\text{SBLL}_t(x_l) - \hat{\alpha}^\text{o}_t(x_l) \right| = o_p\{(nh)^{-1/2}\} (l \in S_x). \quad (2.10)$$

Hence for any $x_l \in [h, 1-h]$

$$(nh)^{1/2} \left\{ \hat{\alpha}^\text{SBLL}_t(x_l) - \alpha_0(x_l) - b_t(x_l)h^2 \right\} \rightarrow N \left\{ 0, v_t^2(x_l) \right\} (l \in S_x). \quad (2.11)$$

in distribution. With the additional Assumption (A2') given in Supplementary A, the estimator $\hat{\alpha}^\text{SBLL}_t(x_l)$ also satisfies

$$\lim_{n \rightarrow \infty} \text{Pr} \left\{ (-2 \log h)^{1/2} \left( \sup_{x_l \in [h, 1-h]} (nh)^{1/2} \left| \hat{\alpha}^\text{SBLL}_t(x_l) - \alpha_0(x_l) \right| - d_n \right) < t \right\} = e^{-2e^{-t}} \quad (2.12)$$

for any $t$ and $l \in S_x$, where $d_n = \log^{1/2}(h^{-2}) + \log \{\|K\|^2/(2\pi\|K\|^2)\}/\log^{1/2}(h^{-2})$. 


Theorem 2.4 provides analytical expressions in constructing asymptotic confidence intervals and SCB under certain conditions. Under Assumptions (A1) – (A6) and (B1) – (B2) described in Supplementary A, for any $\alpha \in (0, 1)$, an asymptotic $100(1 - \alpha)\%$ pointwise confidence interval for $\alpha_0l(x_l)$ over the interval $[h, 1 - h]$ is

$$\hat{\alpha}^{\text{SBLR}}_l(x_l) - \hat{b}_l(x_l)h^2 \pm \hat{\nu}_l(x_l)(nh)^{-1/2}$$

for $l \in \mathcal{S}_x$. Under Assumptions (A1) – (A6) and (B1) – (B3) given in Supplementary A, for any $\alpha \in (0, 1)$, an asymptotic $100(1 - \alpha)\%$ SCB for $\alpha_0l(x_l)$ over the interval $[h, 1 - h]$ is

$$\hat{\alpha}^{\text{SBLR}}_l(x_l) \pm \hat{\nu}_l(x_l)(nh)^{-1/2}[d_n - \log^{-1/2}(h^{-2})\log\{-2^{-1}\log(1 - \alpha)\}] \quad (l \in \mathcal{S}_x).$$

### 2.5 Simulation

We conduct a simulation study to evaluate the performance of the proposed penalized SBLL estimator for the APLM using AGLASSO, abbreviated as SBLL-AGLASSO. We investigate how it performs relative to the ordinary linear least squares estimator with the adaptive lasso penalty (OLS-ALASSO) and the oracle estimator (ORACLE$^*$), that is, the estimator when the variables that have nonzero coefficients and nonzero functions are known prior to statistical analysis. The SBLL-AGLASSO and OLS-ALASSO estimators are computed via the group coordinate descent algorithm (Huang et al., 2012a), implemented using R package grpreg Breheny (2016). In terms of the performances of SCBs, we compare the SBLL-AGLASSO estimator with the oracle estimator (ORACLE$^{**}$), the estimator $\hat{\alpha}_l^\alpha$ defined in (2.8). It is worth pointing out that both “oracle” estimators, ORACLE$^*$ and ORACLE$^{**}$, are benchmarks for selection comparison and estimation comparison, respectively. They are only computable in simulations, not real examples.

Each simulated dataset is generated from the model

$$Y_i = \sum_{k=1}^{p_1} Z_{ik}\beta_k + \sum_{l=1}^{p_2} \alpha_l(X_{il}) + \epsilon_i,$$

where $\beta_1 = 3, \beta_2 = 4, \beta_3 = -2, \beta_4 = \cdots = \beta_{p_1} = 0, \alpha_1(x) = 8\sin(2\pi x)/(2 - \sin(2\pi x)), \alpha_2(x) =$
\[-3 \sin^2(\pi x) + 6 \cos^2(\pi x), \quad \alpha_3(x) = 5(3x - 1)^2, \quad \text{and} \quad \alpha_4(x) = \cdots = \alpha_{p_2}(x) = 0. \] So the numbers of nonzero linear coefficients and nonzero functions are \( s_1 = s_2 = 3. \) All simulations are replicated 200 times.

**Scenario 1 (Independent covariates).** We simulate \( z^*_ik, x_{il} \) independently from the uniform distribution on 0 to 1, and set \( z_{ik} = I(z^*_ik > 0.75), \) for \( i = 1, \ldots, n, \ k = 1, \ldots, p_1, \ l = 1, \ldots, p_2. \) To make an ultra-high-dimensional scenario, we let the sample size \( n = 300 \) and \( n = 500, \) and consider three different dimensions: \( p_1 = p_2 = p, \) where \( p \) is taken to be 1000, 2000 and 5000. The error term \( \epsilon_i \) is simulated from \( N(0, \sigma^2) \) with \( \sigma = 1.5 \) and 2.0.

**Scenario 2 (Dependent covariates).** In order to mimic the dependence structure in our real data, we use SNP data and RNA-seq read counts from the SAM dataset to generate the response variable. For each RNA transcript, we linearly transform the observed values to the interval \([0, 1]\) by subtracting the minimum and dividing by the range. We have \( p_1 = 5000 \) SNP genotypes as linear covariates \( (Z_{ik}, \ k = 1, \ldots, p_1), \) and \( p_2 = 5000 \) RNA transcripts as nonlinear components \( (X_{il}, \ l = 1, \ldots, p_2), i = 1, \ldots, 368. \) The error term \( \epsilon_i \) is chosen from \( N(0, \sigma^2) \) with \( \sigma = 1.0 \) and 1.5. The SNPs with nonzero coefficients and the transcripts with nonzero functions are chosen by simple random sampling from all SNPs and all transcripts, respectively.

To approximate the nonlinear functions, we use cubic B-splines with different numbers of interior knots, from one to six, evenly distributed on the sample quantiles. Because the selection results are very similar as the number of knots varies, in the following, we report the simulation results based on one interior knot. We use the modified bayesian information criteria (BIC; see Lee et al. (2014)) to choose penalty parameters, defined as \( \text{BIC}(\lambda) = \log(RSS_\lambda) + (2n)^{-1} df_\lambda \log(p_1 + p_2 + p_2 N) \times \log(n), \) where \( RSS_\lambda \) is the residual sum of squares associated with tuning parameter \( \lambda = (\lambda_1, \lambda_2) \) for GLASSO estimation and \( \lambda = (\lambda_3, \lambda_4) \) for AGLASSO estimation, \( df_\lambda \) is the number of estimated
nonzero coefficients for the given \( \lambda \), and \( N \) is the number of interior knots for splines. We have also tried the extended BIC (Chen and Chen, 2008) and the results are similar and thus omitted.

We first evaluate the methods on the accuracy of variable selection and prediction. We adopt the following criteria: (1) mean number of selected linear covariates and nonlinear components \(#V\); (2) percentage of cases in which all the nonzero covariates were selected \(\text{IN}\); (3) percentage of occasions on which precisely the correct covariates were selected \(\text{CS}\); (4) mean squared errors \(\text{MSE}\) for linear coefficients \(\beta_1, \beta_2, \beta_3\); (5) average mean squared errors \(\text{AMSE}\) for nonlinear functions \(\alpha_1, \alpha_2, \alpha_3\), which is defined as \(n^{-1} \sum_{i=1}^{n} \left( \hat{\alpha}_{i}^{\text{SBLL}}(x_{il}) - \alpha_{l}(x_{il}) \right)^2\); (6) 10-fold cross-validation mean squared prediction error \(\text{CV-MSPE}\) for the response variable, defined as \(10^{-1} \sum_{m=1}^{10} \left| \kappa_m \right|^{-1} \sum_{i \in \kappa_m} (\hat{y}_i - y_i)^2\), where \(\kappa_1, \ldots, \kappa_{10}\) comprise a random partition of the dataset into 10 disjoint subsets of approximately equal size. Criteria (1) – (3) measure the selection accuracy, Criteria (4) and (5) focus on the estimation accuracy for the model components, and Criterion (6) measures the prediction accuracy.

The variable selection results are provided in Table 2.1. In both scenarios, the SBLL-AGLASSO method performs very well regardless of the data structures. The SBLL-AGLASSO method can effectively identify important linear and nonlinear components by using the proposed semiparametric model while the traditional linear regression model fails to do so when the effects of some covariates are nonlinear. For the SBLL-AGLASSO method, the number of selected variables is very close to the “oracle” (3 and 3, respectively). For both linear and nonlinear components, the “IN” and “CS” grow closer to 100% as the sample size \(n\) increases. Even for the number of variables as large as \(p_1 = 5000\) and \(p_2 = 5000\) in Scenario 2, SBLL-AGLASSO is still able to identify variables nonlinearly associated with the response in all cases and identify variables linearly associated with the response in nearly all cases. From the results in Table 2.1, it is also evident that model misspec-
ification leads to poor variable selection performance for OLS-ALASSO method. Especially for the selection of nonlinear components, which is our main focus for real data analysis, OLS-ALASSO fails to select the right nonlinear components in each simulation. The estimation and prediction results are displayed in Table 2.2. Specifically, we present the CV-MSPEs and the MSEs for linear coefficients $\beta_1$, $\beta_2$ and $\beta_3$ and nonlinear functions $\alpha_1$, $\alpha_2$ and $\alpha_3$. The case with known active covariates (ORACLE*) is also reported and serves as a gold standard. The table clearly indicates that the proposed SBLL-AGLASSO method estimates unknown parameters and functions very well regardless of the dependence between covariates. In addition, the SBLL-AGLASSO estimator provides accurate predictions in the sense that they tend to be very close to predictions produced by the oracle estimator. The OLS-ALASSO performs very poorly in both Scenarios 1 and 2, with much larger MSEs for the linear coefficients and nonlinear functions; the CV-MSPEs of the OLS-ALASSO estimators are much higher (around 7 – 17 times higher in Scenario 1 and Scenario 2) than those of the SBLL-AGLASSO estimators. The poor performance of OLS-ALASSO estimators, in both estimation and prediction, illustrates the importance and necessity of including nonlinear components in our model. Next we investigate the coverage rates of the proposed SCB. For each replication, we consider 20 equally spaced points on $[0, 1]$ and test whether the true functions are covered by the SCB at these points. Table 2.3 shows the empirical coverage probabilities for a nominal 95% confidence level out of 200 replications for Scenarios 1 and 2, respectively. For comparison, we also provide the SCBs using the “oracle smoother” (ORACLE**). From Table 2.3, we observe that coverage probabilities for both SBLL and “oracle” SCBs approach the nominal levels, which provides positive confirmation of Theorem 2.4. In most of the cases in Scenario 1, SBLL-AGLASSO performs as well as the “oracle” SCBs, and the “oracle” SCBs arrive at about the nominal coverage when $n = 500$. In Scenario 2, regardless of the noise level, the coverage rates
Table 2.1: Selection results comparing the SBLL-AGLASSO and OLS-ALASSO.

<table>
<thead>
<tr>
<th>Size</th>
<th>Noise</th>
<th>Method</th>
<th>Linear Part</th>
<th>Nonlinear Part</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>n σ p</td>
<td>#V</td>
<td>IN</td>
</tr>
<tr>
<td>Scenario 1</td>
<td></td>
<td></td>
<td>#V</td>
<td>IN</td>
</tr>
<tr>
<td>300</td>
<td>1.5</td>
<td>1000 SBLL-AGLASSO</td>
<td>3.02</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>0.86</td>
<td>0</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>SBLL-AGLASSO</td>
<td>3.02</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>0.78</td>
<td>0</td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td>SBLL-AGLASSO</td>
<td>3.01</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>0.39</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>SBLL-AGLASSO</td>
<td>3.01</td>
<td>97.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>0.79</td>
<td>0</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>SBLL-AGLASSO</td>
<td>2.98</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>0.66</td>
<td>0</td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td>SBLL-AGLASSO</td>
<td>2.96</td>
<td>95.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>0.33</td>
<td>0</td>
</tr>
<tr>
<td>500</td>
<td>1.5</td>
<td>1000 SBLL-AGLASSO</td>
<td>3.00</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>1.73</td>
<td>12</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>SBLL-AGLASSO</td>
<td>3.00</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>1.50</td>
<td>4.5</td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td>SBLL-AGLASSO</td>
<td>3.01</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>1.37</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>SBLL-AGLASSO</td>
<td>3.00</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>1.64</td>
<td>9</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>SBLL-AGLASSO</td>
<td>3.00</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>1.43</td>
<td>3.5</td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td>SBLL-AGLASSO</td>
<td>3.01</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>1.22</td>
<td>0.5</td>
</tr>
<tr>
<td>Scenario 2</td>
<td></td>
<td></td>
<td>#V</td>
<td>IN</td>
</tr>
<tr>
<td>368</td>
<td>1.0</td>
<td>5000 SBLL-AGLASSO</td>
<td>3.00</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.5</td>
<td>5000</td>
<td>SBLL-AGLASSO</td>
<td>2.88</td>
<td>85.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 2.2: Estimation and cross-validation prediction results comparing the SBLL-AGLASSO, OLS-ALASSO and ORACLE*.

<table>
<thead>
<tr>
<th>Size</th>
<th>Noise</th>
<th>n</th>
<th>σ</th>
<th>p</th>
<th>Method</th>
<th>β1</th>
<th>β2</th>
<th>β3</th>
<th>α1</th>
<th>α2</th>
<th>α3</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>1.5</td>
<td>1000</td>
<td></td>
<td></td>
<td>ORACLE*</td>
<td>0.04</td>
<td>0.05</td>
<td>0.05</td>
<td>0.19</td>
<td>0.16</td>
<td>0.24</td>
<td>2.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SBLL-AGLASSO</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.19</td>
<td>0.16</td>
<td>0.23</td>
<td>2.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>7.29</td>
<td>6.58</td>
<td>4.01</td>
<td>6.26</td>
<td>10.14</td>
<td>11.37</td>
<td>34.35</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ORACLE*</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
<td>0.19</td>
<td>0.18</td>
<td>0.21</td>
<td>2.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SBLL-AGLASSO</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
<td>0.18</td>
<td>0.18</td>
<td>0.21</td>
<td>2.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>7.46</td>
<td>7.49</td>
<td>3.99</td>
<td>6.40</td>
<td>10.13</td>
<td>11.35</td>
<td>34.37</td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ORACLE*</td>
<td>0.05</td>
<td>0.04</td>
<td>0.05</td>
<td>0.18</td>
<td>0.17</td>
<td>0.24</td>
<td>2.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SBLL-AGLASSO</td>
<td>0.05</td>
<td>0.04</td>
<td>0.05</td>
<td>0.18</td>
<td>0.17</td>
<td>0.24</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>8.51</td>
<td>11.04</td>
<td>4.00</td>
<td>6.32</td>
<td>10.21</td>
<td>11.35</td>
<td>35.15</td>
</tr>
<tr>
<td>Scenario 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>368</td>
<td>1.0</td>
<td>5000</td>
<td></td>
<td></td>
<td>ORACLE*</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SBLL-AGLASSO</td>
<td>0.02</td>
<td>0.04</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.04</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>9.00</td>
<td>16.00</td>
<td>4.00</td>
<td>2.87</td>
<td>6.15</td>
<td>3.57</td>
<td>18.73</td>
</tr>
<tr>
<td>1.5</td>
<td>5000</td>
<td></td>
<td></td>
<td></td>
<td>ORACLE*</td>
<td>0.05</td>
<td>0.08</td>
<td>0.05</td>
<td>0.08</td>
<td>0.09</td>
<td>0.08</td>
<td>2.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SBLL-AGLASSO</td>
<td>0.07</td>
<td>0.10</td>
<td>0.65</td>
<td>0.09</td>
<td>0.09</td>
<td>0.08</td>
<td>2.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OLS-ALASSO</td>
<td>9.00</td>
<td>16.00</td>
<td>4.00</td>
<td>2.87</td>
<td>6.15</td>
<td>4.07</td>
<td>20.28</td>
</tr>
</tbody>
</table>
are very close to the nominal level. Figure 2.1 depicts the true function $\alpha_l$, the SBLL-AGLASSO estimator for the true function $\hat{\alpha}_l^{SBLL}$ and 95% SCB for $\hat{\alpha}_l^{SBLL}$, $l = 1, 2, 3$, which are based on a typical run under Scenario 1 with $n = 500$, $p = 1000$ and $\sigma = 1.5$.

Table 2.3: Empirical coverage rates of 95% SCBs based on the SBLL-AGLASSO and ORACLE**.

<table>
<thead>
<tr>
<th>n</th>
<th>$\sigma$</th>
<th>$p$</th>
<th>Coverage (%)</th>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\alpha_1$</td>
<td>$\alpha_2$</td>
</tr>
<tr>
<td>300</td>
<td>1.5</td>
<td>1000</td>
<td>94.0</td>
<td>98.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>94.5</td>
<td>91.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5000</td>
<td>96.0</td>
<td>93.0</td>
</tr>
<tr>
<td>2.0</td>
<td>1000</td>
<td></td>
<td>95.0</td>
<td>98.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>95.5</td>
<td>96.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5000</td>
<td>96.5</td>
<td>94.5</td>
</tr>
<tr>
<td>500</td>
<td>1.5</td>
<td>1000</td>
<td>94.5</td>
<td>94.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>94.0</td>
<td>94.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5000</td>
<td>95.5</td>
<td>98.5</td>
</tr>
<tr>
<td>2.0</td>
<td>1000</td>
<td></td>
<td>97.5</td>
<td>96.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>96.5</td>
<td>98.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5000</td>
<td>97.5</td>
<td>97.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Scenario 2

| 368 | 1.0      | 5000| 97.0         | 98.0         | 99.0         | 98.0         | 98.0         | 97.5         |
|     |          |     |              |              |              |              |              |              |
| 1.5 | 5000     |     | 97.0         | 98.0         | 99.0         | 97.5         | 97.5         | 97.5         |

2.6 Application to SAM Data

We illustrate the application of our proposed method in the ultra-high-dimensional setting by using the SAM data generated by Leiboff et al. (2015). The maize SAM is a small pool of stem cells that generate all the above-ground organs of maize plants. Leiboff et al. (2015) showed that SAM size is correlated with a variety of agronomically important traits in adult plants. The goal
Figure 2.1: Plots of SBLL-AGLASSO estimator (dotted curve) and 95% SCB (dashed curves) of function components $\alpha_l(x_l)$, $l = 1, 2, 3$ (solid curve).

of our analysis is to model and predict SAM size as a function of single nucleotide polymorphism (SNP) genotypes and messenger RNA transcript abundance levels. Following the preprocessing steps described in Section A, we obtain a preprocessed dataset containing log-scale SAM volume measurements, binary SNP genotypes at $p_1 = 148,793$ markers, and log-scale measures of abundance for $p_2 = 22,903$ transcripts for each of $n = 368$ maize inbred lines. Next we perform linear sure independent screening (Fan and Lv, 2008) for SNP genotypes and nonlinear independent screening (Fan et al., 2011) for RNA transcripts to reduce the computational burden. We select the 5000 SNPs and 5000 transcripts that have the strongest associations with response based on marginal regression. So the dataset we analyze consists of log-scale SAM volume measurements, binary SNP genotypes at $p_1 = 5000$ markers, and log-scale measures of abundance for $p_2 = 5000$ transcripts for each of $n = 368$ maize inbred lines. Although thousands of SNP genotypes and RNA transcript abundance levels are included in our dataset, only a few of these covariates may have a non-negligible relationship with SAM volume. Thus, this is a potentially sparse and ultra-high-dimensional regression problem.
It is natural to apply our proposed SBLL-AGLASSO method to model the relationship between SAM tissue volume and SNP genotypes and RNA transcript abundance levels, not only because we have binary and continuous covariates in the dataset, but also because of the selection, estimation, and inference problem we face. To compare the results of APLM to the ordinary linear model, we also analyze the data with OLS-ALASSO. Parallel to the settings in Section 2.5, we use cubic B-splines with four quantile knots for nonlinear function approximation. We use BIC for penalty parameter selection.

As shown in Table 2.4, SBLL-AGLASSO identified 189 SNPs and 4 RNA transcripts associated with log SAM size, while OLS-ALASSO selected 263 SNPs and 105 RNA transcripts. To evaluate the predictive performance of the two methods, we computed 10-fold cross-validation mean squared prediction error (CV-MSPE) for each method. In addition to providing a more parsimonious model in terms of number of covariates, the results at the bottom of Table 2.4 show that SBLL-AGLASSO provided more accurate predictions than OLS-ALASSO. In most folds of cross-validation, SBLL-AGLASSO provides a more precise model, selecting 176.8 SNP genotypes and 5.2 RNA transcripts on average with MSPE 0.115, while on average, OLS-ALASSO selects 249.2 SNP genotypes and 91 RNA transcripts with a higher MSPE of 0.157, as shown in A.1 of the Supplemental Materials. The estimated nonlinear functions for the RNA transcripts selected in the analysis of the full dataset are plotted, along with 95% SCBs, in Figure 2.2.

2.7 Discussion

This paper focuses on the sparse ultra-high-dimensional APLM which strikes a delicate balance between the simplicity of the standard linear regression model and the flexibility of the additive
Figure 2.2: Plots of SBLL-AGLASSO estimator (solid curve) and 95% SCB (dashed curves) for selected transcripts.
Table 2.4: Selected SNPs and RNA transcripts in the SAM data example, and the prediction results

<table>
<thead>
<tr>
<th>RNA Transcripts</th>
<th>SBLL-AGLASSO</th>
<th>OLS-ALASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{3013}$</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$X_{2638}$</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>$X_{2719}$</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>$X_{2791}$</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>$X_9$</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>$X_{4937}$</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SBLL-AGLASSO</th>
<th>OLS-ALASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of SNP Genotypes</td>
<td>189</td>
<td>263</td>
</tr>
<tr>
<td>Number of RNA Transcripts</td>
<td>4</td>
<td>105</td>
</tr>
<tr>
<td>CV MSPE</td>
<td>0.115</td>
<td>0.157</td>
</tr>
<tr>
<td>CV Mean Number of SNPs</td>
<td>176.8</td>
<td>249.2</td>
</tr>
<tr>
<td>CV Mean Number of Transcripts</td>
<td>5.2</td>
<td>91.0</td>
</tr>
</tbody>
</table>

regression model. Our method can be extended to longitudinal data settings through marginal models or mixed-effects models.

Our work differs from previous work both theoretically and practically. From a practical standpoint, we apply our methods in a genome-wide association study (GWAS) using an integrated dataset from genotyped SNPs and messenger RNA transcripts, in which the numbers of genotyped SNPs and RNA transcripts are both much larger than the observation number. The results show that we are able to detect significant SNPs and RNA transcripts efficiently. In addition, the proposed method allows us to further do inference for selected components, which can provide biological insights not previously achievable with existing methods. Theoretically, we consider an APLM that combines the flexibility of nonparametric regression with the parsimony of linear regression. Our approach differs from existing work in three major aspects. First, we consider selection for both the parametric and nonparametric parts of the model simultaneously, while most
existing techniques focus on selection for either the parametric or nonparametric part. Second, to our best knowledge, existing work considering selection and estimation procedures for both linear and nonlinear components does not allow the number nonlinear components to grow in an exponential order of sample size. We are the first to work on simultaneous variable selection and estimation for the APLM when both the numbers of linear and nonlinear components grow in exponential orders of sample size. The theoretical development for model selection and estimation for ultra-high-dimensional covariates in nonparametric components is much more challenging and completely different from the finite-dimension setting. Third, we obtain asymptotic normality for both the linear and nonparametric components, and importantly, we have obtained SCBs for the nonparametric components. A provides a summary of the differences between our work and six papers most related to our work.

This paper leaves open a few research problems. In the SAM data analysis, we model the effects of categorical variables with linear components and the effects of continuous covariates with nonlinear components. Although this makes sense in the SAM application, an important question in using the APLM in practice is how to identify which continuous covariates should appear in the model linearly and which covariates should appear nonlinearly. Lian et al. (2015) proposed a novel double penalization based procedure to distinguish covariates that enter the nonparametric and parametric parts and to identify insignificant covariates simultaneously. For our SAM data analysis, the 148,793 genotyped SNPs are binary variables that naturally enter the APLM linearly. Thus, we need to simultaneously identify significant SNPs, determine which RNA transcripts enter the linear part of the APLM and identify significant RNA transcripts in the ultra-high-dimensional setting. We believe Lian et al. (2015) can be extended to our study using a triple penalization.
Developing an efficient and automatic model identification criterion is challenging for the ultra-high-dimensional setting and warrants future study.

It is worth pointing out the proposed SCB works well when the model selection is correct. More work, however, is needed to understand the properties of the SCB when the selection is wrong. In Appendix A, we conducted a simulation study that mimics the full complexity of our real data to illustrate the performance of the proposed method in a challenging scenario with hundreds of active SNPs that are strongly correlated with the active transcripts. In this scenario, perfect selection is unlikely without extremely large sample size or unrealistically large signals. The selection study shows that the SBLL-AGLASSO approach performs well despite the challenges. However, SCB coverage can fall below nominal levels when variable selection is poor.

**Supplementary material**

Appendices in Sections A and B include:

Supplementary A: proofs of the theorems, useful technical lemmas and their proofs;

Supplementary B: pre-processing steps for SAM data, additional simulation studies based on SAM data, and comparison of our work to previously published approaches.

**Acknowledgements**

This work was supported by the Iowa State University Plant Sciences Institute Scholars Program. In addition, Wang’s research was supported by NSF [grant DMS-1542332], and Nettleton’s research was supported by NSF [grant IOS-1238142].
Bibliography


CHAPTER 3. SPARSE MODEL IDENTIFICATION AND LEARNING FOR ULTRA-HIGH-DIMENSIONAL ADDITIVE PARTIALLY LINEAR MODELS

A paper submitted to the *Journal of Multivariate Analysis*

*Xinyi LI, Li WANG, and Dan NETTLETON*

*Department of Statistics, Iowa State University, Ames, IA 50011*

**Abstract**

The additive partially linear model (APLM) combines the flexibility of nonparametric regression with the parsimony of regression models, and has been widely used as a popular tool in multivariate nonparametric regression to alleviate the “curse of dimensionality”. A natural question raised in practice is the choice of structure in the nonparametric part, that is, whether the continuous covariates enter into the model in linear or nonparametric form. In this paper we present a comprehensive framework for simultaneous sparse model identification and learning for ultra-high-dimensional APLMs where both the linear and nonparametric components are possibly larger than the sample size. We propose a fast and efficient two-stage procedure. In the first stage, we decompose the nonparametric functions into a linear part and a nonlinear part. The nonlinear functions are approximated by constant spline bases, and a triple penalization procedure is proposed to select nonzero components using adaptive group LASSO. In the second stage, we refit data with selected covariates using higher order polynomial splines, and apply spline backfitted local linear smoothing.
to obtain asymptotic normality for the estimators. The procedure is shown to be consistent for model structure identification. It can identify zero, linear, and nonlinear components correctly and efficiently. Inference can be made on both linear coefficients and nonparametric functions. We conduct simulation studies to evaluate the performance of the method, and apply the proposed method to a dataset on the Shoot Apical Meristem (SAM) of maize genotypes for illustration.

**Keywords:** Dimension reduction; inference for ultra-high-dimensional data; semiparametric regression; spline-backfitted local polynomial; structure identification; variable selection.

### 3.1 Introduction

In the past three decades, flexible and parsimonious additive partially linear models (APLMs) have been extensively studied and widely used in many statistical applications, from biostatistics to econometrics, from engineering to social science. Examples of recent work on APLMs include Liang et al. (2008), Liu et al. (2011), Ma and Yang (2011), Wang et al. (2011), Ma et al. (2013), Wang et al. (2014) and Lian et al. (2014). APLMs are natural extensions of classical parametric models with good interpretability and are becoming more and more popular in data analysis.

Suppose we observe \( \{(Y_i, Z_{(i)}, X_{(i)})\}_{i=1}^{n} \). For subject \( i = 1, \ldots, n \), \( Y_i \) is a univariate response, \( Z_{(i)} = (Z_{i1}, \ldots, Z_{ip_1})^T \) is a \( p_1 \)-dimensional vector of covariates that may be linearly associated with the response, and \( X_{(i)} = (X_{i1}, \ldots, X_{ip_2})^T \) is a \( p_2 \)-dimensional vector of continuous covariates that may have nonlinear associations with the response.
We assume \(\{(Y_i, Z_{(i)}, X_{(i)})\}_{i=1}^{n}\) is an i.i.d sample from the distribution of \((Y, Z, X)\), satisfying the following model:

\[
Y_i = \mu + Z_{(i)}^\top \alpha + \sum_{l=1}^{p_1} \phi_l(X_{il}) + \varepsilon_i = \mu + \sum_{k=1}^{p_1} Z_{ik} \alpha_k + \sum_{l=1}^{p_2} \phi_l(X_{il}) + \varepsilon_i,
\]  

(3.1)

where \(\mu\) is the intercept, \(\alpha_k, k = 1, \ldots, p_1\), are unknown regression coefficients, \(\{\phi_l(\cdot)\}_{l=1}^{p_2}\) are unknown smooth functions, and each \(\phi_l(\cdot)\) is centered with \(E\phi_l(X_{il}) = 0\) to make model (3.1) identifiable. The \(X_{(i)}\) is a \(p_2\)-dimensional vector of zero mean covariates having density with a compact support. Without loss of generality, we assume that each covariate \(\{X_{il}\}_{l=1}^{p_2}\) can be rescaled into an interval \(\chi = [a, b]\). The \(\varepsilon_i\)'s are iid random errors with mean zero and variance \(\sigma^2\).

The APLM is particularly convenient when \(Z\) is a vector of categorical or discrete variables, and in this case, the components of \(Z\) enter the linear part of model (3.1) automatically, and the continuous variables usually enter the model nonparametrically. In practice, we might have reasons to believe that some of the continuous variables should enter the model linearly rather than nonparametrically. A natural question is how to determine which continuous covariates have a linear effect and which continuous covariates have a nonlinear effect. If the choice of linear components is correctly specified, then the biases in the estimation of these components are eliminated and root-\(n\) convergence rates can be obtained for the linear coefficients. However, such prior knowledge is rarely available, especially when the number of covariates is large. Thus, structure identification, or linear and nonlinear detection, is an important step in the process of building an APLM from high-dimensional data.

When the number of covariates in the model is fixed, structure identification in additive models (AMs) has been studied in the literature. Zhang et al. (2011) proposed a penalization procedure to identify the linear components in AMs in the context of smoothing splines ANOVA. They
demonstrated the consistency of the model structure identification and established the convergence rate of the proposed method specifically under the tensor product design. Huang et al. (2012b) proposed another penalized semiparametric regression approach using a group minimax concave penalty to identify the covariates with linear effects. They showed consistency in determining the linear and nonlinear structure in covariates, and obtained the convergence rate of nonlinear function estimators and asymptotic properties of linear coefficient estimators; but they did not perform variable selection at the same time.

For high-dimensional AMs, Lian et al. (2015) proposed a double penalization procedure to distinguish covariates that enter the nonparametric and parametric parts and to identify significant covariates simultaneously. They demonstrated the consistency of the model structure identification, and established the convergence rate of nonlinear function estimators and asymptotic normality of linear coefficient estimators. Despite the nice theoretical properties, their method heavily relies on the local quadratic approximation in Fan and Li (2001), which is incapable of producing naturally sparse estimates. In addition, employing the local quadratic approximation can be extremely expensive since it requires the repeated factorization of large matrices, which becomes infeasible when the number of covariates is very large.

Note that all the aforementioned papers (Zhang et al., 2011; Huang et al., 2012b; Lian et al., 2015) about structure identification focus on the AM with continuous explanatory variables. However, in many applications, a canonical partitioning of the variables exists. In particular, if there are categorical or discrete explanatory variables, as in the case of the SAM data studies (see the details in Section 3.5) and in many genome-wide association studies, we may want to keep discrete explanatory variables separate from the other design variables and let discrete variables enter the linear part of the model directly. In addition, if there is some prior knowledge of certain parametric
forms for some specific covariates, such as a linear form, we may lose efficiency if we simply model all the covariates nonparametrically.

The above practical and theoretical concerns motivate our further investigation of the simultaneous variable selection and structure selection problem for flexible and parsimonious APLMs, in which the features of the data suitable for parametric modeling are modeled parametrically and nonparametric components are used only where needed. We consider the setting where both the dimension of the linear components and the dimension of nonlinear components are ultra-high. We propose an efficient and stable penalization procedure for simultaneously identifying linear and nonlinear components, removing insignificant predictors, and estimating the remaining linear and nonlinear components. We prove the proposed Sparse Model Identification, Learning and Estimation (referred to as SMILE) procedure is consistent. We propose an iterative group coordinate descent approach to solve the penalized minimization problem efficiently. Our algorithm is very easy to implement since it only involves simple arithmetic operations, and no complicated numerical optimization steps or matrix factorizations or inversions are required. In one simulation example with $n = 500$ and $p_1 = p_2 = 5000$, it takes less than one minute to complete the entire model identification and variable selection process on a regular PC.

After variable selection and structure detection, we would like to provide an inferential tool for the linear and nonparametric components. The spline method is fast and easy to implement; however, the rate of convergence is only established in mean squares sense, and there is no asymptotic distribution or uniform convergence, so no measures of confidence can be assigned to the estimators. In this paper, we propose a two-step “spline-backfitted local linear smoothing” (SBLL) procedure for APLM estimation, model selection and simultaneous inference for all the components. In the first stage, we approximate the nonparametric functions $\phi_l(\cdot)$, $l = 1, \ldots, p_2$, with undersmoothed
constant spline functions. We perform model selection for the APLM using a triple penalized procedure to select important variables and identify the linear vs. nonlinear structure for the continuous covariates, which is crucial to obtain efficient estimators for the non-zero components. We show that the proposed model selection and structure identification for both parametric and nonparametric terms are consistent, and the estimators of the nonzero linear coefficients and nonzero nonparametric functions are both $L_2$-norm consistent. In the second stage, we refit the data with covariates selected in the first step using higher order polynomial splines to achieve root-$n$ consistency of the coefficient estimators in the linear part, and apply a one-step local linear backfitting to the projected nonparametric components obtained from the refitting. Asymptotic normality for both linear coefficient estimators and nonlinear component estimators, as well as simultaneous confidence bands (SCBs) for all nonparametric components, are provided.

The rest of the paper is organized as follows. In Section 3.2, we describe the first-stage spline smoothing and propose a triple penalized regularization method for simultaneous model identification and variable selection. The theoretical properties of selection consistency and rates of convergence for the coefficient estimators and nonparametric estimators are developed. Section 3.3 introduces the spline-backfitted local linear estimators and SCBs for the nonparametric components. The performance of the estimators is assessed by simulations in Section 3.4 and illustrated by application to the SAM data in Section 3.5. Some concluding remarks are given in Section 3.6. Section A of the online Supplemental Materials evaluates the effect of different smoothing parameters on the performance of the proposed method. Technical details are provided in Section B of the Supplemental Materials.
3.2 Methodology

3.2.1 Model Setup

In the following, the functional form (linear vs. nonlinear) for each continuous covariate in model (3.1) is assumed to be unknown. In order to decide the form of \( \phi_l \), for each \( l = 1, \ldots, p_2 \), we can decompose \( \phi_l \) into a linear part and a nonlinear part: 

\[
\phi_l(x) = \beta_l x + g_l(x),
\]

where \( g_l(x) \) is some unknown smooth nonlinear function (see Assumption (A1) in Appendix B). For model identifiability, we assume that 

\[
E(X_{il}) = 0, \quad E(g_l(X_{il})) = 0 \quad \text{and} \quad E(g'_l(X_{il})) = 0.
\]

The first two constraints \( E(X_{il}) = 0 \) and \( E(g_l(X_{il})) = 0 \), are required to guarantee identifiability for the APLM, that is, \( E(\phi_l(X_{il})) = 0 \). The constraint \( E(g'_l(X_{il})) = 0 \) ensures there is no linear form in nonlinear function \( g_l(x) \). Note that these constraints are also in accordance with the definition of nonlinear contrast space in Zhang et al. (2011), which is a subspace of the orthogonal decomposition of RKHS.

In the following, we assume \( Y_i \)s are centered so that we can express the APLM in (3.1) without an intercept parameter as

\[
Y_i = \sum_{k=1}^{p_1} Z_{ik} \alpha_k + \sum_{l=1}^{p_2} X_{il} \beta_l + \sum_{l=1}^{p_2} g_l(X_{il}) + \varepsilon_i. \tag{3.2}
\]

In the following, we define predictor variable \( Z_k \) as irrelevant in model (3.2), if and only if \( \alpha_k = 0 \), and \( X_l \) as irrelevant if and only if \( \beta_l = 0 \) and \( g_l(x_l) = 0 \) for all \( x_l \) on its support. A predictor variable is defined as relevant if and only if it is not irrelevant. Suppose that only an unknown subset of predictor variables is relevant. We are interested in identifying such subsets of relevant predictors consistently while simultaneously estimating their coefficients and/or functions.
We can define the active and inactive index sets for covariates $Z$ and $X$ in a parallel style. For covariates $Z$, we define

Active index set for $Z$: $S_z = \{ k = 1, \cdots, p_1 : \alpha_k \neq 0 \}$,
Inactive index set for $Z$: $N_z = \{ k = 1, \cdots, p_1 : \alpha_k = 0 \}$.

For continuous covariate $X_l$, we say it is a linear covariate if $\beta_l \neq 0$ and $g_l(x_l) = 0$ for all $x_l$ on its support, and $X_l$ is a nonlinear covariate if $g_l(x_l)$ is not zero. Explicitly, we define the following index sets for $X$:

Active pure linear index set for $X$: $S_{x,PL} = \{ l = 1, \cdots, p_2 : \beta_l \neq 0, g_l \equiv 0 \}$,
Active nonlinear index set for $X$: $S_{x,N} = \{ l = 1, \cdots, p_2 : g_l \neq 0 \}$,
Inactive index set for $X$: $N_x = \{ l = 1, \cdots, p_2 : \beta_l = 0, g_l \equiv 0 \}$.

Note that the active nonlinear index set for $X$, $S_{x,N}$, can be decomposed as $S_{x,N} = S_{x,LN} \cup S_{x,PN}$, where $S_{x,LN} = \{ l = 1, \cdots, p_2 : \beta_l \neq 0, g_l \neq 0 \}$ is the index set for covariates whose linear and nonlinear terms in (3.2) are both nonzero, and $S_{x,PN} = \{ l = 1, \cdots, p_2 : \beta_l = 0, g_l \neq 0 \}$ is the index set for active pure nonlinear index set for $X$.

Therefore, the model selection problem for model (3.2) is equivalent to the problem of identifying $S_z, N_z, S_{x,PL}, S_{x,LN}, S_{x,PN}$ and $N_x$. To achieve this, we propose to minimize

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{k=1}^{p_1} Z_{ik} \alpha_k - \sum_{l=1}^{p_2} X_{il} \beta_l - \sum_{l=1}^{p_2} g_l(X_{il}) \right\}^2$$
$$+ \sum_{k=1}^{p_1} p_{\lambda_n1}(|\alpha_k|) + \sum_{l=1}^{p_2} p_{\lambda_n2}(|\beta_l|) + \sum_{l=1}^{p_2} p_{\lambda_n3}(\|g_l\|_2),$$
(3.3)

where $\|g_l\|_2^2 = E\{g_l^2(X_l)\}$, and $p_{\lambda_n1} (\cdot), p_{\lambda_n2} (\cdot)$ and $p_{\lambda_n3} (\cdot)$ are penalty functions explained in detail in Section 3.2.3. The tuning parameters $\lambda_{n1}, \lambda_{n2}$ and $\lambda_{n3}$ decide the complexity of the selected
model. The smoothness of predicted nonlinear functions is controlled by $\lambda_{n3}$, and $\lambda_{n1}$, $\lambda_{n2}$ and $\lambda_{n3}$ go to $\infty$ as $n$ increases to $\infty$.

3.2.2 Spline Basis Approximation

We approximate the smooth functions $\{g_l(\cdot) : l = 1, \ldots, p_2\}$ in (3.2) by polynomial splines for their simplicity in computation. For example, for each $l = 1, \ldots, p_2$, let $v_{0,l}, \ldots, v_{N_n+1,l}$ be knots that partition $[a, b]$ with $a = v_{0,l} < v_{1,l} < \cdots < v_{N_n,l} < v_{N_n+1,l} = b$. The polynomial splines of order $d$, $b^{(d)}_{J,l}(\cdot)$, $J = 1, \ldots, N_n+d$, are polynomial functions with $(d-1)$-degree (or less) on intervals $[v_{J,l}, v_{J+1,l})$, $J = 0, \ldots, N_n-1$, and $[v_{N_n,l}, v_{N_n+1,l}]$, and have $d-2$ continuous derivatives globally.

To ensure $E\{g_l(X_{il})\} = 0$ and $E\{g_l'(X_{il})\} = 0$, we consider the following normalized first-order B-splines, referred to as piecewise constant splines. We define for any $l = 1, \ldots, p_2$ the piecewise constant B-spline function as the indicator function $I_{J,l}(x_l)$ of the $(N_n+1)$ equally-spaced subintervals of the finite interval $[a, b]$ with length $H = H_n = (b - a) / (N_n + 1)$, that is,

$$I_{J,l}(x_l) = \begin{cases} 1 & a + JH \leq x_l < a + (J + 1)H, \\ 0 & \text{otherwise}, \end{cases}$$

$$I_{N_n,l}(x_l) = \begin{cases} 1 & a + N_nH \leq x_l \leq b, \\ 0 & \text{otherwise}, \end{cases}$$

Define the following centered spline basis $b^{(1)}_{J,l}(x_l) = I_{J,l}(x_l) - \frac{\|I_{J,l}\|_2}{\|I_{J-1,l}\|_2}I_{J-1,l}(x_l)$, for $J = 1, \ldots, N_n$ and $l = 1, \ldots, p_2$, with the standardized version given for any $l = 1, \ldots, p_2$,

$$B^{(1)}_{J,l}(x_l) = b^{(1)}_{J,l}(x_l) / \|b^{(1)}_{J,l}\|_2, \forall J = 1, \ldots, N_n.$$  \hspace{1cm} (3.4)

So for any $l = 1, \ldots, p_2$ and $J = 1, \ldots, N_n$, we have $E\{B^{(1)}_{J,l}(X_{il})\} = 0$, $E\{B^{(1)}_{J,l}(X_{il})\}^2 = 1$. So we can obtain the normalized spline basis following these procedures.
We approximate the nonparametric function \( g_l(x_l), l = 1, \ldots, p_2 \), using the above normalized piecewise constant splines

\[
g_l(x_l) \approx g_{ls}(x_l) = \sum_{J=1}^{N_n} \gamma_{J,l} B_{J,l}(x_l) = B_{l}^{(1)\top}(x_l) \gamma_l, \tag{3.5}
\]

where \( B_{l}^{(1)}(x_l) = (B_{1,l}(x_l), \ldots, B_{N_n,l}(x_l))^\top \), and \( \gamma_l = (\gamma_{1,l}, \ldots, \gamma_{N_n,l})^\top \) is a vector of the spline coefficients. By using the centered constant spline basis functions, we can guarantee that \( \frac{1}{n} \sum_{i=1}^{n} g_{ls}(X_{il}) = 0 \), and \( \frac{1}{n} \sum_{i=1}^{n} g_{ls}'(X_{il}) = 0 \) except at the location of the knots.

Denote a length \( N_n \) vector \( B_{il}^{(1)} = (B_{1,l}(X_{il}), \ldots, B_{N_n,l}(X_{il}))^\top \). For any vector \( a \in \mathbb{R}^p \), denote \( \|a\| = (\sum_{l=1}^{p} |a_l|^2)^{1/2} \) as the \( L_2 \) norm of \( a \). Following from (3.5), to minimize (3.3), it is approximately equivalent to consider the problem of minimizing

\[
\sum_{i=1}^{n} \left\{ Y_i - \sum_{k=1}^{p_1} Z_{ik}\alpha_k - \sum_{l=1}^{p_2} X_{il}\beta_l - \sum_{l=1}^{p_2} B_{il}^{(1)} \gamma_l \right\}^2 + \sum_{k=1}^{p_1} p\lambda_{n1}(|\alpha_k|) + \sum_{l=1}^{p_2} p\lambda_{n2}(|\beta_l|) + \sum_{l=1}^{p_2} p\lambda_{n3}(\|\gamma_l\|). \tag{3.6}
\]

### 3.2.3 Adaptive Group LASSO Regularization

We use adaptive LASSO (Zou, 2006) and adaptive group LASSO (Huang et al., 2010) for variable selection and estimation. Other popular choices include methods based on the Smoothly Clipped Absolute Deviation penalty (Fan and Li, 2001) or the minimax concave penalty (Zhang, 2010).

Specifically, we start with group LASSO estimators obtained from the following minimization:

\[
(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}) = \arg\min_{\alpha, \beta, \gamma} \sum_{i=1}^{n} \left\{ Y_i - \sum_{k=1}^{p_1} Z_{ik}\alpha_k - \sum_{l=1}^{p_2} X_{il}\beta_l - \sum_{l=1}^{p_2} B_{il}^{(1)} \gamma_l \right\}^2 + \lambda_{n1} \sum_{k=1}^{p_1} |\alpha_k| + \lambda_{n2} \sum_{l=1}^{p_2} |\beta_l| + \lambda_{n3} \sum_{l=1}^{p_2} \|\gamma_l\|. \tag{3.6}
\]

Then, let \( w_k^\alpha = |\tilde{\alpha}_k|^{-1} I\{|\tilde{\alpha}_k| > 0\} + \infty \times I\{|\tilde{\alpha}_k| = 0\}, w_l^\beta = |\tilde{\beta}_l|^{-1} I\{|\tilde{\beta}_l| > 0\} + \infty \times I\{|\tilde{\beta}_l| = 0\}, \)

\( w_l^\gamma = \|\tilde{\gamma}_l\|^{-1} I\{\|\tilde{\gamma}_l\| > 0\} + \infty \times I\{\|\tilde{\gamma}_l\| = 0\}, \)

where by convention, \( \infty \times 0 = 0 \).
After obtaining the weights, we can further obtain the adaptive group LASSO estimators by minimizing the adaptive group LASSO objective function, which is defined as

\[
L(\alpha, \beta, \gamma; \lambda_{n1}, \lambda_{n2}, \lambda_{n3}) = \sum_{i=1}^{n} \left\{ Y_i - \sum_{k=1}^{p_1} Z_{ik} \alpha_k - \sum_{l=1}^{p_2} X_{il} \beta_l - \sum_{l=1}^{p_2} B_{il}^{(1)} \gamma_l \right\}^2 + \lambda_{n1} \sum_{k=1}^{p_1} w_k^0 |\alpha_k| + \lambda_{n2} \sum_{l=1}^{p_2} w_l^0 |\beta_l| + \lambda_{n3} \sum_{l=1}^{p_2} w_l^7 \|\gamma_l\|. \tag{3.7}
\]

The adaptive group LASSO estimators are minimizers of (3.7), denoted by

\[
(\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = \arg \min_{\alpha, \beta, \gamma} L(\alpha, \beta, \gamma; \lambda_{n1}, \lambda_{n2}, \lambda_{n3}).
\]

The model structure selected is defined by

\[
\hat{S}_z = \{ 1 \leq k \leq p_1 : |\hat{\alpha}_k| > 0 \}, \quad \hat{S}_{x,PL} = \{ l : |\hat{\beta}_l| > 0, \|\hat{\gamma}_l\| = 0, 1 \leq l \leq p_2 \},
\]

\[
\hat{S}_{x,LN} = \{ l : |\hat{\beta}_l| > 0, \|\hat{\gamma}_l\| > 0, 1 \leq l \leq p_2 \},
\]

\[
\hat{S}_{x,PN} = \{ l : |\hat{\beta}_l| = 0, \|\hat{\gamma}_l\| > 0, 1 \leq l \leq p_2 \}. \tag{3.8}
\]

The spline estimators of each component function are

\[
\hat{g}_l(x_i) = \sum_{J=1}^{N_n} \hat{\gamma}_{lJ} B_{Jl}(x_i) - n^{-1} \sum_{i=1}^{n} \sum_{J=1}^{N_n} \hat{\gamma}_{lJ} B_{Jl}(X_{il}).
\]

Accordingly, the spline estimators for the original component functions \(\phi_l\)'s are \(\hat{\phi}_l(x_i) = \hat{\beta}_l x_i + \hat{g}_l(x_i)\).

The following theorems establish the asymptotic properties of the adaptive group LASSO estimators. Theorem 3.1 shows the proposed method can consistently distinguish nonzero components from zero components. Theorem 3.2 gives the convergence rates of the estimators. We only state the main results here. The regularity conditions and proofs are provided in Appendix B – B.

**Theorem 3.1.** Suppose that Assumptions (A1)-(A6) in Appendix B hold. As \(n \to \infty\), we have

\[
\hat{S}_z = S_z, \quad \hat{S}_{x,PL} = S_{x,PL}, \quad \hat{S}_{x,LN} = S_{x,LN} \text{ and } \hat{S}_{x,PN} = S_{x,PN} \text{ with probability approaching one.}
\]
In the following, to avoid confusion, we use \( \alpha_0 = (\alpha_{01}, \cdots, \alpha_{0q})^\top \), \( \beta_0 = (\beta_{01}, \cdots, \beta_{0p})^\top \) to denote the true parameters in model (3.2), and \( g_0 = (g_{01}, \cdots, g_{0p})^\top \) to denote the nonlinear functions in model (3.2). Let \( \alpha_0 = (\alpha_{0S_z}, \alpha_{0N_z})^\top \), where \( \alpha_{0S_z} \) consists of all nonzero components of \( \alpha_0 \), and \( \alpha_{0N_z} = 0 \) without loss of generality; similarly, let \( \beta_0 = (\beta_{0S_x,L}, \beta_{0N_x})^\top \), where \( \beta_{0S_x,L} \) consists of all nonzero components of \( \beta_0 \), and \( \beta_{0N_x} = 0 \) without loss of generality.

**Theorem 3.2.** Suppose that Assumptions (A1) – (A6) in Appendix B hold. Then

\[
\sum_{k \in S_z} |\alpha_k - \alpha_{0k}|^2 = O_p(n^{-1}N_n) + O_p \left( N_n^{-2} \right) + O_p \left( n^{-2} \sum_{j=1}^{3} \lambda_{nj}^2 \right),
\]

\[
\sum_{l \in S_{x,L}} |\beta_l - \beta_{0l}|^2 = O_p(n^{-1}N_n) + O_p \left( N_n^{-2} \right) + O_p \left( n^{-2} \sum_{j=1}^{3} \lambda_{nj}^2 \right),
\]

\[
\sum_{l \in S_{x,N}} \|g_l - g_{0l}\|_2^2 = O_p(n^{-1}N_n) + O_p \left( N_n^{-2} \right) + O_p \left( n^{-2} \sum_{j=1}^{3} \lambda_{nj}^2 \right). 
\]

### 3.3 Two-stage SBLL Estimator and Inference

After model selection, our next step is to conduct statistical inference for the nonparametric component functions of those important variables. Although the one-step penalized estimation in Section 3.2.3 can quickly identify the nonzero nonlinear components, the asymptotic distribution is not available for the resulting estimators.

