ABSTRACT

WANG, DONG. Model Selection and Estimation in Generalized Additive Models and Generalized Additive Mixed Models. (Under the direction of Dr. Daowen Zhang.)

In this dissertation, we propose a method of model selection and estimation in generalized additive models (GAMs) for data from a distribution in the exponential family. The linear mixed model representation of the smoothing spline estimators of the non-parametric functions is constructed, where the inverse of the smoothing parameters are treated as extra variance components and the importance of these nonparametric functions is controlled by the induced variance components. By maximizing the penalized quasi-likelihood with the adaptive LASSO, we could effectively select the important non-parametric functions. Approximate EM algorithms are applied to achieve the goal of model selection and estimation. In addition, we also calculate the approximate pointwise frequentist and Bayesian confidence intervals for selected functions. The eigenvalue-eigenvector decomposition approach is used to approximate the induced random effects from the nonparametric functions in order to reduce the dimensions of matrices and speed up the computation.

In the case of longitudinal data, we apply the generalized additive mixed models (GAMMs) to model the relationship. The subject-specific random effects are introduced to accommodate the correlation among the responses. Similarly, we propose the adaptive LASSO for generalized additive mixed models to perform model selection and estimation.

To evaluate the method, we investigate simulation studies and provide real data applications.
Model Selection and Estimation in Generalized Additive Models and Generalized Additive Mixed Models

by
Dong Wang

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Statistics
Raleigh, North Carolina
2013

APPROVED BY:

__________________________  ______________________________
Dr. Marie Davidian           Dr. Huixia (Judy) Wang

__________________________  ______________________________
Dr. Howard Bondell            Dr. Daowen Zhang
            Chair of Advisory Committee
DEDICATION

To my loving family.
BIOGRAPHY

Dong Wang was born in Xingtai, Hebei, China. Before she came to US, she received her Bachelor and Master degree in Physics. In 2005, she joined the University of Toledo to start her graduate study in Statistics and got a Master degree two years later. In 2007, she entered the Department of Statistics, North Carolina State University to continue her PhD study. She expects to complete all PhD work by December, 2012.
I express my deepest gratitude to my advisor, Dr. Daowen Zhang, for his continuous support and guidance through the past several years. His creative ideas have been an important source for my research. I am so lucky to have him as my advisor. Without his encouragement and guidance, I would not finish this research.

I would like to thank my committee members, Dr. Marie Davidian, Dr. Judy Wang and Dr. Howard Bondell for their valuable comments to my research and Dr. Mohamed Bourham for serving as my committee member. My thanks also go to previous and current Directors of Graduate Programs: Dr. Pam Arroway, Dr. Leonard Stefanski, Dr. Jacqueline Hughes-Oliver, Dr. John Monahan and Dr. Sujit Ghosh. They encourage students to engage in academic activities, provide financial support for graduate students and deliver great lectures as well. Furthermore, I would like to show my appreciation to all faculty and staff members for their hard work for the department.

In the past five years, I am fortunate enough to have a lot of friends in the department. They always give me support, share the happiness with me and comfort me. Those moments will be in my heart forever.

I am deeply grateful to my parents, Hanlin Wang and Yuqing Cheng, my in-laws for their constant love and support. They are always there for me to count on. Their love and understanding helped me get through tough times. I owe great appreciation to my husband, Jianhui. I am full of gratitude to his love, encouragement and patience. Last but not least, I would like to thank my precious baby, Matthew—you are the angel of my life.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>List of Tables</td>
<td>vii</td>
</tr>
<tr>
<td></td>
<td>List of Figures</td>
<td>ix</td>
</tr>
<tr>
<td><strong>Chapter 1</strong></td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Literature Review on Variable Selection Methods</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1</td>
<td>Linear Regression Models and Variable Selection Methods</td>
<td>1</td>
</tr>
<tr>
<td>1.1.2</td>
<td>Nonparametric Models and Variable Selection Methods</td>
<td>6</td>
</tr>
<tr>
<td>1.2</td>
<td>A Smoothing Spline Estimation of a Nonparametric Function</td>
<td>12</td>
</tr>
<tr>
<td>1.3</td>
<td>The Linear Mixed Model Representation</td>
<td>14</td>
</tr>
<tr>
<td>1.4</td>
<td>The Framework of the Dissertation</td>
<td>14</td>
</tr>
<tr>
<td><strong>Chapter 2</strong></td>
<td>Model Selection and Estimation in Generalized Additive Models</td>
<td>16</td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>16</td>
</tr>
<tr>
<td>2.2</td>
<td>Generalized Additive Models and the Representation</td>
<td>17</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Generalized Additive Models</td>
<td>17</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Quasi-likelihood and the Generalized Linear Mixed Model Representation</td>
<td>18</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Eigenvalue-Eigenvector Decomposition</td>
<td>20</td>
</tr>
<tr>
<td>2.3</td>
<td>The Adaptive LASSO for Generalized Additive Models</td>
<td>21</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Methodology</td>
<td>21</td>
</tr>
<tr>
<td>2.3.2</td>
<td>The EM Algorithm</td>
<td>23</td>
</tr>
<tr>
<td>2.4</td>
<td>Simulation Studies</td>
<td>29</td>
</tr>
<tr>
<td>2.5</td>
<td>A Real Data Application</td>
<td>49</td>
</tr>
<tr>
<td>2.6</td>
<td>Summary</td>
<td>52</td>
</tr>
<tr>
<td><strong>Chapter 3</strong></td>
<td>Model Selection and Estimation in Generalized Additive Mixed Models</td>
<td>53</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>53</td>
</tr>
<tr>
<td>3.2</td>
<td>Generalized Additive Mixed Models and the Representation</td>
<td>54</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Generalized Additive Mixed Models</td>
<td>54</td>
</tr>
<tr>
<td>3.2.2</td>
<td>The Generalized Linear Mixed Model Representation</td>
<td>54</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Eigenvalue-Eigenvector Decomposition</td>
<td>55</td>
</tr>
<tr>
<td>3.3</td>
<td>The Adaptive LASSO and the EM Algorithm for Generalized Additive Mixed Models</td>
<td>56</td>
</tr>
<tr>
<td>3.4</td>
<td>Simulation Studies</td>
<td>62</td>
</tr>
<tr>
<td>3.5</td>
<td>A Real Data Application</td>
<td>86</td>
</tr>
<tr>
<td>3.6</td>
<td>Summary</td>
<td>89</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

| Table 2.1 | The frequency of appearance of the covariates in selected models under the MLE and REML type of methods in 100 runs. | 31 |
| Table 2.2 | The average model size in selected models over 100 runs under the MLE and REML type of methods. The true size of underlying model is 4. The standard error of the average model size is in the range (0.017, 0.121). | 31 |
| Table 2.3 | The average correct model size in selected models over 100 runs under the MLE and REML type of methods. The standard error of the average correct model size is in the range (0, 0.096). | 32 |
| Table 2.4 | The average incorrect model size in selected models over 100 runs under the MLE and REML type of methods. The standard error of the average incorrect model size is in the range (0.017, 0.057). | 32 |
| Table 2.5 | The average model error over 100 runs under the MLE and REML type of methods. The standard error of the average model error is in the range (0.004, 0.038). | 33 |
| Table 2.6 | The frequency of appearance of the covariates in selected models under the MLE and REML type of methods in 100 runs for the simulation setting V. | 34 |
| Table 2.7 | The average model size, the average correct model size, and the average incorrect model size over 100 runs under the MLE and REML type of methods. The standard error of the average model size, the average correct model size, and the average incorrect model size is in the range (0.030, 0.078). | 34 |
| Table 2.8 | The average model error over 100 runs under the MLE and REML type of methods. The standard error of the average model error is in the range (0.023, 0.025). | 35 |
| Table 3.1 | The frequency of appearance of the covariates in selected models under the MLE and REML type of methods. | 63 |
| Table 3.2 | The average model size in selected models over 100 runs under the MLE and REML type of methods. The true size of underlying model is 4. The standard error of the average model size is in the range (0, 0.078). | 64 |
| Table 3.3 | The average correct model size in selected models over 100 runs under the MLE and REML type of methods. The standard error of the average correct model size is in the range (0, 0.059). | 64 |
Table 3.4  The average incorrect model size in selected models over 100 runs under the MLE and REML type of methods. The standard error of the average incorrect model size is in the range (0, 0.053).  

Table 3.5  The average model error over 100 runs under the MLE and REML type of methods. The standard error of the average model error is in the range (0.007, 0.032).  

Table 3.6  The average $\hat{\beta}_0$ and $\hat{\sigma}_b^2$ over 100 runs under the MLE and REML type of methods. The standard error of the average $\hat{\beta}_0$ and $\hat{\sigma}_b^2$ is in the range (0.012,0.048 ).  

Table 3.7  The frequency of appearance of the covariates in selected models under the MLE and REML type of methods for simulation setting V (a rare event case).  

Table 3.8  The average model size, the average correct model size and the average incorrect model size over 100 runs under the MLE and REML type of methods for simulation setting V (a rare event case). The standard error of the average model size, the average correct model size, and the average incorrect model size is in the range (0.029, 0.108).  

Table 3.9  The average model error over 100 runs under the MLE and REML type of methods for simulation setting V (a rare event case). The standard error of the average model error is in the range (0.035, 0.049).  

Table 3.10  The frequency of appearance of the covariates in selected models under the MLE and REML type of methods.  

Table 3.11  The average model size in selected models over 100 runs under the MLE and REML type of methods. The true size of underlying model is 4. The standard error of the average model size is in the range (0.014, 0.079).  

Table 3.12  The average correct model size in selected models over 100 runs under the MLE and REML type of methods. The standard error of the average correct model size is (0, 0.063).  

Table 3.13  The average incorrect model size in selected models over 100 runs under the MLE and REML type of methods. The standard error of the average incorrect model size is in the range (0.014, 0.040).  

Table 3.14  The average model error over 100 runs under the MLE and REML type of methods. The standard error of the average model error is in the range (0.008, 0.034).
LIST OF FIGURES

Figure 2.1 The estimated nonparametric functions averaged over 100 simulation runs with the MLE and REML type of methods for simulation setting I: $n = 200, w = 1$. The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method. ........................................ 36

Figure 2.2 Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting I: $n = 200, w = 1$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation. .......................... 37

Figure 2.3 Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting I: $n = 200, w = 1$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation. ......... 38

Figure 2.4 The estimated nonparametric functions averaged over 100 simulation runs with the MLE and REML type of methods for simulation setting II: $n = 200, w = 5$. The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method. ........................................ 39

Figure 2.5 Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting II: $n = 200, w = 5$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation. ..................... 40

Figure 2.6 Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting II: $n = 200, w = 5$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation. ..................... 41

Figure 2.7 The estimated nonparametric functions averaged over 100 simulation runs with the MLE and REML type of methods for simulation setting III: $n = 300, w = 1$. The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method. ........................................ 42
Figure 2.8  Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting III: $n = 300, w = 1$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation. ................................. 43

Figure 2.9  Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting III: $n = 300, w = 1$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation. ................................. 44

Figure 2.10  The estimated nonparametric functions averaged over 100 simulation runs with the MLE and REML type of methods for simulation setting IV: $n = 300, w = 5$. The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method. ................................. 45

Figure 2.11  Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting IV: $n = 300, w = 5$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation. ................................. 46

Figure 2.12  Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting IV: $n = 300, w = 5$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation. ................................. 47

Figure 2.13  The estimated nonparametric functions using the MLE and REML type of methods for simulation setting V: $n = 300, w = 1, \beta_0 = -3.5$. The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of approach. The red dotted lines are estimates using the REML type of approach. ................................. 48

Figure 2.14  The estimated functions corresponding to selected covariates in the South African Heart Disease data using the REML type of method. The red solid lines are estimated functions, the blue dot-dashed lines are the 95% frequentist confidence intervals and the black dotted lines are the 95% Bayesian confidence intervals of the selected functions. ................................. 51
Figure 3.1 The estimated nonparametric functions averaged over 100 simulation runs with MLE and REML type of methods for simulation setting I: \( m = 75, w = 1 \) in model (3.16). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.

Figure 3.2 Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting I: \( m = 75, w = 1 \) in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.

Figure 3.3 Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting I: \( m = 75, w = 1 \) in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.

Figure 3.4 The estimated nonparametric functions averaged over 100 simulation runs with MLE and REML type of methods for simulation setting II: \( m = 75, w = 5 \) in model (3.16). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.

Figure 3.5 Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting II: \( m = 75, w = 5 \) in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.

Figure 3.6 Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting II: \( m = 75, w = 5 \) in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.

Figure 3.7 The estimated nonparametric functions averaged over 100 simulation runs with MLE and REML type of methods for simulation setting III: \( m = 100, w = 1 \) in model (3.16). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.
Figure 3.8 Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting III: $m = 100, w = 1$ in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.  

Figure 3.9 Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting III: $m = 100, w = 1$ in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.  

Figure 3.10 The estimated nonparametric functions averaged over 100 simulation runs with MLE and REML type of methods for simulation setting IV: $m = 100, w = 5$ in model (3.16). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.  

Figure 3.11 Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting IV: $m = 100, w = 5$ in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.  

Figure 3.12 Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting IV: $m = 100, w = 5$ in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.  

Figure 3.13 The estimated nonparametric functions averaged over 100 simulation runs with MLE and REML type of methods for simulation setting V: $m = 100, w = 1, b_i \sim N(0, 0.2), \beta_0 = -5.5$. The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.  

Figure 3.14 The estimated functions corresponding to selected covariates in the Periodontal Disease data using the REML type of method. The red solid lines are estimated functions, the blue dot-dashed lines are the 95% frequentist confidence intervals and the black dotted lines are the 95% Bayesian confidence intervals of the selected functions.
Figure 1  The estimated nonparametric functions averaged over 100 simulation runs with the MLE and the REML type of methods for simulation setting VI: $m = 75, w = 1$ in model (3.18). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method. . . . . . . . . 96

Figure 2  Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting VI: $m = 75, w = 1$ in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation. . . . . . . . . 97

Figure 3  Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting VI: $m = 75, w = 1$ in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation. . . . . . . . . 98

Figure 4  The estimated nonparametric functions averaged over 100 simulation runs with the MLE and the REML type of methods for simulation setting VII: $m = 75, w = 5$ in model (3.18). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method. . . . . . . . . 99

Figure 5  Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting VII: $m = 75, w = 5$ in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation. . . . . . . . . 100

Figure 6  Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting VII: $m = 75, w = 5$ in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation. . . . . . . . . 101

Figure 7  The estimated nonparametric functions averaged over 100 simulation runs with the MLE and the REML type of methods for simulation setting VIII: $m = 100, w = 1$ in model (3.18). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method. . . . . . . . . 102
Figure 8 Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting VIII: \(m = 100, w = 1\) in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.

Figure 9 Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting VIII: \(m = 100, w = 1\) in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.

Figure 10 The estimated nonparametric functions averaged over 100 simulation runs with the MLE and the REML type of methods for simulation setting IX: \(m = 100, w = 5\) in model (3.18). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.

Figure 11 Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting IX: \(m = 100, w = 5\) in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.

Figure 12 Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting IX: \(m = 100, w = 5\) in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Chapter 1

Introduction

Modeling is fundamental in statistics research as it explains the relationship between the responses of interest and a number of predictor variables. But not all of the covariates truly make significant contributions to the response variable. The more covariates the model includes, the less estimate bias the prediction has. Yet, it will generate a large variance in prediction given too many covariates in the model. Besides, a model cannot be easily fit and does not have a meaningful interpretation with a bunch of uninformative variables. Thus, one of the crucial tasks is to identify the informative covariates in the model and eliminate the ones with rare contribution. With a relatively small number of important predictors, a model can have a parsimonious interpretation and the prediction will be more reliable.

1.1 Literature Review on Variable Selection Methods

1.1.1 Linear Regression Models and Variable Selection Methods

Linear regression models characterize the linear relationship between the response variable and some independent variables. Linear regression models are widely applied in practice because they can be easily fit and have a nice interpretation. The general for-
mulation of a linear regression model can be expressed as

\[ y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i, i = 1, \ldots, n, \] (1.1)

where \( y_i \) is the \( i \)th response, \( n \) is the number of observations, and \( \epsilon_i \) is the error term, usually assumed to be from normal distribution \( N(0, \sigma^2) \). The linear relationship between the responses and the covariates is established by the coefficients \( \beta_1, \ldots, \beta_p \), and the contribution of one covariate is fully reflected by the corresponding coefficient. As a result, a variable is informative if the corresponding coefficient estimate is significant from zero.

Many variable selection methods have been proposed for linear regression models. The traditional methods are discrete processes, such as the best subset search, forward selection, backward elimination, and stepwise selection. The descriptions of these methods are presented below.

- **Best subset search:** We do an exhaustive search on all possible subsets of covariates and build models on each subset. The best model is selected according to a selection criterion. The commonly used criteria are Mallows’ \( C_p \) (Mallows, 1973), AIC (Akaike, 1973) and BIC (Schwarz, 1978). Best subset search investigates all possible model compositions, which guarantees to obtain the best model given a selection criterion. However, it could be a very time-consuming or even infeasible procedure if the number of covariates is large.

- **Forward selection:** The model starts with an empty set, with no variables selected. Then we sequentially add one variable which improves the model fit most. The selection procedure will stop if no variable would increase the model fit significantly. Miller (1990) argued that forward selection works very well if the covariates are independent. But independence is a very strong assumption in practice.

- **Backward elimination:** The method starts with the fully saturated model with all possible variables, and eliminates one variable which is most irrelative to the model at each step. The selection completes if all the variables selected in the model are statistically significant. Backward elimination usually performs better than forward selection if the covariates are dependent.
Stepwise selection: The procedure is a combination of forward selection and backward elimination and it allows either adding or dropping a variable in each step. Compared to forward selection and backward elimination, stepwise selection is more time-consuming and cannot deal with the collinearity case.

As discussed in the discrete processes, we either keep or drop a variable in the model at each step. They are intuitive and simple to implement, which makes them perform well in reality. But they are highly variable due to the discreteness (Breiman, 1996; Fan and Li, 2001). With a small perturbation in the data, selected variables could be dramatically different. Besides, it is hard to establish asymptotic theory and make inferences using these methods.