To obtain estimators whose asymptotic distribution can be used for inference, we first refit the data using selected model,

\[
Y_i = \sum_{k \in \tilde{S}_z} Z_{ik} \alpha_k + \sum_{j \in \tilde{S}_{x,PL}} X_{ij} \beta_j + \sum_{l \in \tilde{S}_{x,N}} \phi_l (X_{il}) + \epsilon_i. \quad (3.9)
\]

We approximate the smooth functions \( \{\phi_l (\cdot) : l \in \tilde{S}_{x,N}\} \) in (3.9) by polynomial splines introduced in Section 3.2.2. Let \( \mathcal{B}_{l}^{(d)} \) be the space of polynomial splines of order \( d \), and \( \mathcal{B}_l^0 = \{b \in \mathcal{B}_l^{(d)} : \)
E\{b(X_l)\} = 0, E\{b^2(X_l)\} < \infty$. Working with $B_t^0$ ensures that the spline functions are centered, see for example Xue and Yang (2006); Wang and Yang (2007); Wang et al. (2014). Let $\left\{B_{ij}^{(d)}(\cdot)\right\}_{j=1}^{M_n}$ be a set of standardized spline basis functions for $B_t^0$ with dimension $M_n = N_n + d - 1$, where $B_{ij}^{(d)}(x_l) = b_{ij}^{(d)}(x_l)/\|b_{ij}^{(d)}\|_2$, $J = 1, \ldots, M_n$, so that $E\{B_{ij}^{(d)}(x_l)\} = 0, E\{B_{ij}^{(d)}(x_l)\}^2 = 1$. Specifically, if $d = 1$, $M_n = N_n$ and $B_{ij}^{(1)}(\cdot)$ is the standardized piecewise constant spline function defined in (3.4).

We propose a one-step backfitting using refitted pilot spline estimators in the first stage followed by local linear estimators. The refitted coefficients are defined as

$$(\hat{\alpha}^*, \hat{\beta}^*, \hat{\gamma}^*) = \arg\min_{\alpha, \beta, \gamma} \sum_{i=1}^{n} \left( Y_i - \sum_{k \in \mathcal{S}_x} Z_{ik} \alpha_k - \sum_{j \in \mathcal{S}_x, PL} X_{ij} \beta_j - \sum_{l \in \mathcal{S}_x, N} B_{il}^{(d)} \gamma_l \right)^2. \tag{3.10}$$

Then the refitted spline estimators for nonlinear functions $\phi_l(\cdot)$ is

$$\hat{\phi}_l^* (x_l) = B_{il}^{(d)} (x_l) \gamma_l^*, \quad l \in \mathcal{S}_x, N. \tag{3.11}$$

Next we establish the asymptotic normal distribution for the parametric estimators. To make $\beta_{0,\mathcal{S}_Z}$ estimable at the $\sqrt{n}$ rate, we need a condition to ensure $X$ and $Z$ are not functionally related.

Define

$$\mathcal{F}_+ = \left\{ f(x) = \sum_{l \in \mathcal{S}_x, N} f_l(x_l), \ E\{f_l(x_l)\} = 0, \ \|f_l\|_2 < \infty \right\}$$

the Hilbert space of theoretically centered $L^2$ additive functions. For any $k \in \mathcal{S}_z$, let $z_k$ be the coordinate mapping that maps $Z$ to its $k$-th component so that $z_k(Z) = Z_k$, and let

$$\psi_k^* = \arg\min_{\psi \in \mathcal{F}_+} \|z_k - \psi\|_2^2 = \arg\min_{\psi \in \mathcal{F}_+} E\{Z_k - \psi(X)\}^2$$

be the orthogonal projection of $z_k$ onto $\mathcal{F}_+$. Then we can define $Z_{\mathcal{S}_x}$ from the projection function $\psi_k^*$ by letting $Z_{\mathcal{S}_x} = \{\psi_k^*(X), k \in \mathcal{S}_z\}^\top$. 
Similarly, for any \( l \in S_{x,PL} \), let \( x_l \) be the coordinate mapping that maps \( X \) to its \( l \)-th component, so that \( x_l(X) = X_l \), and let

\[
\psi^*_l = \arg\min_{\psi \in \mathcal{F}_+} \| x_l - \psi \|_2 = \arg\min_{\psi \in \mathcal{F}_+} \mathbb{E}\{ X_l - \psi(X) \}^2
\]

be the orthogonal projection of \( x_l \) onto \( \mathcal{F}_+ \). Let \( \tilde{X}_{S_{z,PL}} = \{ \psi^*_l(X), l \in S_{x,PL} \}^\top \). Define \( Z_{S_z} = (Z_k, k \in S_z)^\top \) and \( X_{S_{x,PL}} = (X_l, l \in S_{x,PL})^\top \). Denote vector \( T \) and \( \tilde{T} \) as

\[
T = (Z_{S_z}, X_{S_{x,PL}})^\top,
\]

\[
\tilde{T} = (Z_{S_z}, \tilde{X}_{S_{x,PL}})^\top.
\]

**Theorem 3.3.** Under the Assumptions (A1)–(A6), (A3') and (A6') given in Appendix B,

\[
(n\Sigma)^{1/2} \begin{pmatrix} \hat{\alpha}_{S_z}^* - \alpha_{0,S_z} \\ \hat{\beta}_{S_{x,PL}}^* - \beta_{0,S_{x,PL}} \end{pmatrix} \xrightarrow{D} N(0, I),
\]

where \( I \) is an identity matrix and \( \Sigma = \sigma^{-2} E[(T - \tilde{T})(T - \tilde{T})^\top] \).

The proof of Theorem 3.3 is similar to the proof of Liu et al. (2011) and Li et al. (2018) and thus omitted.

Note that the linear coefficients estimators reach the asymptotic normality at the rate of \( n^{1/2} \), which indicates that these linear coefficients estimators reach the rate same as the oracle estimator. However, we cannot obtain similar rates for the estimators for the nonlinear functions, and we cannot obtain the asymptotic normality directly. Thus, in the following, we need to reply on the kernel method to obtain the asymptotic normality for the estimators of the nonlinear functions.

Let \( \Omega_n = \{ \hat{S}_z = S_z, \hat{S}_{x,PL} = S_{x,PL} \} \). In the selection step, we estimate \( S_z \) and \( S_{x,PL} \) consistently, that is, \( P(\Omega_n) \rightarrow 1 \). Within the event \( \Omega_n \), that is, \( \hat{S}_z = S_z \) and \( \hat{S}_{x,PL} = S_{x,PL} \), the estimator \((\hat{\alpha}^*_{S_z}, \hat{\beta}^*_{S_{x,PL}})^\top\) is root-\( n \) consistent according to Theorem 3.3. Since \( \Omega_n \) is shown to have probability tending to one, we can conclude that \((\hat{\alpha}^*_{S_z}, \hat{\beta}^*_{S_{x,PL}})^\top\) is also root-\( n \) consistent.
These refitted pilot estimators defined in (3.10) and (3.11) are then used to define new pseudo-
responses \( \hat{Y}_{il} \), which are estimates of the unobservable “oracle” responses \( Y_{il} \). Specifically,

\[
\hat{Y}_{il} = Y_i - \left\{ \sum_{k \in S_z} Z_{ik} \alpha_k^* + \sum_{l' \in S_{x,PL}} X_{il'} \beta_{il'}^* + \sum_{l'' \in S_{x,N} \setminus \{l\}} \phi_{il''}(X_{il''}) \right\},
\]

(3.13)

\[
Y_{il} = Y_i - \left\{ \sum_{k \in S_z} Z_{ik} \alpha_{0k} + \sum_{l' \in S_{x,PL}} X_{il'} \beta_{0l'} + \sum_{l'' \in S_{x,N} \setminus \{l\}} \phi_{0l''}(X_{il''}) \right\}.
\]

(3.14)

Denote \( K(\cdot) \) a continuous kernel function, and let \( K_h(t) = K(t/h) / h \) be a rescaling of \( K \), where \( h \) is usually called the bandwidth. Next, we define the spline-backfitted local linear (SBLL) estimator of \( \phi_l(x_l) \) as \( \hat{\phi}_{SBLL}^l(x_l) \) based on \( \{ X_{il}, \hat{Y}_{il} \}_{i=1}^n \), which attempts to mimic the would-be SBLL estimator \( \hat{\phi}_o^l(x_l) \) of \( \phi_l(x_l) \) based on \( \{ X_{il}, Y_{il} \}_{i=1}^n \) if the unobservable “oracle” responses \( \{ Y_{il} \}_{i=1}^n \) were available:

\[
\left( \hat{\phi}_o^l(x_l), \hat{\phi}_{SBLL}^l(x_l) \right) = (1 \ 0) \left( X_l^{*\top} W_l X_l^* \right)^{-1} X_l^{*\top} W_l (Y_l, \hat{Y}_l),
\]

(3.15)

where the oracle and pseudo-response vectors are \( Y_l = (Y_1l, \ldots, Y_nl)^\top \) and \( \hat{Y}_l = (\hat{Y}_{1l}, \ldots, \hat{Y}_{nl})^\top \), with \( \hat{Y}_{il} \) and \( Y_{il} \) as defined in (3.14), respectively; and the weight and “design” matrices are

\[
W_l = n^{-1} \text{diag} \{ K_h(X_{il} - x_l) \}_{i=1}^n, \quad X_l^{*\top} = \begin{pmatrix} 1 & \ldots & 1 \\ X_{il} - x_l & \ldots & X_{nl} - x_l \end{pmatrix}.
\]

Asymptotic properties of smoothers of \( \hat{\phi}_o^l(x_l), \ l \in S_{x,N} \), can be easily established by using the theoretical property from kernel estimators. Noticeably, compared to the asymptotic properties we obtained in the Theorem 3.3, the rates will be different.
Specifically, let \( \mu_2(K) = \int u^2 K(u) \, du \), and let \( f_l \) be the probability density function of \( X_l \), then under Assumptions (B1) and (B2) in Appendix B,

\[
\sqrt{n} \left\{ \hat{\phi}_l^{SBLL}(x_l) - \phi_0(x_l) - b_l(x_l)h^2 \right\} \xrightarrow{D} N \left\{ 0, \nu_l^2(x_l) \right\}, \ l \in S_{x,N},
\]

where

\[
b_l(x_l) = \mu_2(K)\phi_0''(x_l)/2, \quad \nu_l^2(x_l) = \|K\|^2f_l^{-1}(x_l)\sigma^2.
\]

The following theorem states that the asymptotic uniform magnitude of the difference between \( \phi_{SBLL}^s(x_l) \) and \( \phi_0^o(x_l) \) is of order \( o_p\{(nh)^{-1/2}\} \), which is dominated by the asymptotic uniform size of \( \phi_l^o(x_l) - \phi_0(x_l) \). As a result, \( \phi_{SBLL}^s(x_l) \) will have the same asymptotic distribution as \( \phi_l^o(x_l) \).

We say \( x_l \in \chi \) is a boundary point if and only if \( x_l = a + ch \) or \( x_l = b - ch \) for some \( 0 \leq c < 1 \) and an interior point otherwise. Let \( \chi_h \) be the interior of the support \( \chi \).

**Theorem 3.4.** Suppose the assumptions in Theorem 3.3 hold. In addition, if Assumptions (B1) and (B2) in Appendix B are satisfied, then the SBLL estimator \( \phi^*_l(x_l) \) given in (3.15) satisfies

\[
\sup_{x_l \in \chi} \left| \phi_{SBLL}^s(x_l) - \phi_0^o(x_l) \right| = o_p\{(nh)^{-1/2}\}, \ l \in S_{x,N}.
\]

Hence with \( b_l(x_l) \) and \( \nu_l^2(x_l) \) as defined in (3.17), for any \( x_l \) in its interior support \( \chi_h \),

\[
\sqrt{n} \left\{ \hat{\phi}_l^{SBLL}(x_l) - \phi_0(x_l) - b_l(x_l)h^2 \right\} \xrightarrow{D} N \left\{ 0, \nu_l^2(x_l) \right\}, \ l \in S_{x,N}.
\]

With the additional Assumption (A2'), the estimator \( \hat{\phi}_l^s(x_l) \) also satisfies, for any \( t \) and \( l \in S_{x,N} \),

\[
\lim_{n \to \infty} \Pr \left\{ \sqrt{\log(h^{-2})} \left( \sup_{x_l \in \chi_h} \frac{\sqrt{n}h}{v_l(x_l)} \left| \hat{\phi}_l^{SBLL}(x_l) - \phi_0(x_l) \right| - \tau_n \right) < t \right\} = e^{-2e^{-t}},
\]

where \( \tau_n = \sqrt{\log(h^{-2})} + \log\{\|K'\|/2(2\pi\|K\|)\}/\sqrt{\log(h^{-2})} \).

Theorem 3.4 provides analytical expressions for constructing asymptotic confidence intervals and SCBs under certain conditions. Under Assumptions (A1) – (A6), (A3'), (A6'), (B1) and (B2)
in Appendix B, for any $\alpha \in (0, 1)$, an asymptotic $100(1 - \alpha)\%$ pointwise confidence interval for $\phi_{0l}(x_l)$ over the interval $\chi_h$ is

$$\hat{\phi}_{l}^{SBLL}(x_l) - \hat{b}_l(x_l)h^2 \pm \hat{v}_l(x_l)(nh)^{-1/2}, \ l \in S_{x,N}.$$  

Under Assumptions (A1) – (A6), (A2)', (A3)', (A6'), (B1) and (B2) in the Appendix, for any $\alpha \in (0, 1)$, an asymptotic $100(1 - \alpha)\%$ SCB for $\phi_{0l}(x_l)$ over the interval $\chi_h$ is

$$\hat{\phi}_{l}^{SBLL}(x_l) \pm \hat{v}_l(x_l)(nh)^{-1/2} \left[ \tau_n - \{\log(h^{-2})\}^{-1/2} \log \left\{ -\frac{1}{2} \log(1 - \alpha) \right\} \right], \ l \in S_{x,N}.$$

### 3.4 Implementation and Simulation

In this section we discuss practical implementations for the SMILE procedure. To meet the zero mean requirement specified in Assumption (A4), we use the centralized $X^*_il$ instead of $X_{il}$ directly, for each $l = 1, \ldots, p_2$. At the risk of abusing the notation, we still use symbol $X$ instead of $X^*$ to avoid creating too many new symbols. To implement the proposed procedure, one needs to select the penalty parameters, the knots for a spline at the selection stage and refitting stage, and the bandwidth for a kernel at the backfitting stage.

**Knot selection.** For spline smoothing involved in both selection and refitting, we suggest placing knots on a grid of evenly spaced sample quantiles. Based on extensive simulation experiments in Section A of the Supplementary Materials, we find that the number of knots often has little effect on the model selection results. Therefore, we recommend using a small number of knots at the model selection stage to reduce the computing cost, especially when the sample size is too small compared to the number of covariates. In practice, $2 \sim 5$ interior knots is usually adequate to identify the model structure.
At the refitting stage, Assumption (A6') in the Supplementary Materials suggests the number of interior knots $M_n$ for a refitting spline needs to satisfy:

\[
\left\{n^{1/(2d)} \lor n^{4/(10d-5)}\right\} \ll M_n \ll n^{1/3},
\]

where $d$ is the degree of the polynomial spline basis functions used in the refitting. The widely used quadratic/cubic splines and any polynomial splines of degree $d \geq 2$ all satisfy this condition. Therefore, in practice we suggest take the following rule-of-thumb number of interior knots

\[
\min\{\left\lfloor n^{1/(2d)}\lor n^{4/(10d-5)} \log(n) \right\rfloor, \left\lfloor n/(4s) \right\rfloor\} + 1,
\]

where $s$ is the number of nonlinear components selected at the first stage, and the term $\left\lfloor n/(4s) \right\rfloor$ is to guarantee that we have at least four observations in each subinterval between two adjacent knots to avoid getting (near) singular design matrices in the spline refitting.

**Bandwidth selection.** Note that Condition (B2) in the Supplementary Materials requires that the bandwidths in the backfitting are of order $n^{-1/5}$. Thus, the bandwidth selection can be done using a standard routine in the literature. In our numerical studies, we find that the rule-of-thumb bandwidth selector (Fan and Gijbels, 1996) often works very well in both estimation and SCB construction.

Section A in the Supplementary Materials provides detailed investigations on how the smoothing parameters affect the proposed SMILE method and evaluates the practical performance in finite-sample simulation studies. Next we present our algorithm and discuss how to choose the penalty parameters.
3.4.1 Algorithm

The minimization of (3.7) can be solved by the group coordinate descent algorithm (Huang et al., 2012a), implemented using R package `grpreg` (Breheny, 2016). As for the selection of penalty parameters, we consider two criteria widely used in high-dimensional settings, modified Bayesian information criteria (BIC; see Lee et al. (2014)) and the extended BIC (EBIC; see Chen and Chen (2008, 2009)):

\[
\text{BIC}(\lambda) = \log(\text{RSS}_\lambda) + df_\lambda \times \frac{\log(p_1 + p_2 + p_2 N_n) \times \log(n)}{2n},
\]

\[
\text{EBIC}(\lambda) = \log(\text{RSS}_\lambda) + df_\lambda \times \frac{\log(n)}{n} + df_\lambda \times \frac{\log(p_1 + p_2 + p_2 N_n)}{n},
\]

where \( \text{RSS}_\lambda \) is the residual sum of squares associated with penalty parameters \( \lambda = (\lambda_1, \lambda_2, \lambda_3)\)\(\text{T}\) and \( df_\lambda \) is the number of estimated nonzero coefficients for the given \( \lambda \). The simulation results are similar based on these two criteria, so in the following, we choose \( \lambda_1 \) and \( \lambda_2 \) by modified BIC and \( \lambda_3 \) by EBIC for illustration using an approach described below.

The classical coordinate descent algorithm deals with the optimization problem with one tuning parameter, and there are several ways to address the triple-penalization or multiple-penalization issue. A natural idea is to solve the optimization problem by searching over a three-dimensional grid for tuning parameters, which can be computationally expensive. To pose a balance between computational efficiency and precision, we propose to solve the triple-penalization problem iteratively. First, BIC is minimized with a common penalty parameter, that is, \( \lambda_1 = \lambda_2 = \lambda_3 = \lambda \), say. Starting with this common parameter \( \lambda \), we obtain the estimators \( \hat{\alpha}^{(0)}, \hat{\beta}^{(0)} \) and \( \hat{\gamma}^{(0)} \) as our initial estimators. Then, \( \lambda_1, \lambda_2, \lambda_3 \) are selected one at a time by minimizing the BIC. More precisely, Algorithm 1 outlines the iterative group coordinate descent algorithm.
Algorithm 1 Iterative group coordinate descent algorithm

Input: Data \(\{(Y_i, Z_{i1}, \ldots, Z_{ip}, X_{i1}, \ldots, X_{ip}, B_{i1}^{(1)}, \ldots, B_{ip}^{(1)})\}_{i=1}^{n}\)
\(\alpha^{(0)}, \beta^{(0)}\) and \(\gamma^{(0)}\): initial parameters of interest
\(\delta_0\): convergence criterion

Output: \(\hat{\alpha}, \hat{\beta}\) and \(\hat{\gamma}\): Estimates of \(\alpha, \beta\) and \(\gamma\)

\[
\text{while } \left\| \left( \alpha^{(m+1)\top}, \beta^{(m+1)\top}, \gamma^{(m+1)\top} \right) - \left( \hat{\alpha}^{(m)\top}, \hat{\beta}^{(m)\top}, \hat{\gamma}^{(m)\top} \right) \right\|^2 > \delta_0 \text{ do }
\]

(i) Given \(\hat{\beta}^{(m)}\) and \(\hat{\gamma}^{(m)}\), obtain \(w_1^{\alpha(m+1)}, \ldots, w_{p_1}^{\alpha(m+1)}\) by minimizing objective function (3.6) with \(\lambda_1\) selected via the modified BIC;

(ii) Given \(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)}\) and \(w_1^{\alpha(m+1)}, \ldots, w_{p_1}^{\alpha(m+1)}\), obtain \(\hat{\alpha}^{(m+1)}\) by minimizing objective function (3.7) with \(\lambda_1\) selected via the modified BIC;

(iii) Given \(\hat{\alpha}^{(m+1)}\) and \(\hat{\gamma}^{(m)}\), obtain \(w_1^{\beta(m+1)}, \ldots, w_{p_2}^{\beta(m+1)}\) by minimizing objective function (3.6) with \(\lambda_2\) selected via the modified BIC;

(iv) Given \(\hat{\alpha}^{(m+1)}, \hat{\gamma}^{(m)}\) and \(w_1^{\beta(m+1)}, \ldots, w_{p_2}^{\beta(m+1)}\), obtain \(\hat{\beta}^{(m+1)}\) by minimizing objective function (3.7) with \(\lambda_2\) selected via the modified BIC;

(v) Given \(\hat{\alpha}^{(m+1)}\) and \(\hat{\beta}^{(m+1)}\), obtain \(w_1^{\gamma(m+1)}, \ldots, w_{p_2}^{\gamma(m+1)}\) by minimizing objective function (3.6) with \(\lambda_3\) selected via EBIC;

(vi) Given \(\hat{\alpha}^{(m+1)}, \hat{\beta}^{(m+1)}\) and \(w_1^{\gamma(m+1)}, \ldots, w_{p_2}^{\gamma(m+1)}\), obtain \(\hat{\gamma}^{(m+1)}\) by minimizing objective function (3.7) with \(\lambda_3\) selected via EBIC.

end

Set \(\hat{\alpha} = \hat{\alpha}^{(m+1)}, \hat{\beta} = \hat{\beta}^{(m+1)}\) and \(\hat{\gamma} = \hat{\gamma}^{(m+1)}\).
3.4.2 Simulation Studies

In this section, we investigate the performance of the proposed sparse model identification and learning estimator, abbreviated as SMILE, in terms of model selection, estimation accuracy and inference performance in a simulation study. We compare the SMILE with the sparse APLM estimator with adaptive group LASSO penalty (SAPLM) proposed in Li et al. (2018), the ordinary linear least squares estimator with the adaptive LASSO penalty (SLM), and the oracle estimator (ORACLE), which uses the same estimation techniques as the SMILE except that no penalization or data-driven variable selection is used because all active and inactive index sets are treated as known. Note that the SAPLM ignores the potential linear structure in covariate $X$, and estimates the effects of each component of $X$ with all nonparametric forms; in contrast, the SLM ignores the potential nonlinear structure in covariate $X$ and requires selected components of covariates $Z$ and $X$ to enter the model in a linear form. In terms of the performances of SCBs, we compare the SMILE with SAPLM and ORACLE. In our simulation, ORACLE works as a benchmark for estimation comparison. It is worth pointing out that the ORACLE estimator is only computable in simulations, not real examples.

We generate simulated datasets using the APLM structure

$$Y_i = \sum_{k=1}^{p_1} Z_{ik} \alpha_k + \sum_{l=1}^{p_2} \phi_l(X_{il}) + \varepsilon_i,$$

where $\alpha_1 = 3$, $\alpha_2 = 4$, $\alpha_3 = -2$, $\alpha_4 = \cdots = \alpha_{p_1} = 0$,

$$\phi_1(x) = 9x,$$

$$\phi_2(x) = -1.5 \cos^2(\pi x) + 3 \sin^2(\pi x) - \mathbb{E}\{-1.5 \cos^2(\pi x + 3 \sin^2(\pi x))\},$$

$$\phi_3(x) = 6x + 18x^2 - \mathbb{E}(6x + 18x^2),$$
and \( \phi_4(x) = \cdots = \phi_{p_2}(x) = 0 \). Notice that \( \phi_1(x) \) is actually a linear function. So there are three variables in the active index set for \( Z \), one variable in the active pure linear index set for \( X \), one variable in the active pure nonlinear index set for \( X \), and one variable in the active linear & nonlinear index set for \( X \).

We simulate \( z_{ik}^* \) independently from the \( \text{Unif}[0, 1] \) and \( x_{il} \) independently from the \( \text{Unif}[-.5,.5] \), and set \( z_{ik} = I(z_{ik}^* > 0.75) \), for \( i = 1, \ldots, n \), \( k = 1, \ldots, p_1 \), \( l = 1, \ldots, p_2 \). To make an ultra-high-dimensional scenario, we let the sample size \( n = 300 \) and \( n = 500 \), and consider three different dimensions: \( p_1 = p_2 = p \), where \( p \) is taken to be 1000, 2000 and 5000. The error term \( \varepsilon_i \) is simulated from \( N(0, \sigma^2) \) with \( \sigma = 0.5 \) and 1.0.

To approximate the nonlinear functions, we use the constant B-spline \( (d = 1) \) with four interior knots for selection and use the cubic B-spline \( (d = 4) \) with four interior knots in the refitting step. For both selection and refitting, the knots are on a grid of evenly spaced sample quantiles. To construct the SCBs, in our simulation studies below, we choose the Epanechnikov kernel function with the rule-of-thumb bandwidth described in Section 4.2 in Fan and Gijbels (1996), which usually works well in our experimental investigation. More simulation studies have been conducted with different choices for spline knots and kernel bandwidth selectors; see Section A of the Supplementary Materials.

We evaluate the methods on the accuracy of variable selection, prediction and inference. In detail, we adopt the following criteria for evaluation:

(B-i) Percent of correctly identified covariates in \( Z \) with nonzero linear coefficients ("CorrZ");
(B-ii) Percent of correctly identified covariates in \( Z \) with zero linear coefficients ("CorrZ0");
(B-iii) Percent of correctly identified covariates in \( X \) with nonzero purely linear functions ("CorrL");
(B-iv) Percent of correctly identified covariates in $X$ with nonzero purely nonlinear functions ("CorrN");

(B-v) Percent of correctly identified covariates in $X$ with nonzero linear and nonlinear functions ("CorrLN");

(B-vi) Percent of correctly identified covariates in $X$ with zero functions ("CorrX0");

(C-i) Percent of covariates in $Z$ with nonzero linear coefficients incorrectly identified as having zero linear coefficients ("Zto0");

(C-ii) Percent of covariates in $X$ with nonzero purely linear functions incorrectly identified as having nonlinear functions ("LtoN");

(C-iii) Percent of covariates in $X$ with nonzero purely nonlinear functions incorrectly identified as having linear functions ("NtoL");

(C-iv) Percent of covariates in $X$ with nonzero linear or nonzero nonlinear functions incorrectly identified as having both zero linear and zero nonlinear functions ("Xto0");

(D-i) Mean squared errors (MSE) for linear coefficients $\alpha_1, \alpha_2, \alpha_3$ and $\beta_1$;

(D-ii) Average MSE (AMSE) for $\phi_1, \phi_2$ and $\phi_3$, defined as $\frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{\phi}_l^{\text{SBLI}}(x_{il}) - \phi_l(x_{il}) \right\}^2$;

(D-iii) 10-fold cross-validation mean squared prediction error (CV-MSPE) for the response variable, defined as $\frac{1}{10} \sum_{m=1}^{10} \frac{1}{|\kappa_m|} \sum_{i \in \kappa_m} (\hat{Y}_i - Y_i)^2$, where $\kappa_1, \ldots, \kappa_{10}$ comprise a random partition of the dataset into 10 disjoint subsets of approximately equal size, and $\hat{Y}_i$ is the prediction obtained from all data aside from the subset containing the $i$th observation;

(D-iv) The coverage rates of the proposed 95% SCB for functions $\phi_2$ and $\phi_3$ (Coverage).

All these performance measures are computed based on 1000 replicates. Note that Criteria (B-i)–(B-vi) measure the frequency of getting the correct model structure; Criteria (C-i)–(C-iv) measure the frequency of getting an incorrect model structure; Criteria (D-i)–(D-iii) focus on the estimation...
and prediction accuracy for the model components; and Criterion (D-iv) measures the inferential performance.

The model selection results are provided in Tables 3.1 and 3.2, respectively. The SMILE can effectively identify informative linear and nonlinear components as well as correctly discover the linear and nonlinear structure in covariate \( X \), while the SAPLM neglects linear structure in \( X \) and SLM fails in presenting the nonlinear part of covariate \( X \). For the SMILE, the numbers of correctly selected nonzero covariates in \( Z \), linear, nonlinear, linear-and-nonlinear components in \( X \), nonzero covariates are very close to ORACLE (3 for corrZ, 1 for corrL, corrN and corrLN, \( p − 3 \) for corrZ0 and corrX0, respectively); and the numbers of incorrectly identified components approach to 0 as the sample size \( n \) increases, as shown in Table 3.2. The SMILE is close in the selection of covariates \( Z \) to the SAPLM estimator, and it overwhelms the SAPLM in identifying the linear-and-nonlinear structure of covariate \( X \). From the results in Tables 3.1 and 3.2, it is also evident that model misspecification leads to poor variable selection performance for the SLM. Especially for the selection of covariates in \( X \), which is our main focus for real data analysis, the SLM fails to select the right nonlinear components in each simulation.

The estimation and prediction results are displayed in Table 3.3. Specifically, we present the MSEs for linear coefficients \( \alpha_1, \alpha_2, \alpha_3 \) and \( \beta_1 \) and AMSEs for functions \( \phi_1, \phi_2 \) and \( \phi_3 \) and the CV-MSPEs for predicting \( Y \). The case with known active covariates (ORACLE) is also reported in each setting and serves as a gold standard. The SMILE performs the best in predicting \( Y \) and estimating the coefficients of covariates \( Z \), as indicated by the closest to ORACLE in CV-MSPE and MSEs of \( \alpha_1, \alpha_2 \) and \( \alpha_3 \) in most simulation settings, while the SLM is much higher (around \( 2 \sim 18 \) times higher). As for the linear structure in \( X \), as shown in MSE for \( \beta_1 \) and AMSE for \( \phi_1 \), the performance of SMILE is comparable to the SAPLM and the SLM, even though restricted
Table 3.1: Statistics (B-i)–(B-vi) comparing the SMILE, SAPLM and SLM.

<table>
<thead>
<tr>
<th>Size</th>
<th>Noise</th>
<th>p</th>
<th>Method</th>
<th>corrZ</th>
<th>corrZ0</th>
<th>corrL</th>
<th>corrN</th>
<th>corrLN</th>
<th>corrX0</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>0.5</td>
<td>1000</td>
<td>SMILE</td>
<td>100</td>
<td>99.99960</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.99940</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>98.6</td>
<td>99.99920</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>99.99850</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td>SMILE</td>
<td>100</td>
<td>99.99995</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.99985</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>97.3</td>
<td>99.99950</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>99.99915</td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td></td>
<td>SMILE</td>
<td>100</td>
<td>99.99996</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>96.63333</td>
<td>99.99988</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>99.99974</td>
</tr>
<tr>
<td>1.0</td>
<td>1000</td>
<td></td>
<td>SMILE</td>
<td>100</td>
<td>99.99920</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.99990</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>100</td>
<td>99.99920</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>96.56667</td>
<td>99.99799</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>99.99719</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td>SMILE</td>
<td>99.93333</td>
<td>99.99995</td>
<td>100</td>
<td>99.8</td>
<td>99.8</td>
<td>99.99975</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>100</td>
<td>99.99970</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>95.7</td>
<td>99.99975</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>99.99905</td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td></td>
<td>SMILE</td>
<td>99.86667</td>
<td>99.99996</td>
<td>100</td>
<td>99.5</td>
<td>99.5</td>
<td>99.99996</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>100</td>
<td>99.99990</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>93.73333</td>
<td>99.99982</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>99.99978</td>
</tr>
<tr>
<td>500</td>
<td>0.5</td>
<td>1000</td>
<td>SMILE</td>
<td>100</td>
<td>99.99990</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.99980</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>100</td>
<td>99.99990</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>99.99960</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td>SMILE</td>
<td>100</td>
<td>99.99995</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>100</td>
<td>99.99985</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>99.99985</td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td></td>
<td>SMILE</td>
<td>100</td>
<td>99.99996</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>99.96667</td>
<td>99.99994</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>99.99994</td>
</tr>
<tr>
<td>1.0</td>
<td>1000</td>
<td></td>
<td>SMILE</td>
<td>100</td>
<td>99.99950</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.99970</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>99.96667</td>
<td>99.99940</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>99.99930</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td>SMILE</td>
<td>100</td>
<td>99.99980</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.99990</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>100</td>
<td>99.99990</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>99.93333</td>
<td>99.99990</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>99.99960</td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td></td>
<td>SMILE</td>
<td>100</td>
<td>99.99994</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>99.76667</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>99.99994</td>
</tr>
</tbody>
</table>
Table 3.2: Statistics (C-i)–(C-iv) comparing the SMILE, SAPLM and SLM.

<table>
<thead>
<tr>
<th>Size</th>
<th>Noise</th>
<th>Method</th>
<th>Z Part</th>
<th>X Part</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>0.5</td>
<td>SMILE</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SAPLM</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SLM</td>
<td>1.4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMILE</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SAPLM</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SLM</td>
<td>2.7</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMILE</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SAPLM</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SLM</td>
<td>3.3667</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMILE</td>
<td>3.4333</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SAPLM</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SLM</td>
<td>4.3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMILE</td>
<td>0.1333</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SAPLM</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SLM</td>
<td>6.2667</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMILE</td>
<td>0.0667</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SAPLM</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SLM</td>
<td>4.3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMILE</td>
<td>0.0333</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SAPLM</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SLM</td>
<td>0.2333</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>
to the selection bias; as the sample size $n$ increases, the performance of the SMILE is perfect and matches it with ORACLE. Note that the SAPLM estimator is incapable in estimating $\beta_1$ in this case. The estimation of nonlinear functions $\phi_2$ and $\phi_3$ is also good for the SMILE, and matches with ORACLE as sample size $n$ increases. The inferior performance of the SAPLM and the poor performance of SLM, in both estimation and prediction, illustrates the importance and necessity of identifying correct model structure.

Next we investigate the coverage rates of the proposed SCB. For each replication, we test whether the true functions are covered by the SCB at the simulated values of the covariate in the interval $[-0.5 + h, 0.5 - h]$, where $h$ is the bandwidth. Table 3.4 shows the empirical coverage probabilities for a nominal 95% confidence level out of 500 replications. For comparison, we also provide the SCBs from the SAPLM and ORACLE estimator. From Table 3.4, we observe that coverage probabilities for the SMILE, SAPLM and ORACLE SCBs all approach the nominal levels as the sample size $n$ increases, which provides positive confirmation of Theorem 3.4. In most of the cases, the SMILE performs as well as or better than the SAPLM, and arrives at about the nominal coverage when $n = 500$ and $\sigma = 1.0$. Figure 3.1 depicts the true function $\phi_l$, the corresponding SMILE $\hat{\phi}_{SBLL}^l$ and the 95% SCB for $\phi_l$ based on $\hat{\phi}_{SBLL}^l$, for $l = 2, 3$, which are based on a typical run with $n = 500$, $p = 1000$ and $\sigma = 1.0$.

3.5 Application

We illustrate the application of our proposed method in the ultra-high-dimensional setting by using the SAM data generated by Leiboff et al. (2015). The maize SAM is a small pool of stem cells located in the plant shoot that generate all the above-ground tissues of maize plants. Leiboff et al. (2015) showed that SAM volume is correlated with a variety of agronomically important traits in
Table 3.3: Estimation results comparing the ORACLE, SMILE, SAPLM and SLM.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\sigma$</th>
<th>$p$</th>
<th>Method</th>
<th>MSE ($\times 10^{-2}$)</th>
<th>AMSE ($\times 10^{-2}$)</th>
<th>CV-</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>0.5</td>
<td>1000</td>
<td>ORACLE</td>
<td>0.47 0.48 0.50 1.06 0.09 0.98 0.83 0.28</td>
<td>9.13 8.86 25.99 18.65 1.63 253.82 178.85 4.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SMILE</td>
<td>0.47 0.48 0.50 1.06 0.11 0.94 0.77 0.28</td>
<td>10.51 8.70 41.05 21.29 1.84 252.75 180.40 4.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>0.47 0.47 0.53 1.08 0.19 0.95 0.73 0.27</td>
<td>1.00 1.00 2.00 2.00 1.00 1.00 1.00 1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>0.45 0.44 0.53 1.06 0.09 0.98 0.83 0.28</td>
<td>10.29 8.66 49.64 19.90 1.73 252.44 179.41 4.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ORACLE</td>
<td>0.47 0.48 0.50 1.06 0.09 0.98 0.83 0.28</td>
<td>9.13 8.86 25.99 18.65 1.63 253.82 178.85 4.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SMILE</td>
<td>0.47 0.48 0.50 1.06 0.11 0.94 0.77 0.28</td>
<td>10.51 8.70 41.05 21.29 1.84 252.75 180.40 4.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>0.47 0.47 0.53 1.08 0.19 0.95 0.73 0.27</td>
<td>1.00 1.00 2.00 2.00 1.00 1.00 1.00 1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>0.45 0.44 0.53 1.06 0.09 0.98 0.83 0.28</td>
<td>10.29 8.66 49.64 19.90 1.73 252.44 179.41 4.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ORACLE</td>
<td>0.47 0.48 0.50 1.06 0.09 0.98 0.83 0.28</td>
<td>9.13 8.86 25.99 18.65 1.63 253.82 178.85 4.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SMILE</td>
<td>0.47 0.48 0.50 1.06 0.11 0.94 0.77 0.28</td>
<td>10.51 8.70 41.05 21.29 1.84 252.75 180.40 4.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SAPLM</td>
<td>0.47 0.47 0.53 1.08 0.19 0.95 0.73 0.27</td>
<td>1.00 1.00 2.00 2.00 1.00 1.00 1.00 1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SLM</td>
<td>0.45 0.44 0.53 1.06 0.09 0.98 0.83 0.28</td>
<td>10.29 8.66 49.64 19.90 1.73 252.44 179.41 4.83</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3.4: Coverage rates comparing the ORACLE, SMILE and SAPLM.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\sigma$</th>
<th>$p$</th>
<th>$\phi_2$ Coverage (%)</th>
<th>$\phi_3$ Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>ORACLE</td>
<td>SMILE</td>
</tr>
<tr>
<td>300</td>
<td>0.5</td>
<td>1000</td>
<td>93.7</td>
<td>94.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>92.6</td>
<td>93.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5000</td>
<td>92.3</td>
<td>93.0</td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
<td>1000</td>
<td>96.0</td>
<td>95.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>95.4</td>
<td>95.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5000</td>
<td>95.1</td>
<td>95.6</td>
</tr>
<tr>
<td>500</td>
<td>0.5</td>
<td>1000</td>
<td>92.9</td>
<td>93.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>92.5</td>
<td>92.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5000</td>
<td>92.5</td>
<td>92.6</td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
<td>1000</td>
<td>97.1</td>
<td>96.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>95.2</td>
<td>95.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5000</td>
<td>94.7</td>
<td>95.1</td>
</tr>
</tbody>
</table>

Figure 3.1: Plots of the SMILE (dotted curve) and the 95% SCB (upper and lower dashed curves) of the nonparametric component $\phi_l(x_l)$, $l = 2, 3$ (solid curve).
adult plants. The goal of our analysis is to model and predict SAM volume as a function of single nucleotide polymorphism (SNP) genotypes and messenger RNA transcript abundance levels using data from maize inbred lines. Following the preprocessing steps described in Section B.5 in the Supplementary Materials in Li et al. (2018), linear sure independent screening (Fan and Lv, 2008) for SNP genotypes, and nonlinear independent screening (Fan et al., 2011) for RNA transcripts, the dataset we analyze consists of log-scale SAM volume measurements, binary SNP genotypes at \( p_1 = 5203 \) markers, and log-scale measures of abundance for \( p_2 = 1020 \) transcripts for each of \( n = 368 \) maize inbred lines.

Li et al. (2018) used the APLM to model the relationship between the log SAM volume response and predictors determined by SNP genotypes and RNA transcript abundance levels. Because the SNP genotypes are binary, they naturally entered the linear part of the APLM, and for convenience all the RNA transcripts were included in the nonlinear part of the APLM in Li et al. (2018). As discussed before, failing to account for exactly linear features makes the APLM less efficient statistically and computationally. In the following we apply our proposed SMILE method to distinguish among RNA transcripts entering the nonparametric and parametric parts of the APLM and to identify significant SNP genotypes and RNA transcripts simultaneously.

To compare the results of SMILE to the sparse APLM and the sparse linear regression model, we also analyze the data using the SAPLM and SLM estimators presented in Li et al. (2018). Parallel to the settings in Section 3.4, we use constant B-splines with four quantile knots for model structure identification, and use cubic B-splines with one quantile knot for nonlinear function approximation. We use the iterative algorithm proposed in Section 3.4.1 for penalty parameter selection and estimation.
As shown in Table 3.5, the SMILE identified 169 SNPs, 10 RNA transcripts linearly associated with log SAM size and 2 RNA transcripts that have nonlinear association with log SAM size. In contrast, the SAPLM selected 177 SNPs and 3 RNA transcripts, and the SLM selected 167 SNPs and 32 RNA transcripts. To evaluate the predictive performance of the two methods, we computed 10-fold cross-validation mean squared prediction error (CV-MSPE) for each method. The SMILE estimated nonlinear function for the selected nonlinear RNA transcript is plotted, along with 95% SCBs, in Figure 3.2.

Table 3.5: Selected SNPs and Transcripts by SMILE, SAPLM and SLM.

<table>
<thead>
<tr>
<th>RNA Transcripts Selected</th>
<th>SMILE</th>
<th>SAPLM</th>
<th>SLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{725}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$X_{127}, X_{136}, X_{141}, X_{208}, X_{289}, X_{312}, X_{493}, X_{749}, X_{855}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$X_{153}^<em>, X_{677}^</em>$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$X_{157}, X_{701}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$X_{209}, X_{314}, X_{320}, X_{342}, X_{419}, X_{472}, X_{489}, X_{553}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$X_{589}, X_{601}, X_{615}, X_{783}, X_{785}, X_{793}, X_{846}, X_{863}, X_{940}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>$X_{946}, X_{978}, X_{1002}, X_{1018}$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

| Number of SNP Genotypes | 169 | 177 | 167 |
| Number of Linear RNA Transcripts | 10 | 0 | 32 |
| Number of Functional RNA Transcripts | 2 | 3 | 0 |

| CV MSPE | 0.060 | 0.102 | 0.132 |
| CV Mean Number of SNPs | 153.9 | 175.9 | 83.1 |
| CV Mean Number of Linear Transcripts | 8.7 | 0 | 17.7 |
| CV Mean Number of Nonlinear Transcripts | 1.9 | 3.8 | 0 |

* nonlinear association identified by SMILE for $X_{153}$ and $X_{677}$

3.6 Discussion

This paper focuses on the simultaneous sparse model identification and learning for ultra-high-dimensional APLMs which strikes a delicate balance between the simplicity of the standard linear
regression models and the flexibility of the additive regression models. We proposed a two-stage penalization method, called SMILE, which can efficiently select nonzero components and identify the linear-and-nonlinear structure in the functional terms, as well as simultaneously estimate and make inference for both linear coefficients and nonlinear functions. First, we have devised a groupwise penalization method in the APLM for simultaneous variable selection and structure identification. After identifying important covariates and the functional forms for the selected covariates, we have further constructed SCBs for the nonzero nonparametric functions based on refined spline-backfitted local linear estimators. Our simulation studies and applications demonstrate the proposed SMILE procedure can be more efficient than the penalized linear regression and the penalized APLM without model identification, and can improve predictions.

Our work differs from previous works in practical, theoretical and computational aspects: (i) We perform variable selection and model structure identification simultaneously, for both the linear components in $Z$, and the linear and nonlinear forms for the components of $X$. In contrast, existing works either perform only model structure identification, or perform variable selection only for
components in $X$. (ii) Besides the consistency of model structure identification, we also provide inference tools for both the regression coefficients and the component functions. (iii) Compared to the local quadratic approximation approach used in Lian et al. (2015), which cannot provide exactly zero solutions and is inefficient for fitting large regression problems, our proposed iterative group coordinate descent algorithm takes advantage of sparsity in computation and is able to deal with the triple penalization problem very efficiently. (See Breheny and Huang (2015) for a detailed comparison of these two algorithms.) Our algorithm is easy to implement and can provide analysis results for large data sets with thousands of dimensions within seconds.

Our work deals with independent observations but can be extended to longitudinal data settings through marginal models or mixed-effects models. In addition, although we consider continuous response variables in our work, our approach can be readily extended to generalized additive partially linear models, to deal with different types of responses. Currently, the APLM assumes that the effects of all covariates are additive, which may overlook the potential interaction between covariates. Our method can be extended to models that can accommodate interactions between covariates, for example, APLMs with interaction terms. We leave such extensions to future work.

**Supplementary Materials**

The appendices in Sections B and B contain a study of the effects of the smoothing parameters on the proposed SMILE procedure, as well as technical assumptions, lemmas and the proofs.
Acknowledgements

This work was supported by the Iowa State University Plant Sciences Institute Scholars Program. In addition, Wang’s research was supported by NSF [grant DMS-1542332], and Nettleton’s research was supported by NSF [grant IOS-1238142].

Bibliography


CHAPTER 4. SPARSE LEARNING FOR IMAGE-ON-SCALAR REGRESSION WITH APPLICATION TO IMAGING GENETICS STUDIES

A paper for submission

Xinyi LI\textsuperscript{1}, Li WANG\textsuperscript{1}, GuanNan WANG\textsuperscript{2}, and the Alzheimer's Disease Neuroimaging Initiative

\textsuperscript{1}Department of Statistics, Iowa State University, Ames, IA 50011
\textsuperscript{2}Department of Mathematics, College of William & Mary, Williamsburg, VA 23185

Abstract

Motivated by recent advances in technology for medical imaging and high-throughput genotyping, we consider an imaging genetics approach to discover relationships between the interplay of genetic variation and environmental factors and measurements from imaging phenotypes. We propose an image-on-scalar regression method, in which the spatial heterogeneity of gene-environment interactions on imaging responses is investigated via an ultra-high-dimensional spatially varying coefficient model (SVCM). Bivariate splines on triangulations are used to represent the coefficient functions over an irregular two-dimensional (2D) domain of interest. When applying the image-on-scalar regression, a natural question raised in practice is if the coefficient function is really varying over space. In this paper, we present a unified approach for simultaneous sparse learning and model structure identification (i.e., varying and constant coefficients separation). Our method can identify zero, nonzero constant and spatially varying components correctly and efficiently. The es-
timators of constant coefficients and varying coefficient functions are consistent. The performance of the method is evaluated by a few simulation examples and a brain mapping study based on the Alzheimer’s Disease Neuroimaging Initiative (ADNI) data.

**Keywords:** Bivariate splines, Varying coefficient models, Penalized splines, Spatial Data, Triangulation.

### 4.1 Introduction

High-dimensional data occur very frequently and are especially common in biomedical studies including gene-environment-wide association studies, where one of the important scientific interests is the changes of genetic (G) effects under different environmental (E) conditions, or the Gene-environment (G×E) interaction. Identification of G×E interaction could shed novel insights into the phenotypic plasticity of complex disease phenotypes (Feinberg and Tycko, 2004).

Our study is motivated by some recent successful findings in imaging genetics studies facilitated by the Alzheimer’s Disease Neuroimaging Initiative (ADNI). Recent advances in both high throughput genotyping and brain imaging techniques have capacitated imaging genetics becoming an emerging discipline, which focuses on exploring the genetic influence on structural and functional imaging variations. Compared to traditional case-control designs, imaging genetics outshines in identifying underlying genes by employing imaging measures as phenotypes, such as the fludeoxyglucose positron emission tomography (FDG-PET) images.

Single nucleotide polymorphisms (SNPs) and other polymorphisms in several genes, including Apolipoprotein E (APOE), have been demonstrated to be related to neuroimaging measures in brain disorders, such as mild cognitive impairment (MCI) and Alzheimer’s Disease (AD) (e.g., Kim et al. (2009); Ashford and Mortimer (2002)). However, it is of great interest to identify other
genes that play a role in the development and progression of MCI and AD. In this paper, we handle all the predictors (e.g., environmental factors, genetic factors and their interactions) jointly when investigating the association between imaging responses and scalar predictors. We develop an efficient method for G×E interaction identification to address the high dimensionality of both the imaging and genomic data (Shen et al., 2010; Stein et al., 2010a,b). Although interaction selection has drawn much attention in the literature (Hao and Zhang, 2014; Kong et al., 2017; Li and Liu, 2018), effectively relating hundreds of thousands of predictors to large-scale imaging data remains a challenging task.

In this paper, we develop a novel statistical methodology that integrates the imaging data, environmental data and ultra-high dimensional genetic data in a principled functional regression and variable selection framework. To model the effect of G×E interaction, the varying coefficient model (VCM) introduced by Hastie and Tibshirani (1993) provides a flexible modeling approach. We are particularly interested in the spatially varying-coefficient model (SVCM) as it is powerful for modeling the nonstationarity of regression coefficients over space; see Zhu et al. (2014), Mu et al. (2018) and among others. Here the imaging response variables are associated with the scalar covariates through the functional linear regression, but the regression coefficients can vary from location to location and are modeled as a nonparametric function of spatial coordinates.

We consider imaging measures in a template from n subjects. Let Ω represent a two-dimensional domain, \( s = (s_1, s_2)^T \) denote a point in Ω and \( N_s \) equals the number of voxels in Ω. Without loss of generality, Ω is assumed to be a compact set in \( \mathbb{R}^2 \). For the ith subject, we observe an \( N_s \)-vector of imaging measures \( Y_i(\cdot) \in \Omega \). Let \( s_j = (s_{j1}, s_{j2})^T \) be the center of jth pixel.

For the ith subject, let \( G_i = (G_{i1}, \ldots, G_{ip_1})^T \) be the \( p_1 \)-dimensional genetic factor associated with the variation in response, i.e. SNPs, and \( E_i = (E_{i1}, \ldots, E_{ip_2})^T \) be the \( p_2 \)-dimensional environ-
mental factors (e.g., age, gender, height and medical treatments) that affect the imaging measures. Suppose that \( \{(G_i, E_i, Y_i, s_j)\}_{i=1}^n \) satisfies the following spatially varying coefficient model (SVCM)

\[
Y_i(s_j) = \mu_0(s_j) + G_i^T \mathbf{f}^G(s_j) + E_i^T \mathbf{f}^E(s_j) + (G \times E)_i^T \mathbf{f}^{G \times E}(s_j) + \eta_i(s_j) + \varepsilon_i(s_j), \quad i = 1, \ldots, n, \tag{4.1}
\]

where \( \mathbf{f}^G(\cdot), \mathbf{f}^E(\cdot) \) and \( \mathbf{f}^{G \times E}(\cdot) \) are \( p_1 \), \( p_2 \) and \( p_1p_2 \)-dimensional vectors of some unknown square-integrable bivariate functions, respectively; \( \eta_i \)'s are stochastic processes which characterize individual image variations, with mean zero and covariance function \( G_\eta(s, s') \); \( \varepsilon_i \)'s are random processes with mean zero, and each \( \varepsilon_i \) is independent of \( G_i, E_i \) and \( \eta_i \). We would like to see how the effect of each genetic factor changes under the influence of multiple environmental variables, the proposed high-dimensional SVCM (4.1) becomes a natural approach to consider.

As a flexible yet still interpretable extension of the linear model, VCM has been extensively studied in the statistics literature and widely used in practice. Recent works include Gu et al. (2014), and Huang et al. (2004), among many others. The VCM in the high-dimensional data settings have been studied in Wang et al. (2008), Wang and Xia (2009), Wei et al. (2011), Lian (2012), Xue and Qu (2012), Fan et al. (2014), Lian et al. (2015) and Chu et al. (2016). In the application of VCM to the \( G \times E \) interactions, Ma et al. (2015) also proposed a method for selecting relevant genes from a large number of candidates.

In a typical gene-environment-wide association study, the number of SNPs \( p_1 \) could be extremely large, which poses a substantial challenge for applying the above SVCM directly. Model selection for significant predictors is especially critical when the dimension of covariates is high and possibly exceeds the sample size, but the number of nonzero varying-coefficient components is relatively small. Another important statistical question in fitting SVCMs is whether the coefficient
function is really varying over space. In this paper, we are interested in identifying important risk factors related to the imaging responses and in the meantime aim to identify a more parsimonious semiparametric model structure that allows more efficient estimation. In this paper, we propose a unified framework to address the following questions: (i) how to identify those important main genetic and environmental factors and interactions; (ii) how to estimate the coefficient functions for these important variables; and (iii) how to separate nonzero constant and spatially varying components.

We propose a bivariate spline smoothing method to preserve important features (shape and/or smoothness) of imaging data. Two regularization operations are imposed to simultaneously identify the constant and spatially varying coefficient functions, removing insignificant predictors, and estimating the remaining coefficients and/or coefficient functions. We derive model selection consistency for the proposed method and show that it possesses the oracle property when the dimension of covariates exceeds the sample size.

The rest of the paper is organized as follows. Section 4.2 introduces the basic model setup and the proposed estimation procedure for the coefficient functions. Section 4.3 introduces the penalized bivariate spline procedure for selecting varying coefficient models when the dimension of covariates is high, and provides the theoretical properties for model selection consistency. Section 4.4 demonstrates the performance of the proposed method through some simulation studies. In Section 4.5, we present our empirical analysis of the ADNI data using the newly proposed procedure. All technical proofs are relegated to the Appendix.
4.2 Method

4.2.1 Model Setup

For the notation convenience, for the \(i\)th subject, we put all the genetic factors, environmental factors and their interaction terms together as a \(p\)-dimensional covariate vector \(X(i) = (X_{i1}, X_{i2}, \ldots, X_{ip})^T\), where \(X_{i1} = 1\) and \(p = 1 + p_1 + p_2 + p_1p_2\). Similarly, we collect all the coefficient functions \(f^G(\cdot), f^E(\cdot)\) and \(f^{G\times E}(\cdot)\) together as a \(p\)-dimensional vector of functions, denoted as \(f(\cdot)\).

In order to decide if \(f_k(s)\) is varying, for each \(k = 1, \ldots, p\), we can decompose \(f_k\) into a constant and a nonlinear part: \(f_k(s) = \alpha_k + \beta_k(s)\), where \(\beta_k(s)\) is some unknown smooth nonlinear function. For model identifiability, we assume that \(\int_{\Omega} \beta_k(s) dQ(s) = 0\), and this constraint ensures there is no constant effect in nonlinear function \(\beta_k(s)\). Then, we can express our model (4.1) as

\[
Y_i(s) = \sum_{k=1}^{p} X_{ik}\alpha_k + \sum_{k=1}^{p} X_{ik}\beta_k(s) + \eta_i(s) + \varepsilon_i(s), \quad i = 1, \ldots, n, \quad s \in \Omega.
\] 

In the following, we say \(X_k\) has a constant effect on the response if \(\alpha_k \neq 0\) and \(\beta_k(s) = 0\) for all \(s \in \Omega\), and \(X_k\) has a varying effect on the response if \(\beta_k(s)\) is not zero for some \(s\). Recall that our interest lies in selecting variables with nonzero varying and constant effects. In practice, some of the covariates may be irrelevant to the response variable, with the corresponding varying-coefficient functions being zero almost surely. So we identify the irrelevant covariates and estimate the nonzero coefficient functions for the relevant ones simultaneously.
Explicitly, we define the active constant, active varying and inactive index sets for $X$ shown as follows:

Active constant index set for $X$ : $\mathcal{A}_c = \{k = 1, \cdots, p : \alpha_k \neq 0, \beta_k(\cdot) \equiv 0\},$

Active varying index set for $X$ : $\mathcal{A}_v = \{k = 1, \cdots, p : \beta_k(\cdot) \neq 0\},$

Inactive index set for $X$ : $\mathcal{N} = \{k = 1, \cdots, p : \alpha_k \equiv 0, \beta_k(\cdot) \equiv 0\}.$ \hspace{1cm} (4.3)

Accordingly, we define the active index set for $X$ as $\mathcal{A} = \mathcal{A}_c \cup \mathcal{A}_v.$

4.2.2 Triangulations and Bivariate Spline Approximation

Our estimation method is based on penalized bivariate splines over triangulations. The idea is to approximate the function $\beta_k(\cdot)$ by bivariate splines which are piecewise polynomial functions over a 2D triangulated domain. We use this approximation to construct least squares estimators of the linear and nonlinear components of the model with a penalization term. In the following of this section, we describe the background of triangulations and introduce the penalized spline estimators.

4.2.2.1 Triangulations

Triangulation is an effective strategy to handle data distribution on irregular regions with complex boundaries and/or interior holes. It has attracted substantial recent attention in many applied areas, such as geo-spatial studies, numerical solutions of partial differential equations, image enhancements, and computer aided geometric design. See, for example, and the recent comprehensive books by Lai and Schumaker (2007).
We use $\tau$ to denote a triangle which is a convex hull of three points not located in one line. A collection $\Delta = \{\tau_1, ..., \tau_N\}$ of $N$ triangles is called a triangulation of $\Omega = \bigcup_{j=1}^{N} \tau_j$ provided that if a pair of triangles in $\Delta$ intersect, then their intersection is either a common vertex or a common edge. In general, any kind of polygon shapes can be used for the partition of $\Omega$. In this paper we restrict our attention to triangulations of $\Omega$ because any polygonal domain of arbitrary shape can be partitioned into finitely many triangles; see Ramsay (2002) for a triangulation of the island of Montreal as an example. In the following, we assume that all $s_j$'s are inside triangles of $\Delta$. That is, they are not on edges or vertices of triangles in $\Delta$. Otherwise, we can simply count them twice or multiple times if any observation is located on an edge or at a vertex of $\Delta$.

Given a triangle $\tau \in \Delta$, let $|\tau|$ be its longest edge length, and $\rho_\tau$ be the radius of the largest disk which can be inscribed in $\tau$. Define the shape parameter of $\tau$ as the ratio $\beta_\tau = |\tau|/\rho_\tau$. When $\beta_\tau$ is small, the triangles are relatively uniform in the sense that all angles of triangles in the triangulation $\tau$ are relatively the same. Denote the size of $\Delta$ by $|\Delta| := \max\{|\tau|, \tau \in \Delta\}$, i.e., the length of the longest edge of $\Delta$.

4.2.2.2 Bivariate spline estimators

For a nonnegative integer $r$, let $C^r(\Omega)$ be the collection of all $r$-th continuously differentiable functions over $\Omega$. Given a triangulation $\Delta$, let $S^r_d(\Delta) = \{s \in C^r(\Omega) : s|_\tau \in \mathbb{P}_d(\tau), \tau \in \Delta\}$ be a spline space of degree $d$ and smoothness $r$ over triangulation $\Delta$, where $s|_\tau$ is the polynomial piece of spline $s$ restricted on triangle $\tau$, and $\mathbb{P}_d$ is the space of all polynomials of degree less than or equal to $d$. 
For any $k = 1, \ldots, p$, let $\triangle_k$ be the triangulation for the $k$th component, and let $\{B_{k\ell}\}_{\ell \in J_k}$ be a set of bivariate Bernstein basis polynomials for $S^*_d(\triangle_k)$. Then we can write the function

$$\beta_k(s) \approx \sum_{\ell \in J_k} B_{k\ell}(s)c_{k\ell} = B_k(s)^\top c_k,$$

where $c_k = (c_{k\ell}, \ell \in J_k)^\top$ is the spline coefficient vector. To meet the smoothness requirement of the splines, we need to impose some constraints on the spline coefficients. The smoothness conditions are linear. Denote $H_k$ the constraint matrix on the coefficients $c_k$, which depends on $r_k$ and the structure of the triangulation and enforces smoothness across shared edges of triangles; see Zhou and Pan (2014). Put all smoothness conditions together to write $H_k c_k = 0$. Without loss of generality, we assume $B_1(s) = \cdots = B_p(s)$, and denote it as $B(s) = (B_\ell(s), \ell \in J)^\top$. Similarly, we assume $H = H_1 = \cdots = H_p$.

Then we can approximate the nonparametric function $\beta_k(s)$, $k = 1, \ldots, p$, using the normalized triangulation splines as

$$\beta_k(s) \approx \beta_{nk}(s) = B(s)^\top c_k,$$

where $c_k = (c_{k\ell}, \ell \in J_k)^\top$ is a vector of the spline coefficients.

Note that we normalize the spline basis functions because of the selection procedures, which we will develop and method in details in the next section. For estimation procedures, the standardization step is not necessary for spline basis functions.
4.3 Adaptive group lasso estimator

Therefore, the model selection problem for model (4.2) is equivalent to the problem of identifying \( A_c \) and \( A_v \). To achieve this, given \( \{(X_{(i)}, Y_{(i)}(s_j)) : i = 1, \ldots, n, j = 1, \ldots, N_s\} \), we propose to minimize

\[
\sum_{i=1}^{n} \sum_{j=1}^{N_s} \left\{ Y_{ij} - \sum_{k=1}^{p} X_{ik} \alpha_k - \sum_{k=1}^{p} X_{ik} \beta_k(s_j) \right\}^2 + \sum_{k=1}^{p} p_{\lambda n_1}(|\alpha_k|) + \sum_{k=1}^{p} p_{\lambda n_2}(\|\beta_k\|_2), \quad (4.5)
\]

where \( \|\beta_k\|_2^2 = \int_{\Omega} \beta_k^2(s) dQ(s) \), and \( p_{\lambda n_1}(\cdot) \) and \( p_{\lambda n_2}(\cdot) \) are penalty functions explained in detail in Section 4.3. The tuning parameters \( \lambda_{n1} \) and \( \lambda_{n2} \) decide the complexity of the selected model, and \( \lambda_{n1} \) and \( \lambda_{n2} \) go to \( \infty \) as \( n \) increases to \( \infty \).

We now approximate the varying coefficient functions by using the bivariate basis functions introduced in the Section 4.2.2. For any vector \( a \in \mathbb{R}^p \), denote \( \|a\|_2 = (\sum_{l=1}^{p} |a_l|^2)^{1/2} \) as the \( L_2 \) norm of \( a \). Let \( w_n = (w_{n1}, \ldots, w_{np})^\top \) be a given vector of weights, which needs to be chosen appropriately to achieve selection consistency. Let \( Y_{ij} = Y_i(s_j) \), then combining (4.4) and (4.5), we have

\[
L_n(\alpha, c; \lambda_{n1}, \lambda_{n2}) = \sum_{i=1}^{n} \sum_{j=1}^{N_s} \left\{ Y_{ij} - \sum_{k=1}^{p} X_{ik} \alpha_k - \sum_{k=1}^{p} X_{ik} B^\top(s_j)c_k \right\}^2 + \lambda_{n1} \sum_{k=1}^{p} w_{n,k}^c |\alpha_k| + \lambda_{n2} \sum_{k=1}^{p} w_{n,k}^c \|c_k\|_2, \quad \text{subject to } Hc_k = 0, \quad (4.6)
\]

where \( \lambda_{n1} \) and \( \lambda_{n2} \) is regularization parameters controlling the amount of shrinkage.

We first remove the constraint via QR decomposition of \( H^\top \), one has

\[
H^\top = QR = (Q_1 \; Q_2) \begin{pmatrix} R_1 \\ R_2 \end{pmatrix},
\]

where \( Q \) is an orthogonal matrix and \( R \) is an upper triangle matrix, the submatrix \( Q_1 \) is the first \( r \) columns of \( Q \), where \( r \) is the rank of matrix \( H \), and \( R_2 \) is a matrix of zeros.
Now the constrained optimization problem in (4.6) is reduced to the following non-constrained optimization problem

\[
L_n(\alpha, \gamma; \lambda_{n1}, \lambda_{n2}) = \sum_{i=1}^{n} \sum_{j=1}^{N} \left\{ Y_{ij} - \sum_{k=1}^{p} X_{ik}\alpha_k - \sum_{k=1}^{p} X_{ik}\{B^*(s_j)\}^\top \gamma_k \right\}^2 \\
+ \lambda_{n1} \sum_{k=1}^{p} w_{n,k}^\alpha |\alpha_k| + \lambda_{n2} \sum_{k=1}^{p} w_{n,k}^\gamma \|\gamma_k\|_2,
\]

(4.7)

where \( B^*(s) = Q_2^T B(s) \).

To ensure

\[
\int_{\Omega} \beta_k(s) dQ(s) = 0,
\]

we consider the normalized bivariate spline basis of \( B^*(s) = \{B_0^*(s), \ldots, B_{J_n}^*(s)\} \). That is, for any \( \ell \geq 2 \), define the centralized bivariate spline basis as

\[
B_0^\ell(\cdot) = B_0^*(\cdot) - \frac{\int_{\Omega} B_0^*(s) dQ(s)}{\int_{\Omega} B_0^*(s) dQ(s)} B_0^*(\cdot), \quad \ell = 1, \ldots, J_n,
\]

and rescale it by

\[
B_N^\ell(\cdot) = B_0^\ell(\cdot)/\|B_0^\ell(\cdot)\|_2, \quad \ell = 1, \ldots, J_n,
\]

where

\[
\|f\|_2^2 = \int_{\Omega} f^2(s) dQ(s)
\]

for any function \( f \). At the risk of abusing the notation, below we still use symbol \( B_\ell(\cdot) \) instead of \( B_N^\ell(\cdot) \) to denote the normalized bivariate spline basis to avoid creating too many new symbols.

Denote \( J_n \) the number of normalized basis functions, and \( B(s) = \{B_\ell(s), \ell = 1, \ldots, J_n\} \) the collect of all the normalized basis functions.
Minimizing 4.7, one obtains the estimator $\hat{\alpha}_k$ and the estimator of $\beta_k(\cdot)$ as $\hat{\beta}_k(s) = \hat{\gamma}_k^\top B^*(s)$, $k = 1, \ldots, p$. Therefore, one obtains the estimator of $f_k(\cdot)$ as follows:

$$\hat{f}_k(s) = \hat{\alpha}_k + B(s)\hat{\gamma}_k, \quad k = 1, \ldots, p.$$ 

Then, the model structure selected is defined by

$$\hat{A}_c = \{k : |\hat{\alpha}_k| \neq 0, \|\hat{\gamma}_k\| = 0, 1 \leq k \leq p\},$$

$$\hat{A}_v = \{k : \|\hat{\gamma}_k\| \neq 0, 1 \leq k \leq p\},$$

$$\hat{N} = \{k : |\hat{\alpha}_k| = 0, \|\hat{\gamma}_k\| = 0, 1 \leq k \leq p\}.$$  (4.8)

### 4.3.1 Estimation and Selection Consistency

This section studies the asymptotic properties of the proposed estimators. Throughout, we will use an index 0 to denote the true parameter values and functions in model (4.1). To discuss these properties, we introduce some notation of norms. For any function $g$ over the closure of domain $\Omega$, denote $\|g\|_{L^2(\Omega)}^2 = \int_{s \in \Omega} g^2(s)ds_1ds_2$ the regular $L_2$ norm of $g$, and $\|g\|_{\infty, \Omega} = \sup_{s \in \Omega} |g(s)|$ the supremum norm of $g$. For directions $s_j, j = 1, 2$, let $D^{q}_{s_j} g(s)$ denote the $q$-th order derivative in the direction $s_j$ at the point $s$. Let

$$|g|_{v, \infty, \Omega} = \max_{i+j=v} \|D^i_{s_1} D^j_{s_2} g(s)\|_{\infty, \Omega}$$

be the maximum norms of all the $v$th order derivatives of $g$ over $\Omega$. Let

$$\mathcal{W}^{\ell, \infty}(\Omega) = \{g : |g|_{\ell, \infty, \Omega} < \infty, 0 \leq k \leq \ell\}$$

be the standard Sobolev space.

In the following we first introduce some technical assumptions.
(A1) (Model) The numbers of nonzero components $|A_c|$ and $|A_v|$ are fixed; for any $k \in A_v$, the varying coefficient function $\beta_{0k} \in \mathcal{W}^{d+1,\infty}(\Omega)$ for an integer $d \geq 1$; there exist constants $c_\alpha > 0$ and $c_\beta > 0$ such that $\min_{k \in A_v} |\alpha_{0k}| \geq c_\alpha$, and $\min_{k \in A_v} \|\beta_{0k}\|_2 \geq c_\beta$.

(A2) (Covariates) For any $k = 1, \ldots, p$, there exists a positive constant $C_X$ such that $E|X_k|^8 \leq C_X$. The eigenvalues of $\Sigma_X = E(XX^\top)$ are bounded away from 0 and infinity.

(A3) (Errors) The $\epsilon_i(\cdot)$ is independent square-integrable random process, and $\epsilon_i$ is sub-gaussian with variance proxy $\sigma^2$, i.e., $E\{\exp(t^\top \epsilon_i)\} \leq \exp(\sigma^2/2)$, for any $\|t\|^2 = 1$ and some $\sigma > 0$.

Each component of $\{\eta(s) : s \in \Omega\}$, $\{(\eta(s)\eta(s'))^\top : (s, s') \in \Omega^2\}$ and $\{X\eta(s) : s \in \Omega\}$ are Donsker classes.

For any $k$, $\psi_k(s) \in \mathcal{C}^{(1)}(\Omega)$ and the variance function $0 < c_G \leq G_\eta(s, s) \leq C_G \leq \infty$ for any $s \in \Omega$.

(A4) (Resolution of images). If the location points $s$ are deterministic, we assume that

$$\sup_{s \in \Omega} |Q_{N_s}(s) - Q(s)| = O(N_s^{-1/2}),$$

where $Q_{N_s}(s) = N_s^{-1} \sum_{j=1}^{N_s} I(s_{j1} \leq s_1, s_{j2} \leq s_2)$ is the empirical cumulative distribution function and $Q(s)$ is a distribution with a positive continuous density. We also assume that

as $n \to \infty$, $N_s^{-1} n^{1/(d+1)+\delta} \to 0$ for some $\delta \in [0, 1]$.

(A5) (Triangulations) The triangulation $\triangle$ is $\pi$-quasi-uniform, that is, there exists a positive constant $\pi$ such that the triangulation $\triangle$ satisfies $|\triangle|/\rho_\tau \leq \pi$, for all $\tau \in \triangle$. 
(A6) (Initial estimators) The initial estimators satisfy $r_n \max_{k \not\in A_c} |\tilde{\alpha}_k| = O_p(1)$, $r_n \max_{k \not\in A_v} \|\tilde{\gamma}_k\|_2 = O_p(1)$, $r_n \to \infty$; and there exist positive constants $b_\alpha$ and $b_\gamma$ such that

$$\Pr \left( \min_{k \in A_c} |\tilde{\alpha}_k| \geq c_\alpha b_\alpha \right) \to 1, \quad \Pr \left( \min_{k \in A_v} \|\tilde{\gamma}_k\|_2 \geq c_\gamma b_\gamma \right) \to 1.$$  

(A7) (Parameters for consistency) Suppose that

$$\sqrt{n} N_2 \log(p) \lambda_{n1} + \sqrt{n} N_2 \log(pJ_n) \lambda_{n2} r_{n\gamma} + \frac{nN_s |\triangle| \lambda_{n1} r_{n\alpha} \lambda_{n2} r_{n\gamma}}{\lambda_{n1} r_{n\alpha} \lambda_{n2} r_{n\gamma}} = o(1),$$

$$\frac{\lambda_{n1}^2}{n^2 N_s^2} + \frac{\lambda_{n2}^2}{n^2 N_s^2} = o(1), \quad \frac{n}{J_n^{(\kappa+1)}} \frac{n}{\log(pJ_n)} = o(1).$$

Assumptions (A1) – (A3) are regularity conditions that are commonly used in the high-dimension literature. Assumptions (A4) – (A5) are regular conditions that are widely used in the triangulation bases literature. To obtain the selection consistency of the proposed method, we need an order requirement for a general initial estimator; see Assumption (A6). Theorem E.1 in the appendix demonstrates that the group LASSO estimator defined in (E.9) satisfies Assumption (A6) under some weak conditions; specifically, if $\lambda_{n1}^2 + \lambda_{n2}^2 \asymp nN_s^2 \log(p)$, then the consistent rates for the group LASSO estimator in (A5) have order $r_{n\alpha} \asymp r_{n\gamma} = O\left(\sqrt{n/\log(p)}\right)$. Consequently, the Assumption (A7) indicates that $p = \exp\{o(n)\}$.

The following theorems establish the asymptotic properties of the adaptive group LASSO estimators. We only state the main results here. The proofs are provided in the Appendix. Theorem 4.1 below shows that the proposed procedure is consistent in both variable selection and the separation of varying and constant coefficients.

**Theorem 4.1.** Suppose that Assumptions (A1) – (A7) hold. Then for $\hat{A}_c$, $\hat{A}_v$ in (4.8), as $n \to \infty$, $\Pr(\hat{A}_c = A_c) \to 1$, $\Pr(\hat{A}_v = A_v) \to 1$. 
Theorem 4.2. Suppose that Assumptions (A1) – (A7) hold. Then

$$\sum_{k \in A_c} |\tilde{\alpha}_k - \alpha_0k|^2 = O_p \left( \frac{J_n \log (J_n)}{n} \right) + O_p \left( |\Delta|^2 (\kappa + 1) \right) + O_p \left( \frac{\lambda_{n1}^2 + \lambda_{n2}^2}{n^2 N_s^2} \right),$$

$$\sum_{k \in A_v} \|\tilde{\gamma}_k - \gamma_0k\|^2 = O_p \left( \frac{J_n \log (J_n)}{n} \right) + O_p \left( |\Delta|^2 (\kappa + 1) \right) + O_p \left( \frac{\lambda_{n1}^2 + \lambda_{n2}^2}{n^2 N_s^2} \right).$$

4.4 Simulation Studies

We simulate data at all voxels on a 20 × 20 phantom image for n subjects. At each $s = (s_1, s_2)^T$ in $\Omega$, $Y_i(s)$ was simulated according to

$$Y_i(s_j) = G_i^T f^G(s_j) + E_i^T f^E(s_j) + (G \times E_i)^T f^{G \times E}(s_j) + \eta_i(s_j) + \varepsilon_i(s_j),$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, N_s$, where $G_i = (G_{i1}, \ldots, G_{i|p|})^T$, $f^G(s) = (f^G_1(s), \ldots, f^G_{|p|}(s))^T$, $E_i = (E_{i1}, \ldots, E_{i|p|})^T$, $f^E(s) = (f^E_1(s), \ldots, f^E_{|p|}(s))^T$, $(G \times E_i)^T = ((G \times E)_{i1}, \ldots, (G \times E)_{i|p|})^T$, $f^{G \times E}(s) = (f^{G \times E}_1(s), \ldots, f^{G \times E}_{|p|}(s))^T$.

We take $n = 50$, $p_1 = 100$, $p_2 = 3$, and $A_c^G = \{1\}$, $A_v^G = \{2\}$, $N_c^G = \{3, \ldots, p_1\}$, $A_c^E = \{1\}$, $A_v^E = \{2\}$, $N_c^E = \{3\}$, $A_c^{G \times E} = \{1\}$, $A_v^{G \times E} = \emptyset$, $N_c^{G \times E} = \{2, \ldots, p_1 p_2\}$. The nonzero functions $f^G_1(s)$, $f^G_2(s)$, $f^G_3(s)$, $f^G_4(s)$, $f^E_1(s)$, $f^{G \times E}(s)$, $f^{G \times E}_1(s)$, $f^{G \times E}_2(s)$ are linear, and the nonzero varying functions and errors are randomly generated from the Matérn processes with covariance function

$$C(||s - s'||; \sigma_s^2, \theta, \nu) = \sigma_s^2 \frac{1}{\Gamma(\nu)} \frac{1}{2^{\nu-1}} \left( \frac{\sqrt{2\nu}||s - s'||}{\theta} \right)^\nu F_\nu \left( \frac{\sqrt{2\nu}||s - s'||}{\theta} \right),$$

(4.9)

where $\Gamma$ is the gamma function, $F_\nu$ is the modified Neumann function, $\sigma_s^2$ is the sill and $\theta$ is the range parameter. The Matérn process is very flexible family of stationary processes which produce more realistic structures for biological applications, as compared to Brownian motion or simple low dimensional structures. In our simulation, the nonzero varying functions are generated with parameters ($\sigma_s = 1, \theta = 1/4, \nu = 5/2$) and the errors $\{\eta(s) : s \in \Omega\}$ are generated in the same way.
but with parameters \((\sigma_s = 1, \theta = 1/4, \nu = 3/2)\), which results in errors that are less smooth than the parameter functions. The errors \(\{\varepsilon(s) : s \in \Omega\}\) are generated in Gaussian process with mean zero and sigma taken as 1.5 and 1.25.

We generate \(E_{i2}\) independently from a Bernoulli distribution with success rate 0.5 and \(E_{i3}\) independently from the uniform distribution on \([1, 2]\). In this example, the signal-noise-ratio (SNR) is nearly 3 and 5, where the SNR is defined as the following

\[
\text{SNR} = \frac{N_s^{-1} \sum_{j=1}^{N_s} \text{Var}\{G^\top f^G(s_j) + E^\top f^E(s_j) + (G \times E)^\top f^{G \times E}(s_j)\}}{N_s^{-1} \sum_{j=1}^{N_s} \text{Var}\{\varepsilon_i(s_j)\}},
\]

As for the selection of penalty parameters, we consider two criteria widely used in high-dimensional settings, BIC and modified Bayesian information criteria (mBIC; see Lee et al. (2014)):

\[
\begin{align*}
\text{BIC}(\lambda) &= \log(\text{RSS}_\lambda) + df_\lambda \times \frac{\log(n)}{n}, \\
\text{mBIC}(\lambda) &= \log(\text{RSS}_\lambda) + df_\lambda \times \frac{\log\{(p + 1)J_0\} \times \log(n)}{2n},
\end{align*}
\]

where \(\text{RSS}_\lambda\) is the residual sum of squares associated with penalty parameters \(\lambda = (\lambda_1, \lambda_2)\) and \(df_\lambda\) is the number of estimated nonzero coefficients for the given \(\lambda\). The simulation results are similar based on these two criteria, so in the following, we choose \(\lambda_1\) by modified BIC and \(\lambda_2\) by mBIC for illustration using an approach described in the Algorithm 2.

To approximate the nonlinear varying coefficient functions, we use the normalized bivariate spline basis functions, with smoothness degree \(d = 2\), smoothness constraint \(r = 1\), and 29 triangles.

We evaluate the methods on the accuracy of variable selection, model identification and prediction. In detail, we adopt the following criteria for evaluation:

(B-i) Percent of correctly identified covariates in \(X\) with nonzero constant coefficients (“TC”);
(B-ii) Percent of correctly identified covariates in \(X\) with nonzero varying functions (“TV”);
(B-iii) Percent of correctly identified covariates in \(X\) with zero functions (“TN”).
Algorithm 2 Iterative group coordinate descent algorithm

Input: Data \( \{ (Y_i, X_{i1}, \ldots, X_{ip}, B_i) \}_{i=1}^n \)
\( \alpha_0^{(0)} \) and \( \gamma_0^{(0)} \): initial parameters of interest
\( \delta_0 \): convergence criterion

Output: \( \alpha \) and \( \gamma \): Estimates of \( \alpha \) and \( \gamma \)

while \( \| (\hat{\alpha}^{(m+1)} , \hat{\gamma}^{(m+1)})^\top - (\hat{\alpha}^{(m)} , \hat{\gamma}^{(m)})^\top \|^2 > \delta_0 \) do

(i) Given \( \hat{\alpha}^{(m+1)} \), obtain \( w_1^{(m+1)}, \ldots, w_p^{(m+1)} \) by minimizing objective function (E.9) with \( \tilde{\lambda}_n \) selected via mBIC;

(ii) Given \( \hat{\alpha}^{(m+1)} \) and \( w_1^{(m+1)}, \ldots, w_p^{(m+1)} \), obtain \( \hat{\gamma}^{(m+1)} \) by minimizing objective function (4.7) with \( \lambda_n \) selected via mBIC;

(iii) Given \( \hat{\gamma}^{(m)} \), obtain \( w_1^{c(m+1)}, \ldots, w_p^{c(m+1)} \) by minimizing objective function (E.9) with \( \tilde{\lambda}_n \) selected via the BIC;

(iv) Given \( \hat{\gamma}^{(m)} \) and \( w_1^{c(m+1)}, \ldots, w_p^{c(m+1)} \), obtain \( \hat{\alpha}^{(m+1)} \) by minimizing objective function (4.7) with \( \lambda_n \) selected via the BIC.

end

Set \( \alpha = \alpha^{(m+1)} \) and \( \gamma = \gamma^{(m+1)} \).

(C-i) Number of covariates in \( X \) with nonzero purely constant coefficients incorrectly identified as having nonlinear varying coefficient functions (“CtoV”);

(C-ii) Number of covariates in \( X \) with nonzero varying nonlinear functions incorrectly identified as having constant coefficients (“VtoC”);

(C-iii) Number of covariates in \( X \) with nonzero constant coefficients or nonzero nonlinear varying functions incorrectly identified as having both zero linear and zero nonlinear functions (“FN”);

(D-i) Mean squared errors (MSE) for linear coefficients \( \alpha_1 \);

(D-ii) Average MSE (AMSE) for \( \beta_2 \), defined as \( N^{-1}_s \sum_{j=1}^{N_s} (\tilde{\beta}_2(s_j) - \beta_2(s_j))^2 \);

All these performance measures are computed based on 100 replicates. Note that Criteria (B-i)–(B-iii) measure the frequency of getting the correct model structure; Criteria (C-i)–(C-iii) measure the frequency of getting an incorrect model structure; Criteria (D-i)–(D-ii) focus on the estimation and prediction accuracy for the model components.
The true surfaces and the estimated functions are shown in Figure 4.1. Visually one can see that the estimated functions successfully capture the spatial pattern of the true functions.

The model selection results are provided in Tables 4.1 and 4.2, respectively. The proposed method can effectively identify informative constant and varying components as well as correctly discover the constant and varying spatial structure. The percents of correctly selected nonzero constant and varying components, for genetic markers (G part), environmental factors (E part) and their interactive effects, are very close to perfect conditions (100%); and the numbers of incorrectly identified components are equal or very close to 0. There are no misspecification for the constant
and varying parts, and only a few out of 100 iterations incorrectly select more or less than the true level.

The estimation results are displayed in Table 4.3. Specifically, we present the MSEs for linear coefficients $\alpha_1^G, \alpha_1^E$ and $\alpha_1^{G\times E}$, and AMSEs for functions $\beta_1^G$ and $\beta_1^E$. All the MSEs of linear coefficients are in the $10^{-2}$ level, and all the AMSEs are in the $10^{-1}$ level, indicating the accuracy of the proposed method in estimation.

### Table 4.1: Statistics for Correct Selection

<table>
<thead>
<tr>
<th>Crit</th>
<th>n</th>
<th>SNR</th>
<th>$p_1$</th>
<th>G Part</th>
<th>E Part</th>
<th>G\times E Part</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TC</td>
<td>TV</td>
<td>TN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TC</td>
<td>TV</td>
<td>TN</td>
</tr>
<tr>
<td>mBIC</td>
<td>50</td>
<td>3</td>
<td>100</td>
<td>100</td>
<td>96</td>
<td>99.97</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td></td>
<td>100</td>
<td>97</td>
<td>99.98</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td></td>
<td></td>
<td>100</td>
<td>94</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td></td>
<td>100</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td></td>
<td></td>
<td>100</td>
<td>99.99</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td></td>
<td>100</td>
<td>100</td>
<td>99.99</td>
<td>100</td>
</tr>
<tr>
<td>EBIC</td>
<td>50</td>
<td>3</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.99</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td></td>
<td>100</td>
<td>100</td>
<td>99.99</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td></td>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td></td>
<td>100</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td></td>
<td></td>
<td>100</td>
<td>100</td>
<td>99.99</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.99</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td></td>
<td>100</td>
<td>99.99</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td></td>
<td></td>
<td>100</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td></td>
<td>100</td>
<td>99.99</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td></td>
<td></td>
<td>100</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td></td>
<td>100</td>
<td>100</td>
<td>99.99</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td></td>
<td></td>
<td>100</td>
<td>99.99</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td></td>
<td></td>
<td>100</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
### Table 4.2: Statistics for False Selection

<table>
<thead>
<tr>
<th>Crit</th>
<th>n</th>
<th>SNR</th>
<th>$p_1$</th>
<th>G Part</th>
<th>E Part</th>
<th>G × E Part</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CtoV</td>
<td>VtoC</td>
<td>FN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CtoV</td>
<td>VtoC</td>
<td>FN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CtoV</td>
<td>FN</td>
<td></td>
</tr>
<tr>
<td>mBIC</td>
<td>50</td>
<td>3</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>3</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EBIC</td>
<td>50</td>
<td>3</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>3</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 4.3: Statistics for Estimation

<table>
<thead>
<tr>
<th>Crit</th>
<th>n</th>
<th>SNR</th>
<th>$p_1$</th>
<th>G Part</th>
<th>E Part</th>
<th>G × E Part</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\alpha_1$</td>
<td>$\beta_2(\cdot)$</td>
<td>$\alpha_1$</td>
</tr>
<tr>
<td>mBIC</td>
<td>50</td>
<td>3</td>
<td>100</td>
<td>0.0130</td>
<td>0.0519</td>
<td>0.0418</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0.0161</td>
<td>0.0521</td>
<td>0.0642</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0.0188</td>
<td>0.0581</td>
<td>0.1921</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>100</td>
<td>0.0096</td>
<td>0.0330</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0.0097</td>
<td>0.0361</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0.0154</td>
<td>0.0365</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>3</td>
<td>100</td>
<td>0.0055</td>
<td>0.0208</td>
<td>0.0054</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0.0052</td>
<td>0.0218</td>
<td>0.0049</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0.0060</td>
<td>0.0204</td>
<td>0.0051</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>100</td>
<td>0.0041</td>
<td>0.0157</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0.0039</td>
<td>0.0164</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0.0045</td>
<td>0.0154</td>
</tr>
<tr>
<td>EBIC</td>
<td>50</td>
<td>3</td>
<td>100</td>
<td>0.0129</td>
<td>0.0440</td>
<td>0.0419</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0.0161</td>
<td>0.0452</td>
<td>0.0741</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0.0188</td>
<td>0.0440</td>
<td>0.1921</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>100</td>
<td>0.0096</td>
<td>0.0330</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0.0097</td>
<td>0.0337</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0.0154</td>
<td>0.0342</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>3</td>
<td>100</td>
<td>0.0055</td>
<td>0.0208</td>
<td>0.0060</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0.0052</td>
<td>0.0216</td>
<td>0.0049</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0.0060</td>
<td>0.0204</td>
<td>0.0051</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>100</td>
<td>0.0041</td>
<td>0.0157</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>200</td>
<td>0.0039</td>
<td>0.0165</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>500</td>
<td>0.0045</td>
<td>0.0154</td>
</tr>
</tbody>
</table>
Table 4.4: Distribution of patients by diagnosis status and gender

<table>
<thead>
<tr>
<th></th>
<th>CN</th>
<th>MCI</th>
<th>AD</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>68</td>
<td>111</td>
<td>67</td>
<td>246</td>
</tr>
<tr>
<td>Female</td>
<td>39</td>
<td>44</td>
<td>45</td>
<td>128</td>
</tr>
<tr>
<td>All</td>
<td>107</td>
<td>155</td>
<td>112</td>
<td>374</td>
</tr>
</tbody>
</table>

4.5 ADNI Data Analysis

We illustrate the application of our proposed method in the ultra-high-dimensional setting by using the FDG-PET data of ADNI, which is publicly available and accessible through http://adni.loni.usc.edu/.

This dataset is collected from 374 subjects; and for each subject, it includes a spatially normalized PET image with $79 \times 95$ pixels, demographical information including age and gender, cognitive impairment status, the epsilon 4 allele of APOE genotype, which is the strongest known genetic risk factor for AD, and other SNP genotypes at 620,901 loci. The original PET images are three dimensional, and we focus on the 48th horizontal section of the brain at the Corpus Callosum level from the bottom. As stated in Marcus et al. (2014), the 2D image in this section cuts through the frontal and parietal lobes, in which many areas are treated to be affected by AD with symptoms of dementia. The range of the PET measurements is -0.11 to 2.15, and the dataset contains the patients varying in age from 56 to 89 years old. As for the cognitive impairment status, out of 374 subjects, 107 have normal cognitive functions in the control group, 155 are diagnosed as mild cognitive impairment (MCI) and 112 are diagnosed as AD. Table 4.4 gives the distribution of patients by diagnosis status and gender.

We apply the proposed SVCM to capture important patient-level features that are associated with the variation of the PET images, identify the model structure (i.e., the features with varying
or constant coefficients), and detect the association. We consider demographical variables including age and gender, dummy variables of AD and MCI, genetic information such as SNPs, and their interaction as possible features in our analysis.

To be specific, we first perform a nonparametric independence screening in VCMs (Fan et al., 2014) to SNP genotypes to reduce the computational burden, using piecewise constant bivariate splines over a triangulation with 1106 triangles and 596 vertices. We select 40 SNPs that have the strongest associations with the image response conditional on age, gender, impairment status and APOE genotypes. Next, we apply our proposed method to model the relationship between the brain image response and the selected 40 SNP genotypes gene factors, along with age, gender, impairment status, APOE genotypes and their interactions. We use bivariate splines generated with degree $d = 2$ and smoothness parameter $r = 1$, and over a triangulation with 174 triangles and 28 vertices. To reduce the computational burden, we employ the divide-and-conquer algorithm (Chen and Xie, 2014) in the selection procedure. We split the patients into 11 different subsets, apply the proposed selection procedure to each subset, and select those features that appear more than 6 (i.e more than half number of subsets) times. Then we refit the model with those selected predictors, using piecewise constant bivariate splines over a triangulation with 1106 triangles and 596 vertices.

Table 4.5 lists the selected features and Figure 4.2 presents the estimated coefficient functions for these selected features. The intercept image shows the PET image from a male normal individual, excluding effects from other features. The main effect of AD on PET image is a decrease of 0.02–0.05 on the most area in the frontal lobe and parietal lobe. Compared to AD, the main effect of MCI has an increase of 0.1–0.15 in some spots in the frontal lobe near the longitudinal fissure. The main effect of age shows a decrease of 0.1 in every 20 years, in the frontal lobe and parietal
lobe near the longitudinal fissure. APOE2 has a different pattern with age and AD. There is an increase of 0.03–0.05 in the parietal lobe and frontal lobe near the longitudinal fissure. Most of the interaction terms with SNPs also show a decrease in the frontal lobe and/or parietal lobe, except for the interaction terms Sex by SNPs named “rs8182037” and “rs10955341”. These two terms show that female with high values B allele frequency in “rs8182037” and “rs10955341” have an increase of 0.04–0.05 in the frontal lobe and parietal lobe.

Table 4.5: Selected features for the ADNI data.

<table>
<thead>
<tr>
<th>×</th>
<th>MCI</th>
<th>AD</th>
<th>Age</th>
<th>Sex</th>
<th>APOE1</th>
<th>APOE2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>rs1445493</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>rs1541312</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>rs2131771</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>rs2955551</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs7556318</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs8182037</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>rs10935030</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs10955341</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Acknowledgements

Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.
Figure 4.2: Estimates of the coefficient functions of features for the ADNI data.
Bibliography


URL https://doi.org/10.1080/01621459.2017.1401541


CHAPTER 5. GENERAL CONCLUSIONS

This dissertation is focused on developing non/semi-parametric method from data with complex features, and the common part is that we consider an ultra-high-dimensional setting in theoretical development and applications to data consisting high-dimensional genetic data in all three papers.

The work in Chapter 2 focuses on the sparse ultra-high-dimensional APLM which strikes a delicate balance between the simplicity of the standard linear regression model and the flexibility of the additive regression model. Our method can be extended to longitudinal data settings through marginal models or mixed-effects models.

Our work differs from previous work both theoretically and practically. From a practical standpoint, we apply our methods in a genome-wide association study (GWAS) using an integrated dataset from genotyped SNPs and messenger RNA transcripts, in which the numbers of genotyped SNPs and RNA transcripts are both much larger than the observation number. The results show that we are able to detect significant SNPs and RNA transcripts efficiently. In addition, the proposed method allows us to further do inference for selected components, which can provide biological insights not previously achievable with existing methods. Theoretically, we consider an APLM that combines the flexibility of nonparametric regression with the parsimony of linear regression. Our approach differs from existing work in three major aspects. First, we consider selection for both the parametric and nonparametric parts of the model simultaneously, while most existing techniques focus on selection for either the parametric or nonparametric part. Second, to our best knowledge, existing work considering selection and estimation procedures for both linear
and nonlinear components does not allow the number nonlinear components to grow in an exponential order of sample size. We are the first to work on simultaneous variable selection and estimation for the APLM when both the numbers of linear and nonlinear components grow in exponential orders of sample size. The theoretical development for model selection and estimation for ultra-high-dimensional covariates in nonparametric components is much more challenging and completely different from the finite-dimension setting. Third, we obtain asymptotic normality for both the linear and nonparametric components, and importantly, we have obtained SCBs for the nonparametric components.

This paper leaves open a few research problems. In the SAM data analysis, we model the effects of categorical variables with linear components and the effects of continuous covariates with nonlinear components. Although this makes sense in the SAM application, an important question in using the APLM in practice is how to identify which continuous covariates should appear in the model linearly and which covariates should appear nonlinearly. For our SAM data analysis, the 148,793 genotyped SNPs are binary variables that naturally enter the APLM linearly. Thus, we need to simultaneously identify significant SNPs, determine which RNA transcripts enter the linear part of the APLM and identify significant RNA transcripts in the ultra-high-dimensional setting.

Subsequently, we propose the method in Chapter 3. This paper focuses on the simultaneous sparse model identification and learning for ultra-high-dimensional APLMs which strikes a delicate balance between the simplicity of the standard linear regression models and the flexibility of the additive regression models. We proposed a two-stage penalization method, called SMILE, which can efficiently select nonzero components and identify the linear-and-nonlinear structure in the functional terms, as well as simultaneously estimate and make inference for both linear coefficients and nonlinear functions. First, we have devised a groupwise penalization method in the APLM for
simultaneous variable selection and structure identification. After identifying important covariates and the functional forms for the selected covariates, we have further constructed SCBs for the nonzero nonparametric functions based on refined spline-backfitted local linear estimators. Our simulation studies and applications demonstrate the proposed SMILE procedure can be more efficient than the penalized linear regression and the penalized APLM without model identification, and can improve predictions.

Our work deals with independent observations but can be extended to longitudinal data settings through marginal models or mixed-effects models. In addition, although we consider continuous response variables in our work, or approach can be readily extended to generalized additive partially linear models, to deal with different types of responses. Currently, the APLM assumes that the effects of all covariates are additive, which may overlook the potential interaction between covariates. Our method can be extended to models that can accommodate interactions between covariates, for example, APLMs with interaction terms. We leave such extensions to future work.

In Chapter 4, we perform the sparse learning for image-on-scalar regression using SVCM for the spatial analysis of biomedical imaging data. We have developed a fast and accurate method for variable selection, model identification, and coefficient images functions estimation. Our method selects important environmental factors, genetic factors and their interaction, identifies their forms of coefficient functions (i.e., constant or varying over domain), and provides coefficient maps that highlight and visualize the association of brain region and the interested risk factors adjusted for the other patient-level features. It can handle the scenario where the dimension of the risk factors is much larger than the sample size, i.e., the ultra-high-dimensional case. In addition, it implements estimation for the coefficient function with various degrees and various smoothness over an arbitrary triangulation, so it can handle irregular shaped 2D objects.
In this paper, we mainly consider the selection, model identification and estimation problem for 2D objects, with application to one slice of 3D brain images. In the future, our methodology can be extended to 3D images to fully visualize the relationships between risk factors and brain diseases, and make full use of the potentiality in imaging genetics.
APPENDIX A. SUPPLEMENTARY FOR “ADDITIVE PARTIALLY LINEAR MODELS FOR ULTRA-HIGH-DIMENSIONAL GENETIC DATA IN MAIZE SHOOT APICAL MERISTEM STUDY”

Technical Details

Technical Assumptions

We state the regularity conditions needed for the theoretical results in the main paper. Let \( r \) be a nonnegative integer, and let \( v \in (0, 1] \) be such that \( d = r + v \geq 1 \). Let \( \mathcal{H}^{(d)} \) be the class of functions \( f \) on \([0, 1]\) whose \( r \)th derivative \( f^{(r)} \) exists and satisfies a Lipschitz condition of order \( v \):

\[
|f^{(r)}(s) - f^{(r)}(t)| \leq C_r |s - t|^v, \text{ for } s, t \in [0, 1].
\]

(A1) (Model) The numbers of nonzero components \( s_1 \) and \( s_2 \) are fixed, and there exist constants \( c_\beta > 0 \) and \( c_\alpha > 0 \) such that \( \min_{k \in S_z} |\beta_{0k}| \geq c_\beta, \min_{l \in S_x} \|\alpha_{0l}\|_2 \geq c_\alpha \).

(A2) (Errors) The errors \( \epsilon_1, \ldots, \epsilon_n \) are independent and identically distributed with \( \mathbb{E}(\epsilon_i) = 0 \), \( \text{Var}(\epsilon_i) = \sigma^2 \), \( \mathbb{E}|\epsilon_i|^{2+\delta} \leq M_\delta \) for some positive constant \( M_\delta (\delta > 0.5) \), and have \( b \)-sub-gaussian tails, i.e., \( \mathbb{E}\exp(t\epsilon) \leq \exp(b^2t^2/2) \), for any \( t \geq 0 \) and some \( b > 0 \).

(A3) (Nonlinear functions) \( \mathbb{E}\{\alpha_{0l}(X_l)\} = 0 \) and \( \alpha_{0l} \in \mathcal{H}^{(d)}, \ l \in S_x \).

(A4) (Covariates) Each \( Z_k \ (k = 1, \ldots, p_1) \) has a bounded second moment. Each \( X_l \ (l = 1, \ldots, p_2) \) has a continuous density and the marginal density function \( f_l \) is bounded and has continuous derivatives. The joint density function \( f(x_1, \ldots, x_{s_2}) \) is continuous and bounded below and above. In addition, the eigenvalues of \( \mathbb{E}\{(ZZ^\top) \mid X\} \) are bounded away from 0.
(A5) (Initial estimators) Suppose the initial estimators satisfy conditions that $r_{n1} \max_{k \in \mathcal{N}_x} |\tilde{\beta}_k| = O_p(1)$, $r_{n2} \max_{l \in \mathcal{S}_x} \{\gamma_l\}^2 = O_p(1)$, $r_{n1}, r_{n2} \to \infty$, and there exist positive constants $c_{b1}$ and $c_{b2}$ such that $\Pr(\min_{k \in \mathcal{S}_x} |\tilde{\beta}_k| \geq c_{b1} b_{n1}) \to 1$, $\Pr(\min_{l \in \mathcal{S}_x} \{\gamma_l\}^2 \geq c_{b2} b_{n2}) \to 1$, where $b_{n1} = \min_{k \in \mathcal{S}_x} |\beta_{0k}|$, and $b_{n2} = \min_{l \in \mathcal{S}_x} \|\alpha_{0l}\|^2$.

(A6) (Parameters and spline basis functions) Let $p_1$ and $p_2$ be the number of linear and nonlinear components, respectively. Suppose that the degree of the spline basis functions is at least $d$,

$$\frac{J_n}{n} + \frac{\lambda_{n3}^2 + \lambda_{n4}^2}{n^2} = o(1), \quad \sqrt[n]{n \log(p_1)} + \frac{\sqrt{nJ_n \log(p_2 J_n)}}{\lambda_{n3} r_{n1}} + \frac{n}{\lambda_{n3} r_{n1} J_n^d} + \frac{n}{\lambda_{n4} r_{n2} J_n^d} = o(1).$$

Next for any real numbers $a$ and $b$, let $a \lor b$ and $a \land b$ denote the maximum and minimum of $a$ and $b$, respectively. For any two sequences $\{a_n\}$, $\{b_n\}$, $n = 1, 2, \ldots$, we use $a_n \asymp b_n$ if there are constants $0 < c_1 < c_2 < \infty$ such that $c_1 < a_n/b_n < c_2$ for all $n$ sufficiently large.

Assumptions (A1) – (A4) are regularity conditions that are commonly used in the APLM literature. To obtain the selection consistency of the SBLL-AGLASSO, we need an order requirement for a general initial estimator; see Assumption (A5). Theorem A.1 in Section C demonstrates that the GLASSO estimator defined in (2.5) satisfies Assumption (A5) under some weak conditions, specifically if $\lambda_{n1}^2 + \lambda_{n2}^2 \asymp n\{\log(p_1) \lor J_n \log(p_2 J_n)\}$ and $J_n \asymp n^{1/(2d+1)}$ for $d \geq 1$, then the consistent rates for the GLASSO estimator in Assumption (A5) have an order $r_{n1} \asymp r_{n2} = O[n^{1/2}\{\log(p_1) \lor J_n \log(p_2 J_n)\}]^{-1/2}$. Consequently, Assumption (A6) is equivalent to:

$$\frac{\lambda_{n3}^2 + \lambda_{n4}^2}{n^2} + \frac{\log(p_1) \lor n^{1/2d+1} \log(p_2)}{\lambda_{n3} \land \lambda_{n4}} + \frac{n^{1/2d+1}\{\log(p_1) \lor n^{1/2d+1} \log(p_2)\}^{1/2}}{\lambda_{n3} \land \lambda_{n4}} = o(1), \quad (A.1)$$

If we take $\lambda_{n3} \asymp \lambda_{n4} = O(n^{1/2})$, then the equation in (A.1) indicates that $p_1 = \exp\{o(n^{1/2})\}$ and $p_2 = \exp\{o(n^{(2d-1)/(4d+2)})\}$.

We need the following additional assumptions in order to develop the asymptotic SCBs for the nonparametric components described in Section 2.4 of the main paper.

(A2′) The conditional distribution of $\epsilon$ given $(Z, X)$ is $N(0, \sigma^2)$. 
(B1) *(Conditions on the kernel function)* The kernel function \( K \in \text{Lip}([-1,1], C_K) \) for some constant \( C_K > 0 \), and is bounded, nonnegative, symmetric, and supported on \([-1,1]\) with the second moment \( \mu_2(K) = \int u^2 K(u) \, du \).

(B2) *(Conditions on bandwidth)* The bandwidth of the kernel \( K \) is \( h \sim n^{-1/5} \), i.e., \( c_h n^{-1/5} \leq h \leq C_h n^{-1/5} \) for some positive constants \( C_h, c_h \).

(B3) *(Conditions on spline basis functions)* The degree of the spline basis functions is at least \( d \), and the number of interior knots \( J_n \) satisfies

\[
\frac{n^{4/5}}{J_n^{2d-1}} + \frac{J_n(\lambda_{n3} + \lambda_{n4}) \log n}{n} + \frac{J_n^{1/2}(\lambda_{n3} + \lambda_{n4})}{n^{3/5}} = o(1). 
\]

If we take \( J_n \propto n^{4/(10d-5)} \log(n) \) for \( d \geq 1 \), \( \lambda_{n3} \propto \lambda_{n4} \propto n^{(6d-9)/(10d-5)}/\log(n) \), Assumption (B3) indicates the number of linear covariates \( p_1 = \exp[o\{n^{6d-9}/(10d-5)\}] \) and the number of nonlinear covariates \( p_2 = \exp[o\{n^{6d-9}/(10d-5)\}] \).

**Selection and estimation properties of the group LASSO estimators**

In this section, we consider the selection and estimation properties of the group LASSO estimator \( \tilde{\theta} \) in (2.5). Define

\[
\bar{S} = \{m : \|\tilde{\theta}_m\|_2 \neq 0, 1 \leq m \leq p_1 + p_2\}. \tag{A.2}
\]

**Theorem A.1.** Suppose that Assumptions (A1) – (A4) hold and that the degree of the spline basis functions is at least \( d \).