Accordingly, a family of shrinkage methods were proposed to address the issues. With a penalty term added to the residual sum of squares, we can estimate the linear coefficients by minimizing the penalized sum of squares. It is a continuous procedure and shrinks the coefficients with noninformative variables to zero while keeping the coefficients with a large magnitude for important variables. Those variables with negligible coefficients estimate are unimportant and therefore they should be eliminated from the model. Thus, the shrinkage methods are unified methods for model selection and estimation. They are more stable than the discrete processes and do not suffer from high variability. Besides, valid asymptotic inferences can be derived in these methods.

The spirit of the shrinkage methods is to minimize the objective function expressed as

$$L(\beta, y, x) + \lambda J(\beta),$$

where $L(\beta, y, x)$ is the loss function and $\lambda J(\beta)$ is a shrinkage penalty term. For linear regression models in (1.1), the loss function is written as

$$L(\beta, y, x) = \sum_{i=1}^{n} \{y_i - \sum_{j=1}^{p} \beta_j x_{ij}\}^2.$$  

Some popular methods with various penalty terms were proposed. The $L_q$ shrinkage family includes

- $L_0$: penalty term is $\lambda \sum_{j=1}^{p} I(\beta_j \neq 0)$, proposed by Donoho and Johnstone (1988),
- $L_1$ (LASSO): penalty term is $\lambda \sum_{j=1}^{p} |\beta_j|$, proposed by Tibshirani (1996),
• \(L_2\) (Ridge): penalty term is \(\lambda \sum_{j=1}^{p} \beta_j^2\), proposed by Hoerl and Kennard (1970a, 1970b),

• \(L_\infty\) (Supnorm penalty): penalty term is \(\lambda \max_j |\beta_j|\), proposed by Zhang et al. (2008).

Other shrinkage methods are

• nonnegative garrotee (Breiman, 1995). The estimates can be obtained by minimizing \(\sum_i (y_i - \sum_{i=1}^{p} c_k \hat{\beta}_k x_{ik})^2\), with \(c_k \geq 0, \sum_k c_k < s\), where \(\hat{\beta}_k\) is the ordinary least square estimator of \(\beta_k\).

• the SCAD (Fan and Li, 2001). The penalty function is written as follows

\[
\begin{align*}
\lambda|\beta| & \quad \text{if } |\beta| \leq \lambda, \\
-|\beta|^2 - 2a\lambda\beta + \lambda^2 & \quad \text{if } \lambda < |\beta| \leq a\lambda, \\
\frac{(a+1)\lambda^2}{2} & \quad \text{if } |\beta| > a\lambda,
\end{align*}
\]

where \(a > 2\) and \(\lambda > 0\) are tuning parameters.

• elastic net (Zou and Hastie, 2005). The penalty term is expressed as \(\lambda|\beta_j| + (1 - \lambda)\beta_j^2\), \(\lambda \in [0, 1]\).

Among the shrinkage methods, LASSO enjoys its popularity. Taking a linear regression model of (1.1) as an example, the vector of coefficient estimate \(\hat{\beta}^* = (\hat{\beta}_1^*, \hat{\beta}_2^*, \ldots, \hat{\beta}_p^*)^T\) can be solved with LASSO by minimizing the following objective function

\[
\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|,
\]

where \(\lambda\) is the shrinkage parameter, controlling the amount of shrinkage. A large \(\lambda\) will generate a high shrinkage. LASSO shrinks both the informative coefficients and the noninformative coefficients with the same strength, which is not efficient in some cases.

A generalization method of LASSO is the group LASSO (Yuan and Lin 2006, Bach, 2008), which can be used to select groups of variables. The group LASSO can handle the categorical variable cases where dummy variables are generated to form a group.
Therefore, it is reasonable to select a group of variables instead of individual variables. Given a vector \( \eta \in \mathbb{R}^d \), \((d \geq 1)\) and a positive definite matrix \( K_{d \times d} \), we let \( ||\eta||_K = (\eta^T K \eta)^{1/2} \). Then, the estimate of the group LASSO can be achieved by minimizing the following function

\[
\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} ||\beta_j||_K,
\]

where the positive \( \lambda \) is a tuning parameter.

Another generalization method is the adaptive LASSO, proposed by Zou (2006) and it was extended to Cox’s proportional hazards models by Zhang and Lu (2007). We apply the adaptive LASSO as the variable selection method in our study. Thus, we briefly introduce the adaptive LASSO below. The adaptive LASSO improves LASSO by applying different weights for different coefficients, imposing big effects on noninformative variables and small impacts on informative variables. Essentially, the objective function to be minimized is

\[
\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j|,
\]

where \( \hat{w}_j \) is the weight associated with \( \beta_j \). The weight vector \( \hat{w} = (\hat{w}_1, \ldots, \hat{w}_p)^T = \frac{1}{|\hat{\beta}|^\gamma}, \) where \( \gamma > 0 \), and \( \hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T \) is the ordinary least square estimator. Obviously, bigger weights are associated with smaller parameter estimates, which ensures more impacts on the less informative variables. Zou (2006) stated that the adaptive LASSO enjoys the oracle properties if the weights are data-driven and smartly selected. In addition, he argued that the formulation (1.6) is a convex function, which implies that the multiple local minima issue does not exist and we can achieve the global minimizer. Due to these advantages, the adaptive LASSO becomes a popular method in variable selection.

If we use maximum likelihood estimator \( \hat{\beta} \), the adaptive LASSO will obtain the coefficient estimates by maximizing the following penalized function

\[
\sum_{i=1}^{n} \ell_i(\beta_1, \ldots, \beta_p; y_i) - \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j|,
\]

where \( \ell_i(\beta_1, \ldots, \beta_p; y_i) \) is the log-likelihood contributed by \( y_i \).

There are still many other model selection methods. The Bayesian approaches select the most appropriate model with the highest posterior probability. Hence, choosing a suitable prior distribution is a critical step in Bayesian methods. More details can be
referred to Mitchell and Beauchamp (1988), Pereira and Stern (2001), Yuan and Lin (2005). One relatively new model selection method, called False Selection Rate (FSR) approach, was proposed by Wu, Boos and Stefanski (2007). They put a number of pseudo-variables to the real data set and perform model selection by controlling the FSR which is the proportion of noninformative variables selected in the models.

1.1.2 Nonparametric Models and Variable Selection Methods

Linear relationship between the response variable and the covariates is a strong assumption which is not necessarily valid for each model in practice. There are models where the linearity assumption does not hold. If we still fit such models by linear method, the prediction is misleading and model misspecification occurs. Under this circumstance, a nonparametric model would be more appropriate. Basically, a general nonparametric model takes the form

$$y_i = f(x_{i1}, \ldots, x_{ip}) + \epsilon_i, i = 1, \ldots, n,$$

where $y_i$ is the $i$th observation, $n$ is the number of observations, and $f(.)$ is a non-specified function. Usually, we assume $f(.)$ to be smooth and continuous. If $f(x_{i1}, \ldots, x_{ip}) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$, a nonparametric model becomes a linear regression model. In nonparametric modeling, how to estimate the function $f(.)$ plays a critical role. Kernel estimations, regression splines and smoothing splines are well-known estimation methods.

Kernel estimation methods use linear estimators to predict the value at a particular point $x$. One famous linear estimator is Nadaraya-Watson estimator which can be expressed as

$$\hat{f}_h(x) = \frac{\sum_{i=1}^{n} K_h(x_i - x) y_i}{\sum_{i=1}^{n} K_h(x_i - x)},$$

where $K$ is a kernel function and $h$ is a bandwidth which is a smoothing parameter controlling the size of the local neighborhood. Some commonly used kernel functions include the Gaussian kernel and the symmetric Beta family kernel.

Regression splines estimation is another way to estimate a nonparametric function with a set of basis functions as follows

$$\hat{f}(x) = b_0 + b_1 x^1 + \ldots + b_r x^r + \sum_{j=1}^{m} \beta_j (x - k_j)_+^r,$$
where $r$ is the order of the regression spline, $k_j$ is the $j$th knot, \( \{b_0, \ldots, b_r, \beta_1, \ldots, \beta_m\}^T \) is a set of coefficients, and \((a)_+\) is $a$ if $a$ is greater than 0, and 0 otherwise. The representation of the regression splines in (1.10) is a linear combination of two parts \( \{x^j\}_{j=0}^r \) and \( \{(x - k_j)^{+}\}_{j=1}^m \). Some other popular basis functions are B-spline basis, natural splines, and radial basis functions (Eilers and Marx, 1996; Green and Silverman, 1994). We use the smoothing splines to estimate the nonparametric functions, and hence we will explain it extensively in the following section.

One unavoidable drawback in nonparametric models is the “curse of dimensionality”. The variance of the estimates greatly inflates as the number of independent variables increases. To deal with the issue, Hastie and Tibshirani (1986) proposed additive regression models. Denote by $y_i$ the $i$th observation out of $n$ observation units ($i = 1, \ldots, n$). We can write an additive regression model as

$$
y_i = \alpha + f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \epsilon_i, \ i = 1, \ldots, n,
$$

where $\alpha$ is a constant, $f_j(.)$’s ($j = 1, \ldots, p$) are one-dimensional smooth and continuous functions, and $\epsilon_i$ is assumed to be distributed from a normal distribution.

There are two benefits of the additive models. Firstly, due to the fact that each additive term can be estimated individually, the curse of dimensionality is no longer a problem. Secondly, the additive relationship can clearly explain the contribution of each covariate to the response variable. Compared to the general nonparametric regression models in (1.8), the additive models have restricted applications because the nonparametric function $f(.)$ in nonparametric regression is decomposed into several one-dimensional nonparametric functions in additive models. Whereas, the additive models are popular in the real applications due to their straightforward interpretation and model conciseness.

Another well-known model in the nonparametric framework is the smoothing spline analysis of variance (SS-ANOVA) which was proposed by Wahba (1990). The regression function in SS-ANOVA can be written as

$$
f(x) = b + \sum_{j=1}^{p} f_j(x_j) + \sum_{1 \leq j < k \leq p} f_{jk}(x_j, x_k) + \ldots + f_{1\ldots p}(x_1, \ldots, x_p),
$$

where $b$ is a constant, $f_j$’s are main effects, $f_{jk}$’s are two-way interactions, and so on. SS-ANOVA is a generalization of additive models because it allows the interaction terms and the regression function $f(.)$ is decomposed into several orthogonal functional components.
If the interactions in (1.12) are eliminated, SS-ANOVA is reduced to an additive model.

Variable selection becomes a challenge in nonparametric models. In context of linear models, the importance of variables is completely reflected by their corresponding coefficients. A variable is considered as a noninformative one if its coefficient estimate is negligible. But it is not the case in the nonparametric framework. Since the nonparametric functions are not specified, we can only drop a variable from the model if the function associated with that variable is estimated to be a constant as a whole component.

One of the regression-spline-based model selection methods is the adaptive group LASSO (Huang et al. 2010) and it is applied in additive models. It is an iterative process, combining the group LASSO with the adaptive LASSO. Taking advantage of the representation of regression spline estimator for the nonparametric components, the method performs variable selection by estimating and selecting the groups of coefficients. The selection procedure has two steps. In the first step, the group LASSO is utilized to get the initial estimates and decrease the dimension of the model. In the second step, we apply the adaptive group LASSO to pick up the final nonparametric functions.

Another regression-spline-based method was proposed by Friedman (1991). He presented a multivariate adaptive regression splines (MARS) algorithm to do model selection for high-dimensional cases. With the help of a basis function set, MARS performs variable selection in two steps. The first step is a forward process, adding two basis functions each time until a large number of variables are selected. One basis functions is

$$B_m(x) = \prod_{k=1}^{K_m} [s_{km}(x_{v(k,m)} - t_{km})]^q_+ ,$$

(1.13)

where $K_m$ denotes the number of factors in the $m$th basis function, $s_{km}$ takes two values $\pm 1$ indicating the truncation side, and $v(k, m)$ labels the predictor variables. The model is usually overfit after the first step. The second step is a backward process, punning the model to fit the data better. It drops variables one by one, eliminating the least informative one at every step until it gets the best model.

Due to the inherent discreetness, MARS is not stable. Consequently, some methods with continuous selection procedure need to be provided. Some popular smoothing-spline-based methods were presented. Lin and Zhang (2006) proposed the component selection and smoothing operator (COSSO) which is one of the most popular model selection and estimation methods in nonparametric regression models. Specifically, it
is a regularization with the penalty as the sum of component norms. We assume a nonparametric regression model in the formulation of (1.8). Based on the SS-ANOVA framework, \( f \) can be estimated by minimizing the regularization

\[
\frac{1}{n} \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \tau_n^2 J(f), \quad \text{with} \quad J(f) = \sum_{\alpha=1}^{p} ||P_\alpha f||, \tag{1.14}
\]

where \( \tau_n \) is the smoothing parameter and \( f \) belongs to a reproducing kernel Hilbert space (RKHS), associated with the SS-ANOVA in (1.12). Let \( F \) denote the RKHS, and \( P_\alpha f \) is the orthogonal projection of \( f \) onto \( F_\alpha \) which is one of \( p \) orthogonal subspaces of \( F \). By imposing a penalty of sum of norms of the functional components, COSSO performs model selection in a continuous shrinkage process. However, COSSO tends to oversmooth the functions, which sometimes keeps it from having the oracle property in the nonparametric framework. Motivated by the adaptive LASSO, Storlie et al. (2009) proposed the adaptive COSSO (ACOSSO) which puts different impacts on the functional components. The corresponding regularization is

\[
\frac{1}{n} \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \tau_n^2 J(f), \quad \text{with} \quad J(f) = \sum_{\alpha=1}^{p} w_\alpha ||P_\alpha f||, \tag{1.15}
\]

where \( w_\alpha \)'s are the weights. In this way, more important functional components are exposed to smaller penalty compared to less important ones. Thus, the informative variables can be quickly identified and we don’t lose much information on the estimation of important functions.

Another smoothing-spline-based model selection method is nonnegative garrote component selection (NGCS) (Yuan 2007). NGCS is a generalization of nonnegative garrote (Breiman 1995) and is applied in nonparametric setting of (1.12). It is a shrinkage method and the shrinkage factor \( d = (d_1(\lambda), \ldots, d_p(\lambda))^T \) is obtained by minimizing the following regularization

\[
\frac{1}{2} \sum_{i=1}^{n} (y_i - z_id_j)^2 + n\lambda \sum_{j=1}^{p} d_j
\]

subject to \( d_j \geq 0, j = 1, \ldots, p, \)

where \( z_i \) is the vector of initial estimate at the functional components, i.e., \( z_i = (\hat{f}_{1\text{init}}(x_{i1}), \ldots, \hat{f}_{p\text{init}}(x_{ip})) \),
..., $\hat{f}_{i,\ldots,p}^{init}(x_i)^T$, and $\lambda > 0$ is a tuning parameter. The estimates of the functional components are then given by $d_1(\lambda)\hat{f}_{1,\ldots,p}^{init}(x_{11}), \ldots, d_{1,\ldots,p}(\lambda)\hat{f}_{1,\ldots,p}^{init}(x_i)$. One initial estimate of functional components is smoothing spline estimate which minimizes

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \tau_1 J_1(f_1) + \ldots + \tau_p J_p(f_p) + \ldots + \tau_{1,\ldots,p} J_{1,\ldots,p}(f_{1,\ldots,p}),$$

where $\tau$’s are the tuning parameters and $J$’s are the squared norms in the functional component spaces. It can be shown that the shrinkage parameter is large if the functional component estimate is small, while the shrinkage parameter is small if the functional component estimate is large. In this way, noninformative functional components will shrink to zero very fast but informative components will keep in a large magnitude.

Ravikumar et al. (2009) proposed a new class of sparse additive models (SpAMs) which have the mixed features of the additive nonparametric regression and the sparse linear modeling. They presented a method of fitting SpAMs in high dimensional data. Based on the additive regression model’s framework in (1.11), there are constraints imposed on the functional components $f_j$’s in order to keep the smoothness of the functional components and the sparsity across components. Let $\mathcal{H}_j$ indicate the Hilbert space of the function $f_j(.)$ such that $E\{f_j(.)\} = 0$, and $E\{f_j^2(.)\} < \infty$. Let the inner product $\langle f_j, f_j^* \rangle = E\{f_j(t_j), f_j^*(t_j)\}$, and $f_j(.) = \beta_j g_j(.)$. The SpAMs need to minimize the following objective function

$$\min_{\beta \in \mathbb{R}^p, g_j \in \mathcal{H}_j} E\{y - \sum_{j=1}^{p} \beta_j g_j(t_j)\}^2$$

subject to $\sum_{j=1}^{p} |\beta_j| < L,$

$$E(g_j) = 0, j = 1, \ldots, p,$$

$$E(g_j^2) = 1, j = 1, \ldots, p.$$ (1.17)

The constraint on $\beta_j$’s makes the sparsity of $\beta$’s and therefore some of the functional components $f_j$’s will go to zero.

If we include two-way interactions in additive regression models, the formulation of
the model can be expressed as

$$y_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}) + \sum_{j>k} f_{jk}(x_{ij}, x_{ik}) + \epsilon_i, i = 1, \ldots, n,$$

(1.18)

where $\alpha$ is a constant, $f_j$’s are main effects, and $f_{jk}$’s are two-way interactions. The previous method in SpAMs cannot be applied here since there is no interaction terms in SpAMs. In addition, COSSO cannot perform well in this case because it is not good at handling the high dimensional situation. Radchenko and James (2010) proposed a variable selection using adaptive nonlinear interaction structures in high dimensions (VANISH) in this case. A simple approach to estimate the nonparametric functions is to minimize the penalized least squares as

$$\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda (\sum_{j=1}^{p} ||f_j||^2 + \sum_{j=1}^{p} \sum_{k=j+1}^{p} ||f_{jk}||^2).$$

(1.19)

The method shrinks a lot of main effects and two-way interactions to zero. But it imposes the same shrinkage on all terms including main effects and interactions, which is not efficient. Thus, Radchenko and James (2010) proposed a convex penalty function that imposes a constraint and adjusts the shrinkage on the interactions based on the different conditions whether or not the main effects are included in the model. The corresponding penalized function is

$$\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda_1 (\sum_{j=1}^{p} ||f_j||^2 + \sum_{k:k\neq p}^{p} ||f_{jk}||^2)^{\frac{1}{2}} + \lambda_2 \sum_{j=1}^{p} \sum_{k=j+1}^{p} ||f_{jk}||.$$

(1.20)

It also can be shown that $\lambda_1$ is treated as the weight associated with the penalty of every additional predictor added into the model, and $\lambda_2$ is an extra penalty of the interaction terms.