(i) If \( \{\log(p_1) \lor J_n \log(p_2 J_n)\}/n \to 0 \) and \( (\lambda_{n1}^2 + \lambda_{n2}^2)/n^2 \to 0 \) as \( n \to \infty \), then with probability converging to one, all the nonzero linear parameters \( \beta_{0k}, k \in S_z \), and nonzero additive components \( \alpha_{0l}, l \in S_x \), are selected.
(ii) In addition,
\[
\sum_{k=1}^{p_1} |\tilde{b}_k - b_{0k}|^2 = O_p \left( \frac{\log(p_1) \lor J_n \log(p_2 J_n)}{n} \right) + O \left( J_n^{-2d} \right) + O \left( \frac{\lambda_{n1}^2 + \lambda_{n2}^2}{n^2} \right),
\]
\[
\sum_{l=1}^{p_2} \|\tilde{a}_l - a_{0l}\|_2^2 = O_p \left( \frac{\log(p_1) \lor J_n \log(p_2 J_n)}{n} \right) + O \left( J_n^{-2d} \right) + O \left( \frac{\lambda_{n1}^2 + \lambda_{n2}^2}{n^2} \right).
\]

**A.0.1 Technical Lemmas**

For any random variable $X$, denote $\|X\|_p = (E|X|^p)^{1/p}$ as the $L_p$ norm for random variable $X$; and denote $\|X\|_\varphi = \inf \{ C > 0 : E\varphi(|X|/C) \leq 1 \}$ as the Orlicz norm for random variable $X$, where $\varphi$ is required to be a non-decreasing, convex function with $\varphi(0) = 0$.

**Lemma A.1.** Suppose that Assumptions (A2) and (A4) hold. Let
\[
T_{1k} = n^{-1/2} \sum_{i=1}^{n} Z_{ik} \epsilon_i, \ 1 \leq k \leq p_1; \ T_{2lj} = n^{-1/2} \sum_{i=1}^{n} B_{ij}(X_{il}) \epsilon_i, \ 1 \leq l \leq p_2, 1 \leq j \leq J_n;
\]
and $T_1 = \max_{1 \leq k \leq p_1} |T_{1k}|$, $T_2 = \max_{1 \leq l \leq p_2, 1 \leq j \leq J_n} |T_{2lj}|$. Then we have $E(T_1) \leq C_1 \log^{1/2}(p_1)$ and
\[
E(T_2) \leq C_2 n^{-1/2} \log^{1/2}(p_2 J_n) \left[ \{2C_3 n J_n \log(2p_2 J_n)\}^{1/2} + C_4 J_n^{1/2} \log(2p_2 J_n) + n \right]^{1/2},
\]
where $C_1$, $C_2$, $C_3$, and $C_4$ are positive constants. In particular, when $J_n \log(p_2 J_n)/n \to 0$, we have $E(T_1) = O\{\log^{1/2}(p_1)\}$, $E(T_2) = O\{\log^{1/2}(p_2 J_n)\}$.

**Proof.** Denote
\[
s_{1nk}^2 = \sum_{i=1}^{n} Z_{ik}^2, \ 1 \leq k \leq p_1; \ s_{2nlj}^2 = \sum_{i=1}^{n} B_{ij}(X_{il})^2, \ 1 \leq l \leq p_2, 1 \leq j \leq J_n.
\]
Next let $s_n^2 = \max_{1 \leq k \leq p_1} s_{1nk}^2$, $s_{2nlj}^2 = \max_{1 \leq l \leq p_2, 1 \leq j \leq J_n} s_{2nlj}^2$. By Assumption (A2), conditional on $Z = \{Z_{ik}, 1 \leq i \leq n, 1 \leq k \leq p_1\}$, $\sqrt{n} T_{1k}$ is $b(\sum_{i=1}^{n} Z_{ik}^2)^{1/2}$ - subgaussian; and conditional on $X = \{X_{il}, 1 \leq i \leq n, 1 \leq l \leq p_2\}$, $\sqrt{n} T_{2lj}$ is $b(\sum_{i=1}^{n} B_{ij}(X_{il})^2)^{1/2}$ - subgaussian.

Define $\varphi_p(x) = \exp(x^p) - 1$, $p \geq 1$. Then $\varphi_p^{-1}(m) = \{\log(1 + m)\}^{1/p}$. By Assumption (A2) and the maximal inequality for sub-Gaussian random variables (as stated in Lemmas 2.2.1 and 2.2.2 of
Van Der Vaart and Wellner (1996),

\[
E(T_1 \mid Z) = E\left( \max_{1 \leq k \leq p_1} |T_{1k}| \mid Z \right) = \left\| \max_{1 \leq k \leq p_1} |T_{1k}| \mid Z \right\|_1 \leq \left\| \max_{1 \leq k \leq p_1} |T_{1k}| \mid Z \right\|_{\varphi_1} \\
\leq \left\| \max_{1 \leq k \leq p_1} |T_{1k}| \mid Z \right\|_{\varphi_2} \times (\log 2)^{1/2} \\
\leq K_1 \{\log 2 \log(1 + p_1)\}^{1/2} n^{-1/2} \max_{1 \leq k \leq p_1} \|\sqrt{n} T_{1k} \mid \{Z_{ik}, 1 \leq i \leq n, 1 \leq k \leq p_1\}\|_{\varphi_2} \\
\leq K_1 \{\log 2 \log(1 + p_1)\}^{1/2} n^{-1/2} \times \max_{1 \leq k \leq p_1} \left( 6\delta^2 \sum_{i=1}^{n} Z_{ik}^2 \right)^{1/2} \\
\leq C_5 n^{-1/2} s_1 n (\log p_1)^{1/2}.
\]

Next,

\[
E(T_2 \mid X) = E\left( \max_{1 \leq l \leq p_2, 1 \leq j \leq J_n} |T_{2lj}| \mid X \right) = \left\| \max_{1 \leq l \leq p_2, 1 \leq j \leq J_n} |T_{2lj}| \mid X \right\|_1 \\
\leq \left\| \max_{1 \leq l \leq p_2, 1 \leq j \leq J_n} |T_{2lj}| \mid X \right\|_{\varphi_1} \leq \left\| \max_{1 \leq l \leq p_2, 1 \leq j \leq J_n} |T_{2lj}| \mid X \right\|_{\varphi_2} \times (\log 2)^{1/2} \\
\leq K_2 \{\log 2 \log(1 + p_2 J_n)\}^{1/2} n^{-1/2} \max_{1 \leq l \leq p_2, 1 \leq j \leq J_n} \|\sqrt{n} T_{2lj} \mid X\|_{\varphi_2} \\
\leq K_2 \{\log 2 \log(1 + p_2 J_n)\}^{1/2} n^{-1/2} \times \max_{1 \leq l \leq p_2, 1 \leq j \leq J_n} \left\{ 6\delta^2 \sum_{i=1}^{n} B_{lj}^2 (X_{il}) \right\}^{1/2} \\
= C_{21} n^{-1/2} s_2 n \{\log(p_2 J_n)\}^{1/2}.
\]

Thus,

\[
E(T_1) = E \{E(T_1 \mid Z)\} \leq C_{11} n^{-1/2} (\log p_1)^{1/2} E(s_1 n), \\
E(T_2) = E \{E(T_2 \mid X)\} \leq C_{21} n^{-1/2} \{\log(p_2 J_n)\}^{1/2} E(s_2 n),
\]

where \(K_1, K_2, C_{11}\) and \(C_{21}\) are positive constants. Therefore, by Assumption (A4), we have

\[
E(Z_{ik})^2 \leq C_{13}^2.
\]
The properties of normalized B-splines imply that, for every \(l, j\), there exist positive constants \(C_{13}\) and \(C_4\), such that

\[|B_{lj}(X_{il})| \leq C_{13} J_n^{1/2} \text{ and } E\{B_{lj}(X_{il})\}^2 = 1.\]

Therefore,

\[E(s_{1n}^2) = \max_{1 \leq k \leq p_1} E(s_{1nk}^2) = \max_{1 \leq k \leq p_1} \sum_{i=1}^n E(Z_{ik}^2) \leq n C_{13}^2,\]

\[\sum_{i=1}^n \left| B_{lj}(X_{il}) - E\{B_{lj}(X_{il})\} \right|^2 \leq n J_n C_3.\]

Thus, by Lemma A.1 of Van de Geer (2008), we have

\[E\left[\max_{1 \leq l \leq p_2, 1 \leq j \leq J_n} \left\{ \sum_{i=1}^n \left( B_{lj}(X_{il}) - E B_{lj}(X_{il}) \right) \right\}\right] \leq \{2 C_{3} n J_n \log(2 p_2 J_n)\}^{1/2} + C_4 J_n^{1/2} \log(2 p_2 J_n).\]

Therefore, by triangle inequality,

\[E(s_{2n}^2) \leq \{2 C_{3} n J_n \log(2 p_2 J_n)\}^{1/2} + C_4 J_n^{1/2} \log(2 p_2 J_n) + n.\]

Thus, \(E(s_{1n}) \leq \{E(s_{1n}^2)\}^{1/2} \leq (C_{13} n)^{1/2}\), and

\[E(s_{2n}) \leq \{E(s_{2n}^2)\}^{1/2} \leq \{2 C_{3} n J_n \log(2 p_2 J_n)\}^{1/2} + C_4 J_n^{1/2} \log(2 p_2 J_n) + n\]^{1/2}.

The lemma follows.

**Lemma A.2.** Let \(J_n = O(n^\gamma)\), where \(0 < \gamma < 0.5\). Suppose that \(|\mathcal{A}|\) is bounded by a fixed constant independent of \(n\), \(p_1\) and \(p_2\). Then under Assumption (A4), with probability approaching one as \(n \to \infty\),

\[c_1 \leq \pi_{\min}(C_{\mathcal{A}}) = \min_{b^\top b = 1} b^\top C_{\mathcal{A}} b \leq \pi_{\max}(C_{\mathcal{A}}) \leq c_2,\]

where \(c_1\) and \(c_2\) are two positive constants.

**Proof.** For index set \(\mathcal{A} \subseteq \{1, \ldots, p_1 + p_2\}\), let \(t_1 = |\mathcal{A} \cap \mathcal{S}_z|\) and \(t_2 = |\mathcal{A} \cap \mathcal{S}_z|\). Let vector \(b \in \mathbb{R}^{t_1 + t_2 J_n}\).

By Lemma A.7 of Ma and Yang (2011) and Lemma 3 of Stone (1985), with probability approaching one as \(n \to \infty\),

\[c_1 I_{t_1 + t_2 J_n} \leq C_{\mathcal{A}} \leq c_2 I_{t_1 + t_2 J_n}\]

for certain constants \(c_1\) and \(c_2\). Then the equality follows since the minimum and maximum eigenvalues of \(C_{\mathcal{A}}\) are

\[\pi_{\min}(C_{\mathcal{A}}) = \min_{b^\top b = 1} b^\top C_{\mathcal{A}} b\]

and

\[\pi_{\max}(C_{\mathcal{A}}) = \max_{b^\top b = 1} b^\top C_{\mathcal{A}} b.\]

\(\square\)
Remark A.1. Note that Stone’s Lemma is applied here to obtain the lower bound of $C_A$, which requires $|A|$ to be bounded by a fixed constant. This restriction explains why we assume that the number of nonzero nonparametric components is fixed.

Lemma A.3. Under Assumption (A3), if the degree of the spline basis functions is at least $d$, then there exists a vector $\gamma_0 = (\gamma_{01}^\top, \ldots, \gamma_{0p_2}^\top)^\top$, such that $\|\gamma_{0l}\|_2 \neq 0$, for $l \in S_x$, $\|\gamma_{0l}\|_2 = 0$, $l \in N_x$ and $\|\alpha_{0l} - B^\top_l \gamma_{0l}\|_2 = O(J_n^{-d})$.

Proof. According to the best approximation result in de Boor (2001), for any function $g(\cdot) \in \mathcal{H}^{(d)}$, there exists a polynomial spline function $g_n(\cdot)$ such that $\|g - g_n\|_\infty \leq C_d \|g^{(d)}\|_\infty J_n^{-d}$. The lemma follows directly from the above result. \hfill \Box

Lemma A.4. Suppose that Assumptions (A1) – (A4) hold and that the degree of the spline basis functions is at least $d$. Recall the definition of $S$ and $\tilde{S}$ in (2.6) and (A.2), with probability approaching one, $|\tilde{S}| \leq M_1 |S| = M_1(s_1 + s_2)$ for a finite constant $M_1 > 1$.

Proof. The proof generally follows the proofs of Theorem 1 of Zhang and Huang (2008) and Theorem 2.1 of Wei and Huang (2010). The main differences are the data structure and error term shown in (A.7).

By Lemma A.3, for some constant $C_4 > 0$,

$$\|\delta\|_2 \leq C_4 (ns_2 J_n^{-2d})^{1/2} = C_4 s_2^{1/2} n^{1/2} J_n^{-d}.$$ 

For any positive integers $q_1, q_2$, pick some index sets $A_1 \subseteq \{1, \ldots, p_1\}$ and $A_2 \subseteq \{1, \ldots, p_2\}$ such that the cardinality of $A_1$ is $|A_1| = q_1$ and $|A_2| = q_2$. 
Denote \( A = A_1 \cup A_2 \).

Define an \((q_1 + q_2 J_n) \times 1\) vector

\[
S_A = \left( \lambda_{n1}s_{q_1}^T, \lambda_{n2}\sqrt{J_n}U_1^{A^T}, \ldots, \lambda_{n2}\sqrt{J_n}U_{q_2}^{A^T} \right)^T
\]

where \( s_{q_1} \in \{ \pm 1 \}^{q_1} \), and \( U_k^A \) in a unit ball with dimension \( J_n \), that is, \( U_k^A \in \mathbb{R}^{J_n} \) and \( \| U_k^A \|_2 = 1 \), \( k = 1, \ldots, q_2 \). Let \( P_A = D_A (D_A^T D_A)^{-1} D_A^T \) be the projection matrix of \( D_A \). Define

\[
\chi_{q_1, q_2} = \max_{A = A_1 \cup A_2} \max_{s_{q_1} \in \{ \pm 1 \}^{q_1}} \frac{\| \eta^T (D_A^T D_A)^{-1} S_A - (I - P_A)D\theta_0 \|_2}{\| D_A (D_A^T D_A)^{-1} S_A - (I - P_A)D\theta_0 \|_2},
\]

\[
\Omega_{s_1, s_2} = \left\{ (D, \eta) : \chi_{q_1, q_2} \leq \sigma C_2 \{ q_1 \log(p_1) \vee q_2 J_n \log(p_2 J_n) \}^{1/2}, \forall q_1 \geq s_1, q_2 \geq s_2 \right\},
\]

where \( C_2 > 0 \) is a sufficiently large constant. As shown in the proof of Theorem 1 of Zhang and Huang (2008) and Theorem 2.1 of Wei and Huang (2010), there exists a constant \( M_1 > 1 \), such that if \( (D, \eta) \in \Omega_{s_1, s_2} \), then \( |\tilde{S}| \leq M_1 |S| = M_1 (s_1 + s_2) \).

So it suffices to show that \((D, \eta) \in \Omega_{s_1, s_2} \). Denote \( V_{\cdot} = D_A (D_A^T D_A)^{-1} S_A - (I - P_A)D\theta_0 \). Then by the triangle and Cauchy-Schwarz inequalities,

\[
\frac{\| \eta^T V_{\cdot} \|}{\| V_{\cdot} \|_2} \leq \frac{\| \epsilon^T V_{\cdot} + \delta^T V_{\cdot} \|}{\| V_{\cdot} \|_2} \leq \frac{\| \epsilon^T V_{\cdot} \|}{\| V_{\cdot} \|_2} + \| \delta \|_2.
\]

For the \( \| \epsilon^T V_{\cdot} \|/\| V_{\cdot} \|_2 \) part, define

\[
\chi_{q_1, q_2}^* = \max_{A = A_1 \cup A_2} \max_{s_{q_1} \in \{ \pm 1 \}^{q_1}} \frac{\| \epsilon^T V_{\cdot} \|}{\| V_{\cdot} \|_2},
\]

\[
\Omega_{s_1, s_2}^* = \left\{ (D, \epsilon) : \chi_{q_1, q_2}^* \leq \sigma C_3 \{ q_1 \log(p_1) \vee q_2 J_n \log(p_2 J_n) \}^{1/2}, \forall q_1 \geq s_1, q_2 \geq s_2 \right\},
\]

where \( C_3 > 0 \) is a sufficiently large constant. As shown in the proof of Theorem 1 of Zhang and Huang (2008) and Theorem 2.1 of Wei and Huang (2010), \( P(\Omega_{s_1, s_2}^*) \rightarrow 1 \). For the \( \| \delta \|_2 \) part, for \( n \) sufficiently large and \( J_n = O(n^{1/(2d+1)}) \), \( \| \delta \|_2 \leq C_4 s_2^{1/2} n^{1/2} J_n^{-d} \leq \sigma C_5 \{ q_1 \log(p_1) \vee q_2 J_n \log(p_2 J_n) \}^{1/2} \). It follows that \( P(\Omega_{s_1, s_2}^*) \rightarrow 1 \). This completes the proof. \( \square \)
A.0.2 Proof of Theorem A.1

We prove part (ii) first. Let \( \theta^\top = (\theta_1^\top, \ldots, \theta_{p_1+p_2}^\top) = (\beta_1, \ldots, \beta_{p_1}, \gamma_1^\top, \ldots, \gamma_{p_2}^\top) \). For \( S \) defined in (2.6) and \( \tilde{S} \) defined in (A.2), denote \( S' = S \cup \tilde{S} = \{ m : \| \theta_{0m} \|_2 \neq 0 \text{ or } \| \tilde{\theta}_m \|_2 \neq 0 \} \) and \( d' = |S'| \).

By Lemma A.4, \( d' = O(s_1 + s_2) \). Notice that \( D\theta = D_{S'} \tilde{\theta}_{S'} \) and \( D\theta_0 = D_{S'} \theta_{0,S'} \), by the definition of \( \tilde{\theta} \) and \( S' \),

\[
\| Y - D_{S'} \tilde{\theta}_{S'} \|_2^2 + \sum_{m \in S'} \{ \lambda_{n1} I(m \leq p_1) + \lambda_{n2} I(m > p_1) \} \| \theta_m \|_2^2 \\
\leq \| Y - D_{S'} \theta_{0,S'} \|_2^2 + \sum_{m \in S'} \{ \lambda_{n1} I(m \leq p_1) + \lambda_{n2} I(m > p_1) \} \| \theta_{0m} \|_2^2.
\]

Recall that \( \eta = Y - D\theta_0 \) and let \( \nu = D_{S'}(\tilde{\theta}_{S'} - \theta_{0,S'}) \), so \( \eta - \nu = Y - D_{S'} \tilde{\theta}_{S'} \), and we have

\[
\| Y - D_{S'} \tilde{\theta}_{S'} \|_2^2 - \| Y - D_{S'} \theta_{0,S'} \|_2^2 = \nu^\top \nu - 2\eta^\top \nu.
\]

Thus, from the triangle inequality and the Cauchy-Schwarz inequality,

\[
\| \nu \|_2^2 - 2\eta^\top \nu \leq \sum_{m \in S'} \{ \lambda_{n1} I(m \leq p_1) + \lambda_{n2} I(m > p_1) \} (\| \theta_{0m} \|_2 - \| \tilde{\theta}_m \|_2) \\
\leq \{ d'(\lambda_{n1}^2 + \lambda_{n2}^2) \}^{1/2} \| \tilde{\theta}_{S'} - \theta_{0,S'} \|_2 \\
\leq d'(\lambda_{n1}^2 + \lambda_{n2}^2)/(nc_*) + nc_* \| \tilde{\theta}_{S'} - \theta_{0,S'} \|_2^2/4,
\]

where \( c_* \) is the lower bound of smallest eigenvalue of \( n^{-1}D_{S'}^\top D_{S'} \).
By Lemma A.2 and Lemma A.4, \( c_s \asymp 1 \) with probability approaching one.

Apparently,

\[
\|\nu\|_2^2 \geq n c_s \|\theta_{0,S'} - \theta_{0,S'}\|_2^2. \tag{A.4}
\]

Define \( \eta^* \equiv D_{S'}(D_{S'}^\top D_{S'})^{-1}D_{S'}^\top \eta \) to be the projection of \( \eta \) onto the column space of \( D_{S'} \). Obviously, \( \eta^\top \nu = \eta^* \top \nu \). By the Cauchy-Schwarz inequality, we have

\[
2|\eta^\top \nu| \leq 2\|\eta^*\|_2 \|\nu\|_2 \leq 2\|\eta^*\|_2^2 + \|\nu\|_2^2/2. \tag{A.5}
\]

Combining (A.3), (A.4) and (A.5), we obtain

\[
\|\bar{\theta}_{S'} - \theta_{0,S'}\|_2^2 \leq 8\|\eta^*\|_2^2/(nc_s) + 4d'(\lambda_{n1}^2 + \lambda_{n2}^2)/(n^2c_s^2). \tag{A.6}
\]

With \( \eta_i \) defined to be the \( i \)th element of \( \eta \), we have the following decomposition:

\[
\eta_i = Y_i - \sum_{k=1}^{p_1} Z_{i_k}^\top \beta_{0k} - \sum_{l=1}^{s_2} \sum_{j=1}^{J_n} \gamma_{0lj} B_{lj}(X_{il})
\]

\[
= Y_i - \sum_{k=1}^{s_1} Z_{i_k}^\top \beta_{0k} - \sum_{l=1}^{s_2} \sum_{j=1}^{J_n} \alpha_{0l}(X_{il}) + \sum_{l=1}^{s_2} \sum_{j=1}^{J_n} \gamma_{0lj} B_{lj}(X_{il})
\]

\[
= \epsilon_i + \sum_{l=1}^{s_2} \delta_{il}, \tag{A.7}
\]

where \( \delta_{il} = \alpha_{0l}(X_{il}) - \sum_{j=1}^{J_n} \gamma_{0lj} B_{lj}(X_{il}) \). Let \( \delta_i = \sum_{l=1}^{s_2} \delta_{il} \), and \( \delta = (\delta_1, \ldots, \delta_n)^\top \). Recall that \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top \). Then \( \eta = \epsilon + \delta \). Define \( \delta_{S'} = (\sum_{l+p_1 \in S',1 \leq l \leq p_2} \delta_{il}, i = 1, \ldots, n)^\top \). By (A.7) and the fact that \( |\delta_{il}| = O_p(J_n^{-d}) \),

\[
\|\eta^*\|_2^2 = \|\epsilon^* + \delta_{S'}\|_2^2 \leq 2\|\epsilon^*\|_2^2 + 2\|\delta_{S'}\|_2^2 \leq 2\|\epsilon^*\|_2^2 + O_p(nd'^2 J_n^{-2d}), \tag{A.8}
\]

where \( \epsilon^* \equiv P_{D_{S'}} \epsilon \) and \( \delta_{S'} \equiv P_{D_{S'}} \delta_{S'} \) are the projections of \( \epsilon \) and \( \delta_{S'} \) onto the column space of \( D_{S'} \), respectively. Observe that \( \|\epsilon^*\|_2^2 = \|D_{S'}^\top D_{S'}\|^{-1/2} \|D_{S'}^\top \epsilon\|_2^2 \leq \|D_{S'}^\top \epsilon\|_2^2/(nc_s) \), and \( \max_{A:|A| \leq d'} \|D_A^\top \epsilon\|_2^2 = \max_{A:|A| \leq d'} \sum_{m \in A} \|D_m^\top \epsilon\|_2^2 \leq nd'(T_1^2 + J_n T_2^2) \), where \( T_1 \) and \( T_2 \) are defined in A.1.
According to Lemma A.1, \( \max_{A:|A| \leq d} \| D_A^* \epsilon \|_2^2 = O_p(n d' \{ \log(p_1) \lor J_n \log(p_2 J_n) \}) \). Therefore, we can obtain that

\[
\| \epsilon^* \|_2^2 = O_p(d' c_*^{-1} \{ \log(p_1) \lor J_n \log(p_2 J_n) \}).
\] (A.9)

Combining (A.6), (A.8) and (A.9), we conclude that

\[
\| \hat{\theta}_{S^*} - \theta_{0, S^*} \|_2^2 = O_p \left[ \left( \frac{d' \{ \log(p_1) \lor J_n \log(p_2 J_n) \}}{nc_*^2} \right) + \frac{4d'(\lambda_{n1}^2 + \lambda_{n2}^2)}{n^2 c_*^2} \right]
= O_p \left[ n^{-1} \{ \log(p_1) \lor J_n \log(p_2 J_n) \} \right] + O \left( J_n^{-2d} \right) + O \left( \frac{n^{-2} (\lambda_{n1}^2 + \lambda_{n2}^2)}{n^2 c_*^2} \right),
\]

where the last inequality follows by \( d' = O(s_1 + s_2) \) and \( c_* \times 1 \) with probability approaching one.

By the properties of splines de Boor (2001), \( \| \hat{\alpha}_l - \alpha_{nl} \|_2^2 \asymp \| \hat{\gamma}_l - \gamma_{0l} \|_2^2 \), where \( \alpha_{nl}, l = 1, \ldots, p_2 \), is the best approximation for function \( \alpha_l \). Hence, part (ii) follows from \( \sum_{k=1}^{p_1} |\hat{\beta}_k - \beta_{0k}|^2 = O(\| \hat{\theta}_{S^*} - \theta_{0, S^*} \|_2^2) \) and \( \sum_{l=1}^{p_2} \| \hat{\gamma}_l - \gamma_{0l} \|_2^2 = O(\| \hat{\theta}_{S^*} - \theta_{0, S^*} \|_2^2) \).

We now prove part (i). Under Assumption (A1), if \( \| \theta_{0m} \|_2 \neq 0 \) but \( \| \tilde{\theta}_m \|_2 = 0 \), then \( \| \theta_{0m} - \tilde{\theta}_m \|_2 \geq c_\beta \lor c_\alpha \), which contradicts part (ii) when \( \log(p_1) \lor J_n \log(p_2 J_n) / n \to 0 \), \( \lambda_{n1}^2 / n^2 \to 0 \), and \( \lambda_{n2}^2 / n^2 \to 0 \). The results follow by the definition of \( \tilde{\alpha}_l, 1 \leq l \leq p_2; \tilde{\gamma} - \gamma_0 = (0_{(s_2 J_n) \times s_1} \ I_{s_2 J_n} ) ( \tilde{\theta} - \theta_0 ) \) and \( \tilde{\beta} - \beta_0 = ( I_{s_1} \ 0_{s_1 \times (s_2 J_n)} ) ( \tilde{\theta} - \theta_0 ) \).

Selection and estimation properties of the adaptive group LASSO estimators

In the following, we denote \( \alpha_{nl} (\cdot) = \sum_{j=1}^{J_n} \gamma_{0lj} B_{lj} (\cdot) \) the best spline approximation of \( \alpha_{0lj} (\cdot) \) such that \( \| \alpha_{0lj} - \alpha_{nl} \|_\infty = \sup_{x \in [0,1]} | \alpha_{0lj}(x) - \alpha_{nl}(x) | = O(J_n^{-d}) \). Let \( \gamma_{0l} = (\gamma_{0lj}, j = 1, \ldots, J_n)^T \) be the vector of the coefficients of the best spline approximation in Lemma A.3 in Section C. Denote

\[
\theta_0^T = (\beta_0^T, \gamma_0^T) = (\beta_{01}, \ldots, \beta_{0p_1}, \gamma_{01}, \ldots, \gamma_{0p_2}) = (\theta_{01}, \ldots, \theta_{0m}, \ldots, \theta_{0p_1+p_2})
\]

Proof of Theorem 2.1. To minimize

\[
L_{n2}(\theta; \lambda_{n3}, \lambda_{n4}) = \frac{1}{2} \left\| Y - Z\beta - \sum_{l=1}^{p_2} B_l \gamma_l \right\|_2^2 + \lambda_{n3} \sum_{k=1}^{p_1} |\beta_k| + \lambda_{n4} \sum_{l=1}^{p_2} \omega_{l+p_1} \| \gamma_l \|_2,
\]
by the Karush-Kuhn-Tucker (KKT) condition Boyd and Vandenberghe (2004), it is equivalent to solve the equations

\[
\begin{cases}
Z_k^T(Y - Z\beta - \sum_{l'=1}^{p_2} B_{l'} \gamma_{l'}) = \lambda_n 3 \omega_k \tau_k, & k = 1, \ldots, p_1, \\
B_l^T(Y - Z\beta - \sum_{l'=1}^{p_2} B_{l'} \gamma_{l'}) = \lambda_n 4 \omega_1 + p_1 \gamma_l, & l = 1, \ldots, p_2,
\end{cases}
\]

where

\[
\tau_k \in \begin{cases}
\{\beta_k/|\beta_k|\}, & \text{if } \beta_k \neq 0, \\
[-1, 1], & \text{if } \beta_k = 0,
\end{cases}
\]

\[
s_l \in \begin{cases}
\{\gamma_l/\|\gamma_l\|_2\}, & \text{if } \|\gamma_l\|_2 \neq 0, \\
\{u \in \mathbb{R}^{J_n} : \|u\|_2 \leq 1\}, & \text{if } \|\gamma_l\|_2 = 0.
\end{cases}
\]

Thus, any \( \theta \) satisfying the following KKT conditions:

**Condition A.2.** \( Z_k^T(Y - Z\beta - \sum_{l'=1}^{p_2} B_{l'} \gamma_{l'}) = \lambda_n 3 \omega_k \beta_k |\beta_k|, \) for any \( k \in S_z \).

**Condition A.3.** \( B_l^T(Y - Z\beta - \sum_{l'=1}^{p_2} B_{l'} \gamma_{l'}) = \lambda_n 4 \omega_1 + p_1 \gamma_l, \) for any \( l \in S_x \).

**Condition A.4.** \[ Z_k^T(Y - Z\beta - \sum_{l'=1}^{p_2} B_{l'} \gamma_{l'}) \leq \lambda_n 3 \omega_k, \) for any \( k \in N_z \).

**Condition A.5.** \( \|B_l^T(Y - Z\beta - \sum_{l'=1}^{p_2} B_{l'} \gamma_{l'})\|_2 \leq \lambda_n 4 \omega_1 + p_1, \) for any \( l \in N_x \).

is the unique minimizer of \( L_{n2}(\theta; \lambda_n 3, \lambda_n 4) \).

Define \( \vartheta_0 = \left( D_S^T D_S \right)^{-1} D_S^T Y \), a vector with length \( s_1 + s_2 J_n \). Denote two vectors \( v_1 = (v_{1m}, 1 \leq m \leq s_1 + s_2) \) and \( v_2 = (v_{2m}, 1 \leq m \leq s_1 + s_2) \) with elements:

\[
v_{1m} = \frac{\omega_m \vartheta_{0m} I\{m \in S_z\}}{|\vartheta_{0m}|} + \omega_{m+1} I\{m - s_1 \in S_x\}, \quad v_{2m} = \frac{\omega_{m+p_1} \vartheta_{0m}}{\|\vartheta_{0m}\|_2} I\{m - s_1 \in S_x\}.
\]

Next define \( \tilde{\theta}_0 = (\tilde{\theta}_{0m}, 1 \leq m \leq p_1 + p_2) \), where \( (\tilde{\theta}_{0m}, m \in S) = \left( D_S^T D_S \right)^{-1} (D_S^T Y - \lambda_n 3 v_1 - \lambda_n 4 v_2) \), \( \tilde{\theta}_{0m} = 0 \) for \( m \in N_z \) and \( \tilde{\theta}_{0m} = 0_{J_n} \) for \( m - p_1 \in N_x \). So we can represent \( \tilde{\theta}_0 \equiv (\tilde{\theta}_{0,S_z}, \tilde{\theta}_{0,N_z}, \tilde{\theta}_{0,S_x}, \tilde{\theta}_{0,N_x}) \), and \( \tilde{\theta}_{0,S} \equiv (\tilde{\theta}_{0,S_z}, \tilde{\theta}_{0,S_x}) \). Denote \( \tilde{S}_0 = \{1 \leq m \leq p_1 + p_2 : \|\tilde{\theta}_{0m}\|_2 > 0\} \). Apparently, \( \tilde{S}_0 \subseteq S \). Note that \( D \tilde{\theta}_0 = D_S \tilde{\theta}_{0,S} \) and \( \{D_m, m \in S\} \) are linearly independent, so by the definition of \( \tilde{\theta}_0 \), Conditions A.2 and A.3 hold for \( \tilde{\theta}_0 \) if \( \tilde{S}_0 \subseteq S \). So if \( \tilde{\theta}_0 \) satisfies
Condition A.6. $S_0 \supseteq S$.

Condition A.7. $|Z_k^\top (Y - D\hat{\theta}_0)| \leq \lambda n_3 \omega_k$, for any $k \in N_z$.

Condition A.8. $\|B_l^\top (Y - D\hat{\theta}_0)\|_2 \leq \lambda n_4 \omega_{l+p_1}$, for any $l \in N_x$.

then $\hat{\theta}_0$ is the unique minimizer of $L_{n2}(\theta; \lambda n_3, \lambda n_4)$, in other words, $\hat{\theta}_0 = \hat{\theta}$ with probability approaching one. Therefore, in order to show $\Pr(\hat{S} = S) \to 1$, it is equivalent to show $\hat{\theta}_0$ satisfies Conditions A.6 – A.8 with probability approaching one, as $n \to \infty$.

Further notice that

Condition A.9. $\|\theta_0m\|_2 - \|\hat{\theta}_0m\|_2 < \|\theta_0m\|_2, \forall m \in S$.

implies Condition A.6. Therefore, to show $\hat{\theta}_0$ is the unique minimizer of $L_{n2}(\theta; \lambda n_3, \lambda n_4)$, and consequently, $\Pr(\hat{S} = S) \to 1$, it suffices to show that $\hat{\theta}_0$ satisfies Conditions A.7–A.9, with probability approaching one, as $n \to \infty$.

According to Lemma A.2 and Lemma A.3 in Section C, we obtain that

$$
\Pr(\hat{S} \neq S) \leq \Pr(\|\theta_0m - \hat{\theta}_0m\|_2 \geq \|\theta_0m\|_2, \exists m \in S) + \Pr(|Z_k^\top (Y - D\hat{\theta}_0)| > \lambda n_3 \omega_k, \exists k \in N_z)
$$

$$
+ \Pr(|B_l^\top (Y - D\hat{\theta}_0)|_2 > \lambda n_4 \omega_{l+p_1}, \exists l \in N_x) \to 0,
$$

as $n \to \infty$. This completes the proof.

Proof of Theorem 2.2. The idea of the proof is similar to the proof of part (ii) in Theorem A.1 in Section C, but we look at index set $S$ instead of $S'$. Let $\pi_1$ and $\pi_2$ be the minimum and maximum eigenvalues of $C_S$, respectively, and let $\pi_3 = \max_{m \notin S} \|n^{-1}D_m^\top D_m\|_2$. By Lemma A.2 in Section C, $\pi_1 \asymp 1$, $\pi_2 \asymp 1$ and $\pi_3 \asymp 1$. For any $l = 1, \cdots, s_2$, let $\alpha_0l(X_l) = (\alpha_0l(X_{1l}), \cdots, \alpha_0l(X_{nl}))^\top$, $\delta_l = \alpha_0l(X_l) - B_l \gamma_l$ and $\delta = \sum_{l=1}^{s_2} \delta_l$. According to the proof of Theorem 2.1, with probability
approaching one, we have

$$\hat{\theta}_S = \hat{\theta}_{0,S} = \left( D_S^\top D_S \right)^{-1} \left( D_S^\top Y - \lambda n_3 v_1 - \lambda n_4 v_2 \right)$$

$$= \left( D_S^\top D_S \right)^{-1} \left[ D_S^\top \left( Z_S \beta_{0,S} + \sum_{l=1}^{s_2} (B_l \gamma_{0l} + \delta_l) + \epsilon \right) - \lambda n_3 v_1 - \lambda n_4 v_2 \right]$$

$$= \theta_{0,S} + \left( D_S^\top D_S \right)^{-1} \left\{ D_S^\top (\delta + \epsilon) - \lambda n_3 v_1 - \lambda n_4 v_2 \right\}.$$  

Let $C_S = n^{-1} D_S^\top D_S$ be an $(s_1 + s_2 J_n) \times (s_1 + s_2 J_n)$ matrix, and let $H = I - D_S \left( D_S^\top D_S \right)^{-1} D_S^\top$ be an $n \times n$ matrix, then

$$\hat{\theta}_S - \theta_{0,S} = n^{-1} C_S^{-1} \{ D_S^\top (\delta + \epsilon) - \lambda n_3 v_1 - \lambda n_4 v_2 \}. \quad (A.10)$$

For $\eta = Y - D \theta$, define $\eta_*$ as the projection of $\eta$ to the column space of $D_S$, that is, $\eta_* \equiv P_{D_S} \eta = D_S (D_S^\top D_S)^{-1} D_S^\top \eta$. Then for $\epsilon_* \equiv P_{D_S} \epsilon$, similar to (A.6), (A.8) and (A.9), and by Lemma A.2 in Section C,

$$\| \epsilon_* \|^2 = \| D_S^\top D_S \|^{-1/2} D_S^\top \epsilon \|^2 \leq (n \pi_1)^{-1} \| D_S^\top D_S \| \| \epsilon \|^2 = O_p \{ \pi_1^{-1} (s_1 + s_2 J_n) \},$$

$$\| \eta_* \|^2 \leq 2 \| \epsilon_* \|^2 + O_p (n s_2 J_n^{-2d}) = O_p (J_n) + O_p (n s_2 J_n^{-2d}),$$

$$\| \hat{\theta}_S - \theta_{0,S} \|^2 \leq \frac{8 \| \eta_* \|^2}{n \pi_1} + \frac{4 (\lambda_{n_3}^2 + \lambda_{n_4}^2)}{n^2 \pi_1} = O_p \left( \frac{J_n}{n} \right) + O \left( \frac{1}{J_n^d} \right) + O_p \left( \frac{\lambda_{n_3}^2 + \lambda_{n_4}^2}{n^2} \right).$$

Therefore, the results follow by the facts that

$$\hat{\beta}_{S_x} - \beta_{0,S_x} = \left( I_{s_1}, 0_{s_1 \times (s_2 J_n)} \right) (\hat{\theta}_S - \theta_{0,S}) \quad \hat{\gamma}_{S_x} - \gamma_{0,S_x} = \left( 0_{(s_2 J_n) \times s_1}, I_{s_2 J_n} \right) (\hat{\theta}_S - \theta_{0,S}) \quad (A.11)$$

and $\| \tilde{\alpha}_l - \alpha_{nl} \|^2 \leq \| \gamma_l - \gamma_{0l} \|^2$,

where $\hat{\beta}_{S_x} = (\hat{\beta}_{k}, k \in S_x)^\top$, $\beta_{0,S_x} = (\beta_{0,k}, k \in S_x)^\top$, $\hat{\gamma}_{S_x} = (\hat{\gamma}_l, l \in S_x)^\top$ and $\gamma_{0,S_x} = (\gamma_l, l \in S_x)^\top$.

Technical Lemmas Used in the Proof of Selection Consistency

**Lemma A.1.** For

$$\mathbf{v}_1 = \left( \frac{\omega_1}{\| \theta_{0,1} \|}, \ldots, \frac{\omega_1}{\| \theta_{0,1} \|}, 0_{s_2 J_n} \right)^\top, \quad \mathbf{v}_2 = \left( 0_{s_1}, \frac{\omega_{p_1+1}}{\| \theta_{0,s_1+1} \|}, \ldots, \frac{\omega_{p_1+s_2}}{\| \theta_{0,s_1+s_2} \|} \right)^\top,$$
under Assumption (A5),

\[
\begin{align*}
\|v_1\|^2 &= O_p\left(h_{n1}^2\right) = O_p\left(b_{n1}^{-4}r_{n1}^{-2} + s_1b_{n1}^{-2}\right), \\
\|v_2\|^2 &= O_p\left(h_{n2}^2\right) = O_p\left(b_{n2}^{-4}r_{n2}^{-2} + s_2b_{n2}^{-2}\right).
\end{align*}
\]

(A.12) (A.13)

Proof. Write

\[
\begin{align*}
\|v_1\|^2 &= \sum_{k=1}^{s_1} \omega_k^2 = \sum_{k=1}^{s_1} |\beta_k|^{-2} = \sum_{k=1}^{s_1} \frac{\beta_{0k}^2 - \beta_k^2}{\beta_{0k}^2} + \sum_{k=1}^{s_1} |\beta_{0k}|^{-2}, \\
\|v_2\|^2 &= \sum_{l=1}^{s_2} \omega_l^2+1 = \sum_{l=1}^{s_2} ||\gamma_l||_{2}^2 - \sum_{l=1}^{s_2} \frac{||\gamma_0||_{2}^2 - ||\gamma_l||_{2}^2}{||\gamma_0||_{2}^2 ||\gamma_l||_{2}^2} + \sum_{l=1}^{s_2} ||\gamma_0||_{2}^2.
\end{align*}
\]

Under (A5), there exist positive constants $M_3$ and $M_4$, such that

\[
\begin{align*}
\sum_{k=1}^{s_1} \frac{\beta_{0k}^2 - \beta_k^2}{\beta_{0k}^2} &\leq M_3c_1^{-2}b_{n1}^{-4}||\beta - \beta_0||_2^2 = O_p\left(b_{n1}^{-4}c_1^{-2}r_{n1}^{-2}\right), \\
\sum_{l=1}^{s_2} \frac{||\gamma_0||_{2}^2 - ||\gamma_l||_{2}^2}{||\gamma_0||_{2}^2 ||\gamma_l||_{2}^2} &\leq M_4c_2^{-2}b_{n2}^{-4}||\gamma - \gamma_0||_2^2 = O_p\left(b_{n2}^{-4}c_2^{-2}r_{n2}^{-2}\right),
\end{align*}
\]

and the results follow by $\sum_{k=1}^{s_1} |\beta_k|^{-2} \leq s_1b_{n1}^{-2}$ and $\sum_{l=1}^{s_2} ||\gamma_l||_{2}^2 \leq s_2b_{n2}^{-2}$. \hfill \Box

Lemma A.2. Under Assumptions (A3) – (A6), as $n \to \infty$,

\[
\Pr\left(\|\theta_{0m} - \hat{\theta}_m\|_2 \geq \|\theta_{0m}\|_2, \exists m \in S\right) \to 0.
\]

Proof. Let $Q_m$ be a $d_m \times (s_1 + s_2J_n)$ matrix, $d_m = 1$ for $1 \leq m \leq s_1$ and $d_m = J_n$ for $s_1 + 1 \leq m \leq s_1 + s_2$, with the form

\[
Q_m = \left( Q_{1m} \quad 0_{(s_2J_n) \times (s_2J_n)} \right) I(1 \leq m \leq s_1) + \left( 0_{J_n \times s_1} \quad Q_{2m-s_1} \right) I(s_1 + 1 \leq m \leq s_1 + s_2)
\]

with $Q_{1m} = (0, \ldots, 0, 1, 0, \ldots, 0)$ and $Q_{2m} = (0_{J_n \times J_n}, \ldots, 0_{J_n \times J_n}, I_{J_n}, 0_{J_n \times J_n}, \ldots, 0_{J_n \times J_n})$, where scalar 1 is the $m$-th element of vector $Q_{1m}$ with length $s_1$, and a $J_n \times J_n$ identity matrix $I_{J_n}$ is at the $m$-th block of the $J_n \times (s_2J_n)$ matrix $Q_{2m}$ with the remaining $J_n \times J_n$ matrices equal to $0_{J_n \times J_n}$.
Then from (A.10), 
\[
\hat\theta_m - \theta_{0m} = n^{-1}Q_mC_S^{-1}\left(D_S^\top \epsilon + D_S^\top \delta - \lambda_{n3}v_1 - \lambda_{n4}v_2 \right).
\]
By the triangle inequality,
\[
\|\hat\theta_m - \theta_{0m}\|_2 \leq n^{-1}\|Q_mC_S^{-1}D_S^\top \epsilon\|_2 + n^{-1}\|Q_mC_S^{-1}D_S^\top \delta\|_2 + n^{-1}\|Q_mC_S^{-1}(\lambda_{n3}v_1 + \lambda_{n4}v_2)\|_2.
\]
Recall that \(\pi_1\) and \(\pi_2\) are the minimum and maximum eigenvalues of \(C_S\), respectively. By Lemmas A.1 and A.2, the first term on the right-hand side
\[
\max_{m \in S} n^{-1}\|Q_mC_S^{-1}D_S^\top \epsilon\|_2 \leq n^{-1/2}\|C_S^{-1}\|_2 \|n^{-1/2}D_S^\top \epsilon\|_2 \leq n^{-1/2}\pi_1^{-1}Op\left\{(s_1 + s_2 J_n)^{1/2}\right\} = O_p\left\{n^{-1/2}(s_1 + s_2 J_n)^{1/2}\right\}.
\]
By Lemma A.2, the second term
\[
\max_{m \in S} n^{-1}\|Q_mC_S^{-1}D_S^\top \delta\|_2 \leq n^{-1}\|C_S^{-1}D_S^\top \delta\|_2 \leq \|C_S^{-1}\|_2 \times \|n^{-1/2}D_S^\top D_S\|^{1/2}_2 \|n^{-1/2}\delta\|_2 \leq \pi_1^{-1}\pi_2^{-1/2}Op\left(s_2^{1/2}J_n^{-d}\right) = O_p\left(s_2^{1/2}J_n^{-d}\right).
\]
By Lemma A.2 and Lemma A.1, the third term
\[
\max_{m \in S} n^{-1}\|Q_mC_S^{-1}(\lambda_{n3}v_1 + \lambda_{n4}v_2)\|_2 \leq n^{-1}\pi_1^{-1}\|\lambda_{n3}v_1 + \lambda_{n4}v_2\|_2 \leq n^{-1}\pi_1^{-1}(\|\lambda_{n3}v_1\|_2 + \|\lambda_{n4}v_2\|_2) = O_p\left\{n^{-1}(\lambda_{n3}h_{n1} + \lambda_{n4}h_{n2})\right\}.
\]
Thus, the claim follows by Assumption (A6).

**Lemma A.3.** *Under Assumptions (A3) – (A6), as \(n \to \infty\),*

\[
Pr\left\{|D_m^\top(Y - \bar{\theta})| > \lambda_{n3}\omega_m, \ \exists \ m \in \mathcal{N}_x\right\} \to 0,
\]

\[
Pr\left\{|D_m^\top(Y - \bar{\theta})|_2 > \lambda_{n4}\omega_m, \ \exists \ m - p_1 \in \mathcal{N}_x\right\} \to 0.
\]
Proof. Note that

\[ Y - D_S\hat{\theta}_S = Z_S\beta_{0,S} + \sum_{l=1}^{s_2} a_{ql}(X_l) + \epsilon - D_S\hat{\theta}_S = D_S\theta_0 + \epsilon - D_S\hat{\theta}_S \]

\[ = -D_S \left( D_S^T D_S \right)^{-1} \left\{ D_S^T (\delta + \epsilon) - \lambda_{n3}v_1 - \lambda_{n4}v_2 \right\} + \delta + \epsilon \]

\[ = \left\{ I - D_S \left( D_S^T D_S \right)^{-1} D_S^T \right\} \delta + \epsilon + D_S \left( D_S^T D_S \right)^{-1} (\lambda_{n3}v_1 + \lambda_{n4}v_2) \]

\[ = H\epsilon + H\delta + n^{-1}\lambda_{n3} D_S C_S^{-1}v_1 + n^{-1}\lambda_{n4} D_S C_S^{-1}v_2. \quad \text{(A.14)} \]

For \( 1 \leq m \leq p_1 + p_2 \), by (A.14), we have

\[ D_m^T(Y - D_S\hat{\theta}_S) = D_m^T H\epsilon + D_m^T H\delta + n^{-1}D_m^T D_S C_S^{-1}(\lambda_{n3}v_1 + \lambda_{n4}v_2). \quad \text{(A.15)} \]

By Lemma A.1,

\[ \text{E}(\max_{m \in N_z} \|n^{-1/2}D_m^T H\epsilon\|_2) = O\{\log(p_1)^{1/2}\}, \quad \text{E}(\max_{m \neq p_1 \in N_z} \|n^{-1/2}D_m^T H\epsilon\|_2) = O\{\log^{1/2}(p_2J_n)\}. \]

Then for the \( D_m^T H\epsilon \) part of (A.15), from Assumption (A5), for all \( m \in N_z \), \( \omega_m = |\tilde{\beta}_m|^{-1} = O_p(r_{n1}) \), there exists a positive constant \( c_1 \), such that

\[ \text{Pr}\left(\|D_m^T H\epsilon\| > \lambda_{n3}\omega_m/3, \exists m \in N_z\right) \leq \text{Pr}\left(\|D_m^T H\epsilon\| > c_1\lambda_{n3}r_{n1}, \exists m \in N_z\right) + o(1) \]

\[ = \text{Pr}\left(\max_{m \in N_z} n^{-1/2}\|D_m^T H\epsilon\| > c_1n^{-1/2}\lambda_{n3}r_{n1}\right) + o(1) \]

\[ \leq \frac{n^{1/2}\text{E}\left(\max_{m \in N_z} n^{-1/2}\|D_m^T H\epsilon\|\right)}{c_1\lambda_{n3}r_{n1}} + o(1) \leq n^{1/2}O\{\log(p_1)^{1/2}\}c_1\lambda_{n3}^{-1}r_{n1}^{-1} + o(1) \]

\[ = O\{n^{1/2}(\log(p_1)^{1/2}\lambda_{n3}^{-1}r_{n1}^{-1})\} + o(1) \]

Similarly, \( \text{Pr}\left(\|D_m^T H\epsilon\|_2 > \lambda_{n4}\omega_m/3, \exists m - p_1 \in N_z\right) = O\left(nJ_n^{1/2}\log^{1/2}(p_2J_n)\lambda_{n4}^{-1}r_{n2}^{-1}\right) + o(1) \). Recall the definition of \( N \) in (2.6) and \( \pi_3 = \max_{m \notin S} n^{-1}D_m^T D_m \). By the properties of splines de Boor (2001),

\[ \max_{m \in N} \|D_m^T H\delta\|_2 \leq n^{1/2} \max_{m \notin S} n^{-1/2}D_m^T D_m \|H\|_2 \|\delta\|_2 = O_p(n\pi_3^{1/2}s_2^{1/2}J_n^{-d}) = O_p(ns_2^{1/2}J_n^{-d}). \]
To bound the last term in (A.15), it follows by Lemma A.2 and (A.12) and (A.13) in Lemma A.1 that

\[
\max_{m \in \mathbb{N}} \left\| n^{-1} D_m^T D_S C_S^{-1} (\lambda_n v_1 + \lambda_n v_2) \right\|_2 \\
\leq \max_{m \notin S} \left\| n^{-1/2} D_m \right\|_2 \times \left\| n^{-1/2} D_S C_S^{-1/2} \right\|_2 \times \left\| C_S^{-1/2} \right\|_2 \times \left\| \lambda_n v_1 + \lambda_n v_2 \right\|_2 \\
\leq \frac{\pi_3^{1/2}}{\pi_1^{1/2}} O_p(\lambda_n h_n + \lambda_n h_{n2}) = O_p(\lambda_n h_n + \lambda_n h_{n2}).
\]

\[
\square
\]

Asymptotic Distributions

For any index set \( A \subseteq \{1, \ldots, p_1 + p_2\} \), denote \( \beta_A = (\beta_k, 1 \leq k \leq p_1, k \in A)^T \), \( \hat{\beta}_A = (\hat{\beta}_k, 1 \leq k \leq p_1, k \in A)^T \), \( \gamma_A = (\gamma_l, 1 \leq l \leq p_2, l + p_1 \in A)^T \) and \( \hat{\gamma}_A = (\hat{\gamma}_l, 1 \leq l \leq p_2, l + p_1 \in A)^T \).

Next, denote \( Z_A = (Z_{i,i}, i = 1, \ldots, n)^T \), where \( Z_{i,i} = (Z_{ik}, 1 \leq k \leq p_1, k \in A)^T \). Similarly, denote \( B_A = (B_{i,j}, i = 1, \ldots, n)^T \), where \( B_{i,j} = (B_{ij}(X_{il})), 1 \leq l \leq p_2, l + p_1 \in A, j = 1, \ldots, J\)\( n \). Let

\[
C_S = n^{-1} D_S^T D_S = \begin{pmatrix} n^{-1} Z_S^T Z_S & n^{-1} Z_S^T B_S \\ n^{-1} B_S^T Z_S & n^{-1} B_S^T B_S \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},
\]

(A.16)

\[
U_S = C_S^{-1} = \begin{pmatrix} U_{11} & -U_{11} C_{12} C_{22}^{-1} \\ -U_{22} C_{21} C_{11}^{-1} & U_{22} \end{pmatrix} = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix},
\]

(A.17)

where \( U_{11}^{-1} = C_{11} - C_{12} C_{22}^{-1} C_{21} = n^{-1} Z_S^T (I_n - P_{B_S}) Z_S \), and \( U_{22}^{-1} = C_{22} - C_{21} C_{11}^{-1} C_{12} = n^{-1} B_S^T (I_n - P_{Z_S}) B_S \), with \( P_{B_S} \) and \( P_{Z_S} \) being projection matrices for \( B_S \) and \( Z_S \), respectively.

In the following, we give the proofs of Theorems 2.3 and 2.4.
Proof of Theorem 2.3. By Lemma A.3, (A.10) in Section C, (A.11), (A.16) and (A.17),
\[
\hat{\beta}_{s_2} - \beta_{0,s_2} = \left( I_{s_1} \ 0_{s_1 \times (s_2 J_n)} \right) C_S^{-1} n^{-1} \left\{ D_S^T (\delta + \epsilon) - \lambda_{n3} v_1 - \lambda_{n4} v_2 \right\}
\]
\[
= n^{-1} U_{11} \left( I_{s_1} - C_{12} C_{22}^{-1} \right) \begin{pmatrix} Z_S^T (\delta + \epsilon) \\ B_S^T (\delta + \epsilon) \end{pmatrix} - \lambda_{n3} v_1 - \lambda_{n4} v_2
\]
\[
= n^{-1} U_{11} Z_S^T (I_n - P_{B_S}) \epsilon + n^{-1} U_{11} Z_S^T (I_n - P_{B_S}) \delta
\]
\[
- n^{-1} U_{11} \left( I_{s_1} - Z_S B_S (B_S^T B_S)^{-1} \right) (\lambda_{n3} v_1 + \lambda_{n4} v_2).
\]
(A.18)

For the first term in (A.18), denote \( \hat{\beta}_\epsilon = n^{-1} U_{11} Z_S^T (I_n - P_{B_S}) \epsilon \), then for any \( b \in \mathbb{R}^{s_1} \) with \( \|b\|_2 = 1 \), there exist \( \{a_i\}_{i=1}^n \), such that \( b^T \hat{\beta}_\epsilon = \sum_{i=1}^n a_i \epsilon_i \), where
\[
\sum_{i=1}^n a_i^2 = n^{-1} b^T U_{11} \left\{ n^{-1} Z_S^T (I_n - P_{B_S}) Z_S \right\} U_{11} b,
\]
\[
a_i^2 = n^{-2} b^T U_{11} \left( Z_{i,S} - Z_S B_S (B_S^T B_S)^{-1} B_S^T Z_S \right) U_{11} b,
\]

and conditioning on \((Z_{i,S}, X_{i,S}), i = 1, \ldots, n\), \( a_i \epsilon_i \)'s are independent. By Lemma A.2, we have
\[
\max_{1 \leq i \leq n} a_i^2 \leq n^{-2} C_1 \max_{1 \leq i \leq n} \left\{ \|Z_{i,S}\|_2^2 + \|B_{i,S}^T B_{i,S} (B_S^T B_S)^{-1} B_S^T Z_S\|_2^2 \right\},
\]
\[
\|B_{i,S}^T B_{i,S} (B_S^T B_S)^{-1} B_S^T Z_S\|_2 \leq C_2 n^{-1} \|B_{i,S}^T B_S^T Z_S\|_2,
\]

for some constants \( C_1 \) and \( C_2 \). Under Assumption (A4), \( \|Z_{i,S}\|_2^2 = O_p(1) \); and also for the \( k \)-th component of \( n^{-1} B_{i,S}^T B_{i,S} Z_S \), \( E \left\{ \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{s_2} Z_{i,j} B_{ij} (X_{i}^T) B_{ij} (X_{i}) \right\}^2 = O(J_n) \). Thus, \( \max_{1 \leq i \leq n} a_i^2 = O_p(n^{-2} J_n) \). By Lemma A.2 in Section C, with probability approaching one, we have \( \sum_{i=1}^n a_i^2 \geq c n^{-1} \) for some constant \( c \). Therefore, \( \max_{1 \leq i \leq n} a_i^2 / \sum_{i=1}^n a_i^2 = O_p(n^{-1} J_n) = o_p(1) \).

Since \( Var(\hat{\beta}_\epsilon) = n^{-1} \sigma^2 U_{11} \), by the Linderberg-Feller CLT and the Cramer-Wold Device, one has \( (n^{-1} U_{11})^{-1/2} \hat{\beta}_\epsilon \overset{D}{\rightarrow} N(0, \sigma^2 I_{s_1}) \).

For the second term in (A.18), denote \( \hat{\beta}_\delta = n^{-1} U_{11} Z_S^T (I_n - P_{B_S}) \delta \), then we have \( \|\hat{\beta}_\delta\|_2 \leq n^{-1/2} \|U_{11}\|_2 \|n^{-1/2} Z_S^T\|_2 \|I_n - P_{B_S}\|_2 \|\delta\|_2 = O_p(J_n^{-d}) \). For the third term in (A.18), denote \( \hat{\beta}_\lambda = \)
Thus, the desired result follows by Assumption (A6), (A.12) and (A.13).

**Proof of Theorem 2.4.** Note that

\[ n^{-1} U_{11} \left( I_{s_1} - Z_S B_S (B_S^T B_S)^{-1} \right) (\lambda_{n3} v_1 + \lambda_{n4} v_2). \]

Then by Lemma A.1 in Section C,

\[ \|\hat{\beta}_X\|_2 \leq n^{-1} \| U_{11} \|_2 \| \lambda_{n3} v_1 + \lambda_{n4} v_2\|_2 = O_p(n^{-1} \lambda_{n3}\|v_1\|_2 + n^{-1} \lambda_{n4}\|v_2\|_2). \]  

(A.19)

Thus, the desired result follows by Assumption (A6), (A.12) and (A.13).

**Proof of Theorem 2.4.** Note that

\[ \alpha_i^{SBL}(x_l) - \hat{\alpha}_i(x_l) = (1, 0) (X_l^T W_l X_l)^{-1} X_l^T W_l (\hat{Y}_l - Y_l), \]

where

\[ \hat{Y}_l - Y_l = Z_S (\beta_{0,l} S - \hat{\beta}_S) + \sum_{r \in S \setminus \{l\}} \{\alpha_{0r}(X_r) - \hat{\alpha}_r(X_r)\} \]

\[ = Z_S (\beta_{0,l} S - \hat{\beta}_S) + B_{S \setminus \{l\}} (\gamma_{0,S \setminus \{l\}} - \hat{\gamma}_{S \setminus \{l\}}) + \sum_{r \in S \setminus \{l\}} \{\alpha_{0r}(X_r) - \alpha_{nl}(X_r)\}, \]

\[ \text{diag}(1, h^{-1}) X_l^T W_l X_l \text{ diag}(1, h^{-1}) \]

\[ = \left( \begin{array}{cc} n^{-1} \sum_{i=1}^{n} K_h(X_{il} - x_l) & n^{-1} \sum_{i=1}^{n} \left( \frac{X_{il} - x_l}{h} \right) K_h(X_{il} - x_l) \\ n^{-1} \sum_{i=1}^{n} \left( \frac{X_{il} - x_l}{h} \right) K_h(X_{il} - x_l) & n^{-1} \sum_{i=1}^{n} \left( \frac{X_{il} - x_l}{h} \right)^2 K_h(X_{il} - x_l) \end{array} \right) \\
= f_l(x_l) \begin{pmatrix} 1 & 0 \\ 0 & \mu_2(K) \end{pmatrix} + u_p(1), \]

with \( u_p(\cdot) = o_p(\cdot) \) uniformly for all \( x_l \in [0, 1] \). So

\[ \left( X_l^T W_l X_l \right)^{-1} = \text{diag}(1, h^{-1}) f_l^{-1}(x_l) \left( \begin{array}{cc} 1 & 0 \\ 0 & \mu_2(K) \end{array} \right) + u_p(1) \]

\[ \text{diag}(1, h^{-1}) X_l^T W_l = \frac{1}{n} \times \left( \begin{array}{cccc} K_h(X_{1l} - x_l) & \cdots & K_h(X_{nl} - x_l) \\ \left( \frac{X_{1l} - x_l}{h} \right) K_h(X_{1l} - x_l) & \cdots & \left( \frac{X_{nl} - x_l}{h} \right) K_h(X_{nl} - x_l) \end{array} \right). \]
Thus,
\[
\hat{\alpha}^\text{SSLv}_{i,l}(x_l) - \hat{\alpha}^0_{i,l}(x_l) = f_i^{-1}(x_l) \left[ u_p(1) + \frac{1}{n} \sum_{i=1}^{n} K_h(X_{il} - x_l) \mathbf{Z}_{i,S}^\top (\beta_{0,S_x} - \hat{\beta}_{S_x}) \right. \\
+ \frac{1}{n} \sum_{i=1}^{n} K_h(X_{il} - x_l) \mathbf{B}_{i,S_x \setminus \{l\}}^\top (\gamma_{0,S_{x \setminus \{l\}}} - \hat{\gamma}_{S_{x \setminus \{l\}}} ) \\
+ \frac{1}{n} \sum_{i=1}^{n} \sum_{l' \in S_{x \setminus \{l\}}} K_h(X_{il} - x_l) \left\{ \alpha_{0l'}(X_{il'}) - \alpha_{nl'}(X_{il'}) \right\} \right]. \\
\tag{A.20}
\]

For the first and third summation terms in the right hand side of (A.20), by Theorem 2.3, we have \( \frac{1}{n} \sum_{i=1}^{n} K_h(X_{il} - x_l) \mathbf{Z}_{i,S}^\top (\beta_{0,S_x} - \hat{\beta}_{S_x}) = u_p \left( n^{-1/2} \right) \); and by Lemma A.3 in Section C, \( \frac{1}{n} \sum_{i=1}^{n} \sum_{l' \in S_{x \setminus \{l\}}} K_h(X_{il} - x_l) \left\{ \alpha_{0l'}(X_{il'}) - \alpha_{nl'}(X_{il'}) \right\} = u_p \left( s_2 J_n^{-d} \right) \). As for the second term, by (A.10) and (A.11),

\[
\hat{\gamma}_{S_x} - \gamma_{0,S_x} = \left( 0_{(s_2 J_n) \times s_1} \mathbf{I}_{s_2 J_n} \right) \mathbf{C}_{S}^{-1} n^{-1} \left\{ \mathbf{D}_{S}^\top (\delta + \epsilon) - n \lambda_{3} \mathbf{v}_1 - n \lambda_{4} \mathbf{v}_2 \right\}.
\]

Define an \((s_2 J_n) \times (s_2 J_n)\) diagonal matrix \( \mathbf{I}^0 = \text{diag}(\mathbf{1}_{(l-1)J_n}, \mathbf{0}_{J_n}, \mathbf{1}_{(s_2-l)J_n}) \), \( l \in S_x \). Next by Lemma A.3 in Section C, (A.16) and (A.17), for any \( l \in S_x \),

\[
\sum_{i=1}^{n} K_h(X_{il} - x_l) \mathbf{B}_{i,S_x \setminus \{l\}}^\top (\gamma_{0,S_{x \setminus \{l\}}} - \hat{\gamma}_{S_{x \setminus \{l\}}} ) \right\} = \sum_{i=1}^{n} K_h(X_{il} - x_l) \mathbf{B}_{i,S}^\top \mathbf{I}^0 \mathbf{U}_{22} \left( -\mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{I}_{s_2 J_n} \right) \frac{1}{n} \left\{ (\mathbf{Z}_{S} \ \mathbf{B}_{S})^\top (\delta + \epsilon) - n \lambda_{3} \mathbf{v}_1 - n \lambda_{4} \mathbf{v}_2 \right\}
\]

\[
= \sum_{i=1}^{n} K_h(X_{il} - x_l) \mathbf{B}_{i,S}^\top \mathbf{I}^0 \mathbf{U}_{22} \left( -\mathbf{B}_{S} \mathbf{Z}_{S} \mathbf{Z}_{S}^{-1} \mathbf{I}_{s_2 J_n} \right) (\lambda_{3} \mathbf{v}_1 + \lambda_{4} \mathbf{v}_2).
\]

Following the same idea in the proof of Lemma A.2 in Section C, we have that there exist constants \( 0 < c_{U_2} < C_{U_2} < \infty \), such that with probability approaching one, \( c_{U_2} \mathbf{I}_{s_2 J_n} \leq \mathbf{U}_{22} \leq C_{U_2} \mathbf{I}_{s_2 J_n} \).
Similar to Lemma A.4 in Wang and Yang (2007), for any $1 \leq l \neq l' \leq s_2$,

$$\sup_{x_t \in [0,1]} \max_{1 \leq j \leq J_n} \left| \frac{1}{n} \sum_{i=1}^{n} K_h (X_{il} - x_t) B_{lj} (X_{il'}) - E\{K_h (X_{il} - x_t) B_{l'j} (X_{il'})\} \right| = O_p\{(\log n)^{1/2}(nh)^{-1/2}\},$$

$$\sup_{x_t \in [0,1]} \max_{1 \leq j \leq J_n} \left| \frac{1}{n} \sum_{i=1}^{n} E\{K_h (X_{il} - x_t) B_{l'j} (X_{il'})\} \right| = O_p\left( J_n^{-1/2} \right).$$

We can show that

$$\sup_{x_t \in [0,1]} \frac{1}{n} \sum_{i=1}^{n} K_h (X_{il} - x_t) B_{l,S}^{\top} I_0^T U_{22} \frac{1}{n} B_{S}^{\top} (I_n - P_{Z_S}) \delta$$

$$= O_p\left[ J_n^{-d+1} \{(nh)^{-1/2}(\log n)^{1/2} + J_n^{-1/2}\} \right],$$

and by Proposition 2 in Wang and Yang (2009),

$$\sup_{x_t \in [0,1]} \frac{1}{n} \sum_{i=1}^{n} K_h (X_{il} - x_t) B_{l,S}^{\top} I_0^T U_{22} \frac{1}{n} B_{S}^{\top} (I_n - P_{Z_S}) \epsilon = O_p\{n^{-1/2}(\log n)^{1/2}\}.$$

Similar arguments as in the proof of (A.19), (A.12) and (A.13) lead to

$$\sup_{x_t \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} K_h (X_{il} - x_t) B_{l,S}^{\top} I_0^T U_{22} \left(-B_{S}^{\top} Z_S (Z_S^{\top} Z_S)^{-1} I_{s_2 J_n}\right) (\lambda_{n3} x_1 + \lambda_{n4} x_2) \right|$$

$$= O_p\left[ n^{-1} J_n (\lambda_{n3} h_{n1} + \lambda_{n4} h_{n2}) \{(nh)^{-1/2}(\log n)^{1/2} + J_n^{-1/2}\} \right].$$

Therefore,

$$\sup_{x_t \in [0,1]} |\hat{\alpha}_l^{SBL}(x_t) - \hat{\alpha}_l^{\theta}(x_t)|$$

$$= O_p\left[ n^{-1/2}(\log n)^{1/2} + \{J_n^{-d+1} + J_n (\lambda_{n3} + \lambda_{n4})/n\} \{(nh)^{-1/2}(\log n)^{1/2} + J_n^{-1/2}\} \right].$$

Hence, the result in (2.10) is established. Consequently, the result in (2.11) follows from (2.9), and the result in (2.12) follows from Claeskens and Van Keilegom (2003).
Pre-processing Steps

The original SAM dataset contains data on 368 maize inbred lines. For each inbred line, measurements include log-scale SAM volume, SNP genotypes at 1,279,929 markers, and RNA sequence read counts for 39,656 transcripts. Prior to the analysis presented in Section 6, we applied quality control pre-processing to the SNP data following Yazdani and Dunson (2015) and quality control pre-processing to the RNA-seq read count data following the user’s guide for the R package “edgeR” (https://www.bioconductor.org/packages/devel/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf). For completeness and reproducibility, we describe those pre-processing steps as follows.

SNP Pre-processing

1. Separately, for each of the 10 maize chromosomes, we impute missing SNP genotype data using the linkage-disequilibrium k-nearest imputation algorithm Money et al. (2015).

2. Imputed SNP genotype data contain up to 3 levels: 0 for major allele, 2 for minor allele, and 1 for others. SNPs that include value 1 are not considered for further analysis, which reduces the SNP dimension to 1,111,438.

3. SNP genotypes with minor allele frequency below 2%, which could be indicative of low genotyping efficiency, are removed, leaving 926,918 SNPs for further consideration.

4. Hierarchical clustering is used to group SNPs into 148,793 clusters, where pairwise correlation between SNPs within a cluster is at least 0.5. One tag SNP, which has the highest mean
correlation with other SNPs within the cluster, is selected to represent each cluster, and the 148,793 tag SNPs are used for subsequent analysis.

**RNA-seq Read Count Pre-processing**

1. As is common in RNA-seq analysis, we begin by removing low-abundance transcripts whose levels are difficult to accurately quantify. Specifically, RNA transcripts whose counts per million (CPM) values are less than or equal to 5 for more than 95% of the samples are removed. This step reduces the number of transcripts considered for further processing to 22,903.

2. To ensure that measurements of transcript abundance are comparable across inbred lines, we divide each of the original read counts for any inbred line by the upper quartile of counts for that inbred line. This is the upper-quartile normalization strategy considered by Bullard et al. (2010).

3. The normalized data from step 2 are log transformed. For transcripts with zero values, half the minimum positive value for that transcript was added to zeros prior to transformation.

**A Simulation Study Based on the SAM Data**

In this section, we conduct another simulation study using the SNPs and RNA transcripts selected in real data analysis in Section 2.6 as the active covariates. With the true linear coefficients and nonlinear functions set to be the same as the estimates obtained in real data analysis, we choose the noise level $\sigma$ as 0.01 and 0.02, where 0.01 and 0.02 are in accordance with the errors in real data analysis. We compare SBLL-AGLASSO with OLS-ALASSO and ORACLE* and ORACLE**.
Table A.1: Cross validation results of the SBLL-AGLASSO and OLS-ALASSO for the SAM data example

<table>
<thead>
<tr>
<th>Fold</th>
<th>SBLL-AGLASSO</th>
<th>OLS-ALASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSPE SNPs SNPs</td>
<td>MSPE SNPs SNPs</td>
</tr>
<tr>
<td>1</td>
<td>0.08 179 5</td>
<td>0.11 249 93</td>
</tr>
</tbody>
</table>

Given the complexity of the real data and the fact that the number of true significant SNPs in the model is large, to better evaluate the performance of the above methods, instead of using measures “#V”, “IN” and “CS” as in Section 2.5, we summarize the selection results in terms of power and Type I error type statistics. In detail, the selection related statistics include:

(i) the number of nonzero linear effects correctly identified as nonzero (denoted as “CorrZ”);

(ii) the number of zero linear coefficients correctly identified as zero (denoted as “CorrZ0”);

(iii) the number of nonzero linear effects incorrectly identified as zero (denoted as “Zto0”);

(iv) the number of nonzero nonlinear effects correctly identified as nonzero (denoted as “CorrX”);

(v) the number of zero nonlinear functions correctly identified as zero (denoted as “CorrX0”);

(vi) the number of nonzero nonlinear functions incorrectly identified as zero (denoted as “Xto0”).
The simulation results are shown in A.2. From A.2, one sees that SBLL-AGLASSO performs very well regardless of the setting. SBLL-AGLASSO can effectively identify important linear and nonlinear components by using the proposed semiparametric model, while the traditional OLS-ALASSO fails to do so when the effects of some covariates are nonlinear. As for the selection of the nonlinear part, SBLL-AGLASSO can correctly and precisely identify the nonzero components, as the “CorrX”, “CorrX0” and “Xto0” are all close to those of the ORACLE* method; while OLS-ALASSO fails to detect true linear components and selects fewer relevant variables in all simulations. As for the linear part, the “CorrZ” and “CorrZ0” of SBLL-AGLASSO are higher than those of OLS-ALASSO. As for CV-MSPE, SBLL-AGLASSO performs better than OLS-ALASSO in all settings.

Table A.2: Selection and estimation results based on SNPs and RNA transcripts selected from real data analysis comparing the SBLL-AGLASSO, OLS-ALASSO and ORACLE* methods

<table>
<thead>
<tr>
<th>Noise</th>
<th>Method</th>
<th>Linear Part</th>
<th>Nonlinear Part</th>
<th>CV-</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>σ</td>
<td></td>
<td>CorrZ</td>
<td>CorrZ0</td>
<td>Zto0</td>
<td>CorrX</td>
</tr>
<tr>
<td>0.01</td>
<td>ORACLE*</td>
<td>189</td>
<td>5014</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>SBLL-AGLASSO</td>
<td>164.02</td>
<td>4934.25</td>
<td>24.98</td>
<td>3.61</td>
</tr>
<tr>
<td></td>
<td>OLS-ALASSO</td>
<td>144.64</td>
<td>4929.60</td>
<td>44.36</td>
<td>1.00</td>
</tr>
<tr>
<td>0.02</td>
<td>ORACLE*</td>
<td>189</td>
<td>5014</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>SBLL-AGLASSO</td>
<td>162.27</td>
<td>4926.40</td>
<td>26.73</td>
<td>3.60</td>
</tr>
<tr>
<td></td>
<td>OLS-ALASSO</td>
<td>144.19</td>
<td>4930.25</td>
<td>44.82</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Next, we compare the estimates and SCBs based on the SBLL-AGLASSO and ORACLE** methods for the seven RNA transcripts with nonzero functions. When the selection of SBLL-AGLASSO method is relatively accurate, which is the case for most of the replications in our simulation, the SBLL-AGLASSO estimates are very close to the ORACLE** estimates, and the coverage performance of the SBLL-AGLASSO method is also similar to that of the ORACLE**.
A.1 depicts the true function $\alpha_{0l}$, the SBLL-AGLASSO estimate $\hat{\alpha}_{l}^{\text{SBLL}}$ and 95% SCBs for $\alpha_{0l}$, $l = 1, \ldots, 4$, based on a typical run with $\sigma = 0.01$. From this figure, one sees that our estimates of the nonlinear components are very similar to those of the ORACLE** even when the selection is not perfectly correct. If the selection of the SBLL-AGLASSO is poor, the SBLL-AGLASSO estimators can be biased, and the simultaneous coverage performance of the SBLL-AGLASSO method is worse than that of the ORACLE**, though the SBLL-AGLASSO estimates can still capture the pattern of the true functions. For example, A.2 illustrates the worst scenario in our simulation, where SBLL-AGLASSO only selects 164 out of 189 SNPs with truly nonzero coefficients, and incorrectly picks an additional 68 SNPs that are unassociated with the response in the true model. Careful inspection reveals that only one of the SCBs produced by the SBLL-AGLASSO method fully cover their target function throughout the entire domain. However, even in this challenging scenario, the SBLL-AGLASSO function estimates seem to capture all the main features of each nonlinear function.

**Comparison with Existing Works**

In this section, we provide a summary of the differences among the papers in the literature related to our work; see A.3.

We differ from Wang and Yang (2007) in four aspects. Compared to Wang and Yang (2007), which focuses on the estimation and inference for only nonparametric functions, we study selection and estimation simultaneously for both linear and nonlinear components, and we further develop inferential tools for both linear and nonlinear parts. Meanwhile, for the nonparametric functions, Wang and Yang (2007) only obtained point-wise confidence intervals, while we provide simultaneous confidence bands. In addition, Wang and Yang (2007) assumed the number of nonparametric
Figure A.1: Plots of the true function (black solid curve), ORACLE** estimator (blue solid curve), SBLL-AGLASSO estimator (red solid curve), ORACLE** 95% SCB (upper and lower blue dashed curves) and SBLL-AGLASSO 95% SCB (upper and lower red dashed curves).
Figure A.2: Plots of the true function (black solid curve), ORACLE** estimator (blue solid curve), SBLL-AGLASSO estimator (red solid curve), ORACLE** 95% SCB (upper and lower blue dashed curves) and SBLL-AGLASSO 95% SCB (upper and lower red dashed curves).
Table A.3: Comparison with existing works in the literature.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Model</th>
<th>Dimension</th>
<th>Selection Focus</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhao and Xue (2009)</td>
<td>VCPLM</td>
<td>$p_1$: fixed, $p_2$: fixed</td>
<td>linear &amp; non-parametric</td>
<td>linear: selection consistency, asymptotic normality nonparametric: selection consistency, convergence rate</td>
</tr>
<tr>
<td>Wei and Huang (2010)</td>
<td>LM</td>
<td>$p_1 \gg n$</td>
<td>linear</td>
<td>selection consistency, convergence rate</td>
</tr>
<tr>
<td>Huang et al. (2010)</td>
<td>AM</td>
<td>$p_2 \gg n$</td>
<td>nonparametric</td>
<td>selection consistency, convergence rate</td>
</tr>
<tr>
<td>Wang et al. (2014)</td>
<td>GAPLM</td>
<td>$p_1 \ll n$, $p_2 \ll n$</td>
<td>linear &amp; non-parametric</td>
<td>linear: selection consistency, asymptotic normality nonparametric: selection consistency, convergence rate</td>
</tr>
<tr>
<td>Sherwood and Wang (2016)</td>
<td>APLM</td>
<td>$p_1 \gg n$, $p_2$: fixed</td>
<td>linear</td>
<td>linear: selection consistency, asymptotic normality nonparametric: selection consistency, convergence rate</td>
</tr>
<tr>
<td>Our paper</td>
<td>APLM</td>
<td>$p_1 \gg n$, $p_2 \gg n$</td>
<td>linear &amp; non-parametric</td>
<td>linear &amp; nonparametric: selection consistency, asymptotic normality simultaneous confidence band</td>
</tr>
</tbody>
</table>

AM: additive model; VCPLM: varying coefficient partially linear model; LM: linear model; (G)APLM: (generalized) additive partially linear model. $p_1$: the number of linear components; $p_2$: the number of nonparametric components.
functions is fixed and finite, while we allow the dimensions of linear and nonlinear components both to grow in the exponential order of sample size. Nevertheless, instead of piecewise constant splines used in Wang and Yang (2007), we adopt a higher degree of polynomial spline basis functions, which leads to a smaller bias and a faster convergence rate.

Compared with Wei and Huang (2010) and Huang et al. (2010), which involve either purely linear or nonparametric components, the establishment of the asymptotic normal distribution of the estimators for the parametric terms is quite challenging given that the numbers of covariates for both parametric and nonparametric terms grow in exponential orders of sample size, and the convergence rate for the nonparametric component estimators is slower than root-$n$. A difficulty here is that the covariates in the parametric components and those in the nonparametric components could be dependent. A significant innovation in our work is the establishment of the asymptotic distribution of estimators for the parameters. To resolve the dependence between the covariates in the nonparametric and parametric parts, we consider a projection of the covariates in the parametric part to the spline space generated by the spline basis for the covariates in the nonparametric part. We then study the properties of the projection. The ultra-high dimensionality raises challenging issues for the investigation of the projection properties.

Wang et al. (2014) studied the selection and estimation problems for generalized additive partially linear models, with the number of covariates smaller than the sample size. Our paper is motivated by the SAM project studied in the real data analysis section. To best predict some phenotypes of the SAM, we integrated two data sources in our analysis: the single nucleotide polymorphism (SNP) genotypes and messenger RNA transcript abundance levels from maize inbred lines. Following the preprocessing steps described in Section A, the dataset we analyze consists of log-scale SAM volume measurements, binary SNP genotypes at 5000 markers, and log-scale mea-
sures of abundance for 5000 transcripts for each of $n = 368$ maize inbred lines. Although hundreds of thousands of SNP genotypes and tens of thousands of RNA transcript abundance levels are included in our dataset, only a few of these covariates may have a non-negligible relationship with SAM volume. Thus, this is a potentially sparse and ultrahigh-dimensional regression problem. In our paper, we allow the numbers of linear covariates and nonlinear components greater than the sample size, and the methodology development in this setting is more complicated. In terms of theoretical achievement, Wang et al. (2014) obtained asymptotic normality for linear coefficients and convergence rate for nonlinear components. While our work not only achieves selection consistency, but also develops asymptotic normality for both linear and nonlinear parts as well as simultaneous confidence bands for the nonlinear components.

Our paper is very different from Sherwood and Wang (2016). Although Sherwood and Wang (2016) also studied the APLM and allow the number of linear covariates $p$ to be of a similar order of $n$ or much larger than $n$, the number of nonparametric components is assumed to be fixed. In addition, Sherwood and Wang (2016) only focused on the selection of the linear covariates, and they obtain local consistency for estimators. In our paper, we study model selection for both the parametric and nonparametric parts, we allow both the number of linear covariates $p_1 \gg n$ and the number of nonparametric components $p_2 \gg n$, and we obtain global consistency and simultaneous confidence bands for the nonparametric components.
APPENDIX B. SUPPLEMENTARY FOR “SPARSE MODEL IDENTIFICATION AND LEARNING FOR ULTRA-HIGH-DIMENSIONAL ADDITIVE PARTIALLY LINEAR MODELS”

Additional Material

Effect of Smoothing Parameters on Performance of SMILE

To implement the proposed SMILE procedure, one needs to select the knots for a spline at the selection stage and refitting stage, and the bandwidth for a kernel at the backfitting stage. In this section, we study how these smoothing parameters affect the proposed SMILE method and evaluate the practical performance in the finite-sample simulation studies described in Section 4.2 of the main paper.

In the literature of polynomial spline smoothing, the knots for a spline are generally put on a grid of equally spaced sample quantiles (Ruppert, 2002). Therefore, we only need to investigate the effect of the number of knots on the performance of SMILE.

At the first stage (model selection), we use piecewise constant splines with the number of interior knots $N = 2, 3, \ldots, 8$ in the simulation. Figure B.1 shows the effect of $N$ on the accuracy of model selection based on the criteria defined in the main paper: (B-i)–(B-vi) and (C-i)–(C-iv). From Figure B.1, it appears that the value $N$ has little effect on the selection results. For all combinations of $n$, $p$ and $\sigma$, no matter which $N$ is used, the “corrZ0”, “corrL”, “corrX0” are all 100%, and the “LtoN” and “Nto0” are all 0%. The values of “corrZ”, “corrN”, “corrLN” and “Zto0” and “Xto0” are not exactly the same when using different values of $N$, but they are almost constant for $N = 2, 3, \ldots, 8$. Especially when the sample size $n = 500$, the proposed SMILE is
able to identify the true model structure regardless of \( p = 1000, 2000 \) or 5000. When \( n = 300 \) and \( p = 5000 \), the selection results become slightly worse when we increase to \( N \geq 6 \).

In summary, the values of \( N \) often have little effect on the model selection results. Choosing small values of \( N \) can also help to reduce computational burden. So we recommend using fewer knots at the model selection stage, especially when the sample size is small compared to the number of predictors. In practice, \( N = 2 \sim 5 \) usually would be adequate to identify the model structure.

Next, we study the effect of the smoothing parameters at the refitting stage. For the selected model, we approximate the nonlinear functional components using higher order polynomial splines to obtain more accurate pilot estimators. Then we apply spline backfitted local linear smoothing to obtain the final SBLL estimators and the corresponding SCBs. According to Assumption (A6'), to obtain the SCB with the desired confidence level, the number of interior knots \( M_n \) for a refitting spline needs to satisfy:

\[
\{ n^{1/(2d)} \land n^{4/(10d-5)} \} \ll M_n \ll n^{1/3},
\]

where \( d \) is the degree of the polynomial spline basis functions used in the refitting. The widely used quadratic/ cubic splines and any polynomial splines of degree \( d \geq 2 \) all satisfy this condition. Therefore, in practice we suggest choosing

\[
M_n = \min\{ \lceil n^{1/(2d)} \lceil n^{4/(10d-5)} \rceil \land \lceil n/(4s) \rceil \}, \lceil n/(4s) \rceil \} + 1, \tag{C.1}
\]

where \( s \) is the number of nonlinear components selected at the first stage and the term \( \lceil n/(4s) \rceil \) is to guarantee that we have at least four observations in each subinterval between two adjacent knots to avoid getting (near) singular design matrices in the spline smoothing. A researcher with some knowledge of the shape of the nonlinear component may be able to select a more suitable number of knots. In our simulation studies, we try 4, 6 and 8 interior knots to test the sensitivity of the SBLL estimators and the corresponding SCBs.
Figure B.1: First stage selection results using different number of knots.
For the local linear smoothing in the backfitting, Condition (B2) requires that the bandwidths are of order $n^{-1/5}$. Any bandwidths with this rate lead to the same limiting distribution for $\hat{\phi}_{SBLL}^l$, so the user can consider any standard routine for bandwidth selection. There have been many proposals for bandwidth selection in the literature. In our simulation, we consider three popular bandwidth selectors described in Fan and Gijbels (1996) and Wand and Jones (1995): rule-of-thumb bandwidth (“thumbBw”), plug-in bandwidth selector (“pluginBw”) and leave-one-out cross-validation bandwidth selector (“regCVBwSelC”). Below we present simulation results to compare the performance of three bandwidth selectors. The kernel that we use here is the Epanechnikov kernel: $K(u) = \frac{3}{4}(1-u^2)I(|u| \leq 1)$.

To see how the refitting smoothing parameters affect estimation accuracy, we report the average mean square errors (AMSEs) of the SBLL estimators based on 4, 6 and 8 interior knots in the spline refitting and three different bandwidth selectors in the kernel backfitting. Figure B.2 presents the AMSEs of the resulting SBLL estimators based on different combinations of the refitting smoothing parameters. For both $\phi_1$ and $\phi_2$, the AMSEs are very similar across the different combinations of knots and bandwidth selectors.

Figure B.3 shows the coverage rates of the SCBs based on different combinations of knots and bandwidth selectors. From Figure B.3, it is clear that the number of knots for a spline in the refitting has very little effect on the coverage of the SCBs. One also observes that the performances of the SCBs based on different smoothing parameters become more similar with increasing sample size, whereas the coverage rates of the SCBs using the “thumbBw” selector are the closest to the nominal level in all the simulation settings. Thus we recommend the “thumbBw” selector, especially when the sample size is small.
Figure B.2: Average mean squared errors (AMSEs) of the SBLL estimators of $\phi_2$ and $\phi_3$. 
Figure B.3: Coverage rates of the SCBs for $\phi_2$ and $\phi_3$. 
Technical Details

This section contains some technical assumptions, lemmas and proofs. For any real numbers \(a\) and \(b\), let \(a \lor b\) and \(a \land b\) denote the maximum and minimum of \(a\) and \(b\), respectively. For any two sequences \(\{a_n\}, \{b_n\}, n = 1, 2, \ldots\), we use \(a_n \asymp b_n\) if there are constants \(0 < c_1 < c_2 < \infty\) such that \(c_1 < a_n/b_n < c_2\) for all \(n\) sufficiently large. On any fixed interval \([a, b]\), we denote the space of the second order smooth functions as \(C^{(d)}[a, b] = \{f \mid f^{(d)} \in C[a, b]\}\) and the class of Lipschitz continuous functions for any fixed constant \(C > 0\) as \(\text{Lip}([a, b], C) = \{f \mid \|f(x) - f(x')\| \leq C |x - x'|, \forall x, x' \in [a, b]\}\).

Furthermore, let \(Y = (Y_1, \ldots, Y_n)^\top\) be an \(n\)-dimensional vector, \(Z = (Z_1, \ldots, Z_{p_1})\) be an \(n \times p_1\) matrix, where \(Z_k = (Z_{1k}, \ldots, Z_{nk})^\top, k = 1, \ldots, p_1,\) and \(X = (X_1, \ldots, X_{p_2})\) be an \(n \times p_2\) matrix, where \(X_l = (X_{1l}, \ldots, X_{nl})^\top, l = 1, \ldots, p_2.\) Let \(B^{(d)} = (B^{(d)}_1, \ldots, B^{(d)}_{p_2})\) be a dimension \(n \times (p_2M_n)\) matrix, where \(B^{(d)}_l = (B^{(d)}_{1l}(X_{1l}), \ldots, B^{(d)}_{nl}(X_{nl}))^\top\) is a dimension \(n \times M_n\) matrix of spline basis functions of order \(d\), for \(l = 1, \ldots, p_2.\)

Let \(A \subseteq \{1, \ldots, p_1 + 2p_2\}\) be an index set, and let \(|A|\) denote the cardinality of set \(A.\)

Technical Assumptions

We state the regularity conditions needed for the theoretical results in this paper.

(A1) (Conditions on model) The numbers of nonzero components \(|S_x|, |S_{x,L}|\) and \(|S_{x,N}|\) are fixed, and there exist constants \(c_\alpha > 0, c_\beta > 0\) and \(c_g > 0\) such that \(\min_{k \in S_x} |\alpha_{0k}| \geq c_\alpha, \min_{l \in S_{x,L}} |\beta_{0l}| \geq c_\beta,\) and \(\min_{l \in S_{x,N}} \|g_{0l}\|_2 \geq c_g.\)
(A2) (Conditions on errors) The errors \( \varepsilon_1, \ldots, \varepsilon_n \) are independent and identically distributed with 

\[
E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2, \quad E|\varepsilon_i|^{2+\delta} \leq M_\delta \quad (\delta > 0.5), \quad \text{and have}
\]

\( b \)-sub-gaussian tails, i.e., \( E\{\exp(t\varepsilon)\} \leq \exp(b^2t^2/2) \), for any \( t \geq 0 \) and some \( b > 0 \).

(A3) (Conditions on nonlinear functions) The additive component function \( g_l(\cdot) \in C^{(2)}[a, b] \), \( l = 1, \ldots, p_2 \).

(A4) (Conditions on covariates) Each covariate in the parametric part of the model has a bounded second moment, that is, there is a positive constant \( C_3 \) such that \( E(Z_k)^2 \leq C_3^2 \), \( 1 \leq k \leq p_1 \); also, \( E(X_l) = 0 \), and there is a positive constant \( C_4 \) such that \( E(X_l)^2 \leq C_4^2 \), \( 1 \leq l \leq p_2 \). The joint density function of active pure linear \( X \) is continuous and bounded below and above.

Each covariate in the nonparametric part of the model has a continuous density and there exist constants \( C_1 \) and \( C_2 \) such that the marginal density function \( f_l \) of \( X_l \) has continuous derivatives on its support, and satisfies \( 0 < C_1 \leq f_l(x_l) \leq C_2 < \infty \) on its support for every \( 1 \leq l \leq p_2 \). In addition, the eigenvalues of \( E\{(ZZ^\top)X\} \) are bounded away from 0.

(A5) (Conditions on the initial estimators) The initial estimators satisfy 

\[
r_{n1} \max_{k \in N_x} |\tilde{\alpha}_k| = O_p(1), \quad r_{n2} \max_{l \in N_x} |\tilde{\beta}_l| = O_p(1), \quad r_{n3} \max_{l \in N_x} \|\tilde{\gamma}_l\|_2 = O_p(1), \quad r_{n1}, r_{n2}, r_{n3} \to \infty, \quad \text{and there exist positive}
\]

constants \( c_{b1}, c_{b2} \) and \( c_{b3} \) such that \( \Pr\left(\min_{k \in S_x} |\tilde{\alpha}_k| \geq c_{b1} b_{n1}\right) \to 1, \quad \Pr\left(\min_{l \in S_{x,L}} |\tilde{\beta}_l| \geq c_{b2} b_{n2}\right) \to 1, \quad \Pr\left(\min_{l \in S_{x,N}} \|\tilde{\gamma}_l\|_2 \geq c_{b3} b_{n3}\right) \to 1, \quad \text{where} \quad b_{n1} = \min_{k \in S_x} |\alpha_{0k}|, \quad b_{n2} = \min_{l \in S_{x,L}} |\beta_{0l}|, \quad \text{and} \quad b_{n3} = \min_{l \in S_{x,N}} \|g_0\|_2.
\]

(A6) (Conditions on parameters and spline basis functions) Let \( p_1 \) and \( p_2 \) be the number of linear and nonlinear components, respectively. Suppose that 

\[
n^{-1}N_n + n^{-2} \sum_{j=1}^{3} \lambda_{nj}^2 = o(1),
\]
\[ \frac{\sqrt{n \log(p_1)}}{\lambda_n r_{n1}} + \frac{\sqrt{n \log(p_2)}}{\lambda_n r_{n2}} + \frac{\sqrt{n N_n \log(p_2 N_n)}}{\lambda_n r_{n3}} + \sum_{j=1}^{3} \frac{n}{\lambda_n r_{nj} N_n} = o(1). \]

Assumptions (A1) – (A4) are regularity conditions that are commonly used in the APLM literature. To obtain the selection consistency of the SBLL-AGLASSO, we need an order requirement for a general initial estimator; see Assumption (A5). Theorem D.1 below demonstrates that the group LASSO estimator defined in (3.6) satisfies Assumption (A5) under some weak conditions, specifically if \( \sum_{j=1}^{3} \lambda_n^2 \asymp n \{ \log(p_1) \lor N_n \log(p_2 N_n) \} \) and \( N_n \asymp n^{1/3} \), then the consistent rates for the group LASSO estimator in (A5) have order \( r_{n1} \asymp r_{n2} \asymp r_{n3} = O\{n^{1/2}/\sqrt{\log(p_1) \lor N_n \log(p_2 N_n)}\} \).

Consequently, Assumption (A6) is equivalent to:
\[
\frac{\sum_{j=1}^{3} \lambda_n^2}{n^2} + \frac{\log(p_1) \lor N_n \log(p_2 N_n)}{(\lambda_n \land \lambda_n \land \lambda_n)} + \frac{n^{1/6} \sqrt{\log(p_1) \lor N_n \log(p_2 N_n)}}{(\lambda_n \land \lambda_n \land \lambda_n)} = o(1), \tag{D.1}
\]

If we take \( \lambda_n \asymp \lambda_n \asymp \lambda_n = O(n^{1/2}) \), then (D.1) indicates \( p_1 = \exp\{o(n^{1/2})\} \) and \( p_2 = \exp\{o(n^{1/6})\} \).

We need the following additional assumptions in order to develop the asymptotic SCBs for the nonparametric components.

(A2') (Conditions on errors) The conditional distribution of \( \varepsilon \) given \( (Z, X) \) is \( N(0, \sigma^2) \).

(A3') (Conditions on nonlinear functions) For any \( l \in S_{x,N}, \phi_0 l \in C^{(d)}[a, b] \), for some integer \( d \geq 2 \).

In addition, \( \psi_l^\gamma \) defined in (3.12) satisfies \( \psi_l^\gamma \in C^{(d)}[a, b] \).

(A6') (Conditions on spline basis functions) The order of the spline basis functions is at least \( d \), and the number of interior knots \( M_n \) satisfies: \( \{n^{1/(2d)} \lor n^{4/(10d-5)}\} \ll M_n \ll n^{1/3} \).

(B1) (Conditions on the kernel function) The kernel function \( K \in \text{Lip}([-1, 1], C_K) \) for some constant \( C_K > 0 \), and is bounded, nonnegative, symmetric, and supported on \([-1, 1]\) with the second moment \( \mu_2(K) = \int u^2 K(u) \, du \).
(B2) (Conditions on bandwidth) The bandwidth of the kernel $K$ is $h \sim n^{-1/5}$, i.e., $c_h n^{-1/5} \leq h \leq C_h n^{-1/5}$ for some positive constants $C_h$, $c_h$.

Assumptions (A2'), (A3'), (B1) and (B2) are typical in the local polynomial smoothing literature; see, for instance, Wang and Yang (2009). Assumption (A6') imposes the condition of the number of knots for spline smoothing. For example, if $d = 2$, we can take $M_n \sim n^{4/15} \log n$.

**Selection and estimation properties of the group LASSO estimators**

In this section, we consider the selection and estimation properties of the group LASSO estimator $\hat{\theta} = (\alpha^\top, \beta^\top, \gamma^\top)^\top$ in (3.6).

In the following, denote $\alpha = (\alpha_1, \ldots, \alpha_{p_1})^\top$ with length $p_1$, $\beta = (\beta_1, \ldots, \beta_{p_2})^\top$ with length $p_2$, and $\gamma = (\gamma_{1\top}, \ldots, \gamma_{p_2\top})^\top$ with length $(p_2 N_n)$. Let

$$\theta^\top = (\alpha^\top, \beta^\top, \gamma^\top) = (\alpha_1, \ldots, \alpha_{p_1}, \beta_1, \ldots, \beta_{p_2}, \gamma_{1\top}, \ldots, \gamma_{p_2\top}) = (\theta_1^\top, \ldots, \theta_{m}^\top, \ldots, \theta_{p_1+2p_2}^\top),$$

where $\theta_m = \alpha_m I\{1 \leq m \leq p_1\} + \beta_{m-p_1} I\{p_1 + 1 \leq m \leq p_1 + p_2\} + \gamma_{m-p_1-p_2} I\{p_1 + p_2 + 1 \leq m \leq p_1 + 2p_2\}$, with $I(\cdot)$ being an indicator function. Let

$$D = (Z_1, \ldots, Z_{p_1}, X_1, \ldots, X_{p_2}, B_{1}^{(1)}, \ldots, B_{p_2}^{(1)}) \equiv (D_1, \ldots, D_m, \ldots, D_{p_1+2p_2})$$

be an $n \times (p_1 + p_2 + p_2 N_n)$ matrix, where

$$D_m = Z_m I\{1 \leq m \leq p_1\} + X_m I\{1 \leq m \leq p_1\} + B_{m-p_1-p_2}^{(1)} I\{p_1 + p_2 + 1 \leq m \leq p_1 + 2p_2\},$$

an $n \times d_m$ submatrix of $D$ with $d_m = I(1 \leq m \leq p_1) + I(p_1 + 1 \leq m \leq p_1 + p_2) + N_n I(p_1 + p_2 + 1 \leq m \leq p_1 + 2p_2)$.

Define

$$\bar{S} = \{m : \|\tilde{\theta}_m\| \neq 0, 1 \leq m \leq p_1 + 2p_2\}. \quad (D.2)$$
Next we define the active linear index set for \( X \) as \( S_{x,L} = S_{x,PL} \cup S_{x,LN} \), the inactive linear index set for \( X \) as \( N_{x,L} \), and the inactive nonlinear index set for \( X \) as \( N_{x,N} \). Note that \( N_x = N_{x,L} \cap N_{x,N} \).

Further, let
\[
S = S_z \cup \{l + p_1 : l \in S_{x,L}\} \cup \{l + p_1 + p_2 : l \in S_{x,N}\},
\]
\[
N = N_z \cup \{l + p_1 : l \in N_{x,L}\} \cup \{l + p_1 + p_2 : l \in N_{x,N}\}.
\]

For any index set \( A \subseteq \{1, \ldots, p_1 + 2p_2\} \), define \( D_A = \{D_m : m \in A\} \). Next denote \( C_A = n^{-1}D_A^\top D_A \), and let \( \pi_{\min}(C_A) \) and \( \pi_{\max}(C_A) \) represent the minimum and maximum eigenvalues of \( C_A \), respectively.

**Lemma D.1.** Let \( N_n = O(n^\gamma) \), where \( 0 < \gamma < 0.5 \). Suppose that \( |A| \) is bounded by a fixed constant independent of \( n \), \( p_1 \) and \( p_2 \). Then under Assumption (A4), with probability approaching one as \( n \to \infty \), \( c_1 \leq \pi_{\min}(C_A) \leq \pi_{\max}(C_A) \leq c_2 \), where \( c_1 \) and \( c_2 \) are two positive constants.

**Proof.** Similar to the proof of Lemma A.1 in Li et al. (2018).

**Lemma D.2.** Under Assumption (A3), there exists a vector \( \gamma_0 = (\gamma_0^\top, \ldots, \gamma_{0p_2}^\top)^\top \), such that \( \|\gamma_0\| \neq 0 \), for \( l \in S_{x,N} \), \( \|\gamma_0l\| = 0 \), \( l \in N_{x,N} \) and \( \|g_{0l} - B_l^{(d)^\top} \gamma_0l\|_2 = O(M_n^{-d}) \).

**Proof.** Similar to the proof of Lemma A.2 in Li et al. (2018).

In the following, we denote \( g_{nl}(\cdot) = \sum_{J=1}^{N_n} \gamma_{0Jl} B_{Jl}^{(1)}(\cdot) \) the best constant spline approximation of \( g_{0l}(\cdot) \) such that \( \|g_{0l} - g_{nl}\|_\infty = \sup_{x \in [a,b]} |g_{0l}(x) - g_{nl}(x)| = O(N_n^{-1}) \). Let \( \gamma_0 = (\gamma_{0J}, J = 1, \ldots, N_n)^\top \) be the vector of the coefficients of the best spline approximation in Lemma D.2. Denote
\[
\theta_0^\top = \left(\theta_{01}^\top, \ldots, \theta_{0m}^\top, \ldots, \theta_{0,p_1+2p_2}^\top\right),
\]
where
\[
\theta_0^\top = (\alpha_0^\top, \beta_0^\top, \gamma_0^\top) = (\alpha_{01}, \ldots, \alpha_{0p_1}, \beta_{01}, \ldots, \beta_{0p_2}, \gamma_{01}, \ldots, \gamma_{0p_2}).
\]
Define $\theta_A = (\theta_m^T : m \in A)^T$, $\theta_{0,A} = (\theta_{0m}^T : m \in A)^T$ and $\bar{\theta}_A = (\bar{\theta}_m^T : m \in A)^T$.

**Theorem D.1.** Suppose that Assumptions (A1) – (A4) hold.

(i) If $\{\log(p_1) \vee N_n \log(p_2 N_n)\}/n \to 0$ and $n^{-2} \sum_{j=1}^{3} \bar{\lambda}_{nj}^2 \to 0$ as $n \to \infty$, then with probability converging to one, all the nonzero linear parameters $\alpha_k$ and $\beta_{0l}$, $k \in S_z$, $l \in S_{x,L}$, and nonzero additive components $g_{0l}$, $l \in S_{x,N}$, are selected.