Yau, Kohn, and Wood (2003) brought up a Bayesian method to select variables in a nonparametric setting. Zhang and Wahba (2004) proposed a nonparametric penalized likelihood approach, called Likelihood Basis Pursuit (LBP), in model selection for additive models. Meier, van de Geer and Bühlmann (2009) developed model estimation for high-dimensional generalized additive models by penalizing both the sparsity and the roughness. Other methods for variable selection in additive models include Huang and
1.2 A Smoothing Spline Estimation of a Nonparametric Function

We use smoothing splines to estimate the nonparametric functions. First of all, we briefly introduce the reproducing kernel Hilbert space (RKHS) and the smoothing spline estimator. A Hilbert space, $\mathcal{H}$, is an inner product space. It is also complete and separable with respect to the norm/distance function defined by the inner product. For any $f, g \in \mathcal{H}$ and $a \in \mathbb{R}$, $<.,.>$ is defined as an inner product if and only if it satisfies the following three properties:

- $< f, g > = < g, f >$,
- $< f + g, h > = < f, h > + < g, h >$, and $< af, g > = a < f, g >$,
- $< f, f > \geq 0$ and equal if and only if $f = 0$.

A linear functional $L$ is a mapping of an element in the linear space to a real number. One fundamental theorem in Hilbert space is the Riesz theorem. It claims that for every continuous linear functional $L$ in a Hilbert space $\mathcal{H}$, there exists a unique $g_L \in \mathcal{H}$ such that $L(f) = < g_L, f >$ for $\forall f \in \mathcal{H}$. If a Hilbert space $\mathcal{H}$ is a real valued function space, we can define an evaluation function $L_t(f)$ which maps $f$ to $f(t)$. Furthermore, if the evaluation function is bounded, i.e., $|L(f)| \leq M||f||_\mathcal{H}$ for a constant $M$, $\mathcal{H}$ becomes a reproducing kernel Hilbert space (RKHS). Let’s take the following simple regression model for example

$$y_i = f(t_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $n$ is the number of observations. Without loss of generality, we assume that the knots are distinct and ordered, $0 \leq t_1 < \ldots < t_n \leq 1$, and $\epsilon_i$’s are independent and identically distributed from normal distribution $N(0, \sigma^2)$. In addition, we assume that $f$ is a smooth function in the space

$$W_h = \{g(t)|g(t), g^{(1)}(t), \ldots, g^{(h-1)}(t) \text{ absolutely continuous, } \int_0^1 \{g^{(h)}(t)\}^2 dt < \infty\},$$
where \( g^{(j)}(t) \) denotes the \( j \)th derivative of \( g(t) \). After an inner product is defined by

\[
<f_1, f_2>_W_h = \sum_{j=1}^{h} f_1^{(j-1)}(0) f_2^{(j-1)}(0) + \int_0^1 f_1^{(h)}(u) f_2^{(h)}(u) du,
\]

the space \( W_h \) is proved to be a reproducing kernel Hilbert space. Moreover, \( W_h \) can be decomposed into two subspaces \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \) where \( \mathcal{H}_0 \) denotes the space with constant functions and \( \mathcal{H}_1 \) is the complement of \( \mathcal{H}_0 \). We can get the estimate of the smooth function \( f(.) \) with

\[
\arg\min_{f \in W_h} \frac{1}{n} \sum_{i=1}^{n} \{y_i - f(t_i)\}^2 + \lambda ||P_1 f||^2_{W_h},
\]

(1.21)

where the positive \( \lambda \) is the smoothing parameter and \( P_1(f) \) denotes the projection of \( f \) into space \( \mathcal{H}_1 \). It can be shown that the expression in (1.21) is equivalent to

\[
\arg\min_{f \in W_h} \frac{1}{n} \sum_{i=1}^{n} \{y_i - f(t_i)\}^2 + \lambda \int_0^1 \{f^{(h)}(t)\}^2 dt.
\]

(1.22)

Alternatively, if the likelihood is specified, we can get the smoothing spline estimate by maximizing the penalized log-likelihood

\[
\ell(\beta, f(.); y) - \frac{\lambda}{2} \int_0^1 \{f^{(h)}(t)\}^2 dt,
\]

(1.23)

where \( \ell(.) \) is the log-likelihood. The smoothing spline estimators have several representations. We adopt the representation presented by Kimeldorf and Wahba (1971). The representation of the \( h \)th-order smoothing spline estimator is written as

\[
f(t) = \sum_{j=1}^{h} \delta_j \phi_j(t) + \sum_{i=1}^{n} a_i R_h(t, t_i),
\]

(1.24)

where \( \phi_j(t) = \frac{t^{j-1}}{(j-1)!}, j = 1, \ldots, h \), consisting of the basis of polynomials of order \((h-1)\), and \( \delta_j \)'s are fixed coefficients. In addition, \( R_h(t, t_i) \) can be calculated as

\[
R_h(t, t_i) = \frac{1}{[(h-1)!]^2} \int_0^1 (t - u)^{h-1}(t_i - u)^{h-1} du,
\]

(1.25)
where \((t - u)_+ = t - u\) if \(t \geq u\) and 0 otherwise. With the above expression, we can easily get \(R_h(t_k, t_l)\) for quadratic smoothing spline \((h = 1)\) and cubic smoothing spline \((h = 2)\). When \(h = 1\), \(R_1(t_k, t_l) = \min(t_k, t_l)\), and when \(h = 2\), \(R_2(t_k, t_l) = (3t_k^2t_l - t_k^3)/6\) given \(t_k < t_l\).

1.3 The Linear Mixed Model Representation

Let \(\delta = (\delta_1, \ldots, \delta_h)^T\), \(a = (a_1, \ldots, a_n)^T\), and denote by \(f\) the vector of \(f(t)\) evaluated at point \(t_i\). The representation in (1.24) has the following matrix form

\[
f = T\delta + \Sigma a, \tag{1.26}
\]

where \(T\) is an \(n \times h\) matrix with the \((k, l)\)th element as \(\phi_l(t_k)\), and \(\Sigma\) is a positive definite matrix with the \((k, l)\)th element as \(R_h(t_k, t_l)\). Furthermore, it can be shown that the penalty term \(\lambda \int_0^1 \{f^{(h)}(t)\}^2 dt = \lambda a^T\Sigma a\), which implies that \(a\) can be considered as a random effect from normal distribution \(N(0, \tau\Sigma^{-1})\), with \(\tau = \lambda^{-1}\). Thus, \(\tau\) is treated as an induced variance component from the nonparametric function \(f\). We can see that the smoothing spline estimator in (1.26) has a linear mixed model representation. Therefore, many methods valid in a linear setting could also be adapted. Two points from the linear mixed model representation should be noticed. One is that the nonparametric function \(f\) is an \((h - 1)\)th polynomial if and only if the induced variance component \(\tau\) is zero. The other is that \(f\) is constant if and only if \(\tau\) is zero given \(h = 1\).

1.4 The Framework of the Dissertation

In this dissertation, we focus on the method of model selection and estimation for generalized additive models and generalized additive mixed models. The linear mixed model representation of smoothing spline estimators of nonparametric functions is adopted and the inverse of the smoothing parameters are treated as extra variance components. By estimating the variance components with the adaptive LASSO, we can identify the informative variables. Approximate EM algorithms are presented to get the estimate of the variance components. To estimate the selected nonparametric functions, we use the approximate best linear unbiased predictors (BLUPs).
The framework of the dissertation is organized as follows. In Chapter 2, we propose the method of model selection and estimation in generalized additive models for independent data. A brief introduction of the model is presented at first. Then, the linear mixed model representation of the smoothing spline estimators of the nonparametric functions is derived. Moreover, the eigenvalue-eigenvector decomposition approach is applied to simplify the calculation. The adaptive LASSO is proposed to accomplish the research goal. Approximate EM algorithms are established to get the parameter estimates. To evaluate the proposed method, simulation studies and an application on a real data set are conducted. At the end of the chapter, a short summary is given. The method for model selection and estimation in generalized additive mixed models of longitudinal data is proposed in Chapter 3. First of all, a short introduction of the model structure is presented. Then, we build the model formulation based on the linear mixed model representation of the smoothing spline estimators of the nonparametric functions as well as the eigenvalue-eigenvector decomposition method. Similar to generalized additive models, we propose the adaptive LASSO and derive the approximate EM algorithms to get the estimate of the variance components. The estimation of the selected nonparametric functions can be obtained afterwards. Empirical study results and a real data application are given. At last, a summary is presented to conclude Chapter 3. Chapter 4 provides the conclusion and the future work.
Chapter 2

Model Selection and Estimation in Generalized Additive Models

2.1 Introduction

Linear regression is widely used because it has a straightforward interpretation and can be easily fit. In linear regression models, we usually assume that the responses follow a normal distribution. However, there are lots of situations where the responses are instead from a distribution in the exponential family. Thus, linear regression models are extended to generalized linear models where a link function is needed to relate the response variable to the linear predictors.

Still, the independent variables in generalized linear models are linearly related. But the linear assumption is not always true for each case in reality. To deal with this issue, Hastie and Tibshirani (1990) proposed generalized additive models (GAMs) which are based on the nonlinearity assumption. With a link function, we are able to model the relationship between the mean of the dependent variable and the predictors in an additive way. Compared with generalized linear models, GAMs also allow the response variable to follow a distribution in the exponential family. But the linear relationship in generalized linear models is replaced by the additive relationship in GAMs. In GAMs, the contribution of each covariate is reflected by a nonparametric smooth function and the nonparametric functions contribute to the model in an additive pattern. Therefore, generalized additive models have the mixed features of generalized linear models and additive models. Yet, they still keep the property of easy interpretation.
The objective of this chapter is to propose a method to perform model selection and estimation for generalized additive models. Below is an outline of this chapter. In Section 2.2, a short description of GAMs is given and the generalized linear mixed model representation of GAMs is also presented. Furthermore, the eigenvalue-eigenvector decomposition approach is introduced to simplify the matrix calculation. The adaptive LASSO is applied and the EM algorithms are derived for GAMs in Section 2.3. In Section 2.4 and 2.5, simulation studies are conducted, followed by an application on a real data set. We conclude this chapter with a summary in Section 2.6.

2.2 Generalized Additive Models and the Representation

2.2.1 Generalized Additive Models

Denote by $y_i$ the $i$th response variable out of $n$ observations ($i = 1, \ldots, n$). Assume that $y_i$’s are independently distributed from a distribution in the exponential family. Besides, there are $p$ possible covariates contributing to the response variable. The generic form of generalized additive models can be expressed as

$$g(\mu_i) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}), i = 1, \ldots, n,$$

where $g(.)$ is a link function, $\mu_i = E(y_i|x_{i1}, \ldots, x_{ip})$, $\beta_0$ is a constant, and $n$ is the number of observations. It is assumed that $f_1(.)$, $\ldots$, $f_p(.)$ are unspecified smooth functions centered to have mean zero. We use the link function to construct an additive relationship between the expectation of a dependent variable and several nonparametric one-dimensional functions. Some well-known link functions include identity link for normal data, logit link for binomial/binary data and log link for Poisson data. If the responses follow a normal distribution, we use the identity link function. In this case, a generalized additive model becomes an additive model. Each nonparametric function is a one-dimensional function based on a single covariate and all the nonparametric functions contribute to the model in an additive way.
2.2.2 Quasi-likelihood and the Generalized Linear Mixed Model Representation

Let’s see how to generate quasi-likelihood for a distribution in the exponential family. Suppose \( y_i \)'s are independent and identically distributed from a distribution in the exponential family. The probability density function of \( y_i \) can be written as

\[
f(y_i, \theta_i) = \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\},
\]

where \( \theta_i \) is the natural parameter, \( \phi \) is the known or unknown scale parameter and \( a_i(\phi) = \frac{\phi}{m_i} \) with the positive \( m_i \) being a known weight. In this dissertation, we consider the cases with a known \( \phi \). Thus, we can generate the quasi-likelihood as follows.

It can be easily obtained that the mean and variance components are

\[
\mu_i = E(y_i|x_i) = b'(\theta_i),
\]
\[
\text{var}(y_i|x_i) = a_i(\phi)b''(\theta_i).
\]

Thus, mean and variance are related through \( \theta \). If mean is indicated by \( \mu_i \), variance can be specified as \( a_i(\phi)v(\mu_i) \), where \( v(\mu_i) \) is called the variance function. The quasi-likelihood can be expressed as

\[
L = \exp\left\{ \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{m_i(y_i - u)}{\phi v(u)} du \right\},
\]

(2.3)

If deviance \( d_i = -2 \int_{y_i}^{\mu_i} \frac{m_i(y_i - u)}{v(u)} du \), we can have

\[
L = \exp\left\{ -\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i, \mu_i) \right\}.
\]

(2.4)

Therefore, we can construct the quasi-likelihood in generalized additive models.

There are a number of ways to estimate nonparametric functions. Here, we use the smoothing splines estimation. As for the model setting described in (2.1), there are \( p \) one-dimensional nonparametric functions \( f_1, \ldots, f_p \). Denote by \( \ell\{\beta_0, f_1(\cdot), \ldots, f_p(\cdot); y\} \) the log-quasi-likelihood function. Similar to the expression in (1.23), the penalized log-
quasi-likelihood is written as

\[ \ell\{\beta_0, f_1(\cdot), \ldots, f_p(\cdot); y\} = \frac{\lambda_1}{2} \int_{a_1}^{b_1} \{f_1^{(h)}(x_1)\}^2 dx_1 - \ldots - \frac{\lambda_p}{2} \int_{a_p}^{b_p} \{f_p^{(h)}(x_p)\}^2 dx_p, \tag{2.5} \]

where \(\lambda_j\)'s \((j = 1, \ldots, p)\) are the smoothing parameters associated with the nonparametric functions \(f_j\)'s, controlling the smoothness of the functions and the goodness of fit of the model, and \((a_j, b_j)\) is the range for covariate \(X_j\). As \(\lambda_j\) increases, the smoothing splines tend to oversmooth. Thus, it is important to choose an appropriate value for the smoothing parameter. Given \(\lambda_j\)'s, if we maximize the penalized log-quasi-likelihood in (2.5) with respect to the nonparametric functions, we will get smoothing spline estimators of order \(h\) (Kimeldorf and Wahba, 1971; Wahba, 1990; Zhang and Lin, 2003). We adapt the representation of smoothing spline estimation proposed by Kimeldorf and Wahba (1971).

Take \(f_1(\cdot)\) for example. Let \(x_1 = (x_{11}, x_{21}, \ldots, x_{n1})^T\) denote the vector of \(n\) knots. Denote by \(x_1^0\) the vector of \(r_1\) ordered and different knots in \(x_1\). Without loss of generality, assume \(0 < x_{011}^0 < x_{021}^0 < \ldots < x_{r_11}^0 < 1\). According to the formulation in (1.26), the \(h\)th-order smoothing splines estimation of \(f_1(x_1)\) is

\[ f_1(x_1) = N_1(T_1 \delta_1 + \Sigma_1 a_1), \tag{2.6} \]

where \(N_1\) is an incidence matrix that maps the vector \(x_1^0\) to the original vector \(x_1\), \(T_1\) is a \(r_1 \times h\) matrix with the \((k, l)\)th element as \(\phi_1(x_{k1}^0)\), \(\Sigma_1\) is a positive definite matrix with the \((k, l)\)th element as \(R_h(x_{k1}^0, x_{l1}^0)\), \(\delta_1 = (\delta_{11}, \ldots, \delta_{h1})^T\), and \(a_1 = (a_{11}, \ldots, a_{r_11})^T\), which is assumed to be a random effect from normal distribution \(N(0, \tau_1 \Sigma_1^{-1})\). Notice that \(\tau_1\) is the inverse of smoothing parameter \(\lambda_1\) and it can be considered as an extra variance component induced from \(f_1\).