(ii) In addition,

$$\sum_{k=1}^{p_1} |\bar{\alpha}_k - \alpha_{0k}|_2^2 = O_p \left( \frac{\log(p_1) \vee N_n \log(p_2 N_n)}{n} \right) + O \left( N_n^{-2} + O \left( n^{-2} \sum_{j=1}^{3} \bar{\lambda}_{nj}^2 \right) \right),$$

$$\sum_{l=1}^{p_2} |\bar{\beta}_l - \beta_{0l}|^2 = O_p \left( \frac{\log(p_1) \vee N_n \log(p_2 N_n)}{n} \right) + O \left( N_n^{-2} + O \left( n^{-2} \sum_{j=1}^{3} \bar{\lambda}_{nj}^2 \right) \right),$$

$$\sum_{l=1}^{p_2} \|\bar{g}_l - g_{0l}\|^2 = O_p \left( \frac{\log(p_1) \vee N_n \log(p_2 N_n)}{n} \right) + O \left( N_n^{-2} + O \left( n^{-2} \sum_{j=1}^{3} \bar{\lambda}_{nj}^2 \right) \right).$$

**Proof.** We prove part (ii) first. Let $\bar{\theta} = (\bar{\theta}_1^T, \ldots, \bar{\theta}_{p_1+2p_2}^T) = (\bar{\alpha}_1, \ldots, \bar{\alpha}_{p_1}, \bar{\beta}_1, \ldots, \bar{\beta}_{p_2}, \bar{\gamma}_1, \ldots, \bar{\gamma}_{p_2})$. For $\mathcal{S}$ defined in (D.3) and $\bar{\mathcal{S}}$ defined in (D.2), denote $\mathcal{S}' = \mathcal{S} \cup \bar{\mathcal{S}} = \{m : \|\theta_{0m}\|_2 \neq 0 \text{ or } \|\bar{\theta}_m\|_2 \neq 0\}$ and $d' = |\mathcal{S}'|$. By Lemma D.5, $d' = O(|\mathcal{S}|)$. Notice that $\mathbf{D}\bar{\theta} = \mathbf{D}_{\mathcal{S}}\bar{\theta}_{\mathcal{S}'}$ and $\mathbf{D}\theta_0 = \mathbf{D}_{\mathcal{S}}\theta_{0,\mathcal{S}'}$, by the definition of $\bar{\theta}$ and $\mathcal{S}'$.

$$\|\mathbf{Y} - \mathbf{D}_{\mathcal{S}}\bar{\theta}_{\mathcal{S}'}\|^2 - \|\mathbf{Y} - \mathbf{D}_{\mathcal{S}}\theta_{0,\mathcal{S}'}\|^2$$

$$\leq \sum_{m \in \mathcal{S}'} \{\bar{\lambda}_{n1} I(m \leq p_1) + \bar{\lambda}_{n2} I(p_1 < m \leq p_1 + p_2) + \bar{\lambda}_{n3} I(m > p_1 + p_2)\} \|\theta_{0m}\|$$

$$- \sum_{m \in \mathcal{S}'} \{\bar{\lambda}_{n1} I(m \leq p_1) + \bar{\lambda}_{n2} I(p_1 < m \leq p_1 + p_2) + \bar{\lambda}_{n3} I(m > p_1 + p_2)\} \|\bar{\theta}_m\|.$$
inequality,

\[
\|\nu\|^2 - 2\eta^\top \nu \\
\leq \sum_{m \in S'} \{\bar{\lambda}_{n1} I(m \leq p_1) + \bar{\lambda}_{n2} I(p_1 < m \leq p_1 + p_2) + \bar{\lambda}_{n3} I(m > p_1 + p_2)\}(\|\theta_{0m} - \bar{\theta}_m\|)
\]
\[
\leq \sqrt{d' \sum_{j=1}^{3} \tilde{\lambda}_{nj}^2 \|\bar{\theta}_{S'} - \theta_{0,S'}\|} \leq \frac{d' \sum_{j=1}^{3} \tilde{\lambda}_{nj}^2}{nc_*} + \frac{1}{4} nc_* \|\bar{\theta}_{S'} - \theta_{0,S'}\|^2, \quad (D.4)
\]

where \(c_*\) is the lower bound of eigenvalues of \(n^{-1}D_{S'}^\top D_{S'}\). By LemmaD.1 and Lemma D.5, \(c_* \approx 1\) with probability approaching one. Apparently,

\[
\|\nu\|^2 \geq nc_* \|\bar{\theta}_{S'} - \theta_{0,S'}\|^2. \quad (D.5)
\]

Define \(\eta^* \equiv D_{S'}(D_{S'}^\top D_{S'})^{-1}D_{S'}^\top \eta\) to be the projection of \(\eta\) onto the column space of \(D_{S'}\). Obviously, \(\eta^\top \nu = \eta^* \nu\). By the Cauchy-Schwartz inequality, we have

\[
2|\eta^\top \nu| \leq 2\|\eta^*\|\|\nu\| \leq 2\|\eta^*\|^2 + \frac{1}{2} \|\nu\|^2. \quad (D.6)
\]

Combining (D.4), (D.5) and (D.6), we obtain

\[
\|\bar{\theta}_{S'} - \theta_{0,S'}\|^2 \leq \frac{8\|\eta^*\|^2}{nc_*} + \frac{4d' \sum_{j=1}^{3} \tilde{\lambda}_{nj}^2}{n^2 c_*^2}. \quad (D.7)
\]

With \(\eta_i\) defined to be the ith element of \(\eta\), we have the following decomposition:

\[
\eta_i = Y_i - \sum_{k=1}^{p_1} Z_{ik} \alpha_{0k} - \sum_{l=1}^{p_2} X_{il} \beta_{0l} - \sum_{l=1}^{p_2} \sum_{J=1}^{N_n} \gamma_{0,J,l} B^{(1)}_{J,l}(X_{il})
\]
\[
= Y_i - \sum_{k \in S_x} Z_{ik} \alpha_{0k} - \sum_{l \in S_{x,L}} X_{il} \beta_{0l} - \sum_{l \in S_{x,N}} g_0(X_{il}) + \sum_{l \in S_{x,N}} g_0(X_{il}) - \sum_{l \in S_{x,N}} \sum_{J=1}^{N_n} \gamma_{0,J,l} B^{(1)}_{J,l}(X_{il})
\]
\[
= \varepsilon_i + \sum_{l \in S_{x,N}} \delta_{il}, \quad (D.8)
\]
where \( \delta_{il} = g_{il}(X_{il}) - \sum_{j=1}^{N_{l}} \gamma_{ij,l} B_{jl}^{(l)}(X_{il}) \). Let \( \delta_i = \sum_{l \in S_{x,N}} \delta_{il} \), \( \delta = (\delta_1, \ldots, \delta_n)^{\top} \), and \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\top} \). Then \( \eta = \varepsilon + \delta \). Define \( \delta_{S'} = (\sum_{1 \leq l \leq p_2} \delta_{il}, i = 1, \ldots, n)^{\top} \). By (D.8) and the fact that \( |\delta_{il}| = O_p(N_{n}^{-1}) \),

\[
\|\eta^*\|^2 = \|\varepsilon^* + \delta_{S'}^\top\|^2 \leq 2\|\varepsilon^*\|^2 + 2\|\delta_{S'}\|^2 \leq 2\|\varepsilon^*\|^2 + O_p(nd^2N_n^{-2}), \tag{D.9}
\]

where \( \varepsilon^* \equiv P_{D_{S'}}\varepsilon \) and \( \delta_{S'}^\top \equiv P_{D_{S'}}^\top \delta_{S'} \) are the projections of \( \varepsilon \) and \( \delta_{S'} \) onto the column space of \( D_{S'} \), respectively. Define \( T_1 = \max_{1 \leq k \leq p_1} |n^{-1/2} \sum_{i=1}^{n} Z_i \varepsilon_i| \), \( T_2 = \max_{1 \leq l \leq p_2} |n^{-1/2} \sum_{i=1}^{n} X_i \varepsilon_i| \), and \( T_3 = \max_{1 \leq l \leq p_2, 1 \leq j \leq N_n} |n^{-1/2} \sum_{i=1}^{n} B_{jl}^{(l)}(X_{il}) \varepsilon_i| \). Then we have

\[
\|\varepsilon^*\|^2 = \| \left( D_{S'}^\top D_{S'} \right)^{-1/2} D_{S'}^\top \varepsilon \|^2 \leq \frac{1}{nc_x} \| D_{S'}^\top \varepsilon \|^2,
\]

and

\[
\max_{A; |A| \leq d'} \| D_A^{\top} \varepsilon \|^2 = \max_{A; |A| \leq d'} \sum_{m \in A} \| D_m^{\top} \varepsilon \|^2 \leq nd'(T_1^2 \vee T_2^2 \vee N_nT_3^2).
\]

By Lemma D.6, \( \max_{A; |A| \leq d'} \| D_A^{\top} \varepsilon \|^2 = O_p[n(d' \log(p_1) \vee N_n \log(p_2N_n))] \). Therefore,

\[
\|\varepsilon^*\|^2 = O_p[d'c_x^{-1}\{ \log(p_1) \vee N_n \log(p_2N_n) \}]. \tag{D.10}
\]

Combing (D.7), (D.9) and (D.10), we conclude that

\[
\| \tilde{\theta}_{S'} - \theta_{0,S'} \|^2 = O_p \left[ d' \left\{ \log(p_1) \vee N_n \log(p_2N_n) \right\} \right] + O \left( \frac{d' \sum_{j=1}^{p_2} \lambda_{n,j}^2}{n^2c_x^2} \right) + \frac{4d' \sum_{j=1}^{p_2} \lambda_{n,j}^2}{n^2c_x^2} + O \left( n^{-2} \sum_{j=1}^{3} \lambda_{n,j}^2 \right),
\]

where the last inequality follows by \( d' = O(\|S_x| + |S_{x,L}| + |S_{x,N}|) \) and \( c_x \propto 1 \) with probability approaching one. By the properties of splines (de Boor, 2001), \( \| \tilde{g}_l - g_{il} \|^2 \propto \| \tilde{\gamma}_l - \gamma_{0l} \|^2 \), where \( g_{il}, l = 1, \ldots, p_2 \), is the best approximation for function \( g_l \). Hence, part (ii) follows from \( \sum_{k=1}^{p_1} |\tilde{\alpha}_{k} - \alpha_{0k}|^2 = O(\| \tilde{\theta}_{S'} - \theta_{0,S'} \|^2), \sum_{k=1}^{p_2} |\tilde{\beta}_l - \beta_{0l}|^2 = O(\| \tilde{\theta}_{S'} - \theta_{0,S'} \|^2) \) and \( \sum_{l=1}^{p_2} \| \tilde{\gamma}_l - \gamma_{0l} \|^2 = O(\| \tilde{\theta}_{S'} - \theta_{0,S'} \|^2) \).
We now prove part (i). Under Assumption (A1), if $\|\theta_{0m}\| \neq 0$ but $\|\tilde{\theta}_m\| = 0$, then $\|\theta_{0m} - \tilde{\theta}_m\| \geq c_\alpha \vee c_\beta \vee c_\gamma$, which contradicts part (ii) when $\log(p_1) \vee N_n \log(p_2 N_n)/n \to 0$, $\tilde{\lambda}_{n1}^2/n^2 \to 0$, $\tilde{\lambda}_{n2}^2/n^2 \to 0$ and $\tilde{\lambda}_{n3}^2/n^2 \to 0$. The results follow by

$$\tilde{\alpha} - \alpha_0 = \left( I_{|S_x|} \ 0_{|S_x| \times |S_{x,L}|} \ 0_{|S_x| \times (|S_{x,N}|,N)} \right) (\tilde{\theta} - \theta_0),$$

$$\tilde{\beta} - \beta_0 = \left( 0_{|S_{x,L}| \times |S_x|} \ I_{|S_{x,L}|} \ 0_{|S_{x,L}| \times (|S_{x,N}|,N)} \right) (\tilde{\theta} - \theta_0),$$

$$\tilde{\gamma} - \gamma_0 = \left( 0_{(|S_x,N)|N_n} \times |S_x| \ 0_{(|S_x,N)|N_n \times |S_{x,L}|} \ I_{|S_{x,N}|N_n} \right) (\tilde{\theta} - \theta_0)$$

and the definition of $g_l$, $1 \leq l \leq p_2$. □

Selection and estimation properties of the adaptive group LASSO estimators

In this section, we establish the selection and estimation properties of the adaptive group LASSO estimators as stated in Theorems 3.1 and 3.2.

Proof of Theorem 3.1. By the Karush-Kuhn-Tucker (KKT) condition (Boyd and Vandenberghe, 2004), if $(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})$ is the unique minimizer of $L(\alpha, \beta, \gamma; \lambda_1, \lambda_2, \lambda_3)$, it is equivalent to satisfy

(C1-1) $Z_k^T \left( Y - Z\alpha - X\beta - \sum_{l'=1}^{p_2} B_{l'}^{(1)} \gamma_{l'} \right) = \lambda_{n1} w_k^\alpha \alpha_k / |\alpha_k|$, for any $k \in S_x$,

(C1-2) $X_l^T \left( Y - Z\alpha - X\beta - \sum_{l'=1}^{p_2} B_{l'}^{(1)} \gamma_{l'} \right) = \lambda_{n2} w_l^\beta \beta_l / |\beta_l|$, for any $l \in S_{x,L}$,

(C1-3) $B_l^{(1)^T} \left( Y - Z\alpha - X\beta - \sum_{l'=1}^{p_2} B_{l'}^{(1)} \gamma_{l'} \right) = \lambda_{n3} w_l^\gamma \gamma_l / \|\gamma_l\|$, for any $l \in S_{x,N}$,

(C2) $|Z_k^T \left( Y - Z\alpha - X\beta - \sum_{l'=1}^{p_2} B_{l'}^{(1)} \gamma_{l'} \right)| \leq \lambda_{n1} w_k^\alpha$, for any $k \in N_z$,

(C3) $|X_l^T \left( Y - Z\alpha - X\beta - \sum_{l'=1}^{p_2} B_{l'}^{(1)} \gamma_{l'} \right)| \leq \lambda_{n2} w_l^\beta$, for any $l \in N_x \cup S_{x,PN}$,

(C4) $\|B_l^{(1)^T} \left( Y - Z\alpha - X\beta - \sum_{l'=1}^{p_2} B_{l'}^{(1)} \gamma_{l'} \right) \| \leq \lambda_{n3} w_l^\gamma$, for any $l \in N_x \cup S_{x,PL}$,
Define $\vartheta^o = (\mathbf{D}_S^\top \mathbf{D}_S)^{-1} \mathbf{D}_S^\top \mathbf{Y}$, a vector with length $|S_z| + |S_{x,L}| + |S_{x,N}|N_n$. Denote three vectors, $v_1$, $v_2$ and $v_3$, whose elements are in the form:

$$v_{1m} = \frac{\omega^o_m \bar{\theta}_0 m I\{m \in S_z\} + 0_N I\{m - |S_z| - |S_{x,LN}| \in S_{x,N}\}}{|\bar{\theta}_0 m|},$$

$$v_{2m} = \frac{\omega^o_m - |S_z| \bar{\theta}_0 m I\{m - |S_z| - |S_{x,LN}| \in S_{x,N}\}}{|\bar{\theta}_0 m|},$$

$$v_{3m} = \frac{\omega^o_m - |S_z| - |S_{x,L}| \bar{\theta}_0 m I\{m - |S_z| - |S_{x,LN}| \in S_{x,N}\}}{|\bar{\theta}_0 m|}, \forall m \in S.$$

Next define $\hat{\theta}^o = (\theta_m^o, 1 \leq m \leq p_1 + 2p_2)^\top$, where $\hat{\theta}^o_S \equiv (\theta_m^o, m \in S)^\top = (\mathbf{D}_S^\top \mathbf{D}_S)^{-1} (\mathbf{D}_S^\top \mathbf{Y} - \sum_{j=1}^3 \lambda_{nj} v_j), \bar{\theta}_0^o = 0$ for $m \in N_z$ and $m - p_1 \in N_{x,L}$, and $\theta_m^o = 0_N$ for $m - p_1 - p_2 \in N_{x,N}$. So we can represent $\hat{\theta}^o \equiv (\hat{\theta}^o_{S_z}, \hat{\theta}^o_{N_z}, \hat{\theta}^o_{N_{x,L}, \hat{\theta}^o_{S_{x,L}}}, \hat{\theta}^o_{S_{x,N}}, \hat{\theta}^o_{N_{x,N}})^\top$, and $\hat{\theta}^o_S \equiv (\hat{\theta}^o_{S_z}, \hat{\theta}^o_{S_{x,L}}, \hat{\theta}^o_{S_{x,N}})^\top$. Denote $\hat{S}^o = \{1 \leq m \leq p_1 + 2p_2 : \|\theta_m^o\| > 0\}$. Apparently, $\hat{S}^o \subseteq S$. Notice that $D \hat{\theta}^o = D_S \hat{\theta}^o_S$ and $\{D_m, m \in S\}$ are linearly independent, so by the definition of $\hat{\theta}^o$, (C1-1), (C1-2) and (C1-3) hold for $\hat{\theta}^o$ if $\hat{S}^o \supseteq S$. Therefore, if $\hat{\theta}^o$ satisfies

(C1') $\hat{S}^o \supseteq S$,

(C2') $|Z_k^\top (\mathbf{Y} - D \tilde{\theta}^o)| \leq \lambda_{n1} \omega_k^o$, for any $k \in N_z$,

(C3') $\|X_l^\top (\mathbf{Y} - D \tilde{\theta}^o)\| \leq \lambda_{n2} \omega_l^o$, for any $l \in N_{x,L}$,

(C4') $\|B_l^{(1)} (\mathbf{Y} - D \tilde{\theta}^o\| \leq \lambda_{n3} \omega_l^o$, for any $l \in N_{x,N}$,

then $\tilde{\theta}^o$ is the unique minimizer of $L_n(\theta; \lambda_{n1}, \lambda_{n2}, \lambda_{n3})$, in other words, $\hat{\theta}^o = \tilde{\theta}$ with probability approaching one. Therefore, in order to show $Pr(\hat{S} = S) \rightarrow 1$, it is equivalent to show $\hat{\theta}^o$ satisfies (C1') – (C3') with probability approaching one, as $n \rightarrow \infty$.

Further notice that

(C1'') $\|\theta_{0m}\| - \|\hat{\theta}_{0m}^o\| < \|\theta_{0m}\|, \forall m \in S$. 
implies Condition (C1'). Therefore, to show \( \hat{\theta}^o \) is the unique minimizer of \( L_n(\theta; \lambda_{n1}, \lambda_{n2}, \lambda_{n3}) \), and consequently, \( \Pr(S = S) \to 1 \), it suffices to show that \( \hat{\theta}^o \) satisfies Conditions (C1''), (C2'') and (C3'') with probability approaching one, as \( n \to \infty \).

According to Lemma D.3 and Lemma D.4 below, we obtain that

\[
\Pr(S \neq S) \leq \Pr(\|\theta_{0m} - \hat{\theta}_m^o\| \geq \|\theta_{0m}\|, \exists m \in S) \\
+ \Pr(\|Z_k^T (Y - D\hat{\theta}^o)\| > \lambda_{n1}\omega_k^o, \exists k \in \mathcal{N}_z) \\
+ \Pr(\|X_l^T (Y - D\hat{\theta}^o)\| > \lambda_{n2}\omega_l^\beta, \exists l \in \mathcal{N}_{x,L}) \\
+ \Pr(\|B_{l}^{(1)^T} (Y - D\hat{\theta}^o)\| > \lambda_{n3}\omega_l^\gamma, \exists l \in \mathcal{N}_{x,N}) \to 0,
\]

as \( n \to \infty \). This completes the proof.

The following Lemma D.3 and Lemma D.4 are used in the proof of Theorem 3.1.

**Lemma D.3.** Under Assumptions (A3) - (A6), as \( n \to \infty \),

\[
\Pr(\|\theta_{0m} - \hat{\theta}_m^o\| \geq \|\theta_{0m}\|, \exists m \in S) \to 0.
\]

*Proof.* Let \( Q_m \) be an \( d_m \times (|S_z| + |S_{x,L}| + |S_{x,N}|N_n) \) matrix, \( d_m = 1 \) for \( m \in S_z \) or \( m \in S_{x,L} \), and \( d_m = N_n \) for \( m \in S_{x,N} \), with the form

\[
Q_m = \begin{pmatrix}
Q_{1m} & 0_{(|S_{x,N}|N_n) \times (|S_{x,N}|N_n)} \\
0_{(|S_z| + |S_{x,L}|) \times (|S_{x,N}|N_n)} & I_{S_z \cup S_{x,L}} \\
0_{N_n \times (|S_z| + |S_{x,L}|)} & Q_{2m} - I_{S_{x,N}}
\end{pmatrix} I(m \in S_{x,N})
\]

with \( Q_{1m} = (0, \ldots, 0, 1, 0, \ldots, 0) \) and \( Q_{2m} = (0_{N_n \times N_n}, \ldots, 0_{N_n \times N_n}, I_{N_n}, 0_{N_n \times N_n}, \ldots, 0_{N_n \times N_n}) \), where scalar 1 is the \( m \)-th element of vector \( Q_{1m} \) with length \( |S_z| \), and an \( N_n \times N_n \) identity matrix \( I_{N_n} \) is at the \( m \)-th block of the \( N_n \times (|S_{x,N}|N_n) \) matrix \( Q_{2m} \) with rest \( N_n \times N_n \) matrices of zeros \( 0_{N_n \times N_n} \).
Then from (D.13), \( \hat{\theta}_m - \theta_{0m} = n^{-1} Q_m C_S^{-1} \left( D_S^T \varepsilon + D_S^T \delta - \sum_{j=1}^{3} \lambda_j v_j \right) \). By the triangle inequality,

\[
\| \hat{\theta}_m - \theta_{0m} \| \leq n^{-1} \| Q_m C_S^{-1} D_S^T \varepsilon \| + n^{-1} \| Q_m C_S^{-1} D_S^T \delta \| + n^{-1} \| Q_m C_S^{-1} \left( \sum_{j=1}^{3} \lambda_j v_j \right) \|
\]

Recall that \( \pi_1 \) and \( \pi_2 \) are the minimum and maximum eigenvalues of \( C_S \), respectively. By Lemmas D.6 and D.1, the first term on the right-hand side

\[
\max_{m \in S} n^{-1} \| Q_m C_S^{-1} D_S^T \varepsilon \|_2 \leq \max_{m \in S} n^{-1} \pi_1^{-1} \| Q_m D_S^T \varepsilon \|_2
\]

By Lemma D.1, the second term

\[
\max_{m \in S} n^{-1} \| Q_m C_S^{-1} D_S^T \delta \|_2 \leq \max_{m \in S} n^{-1} \pi_1^{-1} \| Q_m D_S^T \delta \|_2
\]

By Lemma D.1 and Lemma D.7, the third term

\[
\max_{m \in S} n^{-1} \| Q_m C_S^{-1} \left( \sum_{j=1}^{3} \lambda_j v_j \right) \|_2 \leq n^{-1} \pi_1^{-1} \left( \sum_{j=1}^{3} \| \lambda_j v_j \| \right)
\]

Thus, the claim follows by Assumption (A6).

\[\Box\]

**Lemma D.4.** Under Assumptions (A3) – (A6), as \( n \to \infty \),

\[
\Pr \left( \| Z_k^T (Y - D\hat{\theta}^0) \| > \lambda_{n1} \omega_k^\alpha, \ \exists k \in \mathcal{N}_z \right) \to 0,
\]

\[
\Pr \left( \| X_l^T (Y - D\hat{\theta}^0) \| > \lambda_{n2} \omega_l^\beta, \ \exists l \in \mathcal{N}_{x,L} \right) \to 0,
\]

\[
\Pr \left( \| B_l^{(1)}^T (Y - D\hat{\theta}^0) \| > \lambda_{n3} \omega_l^\gamma, \ \exists l \in \mathcal{N}_{x,N} \right) \to 0.
\]
Proof. Note that

\[ Y - D_S \hat{\theta}_S^0 = Z_S \alpha_{0,S} + X_S \beta_{0,S} + \sum_{l \in S, N} \alpha_l (X_l) + \varepsilon - D_S \hat{\theta}_S \]

\[ = D_S \theta_{0,S} + \delta + \varepsilon - D_S \hat{\theta}_S \]

\[ = -D_S \left(D_S^T D_S\right)^{-1} \left(D_S^T (\delta + \varepsilon) - \left(\sum_{j=1}^{3} \lambda_{nj} v_j \right)\right) + \delta + \varepsilon \]

\[ = \left(I - D_S \left(D_S^T D_S\right)^{-1} D_S^T\right) (\delta + \varepsilon) + D_S \left(D_S^T D_S\right)^{-1} \left(\sum_{j=1}^{3} \lambda_{nj} v_j \right) \]

\[ = H \varepsilon + H \delta + n^{-1} D_S C_S^{-1} \left(\sum_{j=1}^{3} \lambda_{nj} v_j \right). \] (D.12)

For \(1 \leq m \leq p_1 + 2p_2\), by (D.12), we have

\[ D_m^T (Y - D_S \hat{\theta}_S^0) = D_m^T H \varepsilon + D_m^T H \delta + n^{-1} D_m^T D_S C_S^{-1} \left(\sum_{j=1}^{3} \lambda_{nj} v_j \right). \]

By Lemma D.6,

\[ E \left(\max_{m \in \mathcal{N}_z} \|n^{-1/2} D_m^T H \varepsilon\|_2\right) = E \left(\max_{k \in \mathcal{N}_z} \|n^{-1/2} Z_k^T H \varepsilon\|_2\right) = O\left(\sqrt{\log(p_1)}\right), \]

\[ E \left(\max_{m - p_1 \in \mathcal{N}_z} \|n^{-1/2} D_m^T H \varepsilon\|_2\right) = E \left(\max_{k \in \mathcal{N}_z} \|n^{-1/2} X_k^T H \varepsilon\|_2\right) = O\left(\sqrt{\log(p_2)}\right), \]

\[ E \left(\max_{m - p_1 - p_2 \in \mathcal{N}_z} \|n^{-1/2} D_m^T H \varepsilon\|_2\right) = E \left(\max_{k \in \mathcal{N}_z} \|n^{-1/2} B_k^{(1)} H \varepsilon\|_2\right) = O\left(\sqrt{\log(p_2 N_n)}\right), \]

then for the \(D_m^T H \varepsilon\) part, from Condition (A5), for all \(k \in \mathcal{N}_z\), \(\omega_k = |\bar{a}_k|^{-1} = O_p(r_{n1})\), there exists a positive constant \(c_1\), such that

\[ \Pr \left(\|Z_k^T H \varepsilon\| > \lambda_n \omega_k^0 / 3, \exists k \in \mathcal{N}_z\right) \leq \Pr \left(\|Z_k^T H \varepsilon\| > c_1 \lambda_n r_{n1}, \exists k \in \mathcal{N}_z\right) + o(1) \]

\[ = \Pr \left(\max_{k \in \mathcal{N}_z} \|n^{-1/2} Z_k^T H \varepsilon\| > c_1 n^{-1/2} \lambda_n r_{n1}\right) + o(1) \]

\[ \leq \frac{n^{1/2} E \left(\max_{k \in \mathcal{N}_z} \|n^{-1/2} Z_k^T H \varepsilon\|\right)}{c_1 \lambda_n r_{n1}} + o(1) \leq n^{1/2} O\left(\log(p_1)^{1/2}\right) c_1^{-1} \lambda_n^{-1} r_{n1}^{-1} + o(1) \]

\[ = O\left(n^{1/2} \sqrt{\log(p_1) \lambda_n r_{n1}^{-1}}\right) + o(1). \]
Similarly, \( \Pr \left( \|X_l^T H \varepsilon\|_2 > \lambda_n \omega^2 / 3, \forall l \in \mathcal{N}_{x,l} \right) = O \left( n^{1/2} \sqrt{\log(p_2) \lambda_{n_2}^{-1} r_{n_2}} \right) + o(1) \), and

\[
\Pr \left( \|B_{l}^{(1)} H \varepsilon\|_2 > \lambda_n \omega \gamma^2 / 3, \forall l \in \mathcal{N}_{x,N} \right) = O \left( n^{1/2} \sqrt{\log(p_2 N_n) \lambda_{n_3}^{-1} r_{n_3}} \right) + o(1). \]

Recall the definition of \( \mathcal{N} \) in (D.3) and \( \pi_3 = \max_{m \notin S} \|n^{-1} D_m^T D_m\|_2 \), by the properties of spline (de Boor, 2001), the \( D_m^T H \delta \) term has

\[
\max_{m \in N} \|D_m^T H \delta\|_2 \leq n^{1/2} \max_{m \notin S} \left( \frac{1}{n} D_m^T D_m \right)^{1/2} \|H\|_2 \|\delta\|_2 = O_p \left( \frac{n^{1/2} \pi_3^2}{N_n} |\mathcal{S}_{x,N}|^{1/2} \right) = O_p \left( \frac{n}{N_n} \right).
\]

and the last term follows by Lemma D.1 and (D.19) and (D.20) in Lemma D.7 that

\[
\begin{align*}
\max_{m \in \mathcal{N}} \|n^{-1} D_m^T D_m \|_2 & \leq \max_{m \notin S} \|n^{-1/2} D_m \|_2 \times \|n^{-1/2} D_m C_S^{-1/2} \|_2 \times |C_S^{-1/2}|_2 \times \|n \lambda_1 v_1 + n \lambda_2 v_2 + n \lambda_3 v_3\|_2 \\
& \leq \pi_3^{1/2} \pi_1^{-1/2} \left( \sum_{j=1}^3 \lambda_{n_j} h_{n_j} \right) = O_p \left( \sum_{j=1}^3 \lambda_{n_j} h_{n_j} \right).
\end{align*}
\]

Proof of Theorem 3.2. The idea of the proof is similar to the proof of part (ii) in Theorem D.1, but we look at index set \( S \) instead of \( S' \). Let \( \pi_1 \) and \( \pi_2 \) be the minimum and maximum eigenvalues of \( C_S \), respectively, and let \( \pi_3 = \max_{m \notin S} \|n^{-1} D_m^T D_m\| \). By Lemma D.1, \( \pi_1 \times 1, \pi_2 \times 1 \) and \( \pi_3 \times 1 \). For any \( l \in S_{x,N} \), let \( g_{0l}(X_l) = (g_{0l}(X_{1l}), \ldots, g_{0l}(X_{nl}))^T \), \( \delta_l = g_{0l}(X_l) - B_{l}^{(1)} \gamma_l \) and \( \delta = \sum_{l \in S_{x,N}} \delta_l \).

According to the proof of Theorem 3.1, with probability approaching one, we have

\[
\hat{\theta}_S = \hat{\theta}_S^0 = \left( D_S^T D_S \right)^{-1} \left( D_S^T Y - \sum_{j=1}^{3} \lambda_{n_j} v_j \right) = \left( D_S^T D_S \right)^{-1} \times \left[ D_S^T \left( \sum_{j=1}^{3} \lambda_{n_j} h_{n_j} \right) + \sum_{l \in S_{x,N}} \left( B_{l}^{(1)} \gamma_0 + \delta_l \right) \right] = \theta_{0,S} + \left( D_S^T D_S \right)^{-1} \left( D_S^T \left( \delta + \varepsilon \right) - \sum_{j=1}^{3} \lambda_{n_j} v_j \right).
\]
Let $C_S = n^{-1}D_S^T D_S$ be an $(|S_z| + |S_{x,L}| + |S_{x,N}|N_n) \times (|S_z| + |S_{x,L}| + |S_{x,N}|N_n)$ matrix, and let $H = I - D_S \left(D_S^T D_S\right)^{-1} D_S^T$ be an $n \times n$ matrix, then

$$\hat{\theta}_S - \theta^*_S = n^{-1}C_S^{-1} \left(D_S^T (\delta + \varepsilon) - \sum_{j=1}^3 \lambda_{nj} v_j \right).$$ (D.13)

For $\eta = Y - D\theta$, define $\eta_*$ as the projection of $\eta$ to the column space of $D_S$, that is, $\eta_* = P_{D_S} \eta = D_S (D_S^T D_S)^{-1} D_S^T \eta$. Then for $\varepsilon_* \equiv P_{D_S} \varepsilon$, similar to (D.7), (D.9) and (D.10), and by Lemma D.1,

$$\|\varepsilon_*\|^2 = \|D_S^T D_S\|^{1/2} D_S^T \varepsilon\|^2 \leq \left(n \pi_1\right)^{-1} D_S^T \varepsilon\|^2 = O_p\left(\pi_1^{-1}\left(|S_z| + |S_{x,L}| + |S_{x,N}|N_n\right)\right),$$

$$\|\eta_*\|^2 \leq 2\|\varepsilon_*\|^2 + O_p(n |S_{x,N}|N_n^{-2}) = O_p\left(\pi_1^{-1}\left(|S_z| + |S_{x,L}| + |S_{x,N}|N_n\right)\right) + O_p(n N_n^{-2}),$$

$$\|\hat{\theta}_S - \theta_{0,S}\|^2 \leq \frac{8\|\eta_*\|^2}{n \pi_1} + \frac{4\left(\lambda^2_{n1} |S_z| + \lambda^2_{n2} |S_{x,L}| + \lambda^2_{n3} |S_{x,N}|\right)}{n^2 \pi_1^2}$$

$$= O_p\left\{\left(|S_z| + |S_{x,L}| + |S_{x,N}|N_n\right) + O\left(\frac{|S_{x,N}|}{\pi_1 N_n^2}\right)\right\} + O_p\left\{\frac{\lambda^2_{n1} |S_z| + \lambda^2_{n2} |S_{x,L}| + \lambda^2_{n3} |S_{x,N}|}{n^2 \pi_1^2}\right\}.$$ (D.14)

Therefore, the results follow by the facts that

$$\hat{\alpha}_{S_z} - \alpha_{0,S_z} = \left(I_{|S_z|} \ 0_{|S_z| \times |S_{x,L}|} \ 0_{|S_z| \times |S_{x,N}|N_n}\right) (\hat{\theta}_S - \theta_{0,S}),$$

$$\hat{\beta}_{S_{x,L}} - \beta_{0,S_{x,L}} = \left(0_{|S_z| \times |S_{x,L}|} \ I_{|S_{x,L}|} \ 0_{|S_{x,L}| \times |S_{x,N}|N_n}\right) (\hat{\theta}_S - \theta_{0,S}),$$

$$\hat{\gamma}_{S_{x,N}} - \gamma_{0,S_{x,N}} = \left(0_{(|S_{x,N}|N_n) \times |S_{x,L}|} \ 0_{(|S_{x,N}|N_n) \times |S_{x,L}|} \ I_{|S_{x,N}|N_n}\right) (\hat{\theta}_S - \theta_{0,S}),$$ (D.14)

and $\|\hat{g}_l - g_{nl}\|^2 \leq \|\hat{g}_l - g_{0l}\|^2$, where $\hat{\beta}_{S_z} = (\hat{\beta}_k, k \in S_z)^T$, $\beta_{0,S_z} = (\beta_{0k}, k \in S_z)^T$, $\hat{\gamma}_{S_{x,N}} = (\hat{\gamma}_l, l \in S_{x,N})^T$ and $\gamma_{0,S_{x,N}} = (\gamma_l, l \in S_{x,N})^T$.

\[\square\]

**Proof of Inferential Property**

In this section, the spline basis functions considered are of order $d$. For any index set $A \subseteq \{1, \ldots, p_1 + p_2\}$, denote $\beta_A = (\beta_k, 1 \leq k \leq p_1, k \in A)^T$, $\hat{\beta}_A = (\hat{\beta}_k, 1 \leq k \leq p_1, k \in A)^T$, $\gamma_A = (\gamma_l, l \in S_{x,N})^T$ and $\hat{\gamma}_A = (\hat{\gamma}_l, l \in S_{x,N})^T$. 
\( \gamma_A = (\gamma_l, 1 \leq l \leq p_2, l + p_1 \in A)^\top \) and \( \tilde{\gamma}_A = (\tilde{\gamma}_l, 1 \leq l \leq p_2, l + p_1 \in A)^\top \). Next, denote 
\( Z_i = (Z_i, i = 1, \ldots, n)^\top \), where \( Z_i = (Z_{ik}, 1 \leq k \leq p_1, k \in A)^\top \), \( X_i = (X_{il}, 1 \leq l \leq p_2, l + p_1 \in A)^\top \). Similarly, denote \( B_A = (B_{i,A}, i = 1, \ldots, n)^\top \), where \( B_{i,A} = (B_{ij}(X_{il}), 1 \leq l \leq p_2, l + p_1 + p_2 \in A, J = 1, \ldots, N_n)^\top \). Define \( T_S = (Z_{S_x}, X_{S_x,PL}) \). By an abuse of notation, let \( D_S = (T_S, B_S) \), and we define 

\[
C_S = n^{-1}D_S^\top D_S = \begin{pmatrix}
    n^{-1}T_S^\top T_S & n^{-1}T_S^\top B_S \\
    n^{-1}B_S^\top T_S & n^{-1}B_S^\top B_S
\end{pmatrix} = \begin{pmatrix}
    C_{11} & C_{12} \\
    C_{21} & C_{22}
\end{pmatrix},
\]

\[
U_S = C_S^{-1} = \begin{pmatrix}
    U_{11} & -U_{11}C_{12}C_{22}^{-1} \\
    -U_{22}C_{21}C_{11}^{-1} & U_{22}
\end{pmatrix} = \begin{pmatrix}
    U_{11} & U_{12} \\
    U_{21} & U_{22}
\end{pmatrix},
\]

where \( U_{11} = C_{11} - C_{12}C_{22}^{-1}C_{21} = n^{-1}T_S^\top (I_n - P_{B_S}) T_S \) and \( U_{22} = C_{22} - C_{21}C_{11}^{-1}C_{12} = n^{-1}B_S^\top (I_n - P_{T_S}) B_S \), with \( P_{B_S} \) and \( P_{T_S} \) being projection matrices for \( B_S \) and \( T_S \), respectively.

In the following, we give the proof of Theorem 3.4.

**Proof of Theorem 3.4.** Note that for \( l \in S_x,N \)

\[
\hat{\tilde{\phi}}^{SBLL}_l(x_l) - \hat{\phi}^0_l(x_l) = (1, 0) (X_l^\top W_l X_l)^{-1} X_l^\top W_l (\hat{Y}_l - Y_l),
\]

where \( \hat{Y}_l - Y_l = Z_{S_x}(\alpha_0, S_x - \alpha_{S_x}) + X_{S_x,PL}(\beta_{0, S_x,PL} - \beta_{S_x,PL}^*) + \sum_{l' \in S_x,N \setminus \{l\}} \left\{ \phi_{0l'}(X_{l'}) - \hat{\phi}_{l'}^*(X_{l'}) \right\} \)

\[
= Z_{S_x}(\alpha_0, S_x - \alpha_{S_x}) + X_{S_x,PL}(\beta_{0, S_x,PL} - \beta_{S_x,PL}^*) + B_{S_x,PL}(\gamma_0, S_x,PL) - \tilde{\gamma}_{S_x,PL}^* + \sum_{l' \in S_x,N \setminus \{l\}} \left\{ \phi_{0l'}(X_{l'}) - \phi_{nl'}(X_{l'}) \right\},
\]
\[
\text{diag}(1, h^{-1}) X_i^T W_i X_i \text{ diag}(1, h^{-1}) \\
= \begin{pmatrix}
\sum_{i=1}^{n} K_h(x_{il} - x_l) & \sum_{i=1}^{n} \left( \frac{x_{il} - x_l}{h} \right) K_h(x_{il} - x_l) \\
\sum_{i=1}^{n} \left( \frac{x_{il} - x_l}{h} \right) K_h(x_{il} - x_l) & \sum_{i=1}^{n} \left( \frac{x_{il} - x_l}{h} \right)^2 K_h(x_{il} - x_l)
\end{pmatrix} = f_l(x_l) \begin{pmatrix}
1 & 0 \\
0 & \mu_2(K)
\end{pmatrix} + u_p(1),
\]

with \( u_p(\cdot) = o_p(\cdot) \) uniformly for all \( x_l \in [a, b] \). So

\[
\left( X_i^T W_i X_i^* \right)^{-1} = \text{diag}(1, h^{-1}) f_l^{-1}(x_l) \begin{pmatrix}
1 & 0 \\
0 & \mu_2(K)
\end{pmatrix} \text{ diag}(1, h^{-1}),
\]

\[
\text{diag}(1, h^{-1}) X_i^T W_l = \frac{1}{n} \left( \begin{array}{cccc}
K_h(x_{il} - x_l) & \ldots & K_h(x_{nl} - x_l) \\
\left( \frac{x_{il} - x_l}{h} \right) K_h(x_{il} - x_l) & \ldots & \left( \frac{x_{nl} - x_l}{h} \right) K_h(x_{nl} - x_l)
\end{array} \right).
\]

Thus,

\[
\tilde{\phi}_{l}^{\text{SBLL}}(x_l) - \hat{\phi}_l(x_l) = f_l^{-1}(x_l) \left[ \frac{1}{n} \sum_{i=1}^{n} K_h(x_{il} - x_l) Z_i^T L_s (\alpha_{0, s} - \hat{\alpha}^*_s) \\
+ \frac{1}{n} \sum_{i=1}^{n} K_h(x_{il} - x_l) X_i^T L_s (\beta_{0, s, p, l} - \hat{\beta}^*_s) \\
+ \frac{1}{n} \sum_{i=1}^{n} K_h(x_{il} - x_l) B_{i, S_{s, N} \setminus \{l\}}^T (\gamma_{0, s, N} \setminus \{l\} - \hat{\gamma}^*_s) \\
+ \frac{1}{n} \sum_{i=1}^{n} \sum_{\ell' \in S_{s, N} \setminus \{l\}} K_h(x_{il} - x_l) (\phi_{0, l'}(x_{il'}) - \phi_{n, l'}(x_{il'})) \right] + u_p(1) \quad \text{(D.17)}
\]

For the first and second summation terms in the right hand side of (D.17), by Theorem 3.3, we have \( n^{-1} \sum_{i=1}^{n} K_h(x_{il} - x_l) Z_i^T L_s (\alpha_{0, s} - \hat{\alpha}^*_s) = u_p \left( n^{-1/2} \right) \) and \( n^{-1} \sum_{i=1}^{n} K_h(x_{il} - x_l) X_i^T L_s (\beta_{0, s, p, l} - \hat{\beta}^*_s) = u_p \left( n^{-1/2} \right) \); and by Lemma D.2, \( n^{-1} \sum_{i=1}^{n} \sum_{\ell' \in S_{s, N} \setminus \{l\}} K_h(x_{il} - x_l) (\phi_{0, l'}(x_{il'}) - \phi_{n, l'}(x_{il'})) = u_p \left\{ |S_{s, N}| M_n^{-d} \right\} \). As for the third term, define \( \zeta_{il} = \phi_{0, l'}(x_{il'}) - \phi_{n, l'}(x_{il'}) \)},
\[ \sum_{j=1}^{M_n} \gamma_0 B_{jl}^{(d)}(X_{il}), \zeta_i = \sum_{t \in S, u} \zeta_{il}, \text{ and } \zeta = (\zeta_1, \ldots, \zeta_n)^\top, \] similar to the induction with (D.13), we have

\[ \tilde{\theta}_0^* = \theta_0^* = n^{-1} C^{-1}_S \{ D_S^\top (\zeta + \epsilon) \}, \quad (D.18) \]

Then \( \tilde{\gamma}^*_0 \zeta_{Sx,N} - \gamma_0 \zeta_{Sx,N} = \left( 0_{(|Sx,N|M_n) \times (|Sx,l|)} I_{(|Sx,N|M_n)} \right) C^{-1}_S n^{-1} \{ D_S^\top (\zeta + \epsilon) \}. \) Define a diagonal matrix \( I^0_l = \text{diag} \{ 1_{(l-1)M_n}, \theta_{M_n}, 1_{(|Sx,N|-l)M_n} \}, l \in S_{x,N}. \) Then

\[ B_i^{(d)}( \gamma_{Sx,N} \setminus \{ l \} - \tilde{\gamma}^*_0 \zeta_{Sx,N} \setminus \{ l \} ) = B_i^{(d)} I^0_l \left( \tilde{\gamma}^*_0 \zeta_{Sx,N} - \gamma_{Sx,N} \right). \]

Next by Lemma D.2, (D.15) and (D.16), for any \( l \in S_{x,N}, \) we have

\[ \frac{1}{n} \sum_{i=1}^{n} K_h(X_{il} - x_l) B_i^{(d)} I^0_l \left( \tilde{\gamma}^*_0 \zeta_{Sx,N} - \gamma_{Sx,N} \right) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} K_h(X_{il} - x_l) B_i^{(d)} I^0_l \left( -C_2 I_{|Sx,N|} \right) \]

\[ \frac{1}{n} B_i^{(d)} (\zeta + \epsilon) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} K_h(X_{il} - x_l) B_i^{(d)} I^0_l U_{22} \frac{1}{n} B_i^{(d)} (I_n - P_T) (\zeta + \epsilon). \]

Following the same idea in the proof of Lemma D.1, we have that there exist constants \( 0 < c_{U_2} < C_{U_2} < \infty, \) such that with probability approaching one, \( c_{U_2} I_{|Sx,N|} \leq U_{22} \leq C_{U_2} I_{|Sx,N|}. \) Similar to Lemma A.4 in Wang and Yang (2007), for any \( l, l' \in S_{x,N} \) and \( l \neq l', \) we have

\[ \sup_{x_i \in X} \max_{1 \leq j \leq M_n} \left| \frac{1}{n} \sum_{i=1}^{n} [K_h(X_{il} - x_l) B_{jl}^{(d)}(X_{il'}) - E(K_h(X_{il} - x_l) B_{jl}^{(d)}(X_{il'}))] \right| \]

\[ = O_p \left( \frac{\log n}{nh} \right), \]

\[ \sup_{x_i \in X} \max_{1 \leq j \leq M_n} \left| n^{-1} \sum_{i=1}^{n} E(K_h(X_{il} - x_l) B_{jl}^{(d)}(X_{il})) \right| = O_p \left( M_n^{-1/2} \right). \]

We can show that

\[ \sup_{x_i \in X} \frac{1}{n} \sum_{i=1}^{n} K_h(X_{il} - x_l) B_i^{(d)} I^0_l U_{22} \frac{1}{n} B_i^{(d)} (I_n - P_T) \delta \]

\[ = O_p \{ M_n^{-d+1} (\sqrt{\log n/(nh)} + M_n^{-1/2}) \}, \]
and by Proposition 2 in Wang and Yang (2009),
\[
\sup_{x_l \in \chi} \frac{1}{n} \sum_{i=1}^{n} K_h(x_l - x_i) B_{i,S}^{(d)\top} \mathbf{I}_n^{-1} B_s^{(d)\top} (\mathbf{I}_n - \mathbf{P}_T) \epsilon = O_p(\sqrt{\log(n)/n}).
\]
Therefore,
\[
\sup_{x_l \in \chi} |\hat{\phi}_l^*(x_l) - \hat{\phi}_l^o(x_l)| = O_p \left\{ \sqrt{\frac{\log n}{n}} + M_n^{-d+1} \left( \sqrt{\frac{\log n}{nh}} + \sqrt{\frac{1}{M_n}} \right) \right\}.
\]
Hence, the result in (3.18) is established. Consequently, the result in (3.19) follows from (3.16), and the result in (3.20) follows from Claeskens and Van Keilegom (2003).

\[\square\]

Technical Lemmas

The following lemmas are used in the proofs of Theorem D.1 and Theorem 3.1.

**Lemma D.5.** Suppose that Assumptions (A1) – (A4) hold. Recall the definition of \(S\) and \(\mathbf{\tilde{S}}\) in (D.3) and (D.2), with probability approaching one, \(|\mathbf{\tilde{S}}| \leq M_1|S| = M_1(|S_2| + |S_{x,L}| + |S_{x,N}|)\) for a finite constant \(M_1 > 1\).

**Proof.** The basic idea of the proof is similar to the proofs of Theorem 1 of Zhang and Huang (2008) and Theorem 2.1 of Wei and Huang (2010). The main differences are the error term shown in (D.8) and we have a more complex data structure. By Lemma D.2, for some constant \(C_4 > 0\), we have
\[
\|\delta\|_2 \leq C_4 \sqrt{n|S_{x,N}|N_n^{-2}} = C_4 |S_{x,N}|^{1/2} n^{1/2} N_n^{-1}.
\]
For any positive integers \(s_1, s_2\) and \(s_3\), pick some index sets \(A_1 \subseteq \{1, \ldots, p_1\} , A_2 \subseteq \{1, \ldots, p_2\} \) and \(A_3 \subseteq \{1, \ldots, p_2\}\) such that the cardinalities of \(A_1, A_2\) and \(A_3\) are \(|A_1| = s_1, |A_2| = s_2\) and \(|A_3| = s_3\), respectively. Denote \(A = A_1 \cup A_2 \cup A_3\).

Define an \((s_1 + s_2 + s_3 N_n) \times 1\) vector \(\mathbf{S}_A = \left( \tilde{\lambda}_{n1} \mathbf{u}_{s_1}^\top, \tilde{\lambda}_{n2} \mathbf{u}_{s_2}^\top, \tilde{\lambda}_{n3} \sqrt{N_n} \mathbf{U}_1^\top, \ldots, \tilde{\lambda}_{s_3} \sqrt{N_n} \mathbf{U}_{s_3}^\top \right)^\top\), where \(\mathbf{u}_{s_1} \in \{\pm 1\}^{s_1}, \mathbf{u}_{s_2} \in \{\pm 1\}^{s_2}\) and \(\mathbf{U}_j\) is in a unit ball with dimension \(N_n\), that is, \(\mathbf{U}_j \in \mathbb{R}^{N_n}\) and
\[ \|U_j\|_2 = 1, \; j = 1, \ldots, s_3. \] Let \( P_A = D_A (D_A^T D_A)^{-1} D_A^T \) be the projection matrix of \( D_A \). Define

\[ \chi_{s_1, s_2, s_3} = \max_{|A_1| = s_1, |A_2| = s_2, A = A_1 \cup A_2 \cup A_3} \max_{u_{s_1} \in \{ \pm 1 \}^{s_1}, u_{s_2} \in \{ \pm 1 \}^{s_2}} |\eta^T (D_A (D_A^T D_A)^{-1} D_A^T)^{-1} S_A - (I - P_A) D \theta_0)| \]

\[ \|D_A (D_A^T D_A)^{-1} S_A - (I - P_A) D \theta_0\|_2, \]

\[ \Omega_{|S_x|, |S_{x,L}|, |S_{x,N}|} = \left\{ (D, \eta) : \chi_{s_1, s_2, s_3} \leq \sigma C 2\sqrt{s_1 \log(p_1)} \lor s_2 \log(p_2) \lor s_3 N_n \log(p_3 N_n), \right. \]

\[ \forall s_1 \geq |S_2|, s_2 \geq |S_{x,L}|, s_3 \geq |S_{x,N}| \right\}, \]

where \( C_2 > 0 \) is some sufficiently large constant. As shown in the proof of Theorem 1 of Zhang and Huang (2008) and Theorem 2.1 of Wei and Huang (2010), there exists a constant \( M_1 > 1 \), such that if \( (D, \eta) \in \Omega_{|S_x|, |S_{x,L}|, |S_{x,N}|} \), then \( |S| \leq M_1 |S| = M_1 (|S_2| + |S_{x,L}| + |S_{x,N}|) \).

So it suffices to show that \( (D, \eta) \in \Omega_{|S_x|, |S_{x,L}|, |S_{x,N}|} \). Denote \( V_A = D_A (D_A^T D_A)^{-1} S_A - (I - P_A) D \theta_0 \), then by the triangle and Cauchy-Schwarz inequalities,

\[ \frac{\|\eta^T V_A\|}{\|V_A\|_2} = \frac{\|\epsilon^T V_A + \delta^T V_A\|}{\|V_A\|_2} \leq \frac{\|\epsilon^T V_A\|}{\|V_A\|_2} + \|\delta\|_2, \]

For the \( \|\epsilon^T V_A\|/\|V_A\|_2 \) part, define

\[ \chi_{s_1, s_2, s_3} = \max_{|A_1| = s_1, |A_2| = s_2, A = A_1 \cup A_2 \cup A_3} \max_{u_{s_1} \in \{ \pm 1 \}^{s_1}, u_{s_2} \in \{ \pm 1 \}^{s_2}} \frac{|\epsilon^T V_A|}{\|V_A\|_2}, \]

\[ \Omega^*_{|S_x|, |S_{x,L}|, |S_{x,N}|} = \left\{ (D, \epsilon) : \chi_{s_1, s_2, s_3}^* \leq \sigma C \sqrt{s_1 \log(p_1)} \lor s_2 \log(p_2) \lor s_3 N_n \log(p_3 N_n), \right. \]

\[ \forall s_1 \geq |S_2|, s_2 \geq |S_{x,L}|, s_3 \geq |S_{x,N}| \right\}, \]

where \( C_3 > 0 \) is some sufficiently large constant. As shown in the proof of Theorem 1 of Zhang and Huang (2008) and Theorem 2.1 of Wei and Huang (2010), \( P(\Omega^*_{|S_x|, |S_{x,L}|, |S_{x,N}|}) \rightarrow 1 \). And for \( \|\delta\|_2 \) part, for \( n \) sufficiently large and \( N_n \gg n^{1/3} \),

\[ \|\delta\|_2 \leq C_4 |S_{x,N}|^{1/2} n^{1/2} N_n^{-1} \leq \sigma C_5 \sqrt{s_1 \log(p_1)} \lor s_2 \log(p_2) \lor s_3 N_n \log(p_2 N_n). \]
It follows that \( P(\Omega | s_{x} , | s_{x,L} , | s_{x,N}) \to 1 \). This completes the proof. \( \Box \)

For any random variable \( X \), denote \( \| X \|_p = (E|X|^p)^{1/p} \) as the \( L_p \) norm for random variable \( X \); and denote \( \| X \|_\varphi = \inf \{ C > 0 : E\{\varphi(|X|/C) \} \leq 1 \} \) as the Orlicz norm for random variable \( X \), where \( \varphi \) is required as a non-decreasing, convex function with \( \varphi(0) = 0 \).