Denote by \(y\) the vector of observations, i.e., \(y = (y_1, y_2, \ldots, y_n)^T\). Applying the smoothing splines estimator to each nonparametric function, GAMs can be written as

\[ g(\mu) = 1\beta_0 + N_1T_1\delta_1 + \ldots + N_pT_p\delta_p + N_1\Sigma_1 a_1 + \ldots + N_p\Sigma_p a_p, \tag{2.7} \]

where \(1\) is a vector of ones, \(\beta_0, \delta_j\)'s \((j = 1, \ldots, p)\) are fixed effects, and \(a_j\)'s are random effects. One special case is the quadratic smoothing spline \((h=1)\) estimator, where \(T_j\)'s are reduced to vectors and \(\delta_j\)'s are reduced to scalars. Since \(\phi_{1j}(x) = \frac{x_1^{1-1}}{(1-1)!} = 1\), \(T_j\)'s


\( (j = 1, \ldots, p) \) are \( r_j \times 1 \) vectors of ones and thus \( N_j T_j \) is an \( n \times 1 \) vector of ones. In this case, the corresponding expression of GAMs can be expressed as

\[
g(\mu) = 1\beta_0 + 1\delta_1 + \ldots + 1\delta_p + N_1\Sigma_1a_1 + \ldots + N_p\Sigma_p a_p
= 1\beta + N_1\Sigma_1a_1 + \ldots + N_p\Sigma_p a_p,
\]

(2.8)

where \( \beta_0, \delta_j \)'s \( (j = 1, \ldots, p) \) are combined into a single term \( \beta \), and \( a_j \)'s are random effects with normal distribution \( N(0, \tau_j\Sigma_j^{-1}) \). As shown in (2.8), generalized additive models can be written in a generalized linear mixed model representation. If the two components are specified as

\[
\mu_a^i = E(y_i|a),
\]

\[
\text{var}(y_i|a) = a_i(\phi)v(\mu_a^i) = \frac{\phi}{m_i}v(\mu_a^i).
\]

The quasi-likelihood can be calculated as

\[
L = \exp\left\{ \sum_{i=1}^{n} \int_{y_i}^{\mu_a^i} m_i(y_i - u) \frac{du}{\phi v(u)} \right\} = \exp\left\{ -\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i, \mu_a^i) \right\}.
\]

(2.9)

\subsection*{2.2.3 Eigenvalue-Eigenvector Decomposition}

As the number of distinct knots \( r_j \)'s \( (j = 1, \ldots, p) \) and the number of covariates \( p \) get large, the matrices involved in the model representation become highly dimensional, which makes the calculations time-consuming. To tackle this problem, we propose the eigenvalue-eigenvector decomposition approach to speed up the calculation. Let’s look at \( f_1 \) first. The positive definite matrix \( \Sigma_1 \) can be decomposed by the eigenvalue-eigenvector approach. Thus, \( \Sigma_1 = \sum_{i=1}^{r_1} \lambda_i e_i e_i^T \), where \( \lambda_i \) is the \( i \)th eigenvalue of \( \Sigma_1 \), \( e_i \) is the corresponding eigenvector, and \( r_1 \) is the dimension of \( \Sigma_1 \). For positive eigenvalues, we have \( \lambda_1 > \lambda_2 > \ldots > \lambda_{r_1} > 0 \). If we pick up the first \( q_1 \) leading eigenvalues and the corresponding eigenvectors, \( \Sigma_1 \) can be approximated as

\[
\Sigma_1 \approx E_1\Lambda_1 E_1^T,
\]

(2.10)

where \( \Lambda_1 = \text{diag}\{\lambda_1, \ldots, \lambda_{q_1}\} \), and \( E \) is constructed by the corresponding eigenvectors, i.e., \( E_1 = [e_1, \ldots, e_{q_1}] \). The number of \( q_1 \) is determined by the proportion of information we want to keep. Let \( p_0 \) denote the proportion. Then, \( q_1 \) is defined by
min\{q_1: \sum_{i=1}^{q_1} \lambda_i / \sum_{i=1}^{r_1} \lambda_i \geq p_0\}. Let b_1 follow normal distribution \(N(0, \tau_1 \Lambda_1)\). Based on the approximation, we can approximate \(\Sigma_1 a_1\) by \(E_1 b_1\), since both of them follow normal distributions with the same mean 0 and the approximately equal variance, which is shown by

\[
\text{var}(\Sigma_1 a_1) = \tau_1 \Sigma_1 \approx \tau_1 \sum_{i=1}^{q_1} \lambda_i e_i e_i^T = \text{var}(E_1 b_1).
\]  

(2.11)

Furthermore, we reparameterize the random effect vector \(b_1\) to \(c_1\) such that \(c_1\) is distributed as normal distribution \(N(0, \tau_1 I_{q_1 \times q_1})\). In this way, we keep the vector \(c_1\) with the same variance structure \(\tau_1 I_{q_1 \times q_1}\), and hence the estimate of \(c_1\) is stable and has the same order of magnitude. Denote by \(B_1\) the matrix associated with \(c_1\). Then, \(B_1 = [\lambda_1^{1/2} e_1, \ldots, \lambda_{q_1}^{1/2} e_{q_1}]\) since

\[
\text{var}(B_1 c_1) = \tau_1 \sum_{i=1}^{q_1} \lambda_i e_i e_i^T = \text{var}(E_1 b_1).
\]  

(2.12)

We apply the same strategy to each random effect \(a_j, j = 1, \ldots, p\). The GAMs in (2.8) can be approximated as

\[
g(\mu) = 1 \beta + N_1 B_1 c_1 + \ldots + N_p B_p c_p.
\]  

(2.13)

Now \(c_j\)'s (\(j = 1, \ldots, p\)) are random effects from normal distribution \(N(0, \tau_j I_{q_j \times q_j})\). We implement the algorithm over the above model representation.

### 2.3 The Adaptive LASSO for Generalized Additive Models

#### 2.3.1 Methodology

We use the quadratic (\(h = 1\)) smoothing spline estimation in model selection. Through the eigenvalue-eigenvector decomposition and reparameterization, the nonparametric function \(f_j(.)\) can be written as

\[
f_j(x) = 1 \delta_j + N_j B_j c_j,
\]  

(2.14)
where \( \delta_j \) is a constant and \( c_j = (c_{1j}, \ldots, c_{q_j})^T \) follows normal distribution \( N(0, \tau_j I_{q_j \times q_j}) \). If the variance component \( \tau_j \) is 0, \( f_j(x) \) is reduced to a constant vector \( 1 \delta_j \). Thus, it does not make any contribution to the model and the corresponding covariate \( X_j \) should be eliminated from the model. As a result, the model selection is reduced to the variance components estimation and selection. If the variance estimate is zero, the corresponding variable is noninformative. On the other hand, the informative variables will have significant estimates of variance components. Inspired by Zou (2006), Zhang and Lu (2007), we propose the adaptive LASSO for GAMs. Let the parameter set \( \theta = (\beta, \tau_1, \ldots, \tau_p)^T \) where \( \tau_1, \ldots, \tau_p \) are induced variance components from nonparametric functions. Let \( c = (c_1^T, \ldots, c_p^T)^T \). Apparently, the random effect \( c \) is distributed as normal distribution \( N(0, \Sigma_c) \), with the variance-covariance matrix as

\[
\Sigma_c = \begin{pmatrix}
\tau_1 I_{q_1 \times q_1} & 0 & \cdots & 0 \\
0 & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & \tau_p I_{q_p \times q_p}
\end{pmatrix}.
\]

The parameter estimates can be achieved by maximizing the penalized log-quasi-likelihood

\[
\ell_p(\theta, \lambda; y) = \ell(\theta; y) - n\lambda \sum_{j=1}^{p} \frac{\tilde{\tau}_j}{\tilde{\tau}_j + \epsilon}, \tag{2.15}
\]

subject to \( \tau_j \geq 0, j = 1, \ldots, p \),

where \( \tilde{\tau}_j \)'s are the estimators under the log-quasi-likelihood \( \ell(\theta; y) \). Small value of \( \epsilon \) (0.001) is chosen to avoid the cases of zero quasi-likelihood \( \ell(\theta; y) \). Small value of \( \epsilon \) (0.001) is chosen to avoid the cases of zero denominators. In order to get the good performance of the proposed method, the selection of the tuning parameter \( \lambda \) is important. We use Bayesian information criteria (BIC) to choose the best value of the tuning parameter on a one-dimensional grid search. BIC can be calculated by

\[
BIC = -2\ell + t \log(n), \tag{2.16}
\]

where \( \ell \) is the log-likelihood, \( t \) is the number of parameters in the selected model, and \( n \) is the number of observations. The best value of \( \lambda \) is the one with the smallest BIC. We will elaborate on the calculation of likelihood in BIC in the following section. After the model selection procedure finishes, we use BLUP to estimate the corresponding nonparametric
function. Given $h = 1$, the estimated value of centered function $f_j(.)$ at the set of ordered and distinct knots $x^o_j$ can be obtained by

$$\hat{f}_j(x^o_j) = B_j\hat{c}_j - \frac{1}{r_j}11^T B_j\hat{c}_j,$$

(2.17)

where $\hat{c}_j$ is obtained by maximizing $\ell(\theta; y)$, and $r_j$ is the number of points in $x^o_j$. In addition, the centered function $f_j$ will give the estimated value at any arbitrary data point $x_{ij}$ by

$$\hat{f}_j(x_{ij}) = N_{ij}\hat{f}_j,$$

(2.18)

where $N_{ij}$ is the $i$th row of the incidence matrix $N_j$.

The variance components are nonnegative. The true value of the variance component for a noninformative function is zero, which is on the boundary area. Maximizing the penalized log-quasi-likelihood in (2.15) directly is not an easy job due to the boundary issue. Moreover, there is no guarantee that the estimates of variance components are positive. To deal with these issues, we can instead use the EM algorithm (Dempster, 1977).

### 2.3.2 The EM Algorithm

The EM algorithm is an iterative process bouncing between the E-stage and the M-stage. At the E-stage, a conditional expectation of log-quasi-likelihood is calculated. At the M-stage, we maximize the conditional expectation from the E-stage to get the parameter estimates. With the EM algorithm, the likelihood always increases during the parameter update. In addition, the procedure of the parameter update can be carried out separately for smaller subsets of the whole parameter set.

#### 2.3.2.1 The MLE type of EM Algorithm

Denote by $Q$ the conditional expectation of the log-quasi-likelihood function. Under the mixed model representation in (2.13), we consider the random effect $c = (c_1^T, \ldots, c_p^T)^T$ as missing data, and $\theta = (\beta, \tau_1, \ldots, \tau_p)^T$. At step $t$, the Q function can be written as

$$Q(\theta|\hat{\theta}^{(t)}) = E\{\log f(y, c; \theta)|y, \hat{\theta}^{(t)}\},$$

(2.19)
where $\hat{\theta}^{(t)}$ is the estimate of $\theta$ at the $t$th step. Since the adaptive LASSO is applied, we need to get the conditional expectation of the penalized log-quasi-likelihood at the E-stage. Inspired by Green (1990), we impose the same penalty term as the one in (2.15) at the E-stage. Thus, the penalized Q function is

$$Q_p(\theta|\hat{\theta}^{(t)}) = Q(\theta|\hat{\theta}^{(t)}) - n\lambda \sum_{j=1}^{p} \frac{\tau_j}{\hat{\tau}_j + \epsilon}$$

$$= \mathbb{E}\{\log f(y, c; \theta)|y, \hat{\theta}^{(t)}\} - n\lambda \sum_{j=1}^{p} \frac{\tau_j}{\hat{\tau}_j + \epsilon}.$$  \hspace{1cm} (2.20)

Both Q function and penalized Q function are calculated with respect to the probability density function $f(c|y, \hat{\theta}^{(t)})$. Under the current parameter estimate $\hat{\theta}^{(t)}$,

$$f(c|y, \hat{\theta}^{(t)}) = \frac{f(y, c; \hat{\theta}^{(t)})}{f(y, \hat{\theta}^{(t)})} \propto f(y, c; \hat{\theta}^{(t)}).$$  \hspace{1cm} (2.21)

By our model specification, $f(y, c; \hat{\theta}^{(t)})$ is given by

$$f(y, c; \hat{\theta}^{(t)}) = f(y|c, \hat{\theta}^{(t)}) \propto |\Sigma_c^{(t)}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i, \mu_c^i) - \frac{1}{2} c^T(\Sigma_c^{(t)})^{-1} c\right\},$$  \hspace{1cm} (2.22)

where $d_i = -2\int_{y_i}^{\mu_c^i} m_c(y_i, u) du$, and

$$\Sigma_c^{(t)} = \begin{pmatrix} \hat{\tau}_1^{(t)} I_{q_1 \times q_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\tau}_p^{(t)} I_{q_p \times q_p} \end{pmatrix}.$$

Let $-K(c) = -\frac{1}{2\phi} \sum_{i=1}^{n} d_i(y_i, \mu_c^i) - \frac{1}{2} c^T(\Sigma_c^{(t)})^{-1} c$. Then we have

$$f(y, c; \hat{\theta}^{(t)}) \propto |\Sigma_c^{(t)}|^{-\frac{1}{2}} \exp\{-K(c)\}.$$  \hspace{1cm} (2.23)

We further expand $-K(c)$ around $\hat{c}^{(t)}$, which is the mode of $f(c|y, \hat{\theta}^{(t)})$. Thus,
\[ f(y, c; \hat{\theta}^{(t)}) \text{ is approximately proportional to} \]
\[ |\Sigma_c^{(t)}|^{-\frac{1}{2}} \exp \left\{ -K(\hat{c}^{(t)}) - [K'(\hat{c}^{(t)})]^T (c - \hat{c}^{(t)}) - \frac{1}{2} (c - \hat{c}^{(t)})^T K''(\hat{c}^{(t)})(c - \hat{c}^{(t)}) \right\} \]
\[ = |\Sigma_c^{(t)}|^{-\frac{1}{2}} \exp \left\{ -K(\hat{c}^{(t)}) \right\} \exp \left\{ -\frac{1}{2} (c - \hat{c}^{(t)})^T K''(\hat{c}^{(t)})(c - \hat{c}^{(t)}) \right\}, \quad (2.24) \]

where \( K' \) indicates the first derivative of \( K \) with respect to \( c \), and \( K'' \) indicates the second derivative of \( K \) with respect to \( c \). Clearly, \( K'(\hat{c}^{(t)}) = 0 \) as \( \hat{c}^{(t)} \) is also the maximizer of \( -K(c) \). Apparently, \( \exp \left\{ -\frac{1}{2} (c - \hat{c}^{(t)})^T K''(\hat{c}^{(t)})(c - \hat{c}^{(t)}) \right\} \) is a Gaussian kernel, which implies that \( c|y, \hat{\theta}^{(t)} \) approximately follows a normal distribution with mean \( \hat{c}^{(t)} \) and variance \( K''(\hat{c}^{(t)})^{-1} \). Due to the same strategy, the likelihood of BIC in (2.16) can be calculated as

\[ \ell(\theta, y) = \int f(y|c; \theta) f(c; \theta) dc, \]
\[ = \int (2\pi)^{-\frac{1}{2}} |\Sigma_c|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\phi} \sum_{i=1}^n d_i(y_i, \mu_i^c) - \frac{1}{2} c^T \Sigma_c^{-1} c \right\} dc, \]
\[ \approx \int (2\pi)^{-\frac{1}{2}} |\Sigma_c|^{-\frac{1}{2}} \exp \left\{ -K(\hat{c}) \right\} \exp \left\{ -\frac{1}{2} (c - \hat{c})^T K''(\hat{c})(c - \hat{c}) \right\} dc, \]
\[ = |\hat{\Sigma}_c|^{-\frac{1}{2}} \exp \left\{ -K(\hat{c}) \right\} |K''(\hat{c})|^{-\frac{1}{2}}, \quad (2.25) \]

where \( r \) is the dimension of \( c \), and \( \hat{c} \) is the mode of \( f(c|y, \hat{\theta}) \).

The penalized \( Q \) function can be written as

\[ Q_p(\theta|\hat{\theta}^{(t)}) = E\{\log f(y, c; \theta)|y, \hat{\theta}^{(t)}\} - n\lambda \sum_{j=1}^p \frac{\tau_j}{\tau_j + \epsilon} \]
\[ = -\frac{1}{2\phi} \sum_{i=1}^n E\{d_i(y_i, \mu_i^c)|y, \hat{\theta}^{(t)}\} - \frac{1}{2} \sum_{j=1}^p (q_j \log \tau_j + 1 \tau_j E(c_j^T c_j|y, \hat{\theta}^{(t)}) \}
\[ - n\lambda \sum_{j=1}^p \frac{\tau_j}{\tau_j + \epsilon}, \quad (2.26) \]

where \( q_j \) is the dimension of \( c_j \), and the conditional expectations are taken with respect to the above approximate normal distribution.

In the M-stage, we maximize \( Q_p \) with respect to \( \tau_j \)'s (\( j = 1, \ldots, p \)) to get the following.
update formula

\[ \hat{\tau}_j^{(t+1)} = \frac{2E(c_j^T c_j|y, \hat{\theta}^{(t)})}{\sqrt{q_j^2 + 4e_j E(c_j^T c_j|y, \hat{\theta}^{(t)}) + q_j}} \approx \frac{2a_j}{\sqrt{q_j^2 + 4a_j e_j + q_j}}, \]  

(2.27)

where

\[ e_j = \frac{2n\lambda}{\bar{\tau}_j + \epsilon}, \]

\[ a_j = (\hat{e}_j^{(t)})^T \hat{c}_j^{(t)} + \text{tr}(\hat{M}_j^{(t)}), \]

\( \hat{c}_j^{(t)} \) is the corresponding vector in \( \hat{c}^{(t)} \), which is the mode of \( f(c|y, \hat{\theta}^{(t)}) \), and \( \hat{M}_j^{(t)} \) is the corresponding matrix in \( \{K''(\hat{c}^{(t)})\}^{-1} \). It should be noticed that the update is an approximate procedure because it is based on the fact that \( c|y, \hat{\theta}^{(t)} \) is an approximate normal distribution.

Since the maximizer of \( \beta \) does not have a closed form, we use the Newton-Raphson’s method to update \( \beta \) as follows:

\[ \hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} - \frac{\frac{d}{d\beta} \sum_{i=1}^{n} E\{d_i(y_i, \mu_i^c)|y, \hat{\theta}^{(t)}\}}{\frac{d^2}{d\beta^2} \sum_{i=1}^{n} E\{d_i(y_i, \mu_i^c)|y, \hat{\theta}^{(t)}\}}. \]  

(2.28)

The integration involved in the Newton-Raphson’s method can be calculated with some numerical integration techniques. The Gaussian quadrature method is applied here to approximate the integration. The expectations in (2.28) are calculated with respect to \( \mu_i^c|y, \hat{\theta}^{(t)} \). Let \( \eta_i = g(\mu_i^c) = \beta + z_i^T \hat{c}^{(t)} \), where \( z_i \) is the \( i \)th row of matrix \( [N_1 B_1, \ldots, N_p B_p] \). Thus, we can obtain that \( \eta_i \) is a random variable from normal distribution \( N(\beta + z_i^T \hat{c}^{(t)}, z_i^T \{K''(\hat{c}^{(t)})\}^{-1} z_i) \). Hence, the expectations are taken with respect to a normal random variable and we apply the Gaussian quadrature method to approximate them.

In summary, the procedure goes as follows:

Step 1: Set initial values to \( \hat{\beta}^{(0)}, \hat{\tau}_1^{(0)}, \ldots, \hat{\tau}_p^{(0)} \).

Step 2: Set \( \lambda = 0 \). At the \( t \)th iteration, get the mode \( \hat{c}^{(t)} \) of \( f(c|y, \hat{\theta}^{(t)}) \) and update \( \hat{\beta}^{(t+1)}, \hat{\tau}_1^{(t+1)}, \ldots, \hat{\tau}_p^{(t+1)} \) until convergence.

Step 3: Initialize \( \hat{\beta}^{(0)}, \hat{\tau}_1^{(0)}, \ldots, \hat{\tau}_p^{(0)} \) with their estimates obtained from Step 2. If \( \hat{\tau}_j \) \( (j = 1, \ldots, p) \) is zero, substitute it by a small positive value as the initial value.

Step 4: For each value of \( \lambda > 0 \), at the \( t \)th iteration, get the mode \( \hat{c}^{(t)} \) of \( f(c|y, \hat{\theta}^{(t)}) \)
and update $\hat{\beta}(t+1), \hat{\tau}_1^{(t+1)}, \ldots, \hat{\tau}_p^{(t+1)}$ until convergence. Using BIC, find the best tuning value of $\lambda$ based on a grid search.

Step 5: For the selected variable $X_j$, estimate $f_j(.)$ by quadratic smoothing splines.

After the model selection procedure completes, we use approximate BLUP to estimate the selected nonparametric functions. For the selected function $f_j$, the estimated function value at an arbitrary point $x_{ij}$ is expressed as

$$
\hat{f}_j(x_{ij}) = (N_{ij}B_j - \frac{1}{r_j}11^T N_{ij}B_j)\hat{c}_j = F_j\hat{c}_j,
$$

(2.29)

where $N_{ij}$ is the $i$th row of the incidence matrix $N_j$, $r_j$ is the number of ordered and distinct points in $x_j$, and $F_j = N_{ij}B_j - r_j^{-1}11^T N_{ij}B_j$. To get a more accurate estimation, we run Step 1 and Step 2 again, but only include the selected covariates and use a higher value of $p_0$ in the eigenvalue-eigenvector decomposition. This is a two-stage estimation because the selection and the estimation are in two separate stages.

Furthermore, we can calculate the pointwise confidence intervals of the selected nonparametric functions. Lin and Zhang (1999) provided the formula of the approximate frequentist and Bayesian covariance matrix of $\hat{\beta}$ and $\hat{c}$.

For our model, the approximate frequentist covariance matrix is written as

$$
cov_F(\hat{\beta}, \hat{c}) = H^{-1}H_0H^{-1},
$$

where $H_0 = (1, Z)^TW(1, Z)$ and $H = \begin{pmatrix} 1^TW1 & 1^TWZ \\ Z^TW1 & Z^TWZ + \Sigma^{-1}_c \end{pmatrix}$.

The approximate Bayesian covariance matrix is written as

$$
cov_B(\hat{\beta}, \hat{c}) = H^{-1}.
$$

Without loss of generality, suppose there are $l$ functions selected, $f_1, \ldots, f_l$. In this case, we have

$$
Z = [N_1B_1, \ldots, N_lB_l],
$$

$$
W = diag[\{\phi m_i^{-1}v(\mu_i)g(\mu_i)^2\}]^{-1},
$$

where $\phi$ is the scale parameter and $m_i$ is a known weight. As a result, the 95% pointwise
frequentist confidence interval for \( f_j(x_{ij}) \) is
\[
[\hat{f}_j(x_{ij}) - 1.96 \times \sqrt{\text{var}_F\{\hat{f}_j(x_{ij})\}}, \hat{f}_j(x_{ij}) + 1.96 \times \sqrt{\text{var}_F\{\hat{f}_j(x_{ij})\}}],
\]
where \( \text{var}_F\{\hat{f}_j(x_{ij})\} = F_j \text{var}_F(\hat{c}_j) F^T_j \), and \( \text{var}_F(\hat{c}_j) \) is the corresponding block of \( \text{cov}_F(\hat{\beta}, \hat{c}) \).

Similarly, the 95\% pointwise Bayesian confidence interval for \( f_j(x_{ij}) \) is
\[
[\hat{f}_j(x_{ij}) - 1.96 \times \sqrt{\text{var}_B\{\hat{f}_j(x_{ij})\}}, \hat{f}_j(x_{ij}) + 1.96 \times \sqrt{\text{var}_B\{\hat{f}_j(x_{ij})\}}],
\]
where \( \text{var}_B\{\hat{f}_j(x_{ij})\} = F_j \text{var}_B(\hat{c}_j) F^T_j \), and \( \text{var}_B(\hat{c}_j) \) is the corresponding block of \( \text{cov}_B(\hat{\beta}, \hat{c}) \).

**2.3.2.2 The REML type of EM Algorithm**

Moreover, we also perform the REML type EM algorithm, where \( \beta \) is considered as an extra random effect. Let \( s = (\beta, c_1^T, \ldots, c_p^T)^T \). Hence, \( s \) is a vector of random effects. The parameter set is \( \theta = (\tau_1, \ldots, \tau_p)^T \). Similar to the derivation above, it can be shown that \( s | y, \hat{\theta}^{(t)} \) approximately follows a normal distribution.

In the M-stage, we maximize \( Q_p \) with the respect to \( \tau_j \)'s (\( j = 1, \ldots, p \)) to get the update formula
\[
\hat{\tau}^{(t+1)}_j = \frac{2E(c_j^T c_j | y, \hat{\theta}^{(t)})}{\sqrt{q_j^2 + 4e_j E(c_j^T c_j | y, \hat{\theta}^{(t)}) + q_j}} \approx \frac{2a_j}{\sqrt{q_j^2 + 4a_je_j + q_j}}, \tag{2.30}
\]
where
\[
e_j = \frac{2n\lambda}{\hat{\tau}_j + c},
\]
\[
a_j = (\hat{c}_j^{(t)})^T \hat{c}_j^{(t)} + \text{tr}(\hat{M}_j^{(t)}),
\]
\( \hat{c}_j^{(t)} \) is the corresponding vector in \( \hat{s}^{(t)} \), which is the mode of \( f(s | y, \hat{\theta}^{(t)}) \), and \( \hat{M}_j^{(t)} \) is the corresponding matrix in \( \{K''(\hat{s}^{(t)})\}^{-1} \).

In summary, the procedure goes as follows:

- **Step 1:** Set initial values to \( \hat{\tau}^{(0)}_1, \ldots, \hat{\tau}^{(0)}_p \).
- **Step 2:** Set \( \lambda = 0 \). At the \( t \)th iteration, get the mode \( \hat{s}^{(t)} \) of \( f(s | y, \hat{\theta}^{(t)}) \) and update \( \hat{\tau}^{(t+1)}_1, \ldots, \hat{\tau}^{(t+1)}_p \) until convergence.
- **Step 3:** Initialize \( \hat{\tau}^{(0)}_1, \ldots, \hat{\tau}^{(0)}_p \) with their estimates obtained from Step 2. If \( \hat{\tau}_j \) (\( j = \ldots, p \))
1, \ldots, p) is zero, substitute it by a small positive value as the initial value.

Step 4: For each value of \( \lambda > 0 \), at the \( t \)th iteration, get the mode \( \hat{s}^{(t)} \) of \( f(s|y, \hat{\theta}^{(t)}) \) and update \( \hat{\tau}_1^{(t+1)}, \ldots, \hat{\tau}_p^{(t+1)} \) until convergence. Using BIC, find the best tuning value of \( \lambda \) based on a grid search.

Step 5: For the selected variable \( X_j \), estimate \( f_j(.) \) by quadratic smoothing splines.

Similar to the discussion in the MLE type of method, the frequentist and Bayesian confidence intervals in the REML method can also be constructed. Notice that both the MLE and the REML type of EM algorithms are approximate procedures because the parameter estimates are calculated based on the approximate normal distribution of \( c|y, \hat{\theta}^{(t)} \), or \( s|y, \hat{\theta}^{(t)} \). As we can see, updates for the variance components are definitely positive since they are calculated by the conditional expectations of quadratic forms, which are shown by (2.27), (2.30). After \( \hat{\tau}_j \)'s converge, the nonparametric functions with negligible estimates of variance components are considered as constant functions and therefore the corresponding variables need to be dropped from the model.

The key step in the EM algorithms is to get the mode of \( f(c|y, \hat{\theta}^{(t)}) \) for the MLE type of method or \( f(s|y, \hat{\theta}^{(t)}) \) for the REML type of method. We implement the Newton-Raphson method to find the mode iteratively. Take the MLE type of method as an example. We need to calculate the first and the second derivative of \(-K(c)\) with respect to \( c \). Denote by \( K'(\hat{c}^{(m)}) \) and \( K''(\hat{c}^{(m)}) \) the first and second derivative of \( K(c) \) with respect to \( c \) evaluated at iteration \( m \). The update of \( c \) at iteration \( m + 1 \) is given by

\[
\hat{c}^{(m+1)} = \hat{c}^{(m)} - K'(\hat{c}^{(m)})K''(\hat{c}^{(m)})^{-1}.
\] (2.31)

After \( \hat{c} \) converges, we plug it into the formula in (2.27) to obtain the update for \( \tau_j \)'s \( (j = 1, \ldots, p) \). A similar scenario can be generated for the REML type of method where the random effect is \( s \).

2.4 Simulation Studies

To evaluate the empirical performances of the algorithms, we carry out simulation studies. Assume there are \( n \) observations \( y_1, \ldots, y_n \) in each Monte Carlo simulation, and they are independently distributed from a binomial distribution with the binomial denominator \( w \). Since the responses follow a binomial distribution, the logit link function is used to construct the generalized additive models. We randomly generate 10 independent
random variables $X_1$ to $X_{10}$, each from a uniform distribution $\text{Unif}[0,1]$. The true model is set to be:

$$\text{logit}(\pi_i) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}),$$

(2.32)

where $\beta_0$ is a constant set to be -0.8 in the simulation, and $\pi_i$ is the binomial proportion defined as $\pi_i = E\{y_i|\beta_0, f_1(.), \ldots, f_4(.)\}/w$. The four important functions are set as follows

\begin{align*}
f_1(x) &= 4(x - 0.2)^2, \\
f_2(x) &= 2x, \\
f_3(x) &= 2 \sin(2\pi x)/(2 - \sin(2\pi x)), \\
f_4(x) &= 0.2 \sin(2\pi x) + 0.4 \cos(2\pi x) + 0.6 \sin^2(2\pi x) + 0.8 \cos^3(2\pi x) + \sin^3(2\pi x).
\end{align*}

(2.33)

Therefore, $X_1$ to $X_4$ are informative covariates and $X_5$ to $X_{10}$ are noninformative ones. The simulation goal is to identify the informative covariates among the 10 covariates and estimate the selected nonparametric functions. The true underlying regression function is $f(x) = f_1(x) + f_2(x) + f_3(x) + f_4(x)$. Due to the special feature of the quadratic ($h=1$) smoothing splines estimator, we use it to perform variable selection and also apply it to model estimation. To speed up the calculation, we apply the eigenvalue-eigenvector decomposition method with $p_0 = 0.95$ in model selection and $p_0 = 0.995$ in estimation.

We investigate the following four simulation settings:

- Setting I: $n = 200, w = 1$,
- Setting II: $n = 200, w = 5$,
- Setting III: $n = 300, w = 1$,
- Setting IV: $n = 300, w = 5$.

Table 2.1 shows the frequency of appearance of the selected covariates under the four settings using the MLE type and the REML type of methods in 100 runs.

From the table, we notice two points. Firstly, if we fix the number of observations, we tend to select the important covariates more frequently when the binomial denominator increases. It makes sense because a binomial distribution is more like a normal distribution with a larger binomial denominator, and as the binomial denominator increases, there is more information in the data about the binomial proportion $\pi_i$. Secondly, the
Table 2.1: The frequency of appearance of the covariates in selected models under the MLE and REML type of methods in 100 runs.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$x_7$</th>
<th>$x_8$</th>
<th>$x_9$</th>
<th>$x_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: $n = 200, w = 1$</td>
<td>MLE</td>
<td>67</td>
<td>61</td>
<td>92</td>
<td>93</td>
<td>8</td>
<td>9</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>47</td>
<td>36</td>
<td>89</td>
<td>98</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>II: $n = 200, w = 5$</td>
<td>MLE</td>
<td>100</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>100</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>III: $n = 300, w = 1$</td>
<td>MLE</td>
<td>85</td>
<td>88</td>
<td>99</td>
<td>100</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>81</td>
<td>81</td>
<td>100</td>
<td>100</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>IV: $n = 300, w = 5$</td>
<td>MLE</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

selection is more accurate for a situation with a larger sample size if we fix the binomial denominator. This is reasonable since more information is available with more data points involved.

In order to evaluate the performance of model selection, we can use another criterion, the average model size of selected models. Table 2.2 contains the average model size over 100 runs for the four cases with both the MLE and the REML type of methods. An interesting fact is that the MLE type of method tends to select bigger models whereas the REML type of method tends to select smaller models.

Table 2.2: The average model size in selected models over 100 runs under the MLE and REML type of methods. The true size of underlying model is 4. The standard error of the average model size is in the range $(0.017, 0.121)$.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: $n = 200$, $w = 1$</td>
<td>MLE</td>
<td>3.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>REML</td>
</tr>
<tr>
<td>II: $n = 200$, $w = 5$</td>
<td>MLE</td>
<td>4.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>REML</td>
</tr>
<tr>
<td>III: $n = 300$, $w = 1$</td>
<td>MLE</td>
<td>3.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>REML</td>
</tr>
<tr>
<td>IV: $n = 300$, $w = 5$</td>
<td>MLE</td>
<td>4.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>REML</td>
</tr>
</tbody>
</table>
Since the average model size in Table 2.2 does not discriminate the correctly selected model size with the incorrectly selected model size, we can use Table 2.3 and 2.4 to indicate the two situations. Both tables show that the MLE type of method selects more covariates than the REML type of method no matter whether the selected covariates are truly informative or not.

Table 2.3: The average correct model size in selected models over 100 runs under the MLE and REML type of methods. The standard error of the average correct model size is in the range (0, 0.096).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average correct model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: ( n = 200, \ w = 1 )</td>
<td>MLE</td>
<td>3.13</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>2.70</td>
</tr>
<tr>
<td>II: ( n = 200, \ w = 5 )</td>
<td>MLE</td>
<td>3.99</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>3.99</td>
</tr>
<tr>
<td>III: ( n = 300, \ w = 1 )</td>
<td>MLE</td>
<td>3.72</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>3.62</td>
</tr>
<tr>
<td>IV: ( n = 300, \ w = 5 )</td>
<td>MLE</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Table 2.4: The average incorrect model size in selected models over 100 runs under the MLE and REML type of methods. The standard error of the average incorrect model size is in the range (0.017, 0.057).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average incorrect model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: ( n = 200, \ w = 1 )</td>
<td>MLE</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.26</td>
</tr>
<tr>
<td>II: ( n = 200, \ w = 5 )</td>
<td>MLE</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.07</td>
</tr>
<tr>
<td>III: ( n = 300, \ w = 1 )</td>
<td>MLE</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.18</td>
</tr>
<tr>
<td>IV: ( n = 300, \ w = 5 )</td>
<td>MLE</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.03</td>
</tr>
</tbody>
</table>
To evaluate the model estimation accuracy, we use the average model errors. The model error is defined by $E\{f(x) + \beta_0 - \hat{f}(x) - \hat{\beta}_0\}^2$. For each Monte Carlo run, we use $\frac{1}{n} \sum_{i=1}^{n} \{f(x_i) + \beta_0 - \hat{f}(x_i) - \hat{\beta}_0\}^2$ to estimate the model error, where $n$ is the number of observations in that run. Table 2.5 reports the average model errors for the cases. The table indicates that for each simulation setting, the MLE method tends to give a better estimation than the REML method. The difference in the average model errors between these two methods decreases as the sample size increases.

Table 2.5: The average model error over 100 runs under the MLE and REML type of methods. The standard error of the average model error is in the range (0.004, 0.038).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average model error</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: $n = 200, \ w = 1$</td>
<td>MLE</td>
<td>0.916</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.976</td>
</tr>
<tr>
<td>II: $n = 200, \ w = 5$</td>
<td>MLE</td>
<td>0.202</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.206</td>
</tr>
<tr>
<td>III: $n = 300, \ w = 1$</td>
<td>MLE</td>
<td>0.567</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.635</td>
</tr>
<tr>
<td>IV: $n = 300, \ w = 5$</td>
<td>MLE</td>
<td>0.141</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.146</td>
</tr>
</tbody>
</table>

We plot the true functions vs. the estimated functions as well as the coverage probabilities of the pointwise confidence intervals in Figure 2.1 to 2.12. We easily observe that the estimated functions in binomial cases are closer to the true ones than those in binary cases. With a fixed binomial denominator, a larger sample size case generates a better estimation. Besides, there are visible biases in estimated functions on boundary areas and at the places where the curvatures of the nonparametric functions are high. As the number of observations or the binomial denominator increases, the biases decreases. The fact is reasonable since we have more information available if the sample size gets larger or the binomial denominator is bigger and therefore the estimated functions can catch the features of the true functions better. The plots for coverage probabilities indicate that the probabilities are significantly high at the places where the biases are small, and those from Bayesian confidence intervals are greater than the ones from frequentist confidence intervals. Moreover, the coverage probabilities from both confidence intervals approach
the nominal value quickly if the binomial denominator increases.

In summary, the proposed methods perform well in selecting the informative covariates and giving good estimation of the selected functions.

In above simulation studies, there are about half responses of 1’s. Next, let’s look at the performance of the proposed methods for a special setting (Setting V) where only a small portion of responses are 1’s. We still let the sample size equal to 300 and the binomial denominator as 1. We set $\beta_0$ as -3.5, so that around one third of $y$’s are 1’s.

• Setting V: $n = 300$, $w = 1$, $\beta_0 = -3.5$.

Table 2.6 shows the frequency table for this special case. The average model size, the average correct model size and the average incorrect model size are reported in Table 2.7. Table 2.6 and 2.7 indicate that the REML type of method outperforms the MLE type of method since the REML type of method selects more informative covariates and less uninformative covariates in this special case.