**Lemma D.6.** Suppose that Assumptions (A2) and (A4) hold. Let

\[
T_{1k} = n^{-1/2} \sum_{i=1}^{n} Z_{ik} \varepsilon_i, \quad 1 \leq k \leq p_1, \quad T_{2l} = n^{-1/2} \sum_{i=1}^{n} X_{il} \varepsilon_i, \quad 1 \leq l \leq p_2,
\]

\[
T_{3Jl} = n^{-1/2} \sum_{i=1}^{n} B_{J,l}^{(d)}(X_{il}) \varepsilon_i, \quad 1 \leq l \leq p_2, \quad 1 \leq J \leq N_n,
\]

and \( T_1 = \max_{1 \leq k \leq p_1} |T_{1k}|, T_2 = \max_{1 \leq l \leq p_2} |T_{2l}| \) and \( T_3 = \max_{1 \leq l \leq p_2, 1 \leq J \leq N_n} |T_{3Jl}|. \) Then we have

\[
E(T_1) \leq C_1 \sqrt{\log(p_1)}, \quad E(T_2) \leq C_2 \sqrt{\log(p_2)},
\]

\[
E(T_3) \leq C_3 n^{-1/2} \sqrt{\log(p_2 N_n)} \left( \sqrt{2C_4 n N_n \log(2p_2 N_n)} + C_5 N_n^{1/2} \log(2p_2 N_n) + n \right)^{1/2},
\]

where \( C_1, C_2, C_3, C_4 \) and \( C_5 \) are positive constants.

In particular, when \( N_n \log(p_2 N_n)/n \to 0 \), we have

\[
E(T_1) = O\{\sqrt{\log(p_1)}\}, \quad E(T_2) = O\{\sqrt{\log(p_2)}\}, \quad E(T_3) = O\{\sqrt{\log(p_2 N_n)}\}.
\]

**Proof.** Denote

\[
s^2_{1nk} = \sum_{i=1}^{n} Z_{ik}^2, \quad 1 \leq k \leq p_1, \quad s^2_{2nl} = \sum_{i=1}^{n} X_{il}^2, \quad 1 \leq l \leq p_2,
\]

\[
s^2_{3nJl} = \sum_{i=1}^{n} (B_{J,l}^{(d)}(X_{il}))^2, \quad 1 \leq l \leq p_2, \quad 1 \leq J \leq N_n.
\]

Next let \( s^2_{1n} = \max_{1 \leq k \leq p_1} s^2_{1nk}, s^2_{2n} = \max_{1 \leq l \leq p_2} s^2_{2nl} \) and \( s^2_{3n} = \max_{1 \leq l \leq p_2, 1 \leq J \leq N_n} s^2_{3nJl}. \) By Assumption (A2), conditional on \( Z = \{Z_{ik}, 1 \leq i \leq n, 1 \leq k \leq p_1\} \), \( \sqrt{n} T_{1k} \) is \( b(\sum_{i=1}^{n} Z_{ik}^2)^{1/2} \)
- subgaussian; and conditional on \( X = \{X_{il}, 1 \leq i \leq n, 1 \leq l \leq p_2\} \), \( \sqrt{n} T_{2i} \) is \( b(\sum_{i=1}^{n} X_{il}^2)^{1/2} \)-subgaussian, and \( \sqrt{n} T_{3j} \) is \( b \left[ \sum_{i=1}^{n} \{B_{j}^{(d)}(X_{il})\}^2 \right]^{1/2} \)-subgaussian.

Define \( \varphi_p(x) = \exp(x^p) - 1, p \geq 1 \). Then \( \varphi_p^{-1}(m) = \{\log(1 + m)\}^{1/p} \). By Assumption (A2) and the maximal inequality for sub-Gaussian random variables (as stated in Lemmas 2.2.1 and 2.2.2 of Van Der Vaart and Wellner (1996)),

\[
E(T_1|Z) = E \left( \max_{1 \leq k \leq p_1} |T_{1k}| |Z\right) = \left\| \max_{1 \leq k \leq p_1} |T_{1k}| |Z\right\|_1 \leq \left\| \max_{1 \leq k \leq p_1} |T_{1k}| |Z\right\|_{\varphi_1} \leq K_1 \sqrt{\log 2} \sqrt{\log(1 + p_1)} n^{-1/2} \max_{1 \leq k \leq p_1} \|\sqrt{n} T_{1k}\| \{Z_{ik}, 1 \leq i \leq n, 1 \leq k \leq p_1\}\|_{\varphi_2} \leq K_1 \sqrt{\log 2} \sqrt{\log(1 + p_1)} n^{-1/2} \max_{1 \leq k \leq p_1} \left( 6b^2 \sum_{i=1}^{n} Z_{ik}^2 \right)^{1/2} \leq C_5 n^{-1/2} s_{1n} \sqrt{\log(p_1)}.
\]

Next,

\[
E(T_2|X) = E \left( \max_{1 \leq l \leq p_2} |T_{2l}| |X\right) = \left\| \max_{1 \leq l \leq p_2} |T_{2l}| |X\right\|_1 \leq \left\| \max_{1 \leq l \leq p_2} |T_{2l}| |X\right\|_{\varphi_1} \leq K_2 \sqrt{\log 2} \sqrt{\log(1 + p_2)} n^{-1/2} \max_{1 \leq l \leq p_2} \|\sqrt{n} T_{2l}\| \{X_{il}, 1 \leq i \leq n, 1 \leq l \leq p_2\}\|_{\varphi_2} \leq K_2 \sqrt{\log 2} \sqrt{\log(1 + p_2)} n^{-1/2} \max_{1 \leq l \leq p_2} \left( 6b^2 \sum_{i=1}^{n} X_{il}^2 \right)^{1/2} \leq C_5 n^{-1/2} s_{2n} \sqrt{\log(p_2)}.
\]
\[ E(T_3 | X) = E \left( \max_{1 \leq l \leq p_2, 1 \leq J \leq N_n} |T_{3,lJ}| |X| \right) = \left\| \max_{1 \leq l \leq p_2, 1 \leq J \leq N_n} |T_{3,lJ}| |X| \right\|_1 \]
\[ \leq \left\| \max_{1 \leq l \leq p_2, 1 \leq J \leq N_n} |T_{3,lJ}| |X| \right\|_{\varphi_1} \leq \sqrt{\log(2)} \left\| \max_{1 \leq l \leq p_2, 1 \leq J \leq N_n} |T_{3,lJ}| |X| \right\|_{\varphi_2} \]
\[ \leq K_3 \sqrt{\log 2} \sqrt{\log(1 + p_2 N_n)} n^{-1/2} \max_{1 \leq l \leq p_2, 1 \leq J \leq N_n} \| \sqrt{n} T_{3,lJ} |X| \|_{\varphi_2} \]
\[ \leq K_3 \sqrt{\log 2} \sqrt{\log(1 + p_2 N_n)} n^{-1/2} \max_{1 \leq l \leq p_2, 1 \leq J \leq N_n} \left[ 6b^2 \sum_{i=1}^n \{B_{J,l}^{(d)}(X_{id})\}^2 \right]^{1/2} \]
\[ = C_{21} n^{-1/2} s_{3n} \sqrt{\log(p_2 N_n)}. \]

Thus,
\[ E(T_1) = E \{E(T_1 | Z)\} \leq C_{11} n^{-1/2} \sqrt{\log(p_1)} E(s_{1n}), \]
\[ E(T_2) = E \{E(T_2 | X)\} \leq C_{21} n^{-1/2} \sqrt{\log(p_2)} E(s_{2n}), \]
\[ E(T_3) = E \{E(T_3 | X)\} \leq C_{31} n^{-1/2} \sqrt{\log(p_2 N_n)} E(s_{3n}), \]

where \( K_1, K_2, K_3, C_{11}, C_{21} \) and \( C_{31} \) are positive constants. By Assumption (A4), we have \( E(Z_{ik})^2 \leq C_{13}^2 \) and \( E(X_{id})^2 \leq C_{23}^2 \). The properties of normalized B-splines imply that, for every \( l, J \), there exist positive constants \( C_{13} \) and \( C_4 \), such that \( |B_{J,l}^{(d)}(X_{id})| \leq C_4 N_n^{1/2} \) and \( E \left( B_{J,l}^{(d)}(X_{id}) \right)^2 = 1 \). Therefore,
\[ E(s_{1n}^2) = \max_{1 \leq k \leq p_1} E(s_{1nk}^2) = \max_{1 \leq k \leq p_1} \sum_{i=1}^n E(Z_{ik}^2) \leq n C_{13}^2, \]
\[ E(s_{2n}^2) = \max_{1 \leq l \leq p_2} \sum_{i=1}^n E(s_{2nl}^2) = \max_{1 \leq l \leq p_2} \sum_{i=1}^n E(X_{id}^2) \leq n C_{23}^2, \]
\[ \text{and } \sum_{i=1}^n E \left[ \left\{ B_{J,l}^{(d)}(X_{id}) \right\}^2 - E^2 \left\{ B_{J,l}^{(d)}(X_{id}) \right\} \right] \leq n N_n C_3. \]
Thus, by Lemma A.1 of Van de Geer (2008), we have
\[ E \left[ \max_{1 \leq l \leq p_2, 1 \leq J \leq N_n} \left\| \sum_{i=1}^n \left\{ B_{J,l}^{(d)}(X_{id}) \right\} \right\|^2 - E^2 \left\{ B_{J,l}^{(d)}(X_{id}) \right\} \right] \]
\[ \leq \sqrt{2 C_4 n N_n \log(2 p_2 N_n)} + C_5 N_n^{1/2} \log(2 p_2 N_n). \]
Therefore, by triangle inequality, \( E(s_{3n}^2) \leq \sqrt{2C_4 n N_n \log(2p_2 N_n)} + C_5 N_n^{1/2} \log(2p_2 N_n) + n \). Thus,

\[
E(s_{1n}) \leq (E s_{1n}^2)^{1/2} \leq (C_{13} n)^{1/2}, \quad E(s_{2n}) \leq (E s_{2n}^2)^{1/2} \leq (C_{23} n)^{1/2}, \quad \text{and}
\]

\[
E(s_{3n}) \leq (E s_{3n}^2)^{1/2} \leq \left\{ \sqrt{2C_4 n N_n \log(2p_2 N_n)} + C_5 N_n^{1/2} \log(2p_2 N_n) + n \right\}^{1/2}.
\]

The lemma follows. \( \Box \)

Lemma D.7. For

\[
v_1 = \left\{ \left( \frac{\omega_k^\alpha \overline{\beta}_0, k \in S_z}{\overline{\beta}_0} \right)^T, 0_{|S_z, L|}, 0_{|S_z, N| N_n} \right\}^T,
\]

\[
v_2 = \left\{ 0_{|S_z|}, \left( \frac{\omega_l^\beta \overline{\beta}_0, |S_z| + t}{\overline{\beta}_0, |S_z| + t} \right)^T, 0_{|S_z, N| N_n} \right\}^T,
\]

\[
v_3 = \left\{ 0_{|S_z|}, 0_{|S_z, L|}, \left( \frac{\omega_l^\beta \overline{\beta}_0, |S_z| + t, |S_z, L| + t}{\overline{\beta}_0, |S_z| + t, |S_z, L|} \right)^T \right\}^T,
\]

under Assumption (A5),

\[
\|v_1\|_2^2 = O_p \left( h_{n_1}^2 \right) = O_p \left( b_{n_1}^{-4} c_{b_1}^{-2} r_{n_1}^{-2} + |S_z| b_{n_1}^{-2} \right), \quad (D.19)
\]

\[
\|v_2\|_2^2 = O_p \left( h_{n_2}^2 \right) = O_p \left( b_{n_2}^{-4} c_{b_2}^{-2} r_{n_2}^{-2} + |S_x, L| b_{n_2}^{-2} \right), \quad (D.20)
\]

\[
\|v_3\|_2^2 = O_p \left( h_{n_3}^2 \right) = O_p \left( b_{n_3}^{-4} c_{b_3}^{-2} r_{n_3}^{-2} + |S_x, N| b_{n_3}^{-2} \right). \quad (D.21)
\]

Proof. Write

\[
\|v_1\|_2^2 = \sum_{k \in S_z} \left( \omega_k^\alpha \right)^2 = \sum_{k \in S_z} |\alpha_k^-|^2 = \sum_{k \in S_z} \frac{\alpha_{0k}^2 - \overline{\alpha_k}^2}{\alpha_{0k}^2 \alpha_k^2} + \sum_{k \in S_z} |\alpha_{0k}|^{-2},
\]

\[
\|v_2\|_2^2 = \sum_{l \in S_{x, L}} \left( \omega_l^\beta \right)^2 = \sum_{l \in S_{x, L}} |\overline{\beta}_l|^2 = \sum_{l \in S_{x, L}} \frac{\beta_{0l}^2 - \overline{\beta_l}^2}{\beta_{0l}^2 \beta_l^2} + \sum_{l \in S_{x, L}} |\beta_{0l}|^{-2},
\]

\[
\|v_3\|_2^2 = \sum_{l \in S_{x, N}} \left( \omega_l^\gamma \right)^2 = \sum_{l \in S_{x, N}} |\overline{\gamma}_l|^2 = \sum_{l \in S_{x, N}} \frac{\|\gamma_{0l}\|^2}{\|\gamma_{0l}\|^2 |\overline{\gamma}_l|^2} + \sum_{l \in S_{x, N}} |\gamma_{0l}|^{-2}.
\]
Under (A5), there exist positive constants $M_1$, $M_2$ and $M_3$, such that

$$\sum_{k \in S_x} \frac{|\alpha_{0k}^2 - \bar{\alpha}_{0k}^2|}{\alpha_{0k}^2 \bar{\alpha}_{0k}^2} \leq M_1 c_{b_1}^{-2} b_{n_1}^{-4} \|\bar{\alpha} - \alpha_0\|^2 = O_p \left( b_{n_1}^{-4} c_{b_1}^{-2} r_{n_1}^{-2} \right),$$

$$\sum_{l \in S_{x,L}} \frac{|\beta_{0l}^2 - \bar{\beta}_{0l}^2|}{\beta_{0l}^2 \bar{\beta}_{0l}^2} \leq M_2 c_{b_2}^{-2} b_{n_2}^{-4} \|\bar{\beta} - \beta_0\|^2 = O_p \left( b_{n_2}^{-4} c_{b_2}^{-2} r_{n_2}^{-2} \right),$$

$$\sum_{l \in S_{x,N}} \frac{|\gamma_{0l}^2 - \bar{\gamma}_{0l}^2|}{\gamma_{0l}^2 \bar{\gamma}_{0l}^2} \leq M_3 c_{b_3}^{-2} b_{n_3}^{-4} \|\bar{\gamma} - \gamma_0\|^2 = O_p \left( b_{n_3}^{-4} c_{b_3}^{-2} r_{n_3}^{-2} \right),$$

and the results follow by the facts $\sum_{k \in S_x} |\alpha_k|^2 \leq |S_x| b_{n_1}^{-2}$, $\sum_{l \in S_{x,L}} |\beta_l|^2 \leq |S_{x,L}| b_{n_2}^{-2}$ and $\sum_{l \in S_{x,N}} \|\gamma_l\|^2 \leq |S_{x,N}| b_{n_3}^{-2}$. \qed
APPENDIX C. SUPPLEMENTARY FOR “SPARSE LEARNING FOR IMAGE-ON-SCALAR REGRESSION WITH APPLICATION TO IMAGING GENETICS STUDIES”

Technical Details

Preliminary results

For a real valued vector $a \in \mathbb{R}^n$, take $\|a\| = (\sum_{i=1}^n a_i^2)^{1/2}$ as its Euclidean norm on $\mathbb{R}^n$. For a function $g$, let $\|g\|_{\infty, 2} = \left\{ \int_{\Omega} g^2(s) dQ(s) \right\}^{1/2}$ be the regular $L_2$ norm and $\|g\|_{\infty, \Omega} = \sup_{s \in \Omega} |g(s)|$ the supremum norm for function $g$ over $\Omega$. For a vector valued function $\mathbf{g} = (g_1, \ldots, g_p)^\top$, denote $\|\mathbf{g}\|_{\infty, 2} = \left\{ \sum_{k=1}^p \|g_k\|_{\infty, 2}^2 \right\}^{1/2}$ as the $L_2$ norm of $\mathbf{g}$. For notation simplicity, we drop the subscript $\Omega$ in the rest of the paper.

For $\mathbf{g}^{(1)}(s) = (g_1^{(1)}(s),\ldots, g_p^{(1)}(s))^\top$ and $\mathbf{g}^{(2)}(s) = (g_1^{(2)}(s),\ldots, g_p^{(2)}(s))^\top$, define the empirical inner product as

$$\langle \mathbf{g}^{(1)}, \mathbf{g}^{(2)} \rangle_{X,n} = \frac{1}{nN^s} \sum_{k,k'=1}^p \sum_{i=1}^n \sum_{j=1}^{N^s} X_{ik}X_{ik'} g_k^{(1)}(s_j)g_{k'}^{(2)}(s_j), \quad (E.1)$$

and the theoretical inner product as

$$\langle \mathbf{g}^{(1)}, \mathbf{g}^{(2)} \rangle_X = \sum_{k,k'=1}^p E(X_kX_{k'}) \int_{\Omega} g_k^{(1)}(s)g_{k'}^{(2)}(s) dQ(s), \quad (E.2)$$

and denote the corresponding empirical and theoretical norms $\|\cdot\|_{X,n}$ and $\|\cdot\|_X$.

Denote $\mathcal{H}^1 = \{ g : \int_{\Omega} g(s) dQ(s) = 0, \int_{\Omega} \{g(s)\}^2 dQ(s) < \infty \}$ as the space for centralized functions, and denote $\mathcal{H}^2 = \{ g : \int_{\Omega} g(s) dQ(s) = 0, \int_{\Omega} \{g(s)\}^2 dQ(s) = 1 \}$ as the space for normalized functions.
In the following, we use $c$ and $C$ for constants, although they may have different values in different context.

We start with some preliminary preparation. First we cite the stability lemma from Lai and Wang (2013).

**Lemma E.1.** Let $\{B_\ell\}_{\ell \in J}$ be the normalized Bernstein bases for the normalized spline space $S_d^r(\Delta) \cap \mathcal{H}^2$. Then there exist positive constants $c, C$ depending on the smoothness $r$ and the shape parameter $\varpi$ such that

$$c \sum_{\ell \in J} \gamma_\ell^2 \leq \left\| \sum_{\ell \in J} \gamma_\ell B_\ell \right\|_2^2 \leq C \sum_{\ell \in J} \gamma_\ell^2.$$

The next lemma provides the approximation power of the normalized bivariate splines.

**Lemma E.2.** Under Assumption (A5), for any $g \in W^{k+1,\infty}(\Omega) \cap \mathcal{H}^1$, we have

(i) for bi-integer $(a_1, a_2)$ with $0 \leq a_1 + a_2 \leq d$, there exists a spline fit $g^*(\cdot) \in S_d^r(\Delta) \cap \mathcal{H}^1$ such that $\|\nabla^{a_1}_s \nabla^{a_2}_s (g - g^*)\|_\infty \leq C |\Delta|^{d+1-a_1-a_2} |g|_{d+1,\infty}$, where $C$ is an absolute constant depending on the degree $d$, smoothness $r$ and the shape parameter $\varpi$.

(ii) for bi-integer $(a_1, a_2)$ with $0 \leq a_1 + a_2 \leq d$, there exists a spline fit $g^*(\cdot) \in S_d^r(\Delta) \cap \mathcal{H}^1(d \geq 3r + 2)$ such that $\|\nabla^{a_1}_s \nabla^{a_2}_s (g - g^*)\|_\infty \leq C |\Delta|^{d+1-a_1-a_2} |g|_{d+1,\infty}$, where $C$ is an absolute constant depending on the degree $d$, smoothness $r$ and the shape parameter $\varpi$.

**Proof.** We just show the proof for (ii). The proof for (i) is similar. According to Lemma 3 in Lai and Wang (2013), there exist a constant $C$ and a spline function $g^{**} \in S_d^r(\Delta)$ such that

$$\|\nabla^{a_1}_s \nabla^{a_2}_s (g - g^{**})\|_\infty \leq C/2 |\Delta|^{d+1-a_1-a_2} |g|_{d+1,\infty}.$$

Then define $g^*(s) = g^{**}(s) - \int_\Omega g^{**}(s) dQ(s)$, one has $g^*(s) \in S_d^r(\Delta) \cap \mathcal{H}^1$ and

$$\|\nabla^{a_1}_s \nabla^{a_2}_s (g - g^*)\|_\infty \leq \|\nabla^{a_1}_s \nabla^{a_2}_s (g - g^{**})\|_\infty + \left\| \nabla^{a_1}_s \nabla^{a_2}_s \int_\Omega g^{**}(s) dQ(s) \right\|_\infty.$$
If \( a_1 + a_2 > 0 \), it is obvious that \( \| \nabla_{s_1}^{a_1} \nabla_{s_2}^{a_2} (g - g^*) \|_\infty \leq C |\Delta|^{d+1-a_1-a_2} |g|_{d+1,\infty} \); if \( a_1 + a_2 = 0 \), note that

\[
\left\| \nabla_{s_1}^{a_1} \nabla_{s_2}^{a_2} \int_\Omega g^{**}(s) dQ(s) \right\|_\infty = \sup_{s \in \Omega} \left| \nabla_{s_1}^{a_1} \nabla_{s_2}^{a_2} \int_\Omega g^{**}(s) dQ(s) \right| \\
\leq \sup_{s \in \Omega} \left| \nabla_{s_1}^{a_1} \nabla_{s_2}^{a_2} \int_\Omega (g^{**} - g)(s) dQ(s) \right| + \sup_{s \in \Omega} \left| \nabla_{s_1}^{a_1} \nabla_{s_2}^{a_2} \int_\Omega g(s) dQ(s) \right| \\
= \left\| \nabla_{s_1}^{a_1} \nabla_{s_2}^{a_2} \int_\Omega (g^{**} - g)(s) dQ(s) \right\|_\infty \leq C/2 |\Delta|^{d+1-a_1-a_2} |g|_{d+1,\infty},
\]

the result follows.

\[\square\]

**Lemma E.3.** Suppose that \( d \geq 3r + 2 \). Under Assumptions (A1), (A4) and (A5), for \( \beta_{0k} \in \mathcal{W}^{\mathcal{C}+1,\infty}(\Omega) \cap \mathcal{H}_1 \), there exists a vector \( \gamma_0 = (\gamma_{01}^\top, \ldots, \gamma_{0p}^\top)^\top \), such that \( \|\gamma_{0k}\| \neq 0 \), for \( k \in \mathcal{A}_v \), \( \|\gamma_{0k}\| = 0 \), \( k \notin \mathcal{A}_v \); and there exists a constant \( C \) depending on \( d \) and \( \pi \) such that for any \( k \) and normalized Bernstein basis polynomials \( B_k \) of degree \( d \), \( B_k \in \mathcal{H}_2 \), \( d \geq 0 \),

\[
\sup_{s \in \Omega} \left| \beta_{0k}(s) - B_k(s)^\top \gamma_{0k} \right| \leq C |\Delta|^{d+1} |\beta_{0k}|_{d+1,\infty}
\]

**Proof.** According to the Theorem 1 in Lai and Wang (2013), there exist a vector \( \gamma_0^* \) and a Constant \( C \), such that for the Bernstein basis polynomials \( B_k^* \),

\[
\sup_{s \in \Omega} \left| \beta_{0k}(s) - B_k^*(s)^\top \gamma_{0k}^* \right| \leq C/2 |\Delta|^{d+1} |\beta_{0k}|_{d+1,\infty}.
\]

Define \( C_B = \| B_k^* - \int B_k^* dQ \|_2 \), let \( \gamma_{0k} = C_B \gamma_{0k}^* \) and \( B_k(s) = C_B^{-1} \{ B_k^*(s) - \int_\Omega B_k^*(s) dQ(s) \} \). Then \( B_k \in \mathcal{H}_2 \), and

\[
\sup_{s \in \Omega} \left| \beta_{0k}(s) - B_k(s)^\top \gamma_{0k} \right| \leq \sup_{s \in \Omega} \left| \beta_{0k}(s) - B_k^*(s)^\top \gamma_{0k}^* \right| + \sup_{s \in \Omega} \left| \int_\Omega B_k^*(s)^\top \gamma_{0k}^* dQ(s) \right| \\
\leq \sup_{s \in \Omega} \left| \beta_{0k}(s) - B_k^*(s)^\top \gamma_{0k}^* \right| + \sup_{s \in \Omega} \left| \int_\Omega \{ B_k^*(s)^\top \gamma_{0k}^* - \beta_{0k}(s) \} dQ(s) \right| + \sup_{s \in \Omega} \left| \int_\Omega \beta_{0k}(s) dQ(s) \right| \\
\leq C |\Delta|^{d+1} |\beta_{0k}|_{d+1,\infty}.
\]

The desired result follows. \[\square\]
In the following, we denote $\beta_{nk}(\cdot) = B_k(\cdot)^\top \gamma_{0k}$ as the best bivariate spline approximation of $\beta_{0k}(\cdot)$ such that $\| \beta_{0k} - \beta_{nk} \|_\infty = \sup_s |\beta_{0k}(s) - \beta_{nk}(s)| = O(|\Delta|^{d+1})$. Let $\gamma_{0k}$ be the vector of the coefficients of the best bivariate spline approximation in Lemma E.3. Denote $\theta_0^\top = (\theta_{01}^\top, \ldots, \theta_{0m}^\top, \ldots, \theta_{0,2p}^\top)$, where $\theta_0^\top = (\alpha_0^\top, \gamma_{01}^\top, \ldots, \gamma_{0p}^\top)$.

Lemma E.4. Under Assumptions (A4) and (A5), for any normalized Bernstein basis polynomials $B_\ell$ and $B_{\ell'}$, of degree $d$, $B_\ell, B_{\ell'} \in H^2$, $d \geq 0$, $\ell, \ell' \in \mathcal{M}$, we have

$$\max_{\ell \in \mathcal{M}} \left| \frac{1}{N_s} \sum_{j=1}^{N_s} B_\ell^k(s_j) - \int_{\Omega} B_\ell^k(s) dQ(s) \right| = O(N_s^{-1/2}|\Delta|^{-k+1}), \quad k \geq 1, \quad (E.3)$$

$$\max_{\ell, \ell' \in \mathcal{M}} \left| \frac{1}{N^2} \sum_{j,j'=1}^{N} G_{\eta}(s_j, s_{j'}) B_\ell(s_j) B_{\ell'}(s_{j'}) - \int_{\Omega^2} G_{\eta}(s, s') B_\ell(s) B_{\ell'}(s') dQ(s) dQ(s') \right| = O(N_s^{-1/2}|\Delta|), \quad (E.4)$$

$$\max_{\ell \in \mathcal{M}} \left| \frac{1}{N_s} \sum_{j=1}^{N_s} B_\ell^2(s_j) \sigma^2(s_j) - \int_{T_m} \sigma^2(s) B_\ell^2(s) dQ(s) \right| = O(N_s^{-1/2}|\Delta|^{-1}). \quad (E.5)$$

Proof. For any piecewise constant $B_\ell(s) \in H^2$, we have

$$B_\ell(s) = \left[ \frac{A(\tau_1)}{A(\tau_\ell)\{A(\tau_\ell) + A(\tau_1)\}} \right]^{1/2} \left\{ I(s \in \tau_\ell) - \frac{A(\tau_\ell)}{A(\tau_1)} I(s \in \tau_1) \right\},$$

$$B_\ell^k(s) = \left[ \frac{A(\tau_1)}{A(\tau_\ell)\{A(\tau_\ell) + A(\tau_1)\}} \right]^{1/2} \left\{ I(s \in \tau_\ell) + (-1)^k \frac{A^k(\tau_\ell)}{A^k(\tau_1)} I(s \in \tau_1) \right\}.$$
Thus, (E.3) holds.

For any normalized Bernstein basis polynomials \( B^k(s) \), by the properties of bivariate spline basis functions in Lai and Schumaker (2007), we have

\[
\left| \frac{1}{N_s} \sum_{j=1}^{N_s} B^k_j(s_j) - \int_{\Omega} B^k(s) dQ(s) \right| = \left| \int_{\Omega} B^k(s) dQ(N_s) - \int_{\Omega} B^k(s) dQ(s) \right|
\]

\[
= \left[ \frac{A(\tau_1)}{A(\tau)\{ A(\tau) + A(\tau_1) \}} \right]^{\frac{1}{2}} \times \left| \frac{1}{N_s} \sum_{j=1}^{N_s} I(s_j \in \tau_\ell) - A(\tau_\ell) \right| + \left( -\frac{A(\tau_1)}{A(\tau_1) - 1} \right) \left( \frac{1}{N_s} \sum_{j=1}^{N_s} I(s_j \in \tau_\ell) - A(\tau_1) \right)
\]

According to Assumption (A5), we have

\[
\max_{\ell \in M} \left| \frac{1}{N_s} \sum_{j=1}^{N_s} B^k_j(s_j) - \int_{\Omega} B^k(s) dQ(s) \right| \leq CN_s^{-1/2}|\Delta|^{-k+1}.
\]

For any \( j = 1, \ldots, N \), let \( V_j \) be the \( j \)th voxel, and it is clear that

\[
\left| \frac{1}{N_s} \sum_{j=1}^{N_s} B^k_j(s_j) - \int_{\Omega} B^k(s) dQ(s) \right| \leq \sum_{j=1}^{N_s} \int_{V_j} \| B^k_j(s_j) - B^k(s) \| dQ(s) + \int_{\Omega \setminus V_j} B^k(s) dQ(s).
\]

For any normalized Bernstein basis polynomials \( B^k(s) \) with \( d \geq 1 \), we have

\[
\int_{\Omega \setminus V_j} B^k(s) dQ(s) = O(N_s^{-1/2}|\Delta|^{-k+1}),
\]

\[
\sum_{j=1}^{N_s} \int_{V_j} \| B^k_j(s_j) - B^k(s) \| dQ(s) \leq \sum_{j: s_j \in \tau_\ell} \int_{V_j} |B^k_j(s_j) - B^k(s)| dQ(s)
\]

\[
\leq C(N |\Delta|^2) \times N^{-1} \times (N_s^{-1/2}|\Delta|^{-1}) \times |\Delta|^{-k}
\]

\[
\leq CN_s^{-1/2}|\Delta|^{-k+1}.
\]

Thus, (E.3) holds.
Next, for any $\ell, \ell' \in M$, 
\[
\frac{1}{N_s^2} \sum_{j=1}^{N_s} \sum_{j'=1}^{N_s} \sum_{j=1}^{N_s} \sum_{j'=1}^{N_s} G_\eta(s_j, s_{j'}) B_\ell(s_j) B_{\ell'}(s_j') - \int_{T_m \times T_m} G_\eta(s, s') B_\ell(s) B_{\ell'}(s') dQ(s) dQ(s') \\
= \sum_{j=1}^{N_s} \sum_{j'=1}^{N_s} \sum_{j=1}^{N_s} \sum_{j'=1}^{N_s} \int_{V_j \times V_{j'}} \{G_\eta(s_j, s_{j'}) B_\ell(s_j) B_{\ell'}(s_j') - G_\eta(s, s') B_\ell(s) B_{\ell'}(s')\} dQ(s) dQ(s') \\
+ \int_{\Omega \setminus \bigcup_{j,j'} V_j \times V_{j'}} \{G_\eta(s_j, s_{j'}) B_\ell(s_j) B_{\ell'}(s_j') - G_\eta(s, s') B_\ell(s) B_{\ell'}(s')\} dQ(s) dQ(s').
\]
As $N_s \to \infty$, 
\[
\int_{\Omega \setminus \bigcup_{j,j'} V_j \times V_{j'}} \{G_\eta(s_j, s_{j'}) B_\ell(s_j) B_{\ell'}(s_j') - G_\eta(s, s') B_\ell(s) B_{\ell'}(s')\} dQ(s) dQ(s') = O(N_s^{-1/2}|\triangle|).
\]
and 
\[
\left| \sum_{j=1}^{N_s} \sum_{j'=1}^{N_s} \int_{V_j \times V_{j'}} \{G_\eta(s_j, s_{j'}) B_\ell(s_j) B_{\ell'}(s_j') - G_\eta(s, s') B_\ell(s) B_{\ell'}(s')\} dQ(s) dQ(s') \right| \
\leq \sum_{\{(j,j'): s_j \in \tau_\ell, s_{j'} \in \tau_{\ell'}\}} \int_{V_j \times V_{j'}} \omega_{jj'}(G_\eta K_\ell, 2/\sqrt{N}) dQ(s) dQ(s'),
\]
where $K_\ell(s, s') = B_\ell(s) B_{\ell'}(s')$ and
\[
\omega_{jj'}(g, \rho) = \sup_{(s_1, s_1'), (s_2, s_2') \in V_j \times V_{j'}} \frac{|g(s_1, s_1') - g(s_2, s_2')|}{\|s_1 - s_2\|^2 + \|s_1' - s_2'\|^2} \leq \rho^2
\]
is the modulus of continuity of $g$ on $V_j \times V_{j'}$. Therefore, by Assumption (A4), we have
\[
\left| \sum_{j=1}^{N_s} \sum_{j'=1}^{N_s} \int_{V_j \times V_{j'}} \{G_\eta(s_j, s_{j'}) B_\ell(s_j) B_{\ell'}(s_j') - G_\eta(s, s') B_\ell(s) B_{\ell'}(s')\} dQ(s) dQ(s') \right| \
\leq (N|\triangle|^2)^2 \times N^{-2} \times (N_s^{-1/2}|\triangle|^{-1}) \times |\triangle|^{-2} = O(N_s^{-1/2}|\triangle|).
\]
Thus, (E.4) follows.

\[
\square
\]

**Lemma E.5.** For any $\ell \in M$, $1 \leq k, k' \leq p$, normalized Bernstein basis polynomials $B_\ell$ of degree $d$, $B_\ell \in H^2$, $d \geq 0$, let $H_{\ell,k,k'} = E(X_k X_{k'}) \int_\Omega B_\ell^2(s) dQ(s)$. Suppose Assumptions (A2) and (A5)
hold and \( N_s|\triangle|^2 \to \infty \) as \( N_s \to \infty \), then with probability 1,

\[
\max_{\ell \in \mathcal{M}} \max_{1 \leq k, k' \leq p} \left| \frac{1}{N_s} \sum_{i=1}^{n} \sum_{j=1}^{N_s} B_{\ell}^2(s_j) X_{ik} X_{ik'} - H_{\ell, k, k'} \right| = O \left\{ n^{-1/2} (\log n)^{1/2} + N_s^{-1/2} |\triangle|^{-1} \right\}.
\]

**Proof.** The proof is similar to the proof of Lemma B.5 in Wang et al. (2018). The main difference is in the spline bases used. So we just give a sketch of proof here.

Let \( \varsigma_{i, \ell} \equiv \varsigma_{i, \ell, k, k'} = \frac{1}{N_s} \sum_{j=1}^{N_s} B_{\ell}^2(s_j) X_{ik} X_{ik'} \). If \( N_s|\triangle|^2 \to \infty \) as \( N_s \to \infty \), then by (E.3), we have

\[
E_{\varsigma_{i, \ell}} = N^{-1} \sum_{j=1}^{N_s} B_{\ell}^2(s_j) E(X_k X_{k'}) \leq C,
\]

and \( E(\varsigma_{i, \ell})^2 = \{ N^{-1} \sum_{j=1}^{N_s} B_{\ell}^2(s_j) \}^2 E(X_k X_{k'})^2 \leq |\triangle|^2 \).

Next define a sequence \( D_n = n^\alpha \) with \( \alpha \in (1/3, 1/2) \). We make use of the following truncated and tail decomposition

\[
X_{ikk'} = X_{ik} X_{ik'} = X_{ikk', 1} + X_{ikk', 2},
\]

where \( X_{ikk', 1} = X_{ik} X_{ik'} I \{|X_{ik} X_{ik'}| > D_n\} \), \( X_{ikk', 2} = X_{ik} X_{ik'} I \{|X_{ik} X_{ik'}| \leq D_n\} \). One can show that since \( N_s|\triangle|^2 \to \infty \) as \( N_s \to \infty \), for any \( r \geq 2 \),

\[
|E_{\varsigma_{i, \ell}}| = |E(X_{ikk', 1})| \left\{ \frac{1}{N_s} \sum_{j=1}^{N_s} B_{\ell}^2(s_j) \right\} \\
\leq D_n^{-2} E|X_{ik} X_{ik'}|^3 \left\{ \int_{\Omega} B_{\ell}^2(s) dQ(s) + O(N_s^{-1/2} |\triangle|^{-1}) \right\} \leq c_b D_n^{-2},
\]

and

\[
\sum_{n=1}^{\infty} \Pr \left\{ \max_{\ell \in \mathcal{M}, 1 \leq k, k' \leq p} \left| \sum_{i=1}^{n} \left( \sum_{j=1}^{N_s} B_{\ell}^2(s_j) \right) \right|^2 \geq \delta n^{-1/2} (\log n)^{1/2} \right\} \leq 2 \sum_{n=1}^{\infty} \sum_{\ell \in \mathcal{M}} \sum_{1 \leq k, k' \leq p} n^{-2-\tau} < \infty.
\]

Thus, \( \sup_{\ell, k, k'} \left| n^{-1} \sum_{i=1}^{n} \varsigma_{i, \ell} \right| = O_{a.s.} \left\{ n^{-1/2} (\log n)^{1/2} \right\} \) as \( n \to \infty \) by Borel-Cantelli Lemma. Furthermore,

\[
\max_{\ell, k, k'} \left| n^{-1} \sum_{i=1}^{n} \varsigma_{i, \ell} - E_{\varsigma_{i, \ell}} \right| \leq \max_{\ell, k, k'} \left| \sum_{i=1}^{n} \varsigma_{i, \ell} \right| + \max_{\ell, k, k'} \left| \sum_{i=1}^{n} \varsigma_{i, \ell} \right| + \max_{\ell, k, k'} \left| E_{\varsigma_{i, \ell}} \right| \\
= O_{a.s.}(n^{-k}) + O_{a.s.} \left\{ n^{-1/2} (\log n)^{1/2} \right\} + O \left( D_n^{-2} \right) \\
= O_{a.s.} \left\{ n^{-1/2} (\log n)^{1/2} \right\}.
\]
And the results follow by

\[
\max_{\ell \in \mathcal{M}} \max_{1 \leq k,k' \leq p} \left| \frac{1}{nN_s} \sum_{i=1}^{n} \sum_{j=1}^{N_s} B^2_\ell(s_j) X_{ik} X_{ik'} - H_{\ell,k,k'} \right|
\]

\[
= \max_{\ell \in \mathcal{M}} \max_{1 \leq k,k' \leq p} \left| n^{-1} \sum_{i=1}^{n} \mathbb{E} \mathbb{E}_{i} - E_{i,\ell} \right| + \left| E X_{ik} X_{ik'} \right| \max_{\ell \in \mathcal{M}} \left| \frac{1}{N_s} \sum_{j=1}^{N_s} B^2_\ell(s_j) - \int_{\Omega} B^2_\ell(s) dQ(s) \right|
\]

\[
= O_P \left\{ n^{-1/2} (\log n)^{1/2} \right\} + O(N_s^{-1/2} |\Delta|^{-1}).
\]

\[\square\]

The following lemma provide the uniform convergence rate at which the empirical inner product in \((E.1)\) approximates the theoretical inner product in \((E.2)\).

**Lemma E.6.** Let \(g^{(1)}_k(s) = \sum_{\ell \in \mathcal{M}} c^{(1)}_{k\ell} B_\ell(s)\), \(g^{(2)}_k(s) = \sum_{\ell \in \mathcal{M}} c^{(2)}_{k\ell} B_\ell(s)\) be any spline functions in \(S^s_d(\Delta) \cap \mathcal{H}^1\). Denote the collection of vectors of functions \(g(s) = (g_1(s),\ldots,g_p(s))^{\top}\) with \(g_k \in S^s_d(\Delta) \cap \mathcal{H}^1\), for \(k = 1,\ldots,p\). Suppose Assumptions (A2) and (A5) hold and \(N_s |\Delta|^2 \to \infty\) as \(N_s \to \infty\), then

\[
R_n = \sup_{g^{(1)},g^{(2)} \in S^s_d(\Delta) \cap \mathcal{H}^1} \left| \left\langle g^{(1)}, g^{(2)} \right\rangle X, n - \frac{\left\langle g^{(1)}, g^{(2)} \right\rangle X}{\| g^{(1)} \|_X \| g^{(2)} \|_X} \right| = O_P \left\{ n^{-1/2} (\log n)^{1/2} + N_s^{-1/2} |\Delta|^{-1} \right\}.
\]

**Proof.** It is easy to see

\[
\left\langle g^{(1)}, g^{(2)} \right\rangle X, n = \frac{1}{nN_s} \sum_{i=1}^{n} \sum_{j=1}^{N_s} \left\{ \sum_{k=1}^{p} \sum_{\ell \in \mathcal{M}} c^{(1)}_{k\ell} X_{ik} B_\ell(s_j) \right\} \left\{ \sum_{k' \prime \ell' \in \mathcal{M}} c^{(2)}_{k'\ell'} X_{ik'} B_{\ell'}(s_j) \right\}
\]

\[
= \sum_{k,k',\ell,\ell'} c^{(1)}_{k\ell} c^{(2)}_{k'\ell'} \frac{1}{nN_s} \sum_{i=1}^{n} \sum_{j=1}^{N_s} X_{ik} X_{ik'} B_\ell(s_j) B_{\ell'}(s_j).
\]

Note that

\[
\| g^{(1)} \|_X^2 = \sum_{k,k',\ell,\ell'} c^{(1)}_{k\ell} c^{(1)}_{k'\ell'} E(X_{k} X_{k'}) \int_{\Omega} B_\ell(s) B_{\ell'}(s) dQ(s),
\]

\[
\| g^{(2)} \|_X^2 = \sum_{k,k',\ell,\ell'} c^{(2)}_{k\ell} c^{(2)}_{k'\ell'} E(X_{k} X_{k'}) \int_{\Omega} B_\ell(s) B_{\ell'}(s) dQ(s).
\]
By the definition in (E.1), we have
\[ \|g(v)\|_X^2 = \sum_{k,k'=1}^p E(X_kX_{k'}) \int \Omega g_k(z)g_{k'}(z)dz = \int \Omega g^{(v)\top}(z)\Sigma_X g^{(v)}(z)dz. \]

According to Assumptions (A2) and (A5),
\[ \|g(v)\|_X^2 \asymp \int \Omega g^{(v)}(z)g^{(v)}(z)dz \asymp \sum_{k=1}^p \|g_k\|_2. \]

Then it follows from Assumptions (A1), (A2) and Lemmas E.1 that,
\[ c_v \sum_{k,\ell} \left\{ c_{k\ell}^{(v)} \right\}^2 \leq \|g(v)\|_X^2 \leq c_v \sum_{k,\ell} \left\{ c_{k\ell}^{(v)} \right\}^2, \]
and
\[ C_1 \left[ \sum_{k,\ell} \left\{ c_{k\ell}^{(1)} \right\}^2 \sum_{k',\ell'} \left\{ c_{k'\ell'}^{(2)} \right\}^2 \right]^{1/2} \leq \|g^{(1)}\|_X \|g^{(2)}\|_X \leq C_2 \left[ \sum_{k,\ell} \left\{ c_{k\ell}^{(1)} \right\}^2 \sum_{k',\ell'} \left\{ c_{k'\ell'}^{(2)} \right\}^2 \right]^{1/2}. \]

With the above preparation, we have
\[ R_n \leq \frac{\sum_{k,\ell} \sum_{k',\ell'} c_{k\ell}^{(1)} c_{k'\ell'}^{(2)}}{C_1 \left[ \sum_{k,\ell} \left\{ c_{k\ell}^{(1)} \right\}^2 \sum_{k',\ell'} \left\{ c_{k'\ell'}^{(2)} \right\}^2 \right]^{1/2}} \times \max_{\ell \in \mathcal{M}} \max_{1 \leq k,k' \leq p} \left| \frac{1}{nN_s} \sum_{i=1}^n \sum_{j=1}^N B_{s_j}^2(s_j)X_{ik}X_{ik'} - E(X_kX_{k'}) \int \Omega B_{s_j}^2(s)Q(s) \right|. \]

It follows that
\[ R_n \leq C_1^{-1} \max_{\ell \in \mathcal{M}} \max_{1 \leq k,k' \leq p} \left| \frac{1}{nN_s} \sum_{i=1}^n \sum_{j=1}^N B_{s_j}^2(s_j)X_{ik}X_{ik'} - E(X_kX_{k'}) \int \Omega B_{s_j}^2(s)Q(s) \right|. \quad (E.6) \]

The desired result follows from (E.6) and Lemma E.5.

As a direct result of Lemma E.6, we can see that
\[ \sup_{g \in S_v(\Delta) \cap H^1} \left| \|g\|_{X,n}^2 \|g\|_X^2 - 1 \right| = O_P\{n^{-1/2}(\log n)^{1/2} + N_s^{-1/2}|\Delta|^{-1}\}. \quad (E.7) \]

**Lemma E.7.** Under Assumptions (A2) and (A5), if \( N_s|\Delta|^2 \to \infty \) as \( N_s \to \infty \), then there exist constants \( 0 < c_\Gamma < C_\Gamma < \infty \), such that with probability approaching 1 as \( n \to \infty \),
\[ c_\Gamma \leq \pi_{\min}(\Gamma_n) \leq \pi_{\max}(\Gamma_n) \leq C_\Gamma, \]
\[ \Gamma_n = \frac{1}{nN_s} \sum_{i=1}^n \sum_{j=1}^N (X_{(i)}X_{(i)}^\top) \otimes \{B(s_j)B^\top(s_j)\}, \quad X_{(i)} = (X_{i1}, \ldots, X_{ip})^\top \] is a vector for the \( i \)th row of \( X \) and \( I \subset \{1, \ldots, p\} \) is a finite index set.

**Proof.** It is easy to see that for any vector \( a \in \mathbb{R}^{|Z| J_n} \),
\[
a^\top \Gamma_n a = \frac{1}{nN_s} a^\top \sum_{i=1}^n \sum_{j=1}^N (X_{(i)}X_{(i)}^\top) \otimes \{B(s_j)B^\top(s_j)\} a = \|g_\gamma\|^2_{X,n}, \quad (E.8)
\]

\( g_a = (g_{a_1}, \ldots, g_{a_p})^\top \) with \( g_{a_k} = \sum_{\ell \in \mathcal{M}} a_{k\ell} B_{\ell} \). By (E.7), we have
\[
c(1 - R_n)\|a\|^2 \leq (1 - R_n)\|g_a\|^2_X \leq \|g_a\|^2_{X,n} = (1 + R_n)\|g_a\|^2_X \leq C(1 + R_n)\|a\|^2,
\]
in which we have used the stability conditions in Lemma E.1. \( \Box \)

Let \( Z_{(ij)} = \left( X_{(i)}^\top, X_{(i)}^\top \otimes B^\top(s_j) \right)^\top \) be a \((p + pJ_n) \times 1\) vector, in which \( B(\cdot) \) is a series of normalized spline basis functions. Then \( Z_{(i)} = (Z_{(i)}^\top, \ldots, Z_{(iN_s)}^\top)^\top = \left( X_{(i)}^\top \otimes 1_{N_s}, X_{(i)}^\top \otimes B \right)^\top \) is a \((p + pJ_n)N_s \times 1\) vector, in which \( 1_{N_s} \) is a length \( N_s \) vector of ones, \( B = (B(s_1), \ldots, B(s_{N_s}))^\top \).

Let \( Z = (Z_{(1)}, \ldots, Z_{(n)})^\top \), and let \( Z_m \) be a column vector or a matrix which represents the \( m \)th column element of \( Z \), where \( Z_m = X_m \) if \( 1 \leq m \leq p \), and \( Z_m = X_{m-p} \otimes B \) if \( p + 1 \leq m \leq 2p \).

For any index set \( I \subset \{1, \ldots, 2p\} \), denote
\[
\theta_I = (\theta_m^\top, m \in I)^\top \quad \text{and} \quad Z_I = (Z_m, m \in I),
\]
where \( \theta_m = \alpha_k \) if \( 1 \leq m \leq p \), and \( \theta_m = \gamma_{m-p} \) if \( p + 1 \leq m \leq 2p \).

Let \( C_I = Z_I^\top Z_I / (N_s) \). Let \( \pi_{\min}(C_I) \) and \( \pi_{\max}(C_I) \) be the minimum and maximum eigenvalues of \( C_I \), respectively. Define \( Z_{1,I} = X_I \otimes 1_{N_s} \) and \( Z_{2,I} = X_I \otimes B \).

**Lemma E.8.** Suppose that \( |I| \) is bounded by a fixed constant independent of \( n \) and \( p \). Then under Assumptions (A2), (A4) and (A5), if \( N_s|\Delta|^2 \to \infty \) as \( N_s \to \infty \), with probability approaching one as \( n \to \infty \),
\[
c \leq \pi_{\min}(C_I) \leq \pi_{\max}(C_I) \leq C,
\]
where \( c \) and \( C \) are two positive constants.

**Proof.** For any \( a = (a_1^\top, a_2^\top)^\top \in \mathbb{R}^{|Z| (J_n+1)} \), in which \( a_1 \in \mathbb{R}^{|Z| J_n} \) and \( a_2 \in \mathbb{R}^{|Z| J_n} \), we have
\[
a^\top C_I a = (nN_s)^{-1} a^\top Z_I^\top Z_I a = (nN_s)^{-1}(Z_{1,I} a_1 + Z_{2,I} a_2)^\top (Z_{1,I} a_1 + Z_{2,I} a_2).\]
So it suffices to get the order of \((nN_s)^{-1}a_1^\top Z_{1,i}a_1\) and \((nN_s)^{-1}a_2^\top Z_{2,i}a_2\). For the first term, we have

\[
(nN_s)^{-1}a_1^\top Z_{1,i}a_1 = n^{-1}a_1^\top X_{1,i}a_1 = O_p(a_1^\top \Sigma a_1) = O_p(1).
\]

And the second term \((nN_s)^{-1}a_2^\top Z_{2,i}a_2 \propto C\) follows from Lemma E.7.

**Remark C.1.** Note that Stone’s Lemma is applied here to obtain the lower bound of \(C_T\), which requires \(|I|\) to be bounded by a fixed constant. This restriction explains why we assume that the number of nonzero nonparametric components is fixed.

For any random variable \(X\), denote \(\|X\|_p = (E|X|^p)^{1/p}\) as the \(L_p\) norm for random variable \(X\); and denote \(\|X\|_\varphi = \inf \{C > 0 : E\varphi(|X|/C) \leq 1\}\) as the Orlicz norm for random variable \(X\), where \(\varphi\) is required to be a non-decreasing, convex function with \(\varphi(0) = 0\).