Table 2.6: The frequency of appearance of the covariates in selected models under the MLE and REML type of methods in 100 runs for the simulation setting V.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$x_7$</th>
<th>$x_8$</th>
<th>$x_9$</th>
<th>$x_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>V: $n = 300$, $w = 1$, $\beta_0 = -3.5$</td>
<td>MLE</td>
<td>84</td>
<td>73</td>
<td>100</td>
<td>100</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>89</td>
<td>82</td>
<td>99</td>
<td>100</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2.7: The average model size, the average correct model size, and the average incorrect model size over 100 runs under the MLE and REML type of methods. The standard error of the average model size, the average correct model size, and the average incorrect model size is in the range (0.030, 0.078).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average model size</th>
<th>Average correct model size</th>
<th>Average incorrect model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>V: $n = 300$, $w = 1$, $\beta_0 = -3.5$</td>
<td>MLE</td>
<td>3.76</td>
<td>3.57</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>3.8</td>
<td>3.7</td>
<td>0.1</td>
</tr>
</tbody>
</table>

In Table 2.8, the average model errors are provided. We observe that in this case, the
Table 2.8: The average model error over 100 runs under the MLE and REML type of methods. The standard error of the average model error is in the range (0.023, 0.025).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average model error</th>
</tr>
</thead>
<tbody>
<tr>
<td>V: n = 300, w = 1, β₀ = −3.5</td>
<td>MLE</td>
<td>0.579</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.524</td>
</tr>
</tbody>
</table>

estimation from the REML type of method is more accurate than that from the MLE type of method. In order to get more information about estimation performance, the plot of the true functions vs. the estimated ones is given in Figure 2.13. The plot shows that the estimation from the MLE type of method is not as good as the one from the REML type of method, especially when the values of covariates are small.
Figure 2.1: The estimated nonparametric functions averaged over 100 simulation runs with the MLE and REML type of methods for simulation setting I: \( n = 200, w = 1 \). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.
Figure 2.2: Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting I: $n = 200, w = 1$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 2.3: Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting I: $n = 200, w = 1$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 2.4: The estimated nonparametric functions averaged over 100 simulation runs with the MLE and REML type of methods for simulation setting II: $n = 200, w = 5$. The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.
Figure 2.5: Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting II: $n = 200, w = 5$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 2.6: Empirical coverage probabilities of 95\% pointwise confidence intervals with the REML type of method for simulation setting II: $n = 200, w = 5$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 2.7: The estimated nonparametric functions averaged over 100 simulation runs with the MLE and REML type of methods for simulation setting III: $n = 300, w = 1$. The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.
Figure 2.8: Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting III: $n = 300, w = 1$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 2.9: Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting III: $n = 300, w = 1$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 2.10: The estimated nonparametric functions averaged over 100 simulation runs with the MLE and REML type of methods for simulation setting IV: $n = 300$, $w = 5$. The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.
Figure 2.11: Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting IV: $n = 300, w = 5$. The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 2.12: Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting IV: \( n = 300, w = 5 \). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 2.13: The estimated nonparametric functions using the MLE and REML type of methods for simulation setting V: \( n = 300, w = 1, \beta_0 = -3.5 \). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of approach. The red dotted lines are estimates using the REML type of approach.
2.5 A Real Data Application

We apply the method for model selection and estimation to the South African Heart Disease data which was used as one of examples in The elements of statistical learning (Hastie et al, 2001) and it can be obtained from http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

The partial data set is taken from a retrospective study of males in the Western Cape, South Africa. There are 462 people, 7 continuous variables and 1 categorical variable in the data set. The binary response variable \( y_i \) is 1 if the \( i \)th person has heart disease, and 0 otherwise. One of the study goals is to identify the intensity of heart disease risk factors for people in that area.

The description of the variables is shown as follows.

- \( \text{sbp} \): systolic blood pressure (mmHg)
- \( \text{tobacco} \): cumulative tobacco consumption (kg)
- \( \text{ldl} \): low density lipoprotein cholesterol (mmol/L)
- \( \text{famhist} \): family history of heart disease (Present, Absent)
- \( \text{typea} \): type-A behavior score (coronary-prone behavior score)
- \( \text{obesity} \): body mass index (BMI)
- \( \text{alcohol} \): current alcohol consumption
- \( \text{age} \): age at onset (years)

The model applied in the data set is a generalized semi-additive model since there is a categorical variable: famhist. The model can be written as

\[
\logit(\mu_i) = \beta_0 + \beta_1 \cdot \text{famhist} + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_7(x_{i7}), i = 1, \ldots n,
\]

where famhist is 1 for people with family history and 0 otherwise, and \( x_{1}, \ldots, x_{7} \) are the seven continuous covariates in the above description. We use all variables in the analysis. By the eigenvalue-eigenvector decomposition method, we use \( p_0 = 0.95 \) in model selection and \( p_0 = 0.999 \) in model estimation. With the adaptive LASSO, we select the same 5 important variables among 8 variables by the MLE type of method and the REML type of method. The selected variables are tobacco, ldl, famhist, typea, and age.
Among 462 observations, there are around one third of observations being as 1’s, which is similar to the simulation setting V. The simulation results tell that the MLE type of method does not make a good estimation in this case. Therefore, we only use the REML type of method to perform estimation. The estimated functions and their 95% pointwise confidence intervals are presented in Figure 2.14.

As indicated by the estimation plot, the estimated nonparametric functions have an overall increasing pattern. The plot implies that people would have a higher probability to get heart disease if they smoke more, have larger low density lipoprotein cholesterol, have higher type-A behavior score, or get older. Moreover, the estimate of the coefficient in famhist is 0.908, which indicates that at given values of tobacco, ldl, typea, and age, the estimated odds of having heart disease in people with family history is \( \exp(0.908) = 2.479 \) times that in those without family history.
Figure 2.14: The estimated functions corresponding to selected covariates in the South African Heart Disease data using the REML type of method. The red solid lines are estimated functions, the blue dot-dashed lines are the 95% frequentist confidence intervals and the black dotted lines are the 95% Bayesian confidence intervals of the selected functions.
2.6 Summary

In this chapter, we focus on model selection and estimation for generalized additive models. We propose the adaptive LASSO in GAMs with the eigenvalue-eigenvector decomposition approach to simplify the matrix calculation. The adaptive LASSO is an approach that shrinks the estimates of the induced variance components. Those covariates with negligible estimates of variance components are associated with noninformative nonparametric functions. In addition, we develop the MLE and REML type of EM algorithms to get the variance parameter estimates. Thus, we can accomplish the task of model selection by variance components estimation and selection. Notice that we approximate the conditional distribution $c|y, \hat{\theta}(t)$ in the MLE type of method and $s|y, \hat{\theta}(t)$ in the REML type of method as a normal distribution. Hence the EM algorithms are approximate procedures. After the informative covariates are selected in the model, we will run the EM algorithm again, but to the model only with the selected functions. In this way, we estimate the corresponding nonparametric functions by the approximate BLUPs. In addition, we provide the pointwise confidence intervals for the selected functions.

We investigate the empirical performances of the proposed methods by simulation studies. We present the frequency of appearance of the selected covariates, the average model size, the average correct model size and the average incorrect model size. To evaluate the model estimation accuracy, we provide the average model errors and the plots of true functions vs. the estimated ones. We calculate the pointwise confidence intervals and plot the coverage probabilities for the selected functions. The simulation results show that the proposed methods can give a good selection and estimation for GAMs. Furthermore, we apply the model selection and estimation method in a real data study.

In practice, we often observe the correlated data, such as in longitudinal studies, clinical trials, and so on. Thus, we will discuss how to apply the adaptive LASSO in the variable selection and estimation for generalized additive mixed models (Lin and Zhang, 1999), which are an extension of generalized additive models for correlated data.
Chapter 3

Model Selection and Estimation in Generalized Additive Mixed Models

3.1 Introduction

In many fields of research, we often encounter overdispersed and correlated data due to repeated measurements, such as in clinical trials, longitudinal studies, and so on. Therefore, subject-specific random effects are introduced into the models to accommodate the correlation among the responses. Based on the principle, Lin and Zhang (1999) proposed generalized additive mixed models (GAMMs) which extend generalized additive models to correlated data cases. One of the research interests in generalized additive mixed models is to identify the informative variables so that the models can have a meaningful interpretation and the prediction results are stable and accurate.

The chapter is organized as follows. A brief introduction of the generalized additive mixed models and the corresponding generalized linear mixed model representation are presented in Section 3.2. In addition, the eigenvalue-eigenvector decomposition approach is utilized to simplify the model representation. In Section 3.3, the adaptive LASSO is proposed and the approximate EM algorithms are derived for model selection and estimation in GAMMs. In Section 3.4, simulation studies are conducted to evaluate the performances of the proposed methods. An application on a real data set is shown in Section 3.5. In the end, we provide a summary of the chapter.
3.2 Generalized Additive Mixed Models and the Representation

3.2.1 Generalized Additive Mixed Models

In the case of repeated measurements, each subject is measured several times. Thus, the measurements are correlated within each subject. Generalized additive mixed models can be applied in this case. Compared with generalized additive models, generalized additive mixed models introduce the subject-specific random effects. Suppose there are $m$ subjects in a data set with $n_i$ measurements for the $i$th ($i = 1, 2, \ldots, m$) subject. The number of measurements is $n = \sum_{i=1}^{m} n_i$. The subject-specific random effect $b_i$ is assumed to be distributed from normal distribution $N(0, D)$, where $D$ is an unstructured variance structure. Then, the random effect $b = (b_1^T, \ldots, b_m^T)^T$ is distributed as normal distribution $N(0, G)$, where $G = \text{diag}\{D, D, \ldots, D\}$. Then, given $b_i$, the responses $y_{ij}$ are conditionally independent with means $E(y_{ij} | b_i) = \mu_{ij}^b$, and variances $\text{var}(y_{ij} | b_i) = \phi m_{ij}^{-1} v(\mu_{ij}^b)$, where $\phi$ is a known or unknown scale parameter and $m_{ij}$ is a known weight. In the cases we consider, $\phi$ is a known quantity. The generic form of generalized additive mixed models can be expressed as

$$
g(\mu_{ij}^b) = \beta_0 + f_1(x_{ij1}) + f_2(x_{ij2}) + \cdots + f_p(x_{ijp}) + z_{ij}^T b_i,
$$

where $g(.)$ is the link function, $\beta_0$ is a constant, $f_1(.) , \ldots , f_p(.)$ are smooth and continuous nonparametric functions centered to have mean zero, and $z_{ij}$ is the corresponding vector with the random effect $b_i$.

3.2.2 The Generalized Linear Mixed Model Representation

Based on the discussion of Section 2.2.2, we can establish the generalized linear mixed model representation for GAMMs. With the $h$th-order smoothing splines estimation for nonparametric functions, the GAMMs can be written as

$$
g(\mu^b) = 1\beta_0 + N_1 T_1 \delta_1 + \cdots + N_p T_p \delta_p + N_1 \Sigma_1 a_1 + \cdots + N_p \Sigma_p a_p + Zb,
$$

where $\mu = (\mu_{11}^b, \ldots, \mu_{1n_1}^b, \ldots, \mu_{m1}^b, \ldots, \mu_{mn_m}^b)^T$, $\beta_0, \delta_k$'s ($k = 1, \ldots, p$) are fixed effects, $a_k$'s are random effects induced from nonparametric functions with normal distribution.
$N\{0, \tau_k \Sigma_k^{-1}\}$, and the real random effect $b$ is from normal distribution $N(0, G)$. The matrix $Z$ is the design matrix associated with $b$ and it can be constructed by $Z = \text{diag}\{Z_i\}$ with $Z_i = (z_{i1}, \ldots, z_{im_i})^T$.

For the quadratic smoothing splines ($h = 1$), $T_k$ is a $r_k \times 1$ vector of ones, $\delta_k$ is a scalar parameter, and thus $N_k T_k \delta_k$ is reduced to $1\delta_k$. In this setting, the generalized linear mixed model representation of GAMMs is

$$g(\mu^b) = 1\beta_0 + 1\delta_1 + \cdots + 1\delta_p + N_1 \Sigma_1 a_1 + \cdots + N_p \Sigma_p a_p + Zb$$

$$= 1\beta + N_1 \Sigma_1 a_1 + \cdots + N_p \Sigma_p a_p + Zb,$$  \hspace{1cm} (3.3)

where $\beta_0, \delta_1, \ldots, \delta_p$ are combined into a single parameter $\beta$.

### 3.2.3 Eigenvalue-Eigenvector Decomposition

A repeated-measurement case usually has a large number of observations, which is computationally challenging in performing the algorithm. Thus, we propose the eigenvalue-eigenvector decomposition approach to approximate the induced random effects $a_k$’s $(k = 1, \ldots, p)$. After the eigenvalue-eigenvector decomposition approach, the approximated induced random effects have different variance structures, which makes the estimation unstable. Therefore we need to reparameterize them in order to have stable estimates. Following the discussion in Section 2.2.3, we can approximate the representation in (3.3) as follows

$$g(\mu^b) = 1\beta + N_1 B_1 d_1 + \cdots + N_p B_p d_p + Zb,$$  \hspace{1cm} (3.4)

where $B_k = [\lambda_1^{1/2} e_1, \ldots, \lambda_{q_k}^{1/2} e_{q_k}]$ with $q_k$ as the number of selected leading eigenvalues of $\Sigma_k$, $e_i$ and $\lambda_i$ are the $i$th leading eigenvector and eigenvalue. The random effect $d_k$ follows normal distribution $N\{0, \tau_k I_{q_k \times q_k}\}$. We work with the representation of (3.4) to perform the model selection and estimation.
3.3 The Adaptive LASSO and the EM Algorithm for Generalized Additive Mixed Models

The representation in (3.4) is quite similar to the one in (2.13), which implies that the adaptive LASSO can be applied in GAMMs to perform the model selection.

3.3.0.1 The MLE type of EM Algorithm

Let the parameter set \( \theta = (\beta, \tau_1, \cdots, \tau_p, e^T)^T \), where \( e \) is the vector of unique elements in \( D \). Notice that the random effects from GAMMs include both the ones induced from smoothing splines estimators of nonparametric functions and the ones from subject-specific random effects. Let \( d = (d_1^T, \cdots, d_p^T)^T \). Thus, the random effect \( d \) is distributed from normal distribution \( N(0, \Sigma_d) \), with \( \Sigma_d = \text{diag}\{\tau_k I_{q_k} \times q_k \} \) \((k = 1, \cdots, p)\). In addition, the real random effect \( b \) has the variance-covariance matrix as \( \text{var}(b) = G \).

Let \( c = (d^T, b^T)^T \). With the adaptive LASSO, the parameter estimates can be obtained by maximizing the penalized log-quasi-likelihood

\[
\ell_p(\theta; \lambda; y) = \ell(\theta; y) - n\lambda \sum_{k=1}^{p} \frac{\tau_k}{\tilde{\tau}_k + \epsilon},
\]

subject to \( \tau_k \geq 0, k = 1, \cdots, p \),

where \( \tilde{\tau}_k \)'s are the estimators of \( \tau_k \) under \( \ell(\theta; y) \), and \( \lambda \) is a tuning parameter. A small value of \( \epsilon \) is chosen to avoid the cases of zero denominators. We use the Bayesian information criteria (BIC) to pick up the best value of the tuning parameter \( \lambda \). For GAMMs, BIC can be calculated as

\[
BIC = -2\ell(\theta; y) + t\log(n),
\]

where \( t \) is the number of parameters in the selected model, and \( n = \sum_{i=1}^{m} n_i \) is the number of measurements. We will discuss the derivation on the likelihood in BIC in the following section.

As discussed before, the true value of the variance component for a noninformative function is zero, which is on the boundary area. Hence, maximizing the penalized log-quasi-likelihood in (3.5) directly is not an easy job due to the boundary issue. Moreover, there is no guarantee that the estimates of variance components through the direct maxi-
mization are positive. To tackle the problems, we instead use the EM algorithm. Consider the random effect $c$ as missing data. At step $t$, the penalized Q function can be written as

$$Q_p(\theta; \hat{\theta}^{(t)}, y, \lambda) = E\{\log f(y, c; \theta) | y, \hat{\theta}^{(t)}\} - n\lambda \sum_{k=1}^{p} \frac{\tau_k}{\tau_k + \epsilon}. \quad (3.7)$$

The above conditional expectation is with respect to the probability density function $f(c | y; \hat{\theta}^{(t)})$. Given the current parameter estimate $\hat{\theta}^{(t)}$,

$$f(c | y; \hat{\theta}^{(t)}) = \frac{f(y, c; \hat{\theta}^{(t)})}{f(y; \hat{\theta}^{(t)})} \propto f(y, c; \hat{\theta}^{(t)}). \quad (3.8)$$

Denote by $\phi$ the known scale parameter. We have

$$f(y, c; \hat{\theta}^{(t)}) \propto |\Sigma_c^{(t)}|^{-\frac{1}{2}} \exp \{-\frac{1}{2\phi} \sum_{i=1}^{m} \sum_{j=1}^{n_i} d_{ij}(y_{ij}, \mu_{ij}^c) - \frac{1}{2} c^T (\Sigma_c^{(t)})^{-1} c\}, \quad (3.9)$$

where $d_{ij} = -2\int_{y_{ij}}^{\mu_{ij}} \frac{m_j(y_{ij} - u)}{\phi v(u)} \, du$ and $\Sigma_c^{(t)}$ is the variance-covariance matrix of the random effect $c$, which can be expressed as

$$\Sigma_c^{(t)} = \begin{pmatrix}
\hat{\tau}_1^{(t)} I_{q_1 \times q_1} & 0 & 0 & \ldots & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \ldots & 0 & \hat{\tau}_p^{(t)} I_{q_p \times q_p} & 0 \\
0 & \ldots & 0 & \hat{\tau}_p^{(t)} I_{q_p \times q_p} & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \ldots & 0 & \hat{\tau}_p^{(t)} I_{q_p \times q_p} & 0 \\
\hat{D}^{(t)} & \ldots & \ldots & \ldots & 0 \\
0 & \ldots & 0 & \hat{D}^{(t)} & 0
\end{pmatrix}.$$ 

Let $-K(c) = -\frac{1}{2\phi} \sum_{i=1}^{m} \sum_{j=1}^{n_i} d_{ij}(y_{ij}, \mu_{ij}^c) - \frac{1}{2} c^T (\Sigma_c^{(t)})^{-1} c$ as a function of $c$. Then, we have

$$f(y, c; \hat{\theta}^{(t)}) \propto |\Sigma_c^{(t)}|^{-\frac{1}{2}} \exp \{-K(c)\}.$$ 

We further expand $-K(c)$ around $\hat{c}^{(t)}$, which is the mode of $f(c | y; \hat{\theta}^{(t)})$. In this way,
\[ f(y, c; \hat{\theta}(t)) \] can be approximately written as
\[
|\Sigma_c^{(t)}|^{-\frac{1}{2}} \exp\left\{ -K(\hat{c}(t)) - \frac{1}{2}(c - \hat{c}(t))^T K''(\hat{c}(t))(c - \hat{c}(t)) \right\}
= |\Sigma_c^{(t)}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(c - \hat{c}(t))^T K''(\hat{c}(t))(c - \hat{c}(t)) \right\},
\]
(3.10)
where \( K' \) indicates the first derivative of \( K \) with respect to \( c \), and \( K'' \) indicates the second derivative of \( K \) with respect to \( c \). Clearly, \( K'(\hat{c}(t)) = 0 \) as \( \hat{c}(t) \) is also the maximizer of \( -K(c) \). Apparently, \( \exp\left\{ -\frac{1}{2}(c - \hat{c}(t))^T K''(\hat{c}(t))(c - \hat{c}(t)) \right\} \) is a Gaussian kernel, which implies that \( c|y, \hat{\theta}(t) \) approximately follows a normal distribution with mean \( \hat{c}(t) \) and variance \( \{K''(\hat{c}(t))\}^{-1} \). Based on the same spirit as above, we can calculate the likelihood in BIC (3.6) as follows

\[
\ell(\theta, y) = \int f(y|c; \theta)f(c; \theta)dc,
= \int (2\pi)^{-\frac{r}{2}} |\Sigma_c|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\hat{\phi}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} d_{ij}(y_{ij}, \mu_{ij}^c) - \frac{1}{2} c^T \Sigma_c^{-1} c \right\} dc,
\approx \int (2\pi)^{-\frac{r}{2}} |\Sigma_c|^{-\frac{1}{2}} \exp\left\{ -K(\hat{c}) \right\} \exp\left\{ -\frac{1}{2}(c - \hat{c})^T K''(\hat{c})(c - \hat{c}) \right\} dc,
= |\Sigma_c|^{-\frac{1}{2}} \exp\left\{ -K(\hat{c}) \right\} |K''(\hat{c})|^{-\frac{1}{2}},
\]
where \( r \) is the dimension of \( c \), and \( \hat{c} \) is the mode of \( f(c|y, \hat{\theta}) \).