**Lemma E.9.** Suppose that Assumptions (A2), (A3) and (A5) hold. Let

\[
T_{1k}^* = (nN_s)^{-1/2} \sum_{i=1}^n X_{ik} \sum_{j=1}^{N_i} \varepsilon_{ij}, \quad T_{2k}^* = (nN_s)^{-1/2} \sum_{i=1}^n X_{ik} \sum_{j=1}^{N_i} \varepsilon_{ij} B(s_j), \quad 1 \leq k \leq p.
\]

Define \(T_1^* = \max_{1 \leq k \leq p} |T_{1k}^*|\) and \(T_2^* = \max_{1 \leq k \leq p} |T_{2k}^*|\). Then we have

\[
E(T_1^*) \leq C\sigma \sqrt{\log(p)}, \quad E(T_2^*) \leq C\sigma \sqrt{\log(pJ_n)}.
\]

In particular, when \(J_n/p \to 0\), we have \(E(T_1^*) = O\{\sigma \sqrt{\log(p)}\}\) and \(E(T_2^*) = O\{\sigma \sqrt{\log(p)}\}\).

**Proof.** Let \(S_k^2 = \sum_{i=1}^n X_{ik}^2 \sigma^2\) and \(S^2 = \max_{1 \leq k \leq p} S_k^2\). For \(\sqrt{nN_s} T_{1k}^* = \sum_{i=1}^n X_{ik} \sum_{j=1}^{N_i} \varepsilon_{ij} = \sum_{i=1}^n X_{ik} 1_{N_i}^\top \varepsilon_i\), by Assumption (A3), conditional on \(X\), we have

\[
E\{\exp(t\sqrt{nN_s} T_{1k})\} = E\left\{\exp\left(t \sum_{i=1}^n X_{ik} 1_{N_i}^\top \varepsilon_i\right)\right\} = \prod_{i=1}^n E\left\{\exp\left(tX_{ik} 1_{N_i}^\top \varepsilon_i\right)\right\} \leq \prod_{i=1}^n \exp(N_i t^2 X_{ik}^2 \sigma^2 / 2) = \exp\left(\frac{t^2 N_s}{2} \sum_{i=1}^n X_{ik}^2 \sigma^2\right),
\]
Thus, $\sqrt{nN_s T_{1k}^\varepsilon}$ is $(\sqrt{nS_k})$-sub-Gaussian. Similarly, conditional on $X$ and $B$, for $\sqrt{nN_s T_{2k}^\varepsilon} = \sum_{i=1}^n X_{ik} \sum_{j=1}^{N_s} \varepsilon_{ij} B(s_j) = \sum_{i=1}^n X_{ik} B^\top \varepsilon_i$, we have

$$E\{\exp(t^\top \sqrt{nN_s T_{2k}^\varepsilon})\} = E\left\{\exp\left(t^\top \sum_{i=1}^n X_{ik} B^\top \varepsilon_i\right)\right\} = \prod_{i=1}^n E\left\{\exp\left(X_{ik} t^\top B^\top \varepsilon_i\right)\right\} \leq \prod_{i=1}^n \exp(\|Bt\|^2 X_{ik}^2 \sigma^2 / 2) \leq \exp\left(\frac{\|t\|^2 \|B\| \sum_{i=1}^n X_{ik}^2 \sigma^2}{2}\right).$$

Thus, $\sqrt{nN_s T_{2k}^\varepsilon}$ is $(\|B^\top B\|^{1/2} S_k)$-sub-Gaussian.

Define $\varphi_p(x) = \exp(x^p) - 1$, $p \geq 1$. Then $\varphi_p^{-1}(m) = \{\log(1 + m)\}^{1/p}$. By Assumption (A3), Lemma E.8 and the maximal inequality for sub-Gaussian random variables (as stated in Lemmas 2.2.1 and 2.2.2 of Van Der Vaart and Wellner (1996)),

$$E(T_1^\varepsilon | X) = E\left(\max_{1 \leq k \leq p} |T_{1k}^\varepsilon| | X\right) = (nN_s)^{-1/2} \max_{1 \leq k \leq p} \left\|\sqrt{nN_s T_{1k}^\varepsilon} \right\|_1 \leq (nN_s)^{-1/2} \left(\max_{1 \leq k \leq p} \left\|\sqrt{nN_s T_{1k}^\varepsilon} \right\|_1\right) \varphi_1 \leq K_1 \left(\log(2)\right)^{1/2} \left(\log(1 + p) (nN_s)^{-1/2} \max_{1 \leq k \leq p} \left\|\sqrt{nN_s T_{1k}^\varepsilon} \right\|_1\right) \varphi_2 \leq (nN_s)^{-1/2} K_1 \left(\log(2)\right)^{1/2} \left(\log(1 + p) \max_{1 \leq k \leq p} \left(6N_s S_k^2\right)^{1/2} \right) = n^{-1/2} K_1' S \sqrt{\log(p)},$$

$$E(T_2^\varepsilon | X) = E\left(\max_{1 \leq k \leq p} |T_{2k}^\varepsilon| | X\right) = (nN_s)^{-1/2} \max_{1 \leq k \leq p} \left\|\sqrt{nN_s T_{2k}^\varepsilon} \right\|_1 \leq (nN_s)^{-1/2} \left(\max_{1 \leq k \leq p} \left\|\sqrt{nN_s T_{2k}^\varepsilon} \right\|_1\right) \varphi_1 \leq (nN_s)^{-1/2} \left(\log(2)\right)^{1/2} \left(\max_{1 \leq k \leq p} \left\|\sqrt{nN_s T_{2k}^\varepsilon} \right\|_1\right) \varphi_2 \leq (nN_s)^{-1/2} K_2 \left(\log(2)\right)^{1/2} \left(\log(1 + pJ_n) \max_{1 \leq k \leq p} \left(6\|B^\top B\| S_k^2\right)^{1/2} \right) = n^{-1/2} K_2' S \sqrt{\log(pJ_n)},$$
where $K_1$, $K'_1$, $K_2$ and $K'_2$ are positive constants that depend only on $\varphi$ and $\sigma$. Thus,

$$E(T_1^e) = E\{E(T_1^e|X)\} \leq n^{-1/2}K'_1 \sqrt{\log(p)} E(S),$$

$$E(T_2^e) = E\{E(T_2^e|X, B)\} \leq n^{-1/2}K'_2 \sqrt{\log(pJ_n)} E(S).$$

Based on Assumption (A2), we have $E S^2 = \max_{1 \leq k \leq p} ES_k^2 = \max_{1 \leq k \leq p} \sum_{i=1}^n \sigma^2 E X_{ik}^2 \leq n\sigma^2 C^2$. The lemma follows. \hfill \square

**Lemma E.10.** Suppose that Assumptions (A2), (A3) and (A5) hold. Let

$$T_{1k}^n = (nN_s)^{-1/2} \sum_{i=1}^n X_{ik} \sum_{j=1}^{N_s} \eta_h(s_j), \quad T_{2k}^n = (nN_s)^{-1/2} \sum_{i=1}^n X_{ik} \sum_{j=1}^{N_s} \eta_h(s_j) B(s_j), \quad 1 \leq k \leq p.$$ 

Define $T_1^n = \max_{1 \leq k \leq p} |T_{1k}^n|$ and $T_2^n = \max_{1 \leq k \leq p} |T_{2k}^n|$. Then we have

$$E(T_1^n) \leq C \sqrt{N_s \log(p)}, \quad E(T_2^n) \leq C \sqrt{N_s \log(pJ_n)}.$$

**Proof.** Based on the large number theory and Assumption (A2), we have that $n^{-1} \sum_{i=1}^n X_{(i)}^{(2)}$ converges to $\Sigma_X$ almost surely. As stated in Assumption (A3), $\{X|((s) : s \in \Omega\}$ is a Donsker class, so $n^{-1/2} \sum_{i=1}^n X_{(i)} \eta_h(s)$ converges to a Gaussian process with zero mean and covariance function $\Sigma_X G_\eta(s, s')$ as $n \to \infty$ (van der Vaar and Wellner, 1996). Thus, we have $N_s^{-1/2} T_{1k}^n$, which is equal to $n^{-1/2} N_s^{-1} \sum_{i=1}^n X_{(i)} \sum_{j=1}^{N_s} \eta_h(s_j)$, converges to a normally distributed random variable with mean zero and variance $\Sigma_X \left\{ N_s^{-2} \sum_{j, j'=1}^{N_s} G_\eta(s_j, s_{j'}) \right\}$. Since $\Sigma_X \left\{ N_s^{-2} \sum_{j, j'=1}^{N_s} G_\eta(s_j, s_{j'}) \right\}$ converges to $\Sigma_X \left\{ \int_{\Omega^2} G_\eta(s, s') dsds' \right\}$ as $n \to \infty$, then we have $\Sigma_X \left\{ N_s^{-2} \sum_{j, j'=1}^{N_s} G_\eta(s_j, s_{j'}) \right\}$ is bounded above. It follows from the law of the large number, Assumptions (A2) and (A3), and similar to the proof of Lemma E.9 that $E(T_1^n) \leq C \sqrt{N_s \log(p)}$. The proof for $E(T_2^n)$ is similar thus omitted. \hfill \square
Define the group LASSO estimator \( \hat{\theta} = (\hat{\alpha}^\top, \hat{\gamma}^\top)^\top \) as the minimizer of the following:

\[
L_n(\alpha, \gamma; \lambda_{n1}, \lambda_{n2}) = \sum_{i=1}^{n} \sum_{j=1}^{N_s} \left\{ Y_{ij} - \sum_{k=1}^{p} X_{ik} \alpha_k - \sum_{k=1}^{p} X_{ik} B^\top(s_j) \gamma_k \right\}^2 + \lambda_{n1} \sum_{k=1}^{p} |\alpha_k| + \lambda_{n2} \sum_{k=1}^{p} ||\gamma_k||,
\]

where \( \lambda_{n1} \) and \( \lambda_{n2} \) are the regularization parameters controlling the amount of shrinkage.

In this section, we consider the selection and estimation properties of the group LASSO estimator \( \hat{\theta} \). Define \( \alpha_k = \hat{\theta}_k \) and \( \gamma_k = \hat{\theta}_{k+p} \). Accordingly, we can define

\[
\hat{A}_c = \{ k : |\alpha_k| \neq 0, 1 \leq k \leq p \}, \quad \hat{A}_v = \{ k : ||\gamma_k|| \neq 0, 1 \leq k \leq p \}, \quad \hat{A} = \hat{A}_c \cup \hat{A}_v.
\]

For any index set \( \mathcal{I} \), we use \( |\mathcal{I}| \) to denote the cardinality of \( \mathcal{I} \).

**Theorem E.1.** Suppose that Assumptions (A1) – (A5) hold.

(i) With probability approaching one, \( |\hat{A}_c| \leq M_1 |A_c| \) and \( |\hat{A}_v| \leq M_1 |A_v| \) for some finite constant \( M_1 > 1 \).

(ii) If \( \log(p J_n)/n \to 0 \) and \( (\lambda_{n1}^2 + \lambda_{n2}^2)/(n^2 N_s^2) \to 0 \) as \( n \to \infty \), then with probability converging to one, all the nonzero parameters \( \alpha_{0k}, k \in A_c \), and \( \beta_{0k}(\cdot), k \in A_v \), are selected.

(iii) In addition,

\[
\sum_{k \in \hat{A}_c} |\hat{\alpha}_k - \alpha_{0k}|^2 = O_p \left\{ \frac{\log(p J_n)}{n} \right\} + O_p \left\{ |\Delta|^{2(\kappa+1)} \right\} + O_p \left( \frac{\lambda_{n1}^2 + \lambda_{n2}^2}{n^2 N_s^2} \right),
\]

\[
\sum_{k \in \hat{A}_v} ||\hat{\gamma}_k - \gamma_{0k}||^2 = O_p \left\{ \frac{\log(p J_n)}{n} \right\} + O_p \left\{ |\Delta|^{2(\kappa+1)} \right\} + O_p \left( \frac{\lambda_{n1}^2 + \lambda_{n2}^2}{n^2 N_s^2} \right).
\]
Proof. The proof generally follows the proofs of Theorems B.1 and B.5 of Li et al. (2018). The main differences are the model structure and error terms as shown below:

\[ \vartheta_{ij} = Y_{ij} - \sum_{m=1}^{2p} Z_{(ij),m}^T \theta_{0m} = Y_{ij} - \sum_{k=1}^{p} X_{ik} \alpha_{0k} - \sum_{k=1}^{p} X_{ik} B^T(s_j) \gamma_{0k} \]

\[ = Y_{ij} - \sum_{k=1}^{p} X_{ik} \alpha_{0k} - \sum_{k=1}^{p} X_{ik} \beta_{0k}(s_j) + \sum_{k=1}^{p} X_{ik} \{ \beta_{0k}(s_j) - B^T(s_j) \gamma_{0k} \} \]

\[ = \eta_{ij} + \varepsilon_{ij} + \sum_{k \in A_v} X_{ik} \delta_{jk}, \quad \text{(E.11)} \]

where \( \delta_{jk} = \beta_{0k}(s_j) - B^T(s_j) \gamma_{0k} \). Denote \( \zeta_{ij} = \sum_{k \in A_v} X_{ik} \delta_{jk} \), \( \zeta_i = (\zeta_{i1}, \ldots, \zeta_{in})^T \), and \( \zeta = (\zeta_1^T, \ldots, \zeta_n^T)^T \). Let \( \vartheta = (\vartheta_{11}, \ldots, \vartheta_{nN})^T \) and \( \varepsilon = (\varepsilon_{11}, \ldots, \varepsilon_{nN})^T \). Then \( \vartheta = (\eta + \varepsilon) + \zeta \).

For any positive integers \( q_1 \) and \( q_2 \), pick some index sets \( I_1, I_2 \subseteq \{1, \ldots, p\} \) such that the cardinalities are \( |I_1| = q_1 \) and \( |I_2| = q_2 \), respectively. Define \( S_I = (\tilde{\lambda}_n^1 u_{q_1}^T, \tilde{\lambda}_n^2 \sqrt{J} U_{I_1}^T, \ldots, \tilde{\lambda}_n^2 \sqrt{J} U_{I_2}^T)^T \), where \( u_{q_1} \in \mathbb{R}^{q_1} \), \( U_k^T \) is in a unit ball with dimension \( J_n \), that is, \( U_k^T \in \mathbb{R}^{J_n} \) and \( \| U_k^T \|_2 = 1, k = 1, \ldots, q_2 \). Let \( P_I = Z_I (Z_I^T Z_I)^{-1} Z_I^T \) be the projection matrix of \( Z_I \). Denote \( V_I = Z_I (Z_I^T Z_I)^{-1} S_I - (I - P_I) Z \theta_0 \). Define

\[ \chi_{q_1,q_2} = \max_{|I_1| = q_1, |I_2| = q_2, I_1 \cup I_2} \max_{u_{q_1} \in \{\pm 1 \}^{q_1}, \| u_{q_1}^T \|_2 = 1, 1 \leq k \leq q_2} \frac{|\vartheta^T V_I|}{\| V_I \|_2}, \]

\[ \Xi_{|A_c|,|A_v|} = \{ (Z, \vartheta) : \chi_{q_1,q_2} \leq (\sqrt{N} s + \sigma) C_1 \sqrt{\frac{q_1 \log(p)}{q_2 \log(pJ_n)}}, \forall q_1 \geq |A_c|, q_2 \geq |A_v| \}, \]

where \( C_1 > 0 \) is a sufficiently large constant. As illustrated in the proof of Theorem 2.1 of Wei and Huang (2010), if \( (Z, \vartheta) \in \Xi_{|A_c|,|A_v|} \), then \( |\tilde{A}_c| \leq M_1 |A_c| \) and \( |\tilde{A}_v| \leq M_1 |A_v| \) for some constant \( M_1 > 1 \).

So it suffices to show that \( (Z, \vartheta) \in \Xi_{|A_c|,|A_v|} \). Then by the triangle inequality and Cauchy-Schwarz inequality,

\[ \frac{|\vartheta^T V_I|}{\| V_I \|_2} = \frac{|(\eta + \varepsilon)^T V_I + \zeta^T V_I|}{\| V_I \|_2} \leq \frac{|(\eta + \varepsilon)^T V_I|}{\| V_I \|_2} + \| \zeta \|_2. \]
For the $|(\eta + \varepsilon)^{T} V_{I}|/\|V_{I}\|_{2}$ part, define

$$\chi^{*}_{q_{1}, q_{2}} = \max_{|I_{1}| = q_{1}, |I_{2}| = q_{2}, I_{1} \cup I_{2} = I} \max_{u_{q_{1}} \in \{\pm 1\}^{q_{1}}} \frac{|(\eta + \varepsilon)^{T} V_{I}|}{\|V_{I}\|_{2}},$$

$$\Xi^{*}_{|A_c|, |A_v|} = \{(Z, \eta + \varepsilon) : \chi^{*}_{q_{1}, q_{2}} \leq (\sqrt{N_{s}} + \sigma)C_{2} \sqrt{q_{1} \log(p) \vee q_{2} \log(pJ_{n})}, \forall q_{1} \geq |A_{c}|, q_{2} \geq |A_{v}|\},$$

where $C_{2} > 0$ is a sufficiently large constant. As shown in the proof of Theorem 2.1 in Wei and Huang (2010), $P(\Xi^{*}_{|A_c|, |A_v|}) \to 1$. For the $\|\zeta\|_{2}$ part, for $n$ sufficiently large and $J_{n} \asymp n^{1/(\kappa + 2)}$, $(nN_{s} \ll J_{n}^{d+1}(N_{s} + \sigma^{2}) \log(pJ_{n}))$

$$\|\zeta\|_{2} \leq C \sqrt{nN_{s}|A_{v}| |\Delta|^{d+1}} \leq (\sqrt{N_{s}} + \sigma)C_{2} \sqrt{|A_{v}| \log(pJ_{n})}.$$  

It follows that $P(\Xi_{|A_c|, |A_v|}) \to 1$. This completes the proof of part (i) of Theorem E.1.

We prove part (iii) first. Let $\tilde{\theta}^{T} = (\tilde{\theta}_{1}^{T}, \ldots, \tilde{\theta}_{2p}) = (\tilde{\alpha}_{1}, \ldots, \tilde{\alpha}_{p}, \tilde{\gamma}_{1}, \ldots, \tilde{\gamma}_{p})$. For $A$ defined in (4.3) and $\tilde{A}$ defined in (E.10), denote $A_{c} = A_{c} \cup \tilde{A}_{c}, A_{v} = A_{v} \cup \tilde{A}_{v}$ and $A' = A_{c} \cup A_{v}$ and $d' = |A'|$. By part (i), $d' = O(|A|)$. Notice that $Z\tilde{\theta} = Z_{A'}\tilde{\theta}_{A'}$ and $Z\theta_{0} = Z_{A'}\theta_{0,A'}$, by the definition of $\tilde{\theta}$, $A_{c}$ and $A_{v}$,

$$\|Y - Z_{A'}\tilde{\theta}_{A'}\|^{2} + \sum_{k \in A_{c}} \bar{\lambda}_{n1}|\bar{\alpha}_{k}| + \sum_{k \in A_{v}} \bar{\lambda}_{n2}|\bar{\gamma}_{k}|$$

$$\leq \|Y - Z_{A'}\theta_{0,A'}\|^{2} + \sum_{k \in A_{c}} \bar{\lambda}_{n1}|\alpha_{0k}| + \sum_{k \in A_{v}} \bar{\lambda}_{n2}|\gamma_{0k}|.$$  

Recall that $\vartheta = Y - Z\theta_{0}$ and let $\nu = Z_{A'}(\vartheta_{A'} - \theta_{0,A'})$, so $\vartheta - \nu = Y - Z_{A'}\tilde{\theta}_{A'}$, and we have

$$\|Y - Z_{A'}\tilde{\theta}_{A'}\|^{2} - \|Y - Z_{A'}\theta_{0,A'}\|^{2} = \nu^{T}\nu - 2\vartheta^{T}\nu.$$  

Thus, from the triangle inequality and the Cauchy-Schwarz inequality,

$$\|\nu\|^{2} - 2\vartheta^{T}\nu \leq \sum_{k \in A_{c}} \bar{\lambda}_{n1}|\alpha_{0k} - |\bar{\alpha}_{k}|| + \sum_{k \in A_{v}} \bar{\lambda}_{n2}(|\gamma_{0k}| - |\bar{\gamma}_{k}|)$$

$$\leq \sqrt{d'(\bar{\lambda}_{n1}^{2} + \bar{\lambda}_{n2}^{2})}\|\vartheta_{A'} - \theta_{0,A'}\| \leq \frac{d'(\bar{\lambda}_{n1}^{2} + \bar{\lambda}_{n2}^{2})}{nN_{s}c_{s}} + \frac{1}{4}nN_{s}c_{s}\|\tilde{\theta}_{A'} - \theta_{0,A'}\|^{2}.$$  

where $c_\ast$ is the lower bound of eigenvalue of $(nN_s)^{-1}Z_A^\top Z_A$. By Lemma E.8 and part (i), $c_\ast \approx 1$ with probability approaching one. Apparently,

$$||\nu||^2 \geq nN_sc_\ast ||\tilde{\theta}_A' - \theta_0A'||^2. \quad (E.13)$$

Define $\theta^* \equiv Z_A'(Z_A^\top Z_A)^{-1}Z_A^\top \theta$ be the projection of $\theta$ onto the column space of $Z_A'$. Obviously, $\theta^\top \nu = \eta^\top \nu$. By the Cauchy-Schwarz inequality, we have

$$2|\theta^\top \nu| \leq 2||\theta^*||||\nu|| \leq 2||\theta^*||^2 + \frac{1}{2}||\nu||^2. \quad (E.14)$$

Combining (E.12), (E.13) and (E.14), we obtain

$$||\tilde{\theta}_A' - \theta_0A'||^2 \leq \frac{8||\theta^*||^2}{nNSC_\ast} + \frac{4d'(\lambda_{12}^2 + \lambda_{22}^2)}{n^2N^2SC_\ast^2}. \quad (E.15)$$

Recall that $\theta = (\theta_1, \ldots, \theta_n)^\top$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, and we have $\theta = (\eta + \varepsilon) + \zeta$. By the fact that $|\delta_k| = O(|\Delta|^{d+1})$, (E.11) and Assumption (A2),

$$||\theta^*||^2 = ||\eta^* + \varepsilon^* + \zeta_A'||^2 \leq 2||\eta^* + \varepsilon^*||^2 + 2||\zeta_A'||^2 \leq 2||\eta^* + \varepsilon^*||^2 + CnNsd^2|\Delta|^2(\kappa + 1), \quad (E.16)$$

where $\eta^* \equiv P_{Z_A'}\eta$, $\varepsilon^* \equiv P_{Z_A'}\varepsilon$ and $\zeta_A' \equiv P_{Z_A'}\zeta_A'$ are the projections of $\eta$, $\varepsilon$ and $\zeta_A'$ onto the column space of $Z_A'$, respectively. Note that $||\eta^* + \varepsilon^*||^2 = ||Z_A^\top(Z_A^\top Z_A)^{-1/2}Z_A^\top(\eta + \varepsilon)||^2 \leq (nNsc_\ast)^{-1}||Z_A^\top(\eta + \varepsilon)||^2$, and $\max_{|I| \leq d'} ||Z_A^\top(\eta + \varepsilon)||^2 = \max_{|I| \leq d'} \sum_{m \in I} ||Z_m\eta||^2 + ||Z_m\varepsilon||^2 \leq (nNsd')^2 \mathcal{V} \{ (T_1) \mathcal{V} + (T_2) \mathcal{V} \} \cup \{ (T_1') \mathcal{V} + (T_2') \mathcal{V} \}$, where $T_1'$ and $T_2'$ are defined in Lemma E.10, and $T_1$ and $T_2$ are defined in Lemma E.9. According to Lemmas E.10 and E.9, $\max_{|I| \leq d'} ||Z_A^\top(\eta + \varepsilon)||^2 = O_p\{nNsd'(N_s + \sigma^2) \log(pJ_n)\}$. Therefore,

$$||\eta^* + \varepsilon^*||^2 = O_p\{d'(N_s + \sigma^2)c_\ast^{-1} \log(pJ_n)\}. \quad (E.17)$$
Combing (E.15), (E.16) and (E.17), we conclude that
\[
\|\tilde{\theta}_{A'} - \theta_{0,A'}\|^2 = O_p \left\{ \frac{d'(N_s + \sigma^2) \log(pJ_n)}{nN_s c_s^2} \right\} + O_p \left( \frac{d^2|\triangle|^{2(\kappa+1)}}{c_s} \right) + \frac{4d'(\bar{\lambda}_n^2 + \bar{\lambda}_n^2)}{n^2N_s c_s^2} \\
= O_p \{n^{-1} \log(pJ_n)\} + O_p \{|\triangle|^{2(\kappa+1)}\} + O\{(nN_s)^{-2}(\bar{\lambda}_n^2 + \bar{\lambda}_n^2)\},
\]
where the last inequality follows by \(d' = O(q)\) and \(c_s \asymp 1\) with probability approaching one.

We now prove part (ii). Under Assumption (A1), if \(|\alpha_{0k}| \neq 0\) but \(|\tilde{\alpha}_k| = 0\), then \(|\alpha_{0k} - \tilde{\alpha}_k| \geq c_{\alpha}\); further according to Lemma E.1, if \(|\gamma_{0k}| \neq 0\) but \(|\tilde{\gamma}_k| = 0\), then \(|\gamma_{0k} - \tilde{\gamma}_k| \geq c_{1}c_{\beta}\), which contradicts part (iii) when \(\log(pJ_n)/n \to 0\) and \((\bar{\lambda}_n^2 + \bar{\lambda}_n^2)/(nN_s)^2 \to 0\). □

Selection and estimation properties of the adaptive group LASSO estimators

In this section, we give the selection and estimation properties of the adaptive group LASSO estimators.

Define \(A_c^* = \{k : |\alpha_k| \neq 0, 1 \leq k \leq p\}\) and \(\hat{A}_c = \{k : \|\tilde{\gamma}_k\| = 0, 1 \leq k \leq p\}\). Recall the definition of \(A_c\) and \(A_v\) in (4.3), \(\hat{A}_c\) and \(\hat{A}_v\) in (4.8), we have \(A_c = A_c^*/A_v\) and \(\hat{A}_c = \hat{A}_c^*/\hat{A}_v\). Theorem 4.1 follows directly from Lemma E.11 below.

**Lemma E.11.** Suppose that Assumptions (A1) – (A7) hold. As \(n \to \infty\), \(\Pr(\hat{A}_c^* = A_c^*) \to 1\), \(\Pr(\hat{A}_v = A_v) \to 1\).

**Proof.** To minimize
\[
L_n(\theta; \lambda_1, \lambda_2) = L_n(\alpha, \gamma; \lambda_1, \lambda_2) = \sum_{i=1}^{n} \sum_{j=1}^{N_s} \left\{ Y_{ij} - \sum_{k=1}^{p} X_{ik} \alpha_k - \sum_{k=1}^{p} X_{ik} B(s_j)^{\top} \gamma_k \right\}^2 \\
+ \lambda_1 \sum_{k=1}^{p} w_{n,k} c_{\alpha_k} \|\alpha_k\| + \lambda_2 \sum_{k=1}^{p} w_{n,k} \|\gamma_k\|,
\]
by the Karush-Kuhn-Tucker (KKT) condition (Boyd and Vandenberghe, 2004), any \(\theta\) satisfying
the following KKT conditions
\[(C1-1) \ (X_k \otimes 1_{N_s})^{\top} \{ Y - (X \otimes 1_{N_s}) \alpha - (X \otimes B) \gamma \} = \lambda_1 w_{n,k} c_{\alpha_k} |\alpha_k|, \text{ for any } k \in A_c^*,
\]
\((C1-2) \quad (X_k \otimes B)^\top \{Y - (X \otimes 1_{N_k})\alpha - (X \otimes B)\gamma\} = \lambda_{n2}w_{n,k}'\gamma_k/\|\gamma_k\|, \text{ for any } k \in A_v,\)

\((C2) \quad \left| (X_k \otimes 1_{N_k})^\top (Y - (X \otimes 1_{N_k})\alpha - (X \otimes B)\gamma) \right| \leq \lambda_{n1}w_{n,k}'', \text{ for any } k \notin A_v,\)

\((C3) \quad \| (X_k \otimes B)^\top \{Y - (X \otimes 1_{N_k})\alpha - (X \otimes B)\gamma\} \| \leq \lambda_{n2}w_{n,k}'', \text{ for any } k \notin A_v,\)

is the unique minimizer of \(L_n(\theta; \lambda_{n1}, \lambda_{n2}).\)

Recall that \(A = A^*_c \cup A_v\) is the active index set for \(X\). Define \(\boldsymbol{\theta}_0 = \left(Z_A^\top Z_A\right)^{-1}Z_A^\top Y\), a vector with length \(|A^*_c| + |A_v||J_n\). Denote two vectors \(v_1\) and \(v_2\) with elements in the form:

\[
v_{1m} = \frac{\omega_{n,m}}{\|\theta_{0m}\|} I\{m \in A^*_c\} + 0_{J_n} I\{m \notin A^*_c\}, \quad \text{(E.18)} \]

\[
v_{2m} = \frac{\omega_n I\{m - |A^*_c| \notin A_v\}}{\|\theta_{0m}\|} I\{m - |A^*_c| \in A_v\}, \quad \forall m \in A.\]

Next define \(\hat{\theta}_{0,A} \equiv (\hat{\theta}_{0m}, m \in A)^\top = \left(Z_A^\top Z_A\right)^{-1}(Z_A^\top Y - \lambda_{n1}v_1 - \lambda_{n2}v_2)\), which can be decomposed to \(\hat{\theta}_{0,A^*_c} = (\hat{\theta}_{0m}, m \in A^*_c)^\top\), and \(\hat{\theta}_{0,A_v} = (\hat{\theta}_{0m}, m \in A_v)^\top\). So we can represent \(\hat{\theta}_0 \equiv (\hat{\theta}_{0,A^*_c}^\top 0_{p-|A^*_c|}, \hat{\theta}_{0,A_v}^\top 0_{p-|A_v|})^\top\). Let \(\hat{A}^0 = \{1 \leq m \leq 2p : \|\hat{\theta}_{0m}\| > 0\}\). Apparently, \(\hat{A}^0 \subseteq A\).

Notice that \(Z_{\hat{\theta}_0} = Z_A^\top \hat{\theta}_{0,A} \) and \(\{Z_m, m \in A\}\) are linearly independent, so by the definition of \(\hat{\theta}_0\), (C1-1) and (C1-2) hold for \(\hat{\theta}_0\) if \(\hat{A}^0 \supseteq A\). So if \(\hat{\theta}_0\) satisfies

\((C1') \quad \hat{A}_0 \supseteq A,\)

(C2) and (C3), then \(\hat{\theta}_0\) is the unique minimizer of \(L_n(\theta; \lambda_{n1}, \lambda_{n2})\). Therefore, in order to show \(\Pr(\hat{A}_c = A^*_c) \to 1\) and \(\Pr(\hat{A}_v = A_v) \to 1\), it is equivalent to show \(\hat{\theta}_0\) satisfies (C1'), (C2) and (C3) with probability approaching one, as \(n \to \infty\).

Further notice that

\((C1'') \quad \|\theta_{0m}\| - \|\hat{\theta}_{0m}\| < \|\theta_{0m}\|, \forall m \in A\)

implies Condition (C1'). Therefore, to show \(\hat{\theta}_0\) is the unique minimizer of \(L_n(\theta; \lambda_{n1}, \lambda_{n2})\), and consequently, \(\Pr(\hat{A}_c = A^*_c) \to 1\) and \(\Pr(\hat{A}_v = A_v) \to 1\), it suffices to show that \(\hat{\theta}_0\) satisfies Conditions (C1''), (C2) and (C3) with probability approaching one, as \(n \to \infty\).
According to Lemma E.13 and Lemma E.14 in Section C, we obtain that

\[
\Pr(\hat{A} \neq A) \leq \Pr(\|\theta_{0m} - \hat{\theta}_{0m}\| \geq \|\theta_{0m}\|, \exists m \in A) \\
+ \Pr((|X_k \otimes 1_{N_s}^T Y - Z\hat{\theta}_0|) > \lambda w_{n,k}, \exists k \notin A_c^*) \\
+ \Pr((\|X_k \otimes B^T Y - Z\hat{\theta}_0\|_2 > \lambda w_{n,k}, \exists k \notin A_v) \rightarrow 0,
\]

as \( n \rightarrow \infty \). This completes the proof. \( \square \)

**Proof of Theorem 4.2.** The idea of the proof is similar to the proof of part (iii) in Theorem E.1 in Section C, and the main difference is that we look at index set \( A \) instead of \( A' \). Let \( \pi_1 \) and \( \pi_2 \) be the minimum and maximum eigenvalues of \( C_A \), respectively, and let \( \pi_3 = \max_{1 \leq m \leq 2p} \|n^{-1}Z_m^T z_m\|_2 \).

By Lemma E.8 in Section C, \( \pi_1 \approx 1, \pi_2 \approx 1 \) and \( \pi_3 \approx 1 \). According to the proof of Theorem 4.1, with probability approaching one, we have

\[
\tilde{\theta}_A = \theta_{0,A} = (Z_A^T Z_A)^{-1}(Z_A^T Y - \lambda_1 v_1 - \lambda_2 v_2) \\
= (Z_A^T Z_A)^{-1}\{Z_A^T (Z_A \theta_{0,A} + \zeta + \eta + \epsilon) - \lambda_1 v_1 - \lambda_2 v_2\} \\
= \theta_{0,A} + (Z_A^T Z_A)^{-1}\{Z_A^T (\zeta + \eta + \epsilon) - \lambda_1 v_1 - \lambda_2 v_2\}.
\]

Let \( H = I - Z_A (Z_A^T Z_A)^{-1} Z_A^T \) be an \((n_{N_s}) \times (n_{N_s})\) matrix, then

\[
\tilde{\theta}_A - \theta_{0,A} = (n_{N_s})^{-1}C_A^{-1}\{Z_A^T (\zeta + \eta + \epsilon) - \lambda_1 v_1 - \lambda_2 v_2\}. \tag{E.19}
\]

For \( \vartheta = Y - Z\theta \), define \( \vartheta_* \) as the projection of \( \vartheta \) to the column space of \( Z_A \), that is, \( \vartheta_* = P_{Z_A} \vartheta = Z_A (Z_A^T Z_A)^{-1} Z_A^T \vartheta \). Then for \( \eta_* \equiv P_{Z_A} \eta \) and \( \epsilon_* \equiv P_{Z_A} \epsilon \), similar to (E.15), (E.16) and (E.17), and by Lemma E.8 in Section C, \( \|\eta_* + \epsilon_*\|^2 = \|(Z_A^T Z_A)^{-1/2}Z_A^T (\eta + \epsilon)\|^2 \leq (n_{N_s} \pi_1)^{-1}\|Z_A^T (\eta + \epsilon)\|^2 \),
by Lemmas E.10 and E.9,

\[
\|Z_A^T(\eta + \varepsilon)\|^2 = \sum_{k\in A_c} \|X_{ik}1_{N_\delta}(\eta_i + \varepsilon_i)\|^2 + \sum_{k\in A_v} \|X_{ik}B(\eta_i + \varepsilon_i)\|^2 \\
\leq C\{nN_\delta(N_s + \sigma^2)|A_c|\log(|A_c|)\} + C\{nN_\delta(N_s + \sigma^2)J_n|A_v|\log(J_n|A_v|)\}.
\]

Thus,

\[
\|\vartheta_s\|^2 \leq 2\|\eta_s + \varepsilon_s\|^2 + 2\|\zeta_A\|^2 \\
= O_p \left[\pi_1^{-1}(N_s + \sigma^2)\{|A_c|\log(|A_c|) + J_n|A_v|\log(J_n|A_v|)\}\right] + O_p(nN_s|A_v|^2\triangle^{2(\kappa+1)})
\]

and

\[
\|\hat{\theta}_A - \theta_{0,A}\|^2 \leq \frac{8\|\vartheta_s\|^2}{nN_s\pi_1} + \frac{4(\lambda_{11}^2|A_c| + \lambda_{22}^2|A_v|)}{n^2N_s^2\pi_1^2} \left[\frac{(N_s + \sigma^2)\{|A_c|\log(|A_c|) + J_n|A_v|\log(J_n|A_v|)\}}{nN_s}\right] \\
+ O_p \left(|A_v|^2\triangle^{2(\kappa+1)}\right) + O\left(\frac{\lambda_{11}^2|A_c| + \lambda_{22}^2|A_v|}{n^2N_s^2}\right).
\]

Therefore, the result follows by the fact that \(\hat{\alpha}_{A_c} - \alpha_{0,A_c} = (I_{|A_c|} \ 0_{|A_c| \times (|A_c|J_n)})(\hat{\theta}_A - \theta_{0,A})\), \(\hat{\gamma}_{A_v} - \gamma_{0,A_v} = (0_{(|A_v|J_n) \times |A_v|} \ I_{|A_v|J_n})(\hat{\theta}_A - \theta_{0,A})\) and \(\hat{\beta}_{A_v}(s) - \beta_{0,A_v}(s) = B(s)^\top(\hat{\gamma}_{A_v} - \gamma_{0,A_v})\).

**Technical lemmas used in the proof of Theorem 4.1**

**Lemma E.12.** For \(v_1\) and \(v_2\) with elements

\[
v_{1m} = \frac{\omega_{n,m}^c\bar{\theta}_{0m}}{||\bar{\theta}_{0m}||}I\{m \in A_c^*\} + 0_{J_n}I\{m \notin A_c^*\},
\]

\[
v_{2m} = \frac{\omega_{n,m-|A_c^*|}^v\bar{\theta}_{0m}}{||\bar{\theta}_{0m}||}I\{m - |A_c^*| \in A_v\},
\]

for all \(m \in A\). Under Assumptions (A1), (A5) and (A6),

\[
\|v_1\|_2^2 = O_p(h_{n1}^2) = O_p\left(b_\alpha^2c_\alpha^{-4}r_{\alpha}^{-2} + |A_c^*|c_\alpha^{-2}\right), \quad (E.20)
\]

\[
\|v_2\|_2^2 = O_p(h_{n2}^2) = O_p\left(b_\gamma^2c_\gamma^{-4}r_{\gamma}^{-2} + |A_v|c_\gamma^{-2}\right). \quad (E.21)
\]
Proof. We have
\[ \|v_2\|^2 = \sum_{k \in A_v} (w_{n,k}^*)^2 = \sum_{k \in A_v} \|\gamma_k\|^2 = \sum_{k \in A_v} \|\gamma_0k\|^2 - \|\bar{\gamma}_k\|^2 + \sum_{k \in A_v} \|\gamma_0k\|^2. \]

Under Assumptions (A1), (A5) and (A6), Lemma E.1 and result (iii) in Theorem E.1, we have
\[ \sum_{k \in A_v} \|\gamma_0k\|^2 - \|\bar{\gamma}_k\|^2 \leq C \epsilon_\gamma^{-4} b_{\gamma}^{-2} \|\bar{\gamma} - \gamma_0\|^2 = O_p \left( \epsilon_\gamma b_{\gamma}^{-2} c_\gamma^{-4} r_{\gamma}^{-2} \right), \]
and the result follows by \( \sum_{k \in A_v} \|\gamma_0k\|^2 \leq |A_v| c_\gamma^{-2} \). The order for \( \|v_1\| \) can be similarly derived. \( \square \)

Lemma E.13. Under Assumptions (A1)–(A7), as \( n \to \infty \),
\[ \Pr \left( \|\theta_0m - \tilde{\theta}_0m\| \geq \|\theta_0m\|, \exists m \in A \right) \to 0. \]

Proof. Let \( \Upsilon_m \) be a \( d_m \times (|A_v^e| + |A_v| J_n) \) matrix, \( d_m = 1 \) for \( m \in A_v^e \), and \( d_m = J_n \) for \( m \in A_v \), with the form
\[ \Upsilon_m = \begin{pmatrix} \Upsilon_{1,m} & 0_{(|A_v^e| \times (|A_v| J_n))} \\ \Upsilon_{2,m} & 0_{J_n \times |A_v^e|} \end{pmatrix} I(m \in A_v^e) + \begin{pmatrix} 0_{J_n \times |A_v^e|} & \Upsilon_{2,m} \end{pmatrix} I(m \in A_v), \]
where \( \Upsilon_{1,m} = (0, \ldots, 0, 1, 0, \ldots, 0) \) and \( \Upsilon_{2,m} = (0_{J_n \times J_n}, \ldots, 0_{J_n \times J_n}, I_{J_n}, 0_{J_n \times J_n}, \ldots, 0_{J_n \times J_n}) \), and scalar 1 is the \( m \)-th element of the \( |A_v^e| \)-dimensional vector \( \Upsilon_{1,m} \), and the \( J_n \times J_n \) identity matrix \( I_{J_n} \) is at the \( m \)-th block of the \( J_n \times (|A_v| J_n) \) matrix \( \Upsilon_{2,m} \).

Then from (E.19), \( \hat{\theta}_0m - \theta_0m = (nN_s)^{-1} \Upsilon_m C_{A}^{-1} \{ Z_{A^e}(\eta + \varepsilon) + Z_{A^e} \zeta - \lambda_1 n_1 v_1 - \lambda_2 n_2 v_2 \} \). By the triangle inequality,
\[ nN_s \|\hat{\theta}_0m - \theta_0m\| \leq \| \Upsilon_m C_{A}^{-1} Z_{A^e}(\eta + \varepsilon) \| + \| \Upsilon_m C_{A}^{-1} Z_{A^e} \zeta \| + \| \Upsilon_m C_{A}^{-1} (\lambda_1 n_1 v_1 + \lambda_2 n_2 v_2) \|. \]

Recall that \( \pi_1 \) and \( \pi_2 \) are the minimum and maximum eigenvalues of \( C_A \), respectively. By Lemmas E.8 and E.9, the first term on the right-hand side
\[ \max_{m \in A} \| \Upsilon_m C_{A}^{-1} Z_{A^e}(\eta + \varepsilon) \| \leq \frac{1}{\pi_1} \| Z_{A^e}(\eta + \varepsilon) \| \]
\[ = O_p \left( \sqrt{nN_s(N_s + \sigma^2)} \{ |A_v^e| \log(|A_v^e|) + \sqrt{J_n |A_v| \log(J_n |A_v|)} \} \right) \]
\[ = O_p \left( \sqrt{nN_s(N_s + \sigma^2) J_n \log(J_n)} \right). \]
By Lemma E.8, the second term
\[
\max_{m \in A} \| Y_m C_A^{-1} Z_A^\top \zeta \| \leq \pi_1^{-1} \| Z_A^\top Z_A \|^{1/2} \| \zeta \| \leq \pi_1^{-1} (n N_s \pi_2)^{1/2} C \{ (n N_s)^{1/2} | \Delta |^{d+1} \} \\
= O_p(n N_s | \Delta |^{d+1}).
\]

By Lemmas E.8 and E.12, the third term
\[
\max_{m \in A} \| Y_m C_A^{-1} (\lambda_n v_1 + \lambda_n v_2) \| \leq \pi_1^{-1} \| \lambda_n v_1 + \lambda_n v_2 \| = O_p(\lambda_n h_1 + \lambda_n h_2).
\]

Thus, the claim follows by Assumption (A7).

**Lemma E.14.** Under Assumptions (A1)–(A7), as \( n \to \infty \),
\[
\Pr \left\{ |(X_k \otimes 1_{N_s})^\top (Y - Z \hat{\theta}_0)| > \lambda_n w_{n,k}, \ \exists k \notin A_n^c \right\} \to 0,
\]
\[
\Pr \left\{ \| (X_k \otimes B)^\top (Y - Z \hat{\theta}_0) \| > \lambda_n w_{n,k}, \ \exists k \notin A_v \right\} \to 0.
\]

**Proof.** Note that
\[
Y - Z \hat{\theta}_0 = Z_A \theta_{0,A} + \zeta + \eta + \varepsilon - Z_A \hat{\theta}_{0,A}
\]
\[
= -Z_A(Z_A^\top Z_A)^{-1} \{ Z_A^\top (\zeta + \eta + \varepsilon) - \lambda_n v_1 - \lambda_n v_2 \} + \zeta + \varepsilon
\]
\[
= \{ I - Z_A(Z_A^\top Z_A)^{-1} Z_A^\top \} (\zeta + \eta + \varepsilon) + Z_A(Z_A^\top Z_A)^{-1} (\lambda_n v_1 + \lambda_n v_2)
\]
\[
= H_A(\eta + \varepsilon) + H_A \zeta + (n N_s)^{-1} Z_A C_A^{-1} (\lambda_n v_1 + \lambda_n v_2), \tag{E.22}
\]
where \( H_A = I - Z_A(Z_A^\top Z_A)^{-1} Z_A^\top \). Then for any \( 1 \leq m \leq 2p \), by (E.22), we have
\[
Z_m^\top (Y - Z_A \hat{\theta}_{0,A}) = Z_m^\top H_A(\eta + \varepsilon) + Z_m^\top H_A \zeta + (n N_s)^{-1} Z_m^\top Z_A C_A^{-1} (\lambda_n v_1 + \lambda_n v_2). \tag{E.23}
\]

By Lemma E.9, we have \( \mathbb{E}(\max_{k \notin A_n^c} |(X_k \otimes 1_{N_s})^\top H_A(\eta + \varepsilon)|) = O\{ \sqrt{n N_s (N_s + \sigma^2) \log(p)} \} \) and \( \mathbb{E}(\max_{k \notin A_v} \| (X_k \otimes B)^\top H_A(\eta + \varepsilon) \|_2) = O\{ \sqrt{n N_s (N_s + \sigma^2) \log(p J_n)} \} \).
Then for the $\mathbf{Z}_m^T \mathbf{H}_A(\eta + \varepsilon)$ part of (E.23), from Assumption (A6), for all $k \in \mathcal{N}$, $w^c_{n,k} = |\bar{\alpha}_k|^{-1} = O_p(r_{n\alpha})$, there exists a positive constant $c_1$, such that

$$\Pr \left( |\mathbf{X}_k \otimes 1_{N_s}^T \mathbf{H}_A(\eta + \varepsilon)| > \lambda_n w^c_{n,k}/3, \exists k \notin A^*_c \right)$$

$$\leq \Pr \left( |\mathbf{X}_k \otimes 1_{N_s}^T \mathbf{H}_A(\eta + \varepsilon)| > c_1 \lambda_n r_{n\alpha}, \exists k \notin A^*_c \right) + o(1)$$

$$= \Pr \left( \max_{k \notin A^*_c} |(nN_s)^{-1/2}(\mathbf{X}_k \otimes 1_{N_s})^T \mathbf{H}_A(\eta + \varepsilon)| > (nN_s)^{-1/2}c_1 \lambda_n r_{n\alpha} \right) + o(1)$$

$$\leq (nN_s)^{-1/2}(c_1 \lambda_n r_{n\alpha})^{-1}E \left\{ \max_{k \notin A^*_c} |(nN_s)^{-1/2}(\mathbf{X}_k \otimes 1_{N_s})^T \mathbf{H}_A(\eta + \varepsilon)| \right\} + o(1)$$

$$\leq (nN_s)^{-1/2}(c_1 \lambda_n r_{n\alpha})^{-1}O \left\{ \sqrt{nN_s + \sigma^2} \log(p) \lambda_n^{-1} r_{n\alpha}^{-1} \right\} + o(1)$$

$$= O \left\{ \sqrt{nN_s(N_s + \sigma^2) \log(p) \lambda_n^{-1} r_{n\alpha}^{-1}} \right\} + o(1).$$

Similarly,

$$\Pr \left( |\|\mathbf{X}_k \otimes \mathbf{B}\| \mathbf{H}_A(\eta + \varepsilon)| > \lambda_n w^\gamma_{n,k} (nN_s)^{-1/2} c_1 \lambda_n r_{n\alpha} \right)$$

$$= O \left\{ \sqrt{nN_s(N_s + \sigma^2) \log(pJ_n) \lambda_n^{-1} r_{n\gamma}^{-1}} \right\} + o(1).$$

Recall the definition of $\mathcal{N}$ in (4.3) and $\pi_3 = \max_{1 \leq m \leq 2p} \|nN_s^{-1/2} \mathbf{Z}_m^T \mathbf{Z}_m\|_2$, then we have,

$$\max_{k \notin A^*_c} \|\mathbf{X}_k \otimes 1_{N_s}^T \mathbf{H}_A \zeta\| \leq (nN_s)^{-1/2} \max_{k \notin A^*_c} \|\mathbf{X}_k \otimes 1_{N_s}^T \mathbf{H}_A \zeta\|^{-1/2} \|\mathbf{H}_A \zeta\|$$

$$= O_p(nN_s |\Delta|^{1/(k+1)}).$$

Consequently,

$$\Pr \left( |\|\mathbf{X}_k \otimes 1_{N_s}\| \mathbf{H}_A \zeta| > \lambda_n w^c_{n,k}, \exists k \notin A^*_c \right) = O \left\{ \frac{nN_s |\Delta|^{1/(k+1)} \lambda_n r_{n\alpha}}{\lambda_n r_{n\alpha}} \right\} + o(1).$$

Similarly,

$$\Pr \left( |\mathbf{X}_k \otimes \mathbf{B}\| \mathbf{H}_A \zeta| > \lambda_n w^\gamma_{n,k}, \exists k \notin A^*_c \right) = O \left\{ \frac{nN_s |\Delta|^{1/(k+1)} \lambda_n r_{n\gamma}}{\lambda_n r_{n\gamma}} \right\} + o(1).$$
To bound the last term in (E.23), it follows by Lemma E.8 and (E.20) in Lemma E.12 that

\[
\max_{m \in \mathbb{N}} \|(n N_s)^{-1} Z_m^\top Z A^{-1} (\lambda_n v_1 + \lambda_n v_2)\|
\leq \max_{1 \leq m \leq 2p} \|(n N_s)^{-1/2} Z_m\| \|(n N_s)^{-1} Z A^{-1}\|^{1/2} ||C_A^{-1}|| \|\lambda_n v_1 + \lambda_n v_2\|
\leq \pi_3^{1/2} \pi_2^{1/2} \rho^{-1} O_p(\lambda_n h_{n1} + \lambda_n h_{n2}) = O_p(\lambda_n h_{n1} + \lambda_n h_{n2}).
\]

Consequently,

\[
\Pr\left(\|(n N_s)^{-1} (X_k \otimes 1_{N_s})^\top Z A^{-1} (\lambda_n v_1 + \lambda_n v_2)\| \|\lambda_n v_1 + \lambda_n v_2\| > \lambda_n w_{n,k}^c, \exists k \notin A_c^d\right)
\leq (c_1 \lambda_n r_{n\alpha})^{-1} \mathbb{E}\left\{\max_{k \notin A_c^d} \|(n N_s)^{-1/2} (X_k \otimes 1_{N_s})^\top Z A^{-1} (\lambda_n v_1 + \lambda_n v_2)\|\right\} + o(1)
= O\left\{\lambda_n^{-1} r_{n\alpha}^{-1} (\lambda_n h_{n1} + \lambda_n h_{n2})\right\},
\]

and

\[
\Pr\left(\|(n N_s)^{-1} (X_k \otimes B)^\top Z A^{-1} (\lambda_n v_1 + \lambda_n v_2)\| \|\lambda_n v_1 + \lambda_n v_2\| > \lambda_n w_{n,k}^\nu, \exists k \notin A_e\right)
\leq (c_1 \lambda_n r_{n\gamma})^{-1} \mathbb{E}\left\{\max_{k \notin A_e} \|(n N_s)^{-1/2} (X_k \otimes B)^\top Z A^{-1} (\lambda_n v_1 + \lambda_n v_2)\|\right\} + o(1)
= O\left\{\lambda_n^{-1} r_{n\gamma}^{-1} (\lambda_n h_{n1} + \lambda_n h_{n2})\right\}.
\]