The penalized \( Q \) function can be written as

\[
Q_p(\theta|\hat{\theta}(t), y, \lambda) = -\frac{1}{2\hat{\phi}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} E\{d_{ij}(y_{ij}, \mu_{ij}^c)\}
- \frac{1}{2} \sum_{k=1}^{p} \left\{ q_k \log\tau_k + \frac{1}{\tau_k} E(d_k^T d_k|y, \hat{\theta}(t)) \right\} - n\lambda \sum_{k=1}^{p} \frac{\tau_k}{\tau_k + \epsilon}
- \frac{1}{2} \left\{ m\log|D| + \sum_{i=1}^{m} E(b_i^T D^{-1} b_i)|y, \hat{\theta}(t)) \right\},
\]
where \( m \) is the number of subjects, and the conditional expectations are taken with respect to the above approximate normal distribution. In the M-stage, we maximize \( Q_p \).
with respect to variance components to get the update formulae as

\[
\hat{\tau}_k^{(t+1)} = \frac{2E(d_k^T d_k | y, \hat{\theta}^{(t)})}{\sqrt{q^2_k + 4e_k E(d_k^T d_k | y, \hat{\theta}^{(t)}) + q_k}} \approx \frac{2a_k}{\sqrt{q^2_k + 4a_k e_k + q_k}}
\]

\[
\hat{D}^{(t+1)} = \frac{1}{m} \sum_{i=1}^{m} (\hat{d}_i^{(t)})^T \hat{d}_i^{(t)} + \hat{G}_{ii}^{(t)},
\]

(3.11)

where

\[
e_k = \frac{2n\lambda}{\hat{\tau}_k + \epsilon},
\]

\[
a_k = (\hat{d}_k^{(t)})^T \hat{d}_k^{(t)} + \text{tr}(\hat{M}_k^{(t)}),
\]

\[
\hat{d}_k^{(t)}, \hat{b}_i^{(t)}
\]

are the corresponding vectors in \( \hat{c}^{(t)} \), which is the mode of \( f(c | y, \hat{\theta}^{(t)}) \) and \( \hat{M}_k^{(t)} \),

\[
\hat{G}_{ii}^{(t)}
\]

are the corresponding matrices in \( \{K''(\hat{c}^{(t)})\}^{-1} \). Thus, the EM algorithm is an approximate procedure.

Since the maximizer of \( \beta \) does not have a closed form, we use the Newton-Raphson’s method to update \( \beta \) as follows:

\[
\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} - \frac{d}{d\beta} \sum_{i=1}^{m} \sum_{j=1}^{n_i} E\{d_{ij}(y_{ij}, \mu_{ij}^{c}) | y, \hat{\theta}^{(t)}\} \frac{d^2}{d\beta^2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} E\{d_{ij}(y_{ij}, \mu_{ij}^{c}) | y, \hat{\theta}^{(t)}\}.
\]

(3.12)

Based on the approximate normal distribution of \( c | y, \hat{\theta}^{(t)} \) and similar to the derivation in (2.28), the expectations in (3.12) are taken with respect to a random normal variable and hence they can be approximated by the Gaussian quadrature method.

In summary, the procedure goes as follows:

Step 1: Set initial values to \( \hat{\beta}^{(0)}, \hat{\tau}_1^{(0)}, \ldots, \hat{\tau}_p^{(0)} \), and \( \hat{D}^{(0)} \).

Step 2: Let \( \lambda = 0 \). At the \( t \)th iteration, get the mode of \( f(c | y, \hat{\theta}^{(t)}) \) and update \( \hat{\beta}^{(t+1)}, \hat{\tau}_1^{(t+1)}, \ldots, \hat{\tau}_p^{(t+1)}, \hat{D}^{(t+1)} \) until convergence.

Step 3: Initialize \( \hat{\beta}^{(0)}, \hat{\tau}_1^{(0)}, \ldots, \hat{\tau}_p^{(0)}, \hat{D}^{(0)} \) with their estimates obtained from Step 2. If \( \hat{\tau}_k \) is zero, substitute it by a small positive value as the initial value.

Step 4: For each value of \( \lambda > 0 \), at the \( t \)th iteration, get the mode of \( f(c | y, \hat{\theta}^{(t)}) \) and update \( \hat{\beta}^{(t+1)}, \hat{\tau}_1^{(t+1)}, \ldots, \hat{\tau}_p^{(t+1)}, \hat{D}^{(t+1)} \) until convergence. Using BIC, find the best value of the tuning \( \lambda \) based on a grid search.

Step 5: For the selected variable \( X_k \), estimate \( f_k(\cdot) \) by quadratic smoothing splines.

If the estimate of \( \tau_k \) from Step 4 is significant, it indicates that the nonparametric
function $f_k$ is not a constant and hence makes a significant contribution to the model. Then we proceed to estimate it with the quadratic smoothing splines ($h=1$). Given $h = 1$, the estimated centered function $f_k(.)$ at the set of ordered and distinct knots $x_k^0$ can be obtained by

$$
\hat{f}_k(x_k^0) = B_k\hat{d}_k - \frac{1}{r_k}11^TB_k\hat{d}_k,
$$

(3.13)

where $r_k$ is the number of points in $x_k^0$. In addition, at any arbitrary value of $x_{ijk}$

$$
\hat{f}_k(x_{ijk}) = N_{ijk}B_k\hat{d}_k - \frac{1}{r_k}11^TN_{ijk}B_k\hat{d}_k
$$

$$
= F_k\hat{d}_k,
$$

(3.14)

where $N_{ijk}$ is the corresponding row of the incidence matrix $N_k$, and $F_k = N_{ijk}B_k - r_k^{-1}11^TN_{ijk}B_k$. To get a more accurate estimation, we run Step 1 and Step 2 again but only include the selected covariates and use a larger value of $p_0$ in the eigenvalue-eigenvector decomposition approach.

Furthermore, we can calculate the pointwise confidence intervals of the selected non-parametric functions. Without loss of generality, we assume that there are $l$ functions selected, $f_1,\ldots,f_l$. Based on the previous discussion, the approximate frequentist covariance matrix is

$$
\text{cov}_F(\hat{\beta}, \hat{d}) = H^{-1}H_0H^{-1},
$$

where $H_0 = (1,Z_1)^TR^{-1}(1,Z_1)$ and $H = \begin{pmatrix} 1^TR^{-1}_1 & 1^TR^{-1}_1Z_1 \\ Z_1^TR^{-1}_1 & Z_1^TR^{-1}_1Z_1 + \Sigma^{-1}_d \end{pmatrix}$. We let $Z_1 = [N_1B_1,\ldots,N_lB_l]$, $W = \text{diag}\{\phi m^{-1}_{ij}v(\mu_{ij})g'(\mu_{ij})^2\}^{-1}$, and $R = W^{-1} + Z\text{var}(b)Z^T$.

In addition, the approximate Bayesian covariance matrix is

$$
\text{cov}_B(\hat{\beta}, \hat{d}) = H^{-1}.
$$

As a result, the 95% pointwise frequentist confidence interval for $f_k(x_{ijk})$ is

$$
[\hat{f}_k(x_{ijk}) - 1.96 \times \sqrt{\text{var}_F(\hat{f}_k(x_{ijk}))}, \hat{f}_k(x_{ijk}) + 1.96 \times \sqrt{\text{var}_F(\hat{f}_k(x_{ijk}))}],
$$

where $\text{var}_F(\hat{f}_k(x_{ijk})) = F_k\text{var}_F(\hat{d}_k)F_k^T$, and $\text{var}_F(\hat{d}_k)$ is the corresponding block of $\text{cov}_F(\hat{\beta}, \hat{d})$. 

60
The 95% pointwise Bayesian confidence interval for \( f_k(x_{ijk}) \) is

\[
\left[ \hat{f}_k(x_{ijk}) - 1.96 \times \sqrt{\text{var}_B\{\hat{f}_k(x_{ijk})\}}, \ \hat{f}_k(x_{ijk}) + 1.96 \times \sqrt{\text{var}_B\{\hat{f}_k(x_{ijk})\}}, \right]
\]

where \( \text{var}_B\{\hat{f}_k(x_{ijk})\} = F_k \text{var}_B(\hat{d}_k) F_k^T \), and \( \text{var}_B(\hat{d}_k) \) is the corresponding block of \( \text{cov}_B(\hat{\beta}, \hat{d}) \).

### 3.3.0.2 The REML type of EM Algorithm

We also perform the REML type of EM algorithm, where \( \beta \) is considered as an extra random effect. Let \( s = (\beta, d_1^T, \ldots, d_p^T, b_1^T, \ldots, b_m^T)^T \), and then \( s \) is the vector of random effects. The parameter set is \( \theta = (\tau_1, \ldots, \tau_p, e^T)^T \), where \( e \) is the vector of unique elements of \( D \). Similar to the derivation in the MLE type of method, we can show that \( s|y, \hat{\theta}^{(t)} \) is an approximate normal distribution with mean \( \hat{s}^{(t)} \), and variance \( \left\{ K''(\hat{s}^{(t)}) \right\}^{-1} \).

Maximizing the penalized Q function with respect to the variance components leads to the following update formulae

\[
\hat{\tau}_k^{(t+1)} = \frac{2E(d_k^T d_k|y, \hat{\theta}^{(t)})}{\sqrt{q_k^2 + 4e_k E(d_k^T d_k|y, \hat{\theta}^{(t)}) + q_k}} \approx \frac{2a_k}{\sqrt{q_k^2 + 4a_ke_k + q_k}}
\]

\[
\hat{D}^{(t+1)} = \frac{1}{m} \sum_{i=1}^{m} (\hat{b}_i^{(t)})^T \hat{b}_i^{(t)} + \hat{G}_{ii}^{(t)}), \quad (3.15)
\]

where

\[
e_k = \frac{2n\lambda}{\hat{\tau}_k + \epsilon},
\]

\[
a_k = (\hat{d}_k^{(t)})^T \hat{d}_k^{(t)} + \text{tr}(\hat{M}_k^{(t)}),
\]

\( \hat{d}_k^{(t)}, \hat{b}_i^{(t)} \) are the corresponding vectors in \( \hat{s}^{(t)} \), which is the mode of \( f(s|y, \hat{\theta}^{(t)}) \), and \( \hat{M}_k^{(t)} \), \( \hat{G}_{ii}^{(t)} \) are the corresponding matrices in \( \left\{ K''(\hat{s}^{(t)}) \right\}^{-1} \). Hence, the REML type of method is also an approximate procedure.

In summary, the procedure goes as follows:

- **Step 1:** Set initial values to \( \hat{\tau}_1^{(0)}, \ldots, \hat{\tau}_p^{(0)} \), and \( \hat{D}^{(0)} \).

- **Step 2:** Let \( \lambda = 0 \). At the \( t \)th iteration, get the mode of \( f(s|y, \hat{\theta}^{(t)}) \) and update \( \hat{\tau}_1^{(t+1)}, \ldots, \hat{\tau}_p^{(t+1)}, \hat{D}^{(t+1)} \) until convergence.

- **Step 3:** Initialize \( \hat{\tau}_1^{(0)}, \ldots, \hat{\tau}_p^{(0)}, \hat{D}^{(0)} \) with their estimates obtained from Step 2. If \( \hat{\tau}_k \) is zero, substitute it by a small positive value as the initial value.
Step 4: For each value of $\lambda > 0$, at the $t$th iteration, get the mode of $f(s|y, \hat{\theta}^{(t)})$ and update $\tau_1^{(t+1)}, \cdots, \tau_p^{(t+1)}, \hat{D}^{(t+1)}$ until convergence. Using BIC, find the best value of the tuning $\lambda$ based on a grid search.

Step 5: For the selected variable $X_k$, estimate $f_k(.)$ by quadratic smoothing splines.

In addition, the pointwise frequentist and Bayesian confidence intervals for selected functions can be constructed similarly to the ones in the previous MLE type of method. Both EM algorithms guarantee that the estimates of variance components are definitely positive since they are updated by the conditional expectations of quadratic forms.

As mentioned before, the key step in the EM algorithm is to find the mode of $f(c|y; \hat{\theta}^{(t)})$ for the MLE type of method or $f(s|y; \hat{\theta}^{(t)})$ for the REML type of method. We implement the Newton-Raphson method, similar to the iteration step in (2.31).

3.4 Simulation Studies

To evaluate the empirical performance of the algorithm, we carry out simulation studies for longitudinal data. In GAMMs, subject-specific random effects are introduced. Here, we first consider subject-specific random intercept models, where $b_i$ is a scalar random variable from normal distribution $N(0, \sigma_b^2)$. Thus $Z = \text{diag}\{1_{n_i}\}$, where $1_{n_i}$ is an $n_i \times 1$ vector of ones. We randomly generate 10 independent random variables $X_1$ to $X_{10}$, each from a uniform distribution $\text{Unif}[0,1]$. We let random variables $X_1$, $X_2$, $X_3$, $X_6$, $X_9$, and $X_{10}$ as time-variant covariates whose values change over time, and the remaining four variables as one-time covariates. The responses $y_{ij}$'s ($i = 1, \ldots, m, j = 1, \ldots, n_i$) are from a binomial distribution with a certain binomial denominator $w$. Hence, there are $m$ subjects with $n_i$ repeated measurements for the $i$th subject in each simulation run. In order to make a more general situation, we let $n_i$, the number of repeated measurements for the $i$th subject, be a random quantity from a binomial distribution $\text{Bin}(10, 0.5)$. As a result, the number of repeated measurements for each subject varies with the center of 5. For binomial or binary responses, we use logit as the link function, and the true underlying model is expressed as:

$$\logit(\pi_{ij}) = \beta_0 + f_1(x_{ij1}) + f_2(x_{ij2}) + f_3(x_{ij3}) + f_4(x_{ij4}) + b_i,$$ (3.16)

where $\beta_0$ is a constant, set to be -0.8 in the simulation, $b_i$ follows normal distribution $N(0, \sigma_b^2)$, and $\pi_{ij}$ is the proportion defined as $E\{y_{ij}|b_i\}/w$ with $w$ being the binomial
denominator of \( y_{ij} \). The four important functions are set as follows

\begin{align*}
    f_1(x) &= 4(x - 0.2)^2, \\
    f_2(x) &= 2x, \\
    f_3(x) &= 2 \sin(2\pi x)/(2 - \sin(2\pi x)), \\
    f_4(x) &= 0.2 \sin(2\pi x) + 0.4 \cos(2\pi x) + 0.6 \sin^2(2\pi x) + 0.8 \cos^3(2\pi x) + \sin^3(2\pi x).
\end{align*}

Therefore, \( X_1 \) to \( X_4 \) are informative covariates and \( X_5 \) to \( X_{10} \) are noninformative ones. The true underlying regression function is \( f(.) = f_1(.) + f_2(.) + f_3(.) + f_4(.) \). We use the quadratic (h=1) smoothing splines to perform the variable selection and estimation.

To speed up the matrix calculation, we apply the eigenvalue-eigenvector decomposition method with \( p_0 = 0.95 \) in the model selection and \( p_0 = 0.995 \) in the estimation.

Let \( b_i \sim N(0, \sigma_b^2 = 1) \). We investigate the following simulation settings:

- Setting I: \( m=75, w = 1 \),
- Setting II: \( m=75, w = 5 \),
- Setting III: \( m=100, w = 1 \),
- Setting IV: \( m=100, w = 5 \).

Table 3.1: The frequency of appearance of the covariates in selected models under the MLE and REML type of methods.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
<th>( x_5 )</th>
<th>( x_6 )</th>
<th>( x_7 )</th>
<th>( x_8 )</th>
<th>( x_9 )</th>
<th>( x_{10} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: ( m = 75, w = 1 )</td>
<td>MLE</td>
<td>98</td>
<td>88</td>
<td>99</td>
<td>99</td>
<td>6</td>
<td>4</td>
<td>11</td>
<td>8</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>94</td>
<td>75</td>
<td>97</td>
<td>99</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>II: ( m = 75, w = 5 )</td>
<td>MLE</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>III: ( m = 100, w = 1 )</td>
<td>MLE</td>
<td>100</td>
<td>87</td>
<td>98</td>
<td>100</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>100</td>
<td>80</td>
<td>99</td>
<td>100</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>IV: ( m = 100, w = 5 )</td>
<td>MLE</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.1 shows the frequency of appearance of the selected covariates using the MLE and the REML type of methods in 100 runs for the four simulation settings. From Table
3.1, we see that we select more times of the important covariates as the binomial denominator increases. If the binomial denominator is fixed, we tend to select the informative covariates more frequently for the cases with larger sample size.

Table 3.2: The average model size in selected models over 100 runs under the MLE and REML type of methods. The true size of underlying model is 4. The standard error of the average model size is in the range (0, 0.078).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: \textit{m} = 75, \newline \textit{w} = 1</td>
<td>MLE</td>
<td>4.17</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>3.82</td>
</tr>
<tr>
<td>II: \textit{m} = 75, \newline \textit{w} = 5</td>
<td>MLE</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>4.01</td>
</tr>
<tr>
<td>III: \textit{m} = 100, \newline \textit{w} = 1</td>
<td>MLE</td>
<td>4.11</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>3.87</td>
</tr>
<tr>
<td>IV: \textit{m} = 100, \newline \textit{w} = 5</td>
<td>MLE</td>
<td>4.07</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Table 3.3: The average correct model size in selected models over 100 runs under the MLE and REML type of methods. The standard error of the average correct model size is in the range (0, 0.059).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average correct model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: \textit{m} = 75, \newline \textit{w} = 1</td>
<td>MLE</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>3.65</td>
</tr>
<tr>
<td>II: \textit{m} = 75, \newline \textit{w} = 5</td>
<td>MLE</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>4.0</td>
</tr>
<tr>
<td>III: \textit{m} = 100, \newline \textit{w} = 1</td>
<td>MLE</td>
<td>3.85</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>3.79</td>
</tr>
<tr>
<td>IV: \textit{m} = 100, \newline \textit{w} = 5</td>
<td>MLE</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>4.0</td>
</tr>
</tbody>
</table>

In order to evaluate the performance of model selection, we can also use the average size of selected models as a criterion. Table 3.2 provides the average model size over
100 runs for different settings. We easily observe that the MLE type of method tends to select bigger models compared to the REML type of method.

Since the average model size in Table 3.2 does not discriminate the correctly selected model size with the incorrectly selected model size, Table 3.3 and 3.4 summarize the two situations. The two tables clearly show that the MLE type of method is likely to select more covariates than the REML type of method no matter whether the covariates are truly informative or not.

Table 3.4: The average incorrect model size in selected models over 100 runs under the MLE and REML type of methods. The standard error of the average incorrect model size is in the range (0, 0.053).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average incorrect model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: (m = 75, \ w = 1)</td>
<td>MLE</td>
<td>0.33</td>
</tr>
<tr>
<td>I: (m = 75, \ w = 5)</td>
<td>REML</td>
<td>0.17</td>
</tr>
<tr>
<td>II: (m = 100, \ w = 1)</td>
<td>MLE</td>
<td>0.10</td>
</tr>
<tr>
<td>II: (m = 100, \ w = 5)</td>
<td>REML</td>
<td>0.01</td>
</tr>
</tbody>
</table>

To evaluate the model estimation accuracy, we use the average model errors. The model error is defined by \(E\{f(x) + \beta_0 - \hat{f}(x) - \hat{\beta}_0\}^2\). In each run, we use \(\frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \{f(x_{ij}) + \beta_0 - \hat{f}(x_{ij}) - \hat{\beta}_0\}^2\) to estimate the model error, where \(n\) is the number of measurements. Table 3.5 reports the average model errors for the cases. The table shows that the MLE type of method gets a better estimation than the REML type of method. But the difference in estimation between the two methods decreases as the binomial denominator increases.

Moreover, we present the average \(\hat{\beta}_0\) and the average \(\hat{\sigma}_b^2\) for simulation setting I to IV in Table 3.6. The true value for \(\beta_0 = -0.8\), and the true value for \(\sigma_b^2 = 1\). We notice the MLE type of method provides a better estimation than the REML type of method on these two parameters. The biases decrease as the sample size and the binomial denominator increase. But both of them give an underestimate of the variance of true
Table 3.5: The average model error over 100 runs under the MLE and REML type of methods. The standard error of the average model error is in the range (0.007, 0.032).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average model error</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: $m = 75$, $w = 1$</td>
<td>MLE</td>
<td>0.636</td>
</tr>
<tr>
<td>REML</td>
<td>0.732</td>
<td></td>
</tr>
<tr>
<td>II: $m = 75$, $w = 5$</td>
<td>MLE</td>
<td>0.243</td>
</tr>
<tr>
<td>REML</td>
<td>0.248</td>
<td></td>
</tr>
<tr>
<td>III: $m = 100$, $w = 1$</td>
<td>MLE</td>
<td>0.564</td>
</tr>
<tr>
<td>REML</td>
<td>0.637</td>
<td></td>
</tr>
<tr>
<td>IV: $m = 100$, $w = 5$</td>
<td>MLE</td>
<td>0.224</td>
</tr>
<tr>
<td>REML</td>
<td>0.228</td>
<td></td>
</tr>
</tbody>
</table>

Besides, we plot the true functions vs. the estimated functions as well as the frequentist and Bayesian coverage probabilities for setting I to IV in Figure 3.1 to 3.12. It is shown that the estimation from the settings with a larger binomial denominator is much better than that from binary data cases. Meanwhile, the estimation becomes more accurate as the sample size increases. There are biases in estimated functions on the boundary areas and at places where the true functions have high curvatures. But biases decrease as the sample size gets larger and the binomial denominator becomes big-

Table 3.6: The average $\hat{\beta}_0$ and $\hat{\sigma}_b^2$ over 100 runs under the MLE and REML type of methods. The standard error of the average $\hat{\beta}_0$ and $\hat{\sigma}_b^2$ is in the range (0.012,0.048).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average $\hat{\beta}_0$</th>
<th>Average $\hat{\sigma}_b^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: $m = 75$, $w = 1$</td>
<td>MLE</td>
<td>-0.626</td>
<td>0.450</td>
</tr>
<tr>
<td>REML</td>
<td>-1.081</td>
<td>0.408</td>
<td></td>
</tr>
<tr>
<td>II: $m = 75$, $w = 5$</td>
<td>MLE</td>
<td>-0.743</td>
<td>0.789</td>
</tr>
<tr>
<td>REML</td>
<td>-0.937</td>
<td>0.767</td>
<td></td>
</tr>
<tr>
<td>III: $m = 100$, $w = 1$</td>
<td>MLE</td>
<td>-0.723</td>
<td>0.580</td>
</tr>
<tr>
<td>REML</td>
<td>-1.049</td>
<td>0.505</td>
<td></td>
</tr>
<tr>
<td>IV: $m = 100$, $w = 5$</td>
<td>MLE</td>
<td>-0.753</td>
<td>0.803</td>
</tr>
<tr>
<td>REML</td>
<td>-0.896</td>
<td>0.794</td>
<td></td>
</tr>
</tbody>
</table>
ger. In each simulation setting, the coverage probabilities from the Bayesian confidence intervals are greater than those from the frequentist confidence intervals. The coverage probabilities from both confidence intervals are closer to the nominal value if the data set has a larger sample size and a greater binomial denominator. Besides, the coverage probabilities are low at places where the estimation in selected functions is not good.

In the above simulation settings, the true value of $\beta_0$ was chosen in such a way that there are about equal number of successes and failures in the response. Below, we investigate how the proposed methods perform for a special binary case where only about one tenth of the responses are 1’s (a rare event case). We still use the model representation in (3.16). To generate a rare event case, we let $\beta_0 = -5.5$. The number of subjects is set to be 100. Besides, we let the subject-specific random effect $b_i$ follow normal distribution $N(0, 0.2)$. We call this simulation setting as V:

- Setting V: $m=100$, $w = 1$, $b_i \sim N(0, 0.2)$, and $\beta_0 = -5.5$.

The model selection can be presented in Table 3.7.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$x_7$</th>
<th>$x_8$</th>
<th>$x_9$</th>
<th>$x_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>V: $m = 100, w = 1,$ $b_i \sim N(0, 0.2), \beta_0 = -5.5$</td>
<td>MLE</td>
<td>97</td>
<td>46</td>
<td>100</td>
<td>100</td>
<td>20</td>
<td>11</td>
<td>25</td>
<td>6</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>92</td>
<td>37</td>
<td>98</td>
<td>93</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Furthermore, the average model size, the average correct model size and the average incorrect model size are provided in Table 3.8. The tables clearly show that the MLE type of method selects bigger models than the REML type of method. There is not a big difference between the selection from the two methods for the truly informative covariates. But for the truly noninformative covariates, the MLE type of method does not give a reasonable selection.

To evaluate the estimation accuracy, the average model errors are given in Table 3.9. The average model error from the REML type of method is less than that from the MLE type of method, which indicates that the REML type of method gives a better estimation on the selected functions than the MLE type of method in the rare event
Table 3.8: The average model size, the average correct model size and the average incorrect model size over 100 runs under the MLE and REML type of methods for simulation setting \( V \) (a rare event case). The standard error of the average model size, the average correct model size, and the average incorrect model size is in the range \((0.029, 0.108)\).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average model size</th>
<th>Average correct model size</th>
<th>Average incorrect model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V: m = 100, w = 1, ) ( b_i \sim N(0, 0.2), \beta_0 = -5.5 )</td>
<td>MLE</td>
<td>4.34</td>
<td>3.43</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>3.29</td>
<td>3.2</td>
<td>0.09</td>
</tr>
</tbody>
</table>

In addition, the variance estimate of the subject-specific random effect is 0.434 in the MLE type of method and 0.094 in the REML type of method. The MLE type of method gives a dramatic overestimate of the variance component, which also indicates that the MLE type of method cannot provide a good estimation for the rare event case.

Table 3.9: The average model error over 100 runs under the MLE and REML type of methods for simulation setting \( V \) (a rare event case). The standard error of the average model error is in the range \((0.035, 0.049)\).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average model error</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V: m = 100, w = 1, ) ( b_i \sim N(0, 0.2), \beta_0 = -5.5 )</td>
<td>MLE</td>
<td>1.018</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.907</td>
</tr>
</tbody>
</table>

Figure 3.13 presents the true functions vs. the estimated functions in this rare event situation. It is shown that the estimated functions from the REML type of method are closer to the true ones than those from the MLE type of method, especially when the values of covariates are small. Thus, for the rare event binary case, the MLE type of method cannot obtain a good estimation compared to the REML type of method.

A more complicated situation is the random intercept and random slope case where the subject-specific random effect \( b_i \) is a vector from a bivariate normal distribution. Similar to the previous simulation studies for the random intercept only situations, we
set the true underlying model as

$$\text{logit}(\pi_{ij}) = \beta_0 + f_1(x_{ij1}) + f_2(x_{ij2}) + f_3(x_{ij3}) + f_4(x_{ij4}) + b_i + t_{ij} b_{i2},$$

(3.18)

where $t_{ij}$ was generated from $(0, 1)$, and $b_i = (b_{i1}, b_{i2})^T$ from a bivariate normal distribution with mean 0, variances’ 0.5 and covariance 0.25. Same as before, we let $n_i$ denote the number of repeated measurements for the $i$th subject, generated from a binomial distribution Bin (10, 0.5). Besides, we apply the eigenvalue-eigenvector decomposition method with $p_0 = 0.95$ in the model selection and $p_0 = 0.995$ in the estimation. We set the simulation settings as:

- Setting VI: $m=75, w = 1$,
- Setting VII: $m=75, w = 5$,
- Setting VIII: $m=100, w = 1$,
- Setting IX: $m=100, w = 5$.

The frequency table is presented in Table 3.10 for the models with random intercept and random slope. We notice that the selection is better for the case with a larger sample size and a bigger binomial denominator, which agrees with the results from the random intercept only models. Besides, the MLE type of method selects more variables than the REML type of method.

Table 3.10: The frequency of appearance of the covariates in selected models under the MLE and REML type of methods.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$x_7$</th>
<th>$x_8$</th>
<th>$x_9$</th>
<th>$x_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>VI: $m = 75$, $w = 1$</td>
<td>MLE</td>
<td>97</td>
<td>85</td>
<td>96</td>
<td>91</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>95</td>
<td>73</td>
<td>97</td>
<td>95</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>VII: $m = 75$, $w = 5$</td>
<td>MLE</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>VIII: $m = 100$, $w = 1$</td>
<td>MLE</td>
<td>100</td>
<td>91</td>
<td>99</td>
<td>97</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>97</td>
<td>80</td>
<td>100</td>
<td>99</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IX: $m = 100$, $w = 5$</td>
<td>MLE</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 3.11 to 3.13 contain the average model size, the average correct model size and the average incorrect model size. Overall, the MLE type of method tends to select more covariates than the REML type of method, which agrees with the results from the frequency table.

Table 3.11: The average model size in selected models over 100 runs under the MLE and REML type of methods. The true size of underlying model is 4. The standard error of the average model size is in the range (0.014, 0.079).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>VI: (m = 75, \ w = 1)</td>
<td>MLE</td>
<td>3.86</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>3.73</td>
</tr>
<tr>
<td>VII: (m = 75, \ w = 5)</td>
<td>MLE</td>
<td>4.03</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>4.02</td>
</tr>
<tr>
<td>VIII: (m = 100, \ w = 1)</td>
<td>MLE</td>
<td>3.99</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>3.84</td>
</tr>
<tr>
<td>IX: (m = 100, \ w = 5)</td>
<td>MLE</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>4.03</td>
</tr>
</tbody>
</table>

Table 3.12: The average correct model size in selected models over 100 runs under the MLE and REML type of methods. The standard error of the average correct model size is (0, 0.063).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average correct model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>VI: (m = 75, \ w = 1)</td>
<td>MLE</td>
<td>3.69</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>3.6</td>
</tr>
<tr>
<td>VII: (m = 75, \ w = 5)</td>
<td>MLE</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>4.0</td>
</tr>
<tr>
<td>VIII: (m = 100, \ w = 1)</td>
<td>MLE</td>
<td>3.87</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>3.76</td>
</tr>
<tr>
<td>IX: (m = 100, \ w = 5)</td>
<td>MLE</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>4.0</td>
</tr>
</tbody>
</table>

To study the estimation performances of proposed methods, the average model errors
Table 3.13: The average incorrect model size in selected models over 100 runs under the MLE and REML type of methods. The standard error of the average incorrect model size is in the range (0.014, 0.040).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average incorrect model size</th>
</tr>
</thead>
<tbody>
<tr>
<td>VI: m = 75, w = 1</td>
<td>MLE</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.13</td>
</tr>
<tr>
<td>VII: m = 75, w = 5</td>
<td>MLE</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.02</td>
</tr>
<tr>
<td>VIII: m = 100, w = 1</td>
<td>MLE</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.08</td>
</tr>
<tr>
<td>IX: m = 100, w = 5</td>
<td>MLE</td>
<td>0.03</td>
</tr>
</tbody>
</table>

are summarized in Table 3.14. The table shows that the MLE type of method outperforms the REML type of method in terms of the average model error since the MLE type of method generates a smaller average model error than the REML type of method in each simulation setting.

Table 3.14: The average model error over 100 runs under the MLE and REML type of methods. The standard error of the average model error is in the range (0.008, 0.034).

<table>
<thead>
<tr>
<th>Setting</th>
<th>Method</th>
<th>Average model error</th>
</tr>
</thead>
<tbody>
<tr>
<td>VI: m = 75, w = 1</td>
<td>MLE</td>
<td>0.680</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.768</td>
</tr>
<tr>
<td>VII: m = 75, w = 5</td>
<td>MLE</td>
<td>0.253</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.260</td>
</tr>
<tr>
<td>VIII: m = 100, w = 1</td>
<td>MLE</td>
<td>0.546</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.643</td>
</tr>
<tr>
<td>IX: m = 100, w = 5</td>
<td>MLE</td>
<td>0.226</td>
</tr>
<tr>
<td></td>
<td>REML</td>
<td>0.229</td>
</tr>
</tbody>
</table>

Besides, we plot the true functions vs. the estimated functions as well as the coverage probabilities for Setting VI to IX in Figure 1 to 12 in the Appendices. Clearly, the good estimation goes with the good selection. There are noticeable biases in estimated functions at boundary areas and places where the nonparametric functions have high
curvatures. The biases decrease as the sample size gets larger and the binomial denominator becomes bigger. In addition, Bayesian confidence intervals are wider and therefore the coverage probabilities from Bayesian intervals are higher than those from frequentist intervals. As the biases in estimated functions become less visible, the coverage probabilities from both confidence intervals get closer to the nominal value.
Figure 3.1: The estimated nonparametric functions averaged over 100 simulation runs with MLE and REML type of methods for simulation setting I: $m = 75, w = 1$ in model (3.16). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.
Figure 3.2: Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting I: $m = 75, w = 1$ in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 3.3: Empirical coverage probabilities of 95\% pointwise confidence intervals with the REML type of method for simulation setting I: $m = 75, w = 1$ in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 3.4: The estimated nonparametric functions averaged over 100 simulation runs with MLE and REML type of methods for simulation setting II: $m = 75, w = 5$ in model (3.16). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.
Figure 3.5: Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting II: $m = 75, w = 5$ in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 3.6: Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting II: $m = 75, w = 5$ in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 3.7: The estimated nonparametric functions averaged over 100 simulation runs with MLE and REML type of methods for simulation setting III: $m = 100, w = 1$ in model (3.16). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.
Figure 3.8: Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting III: $m = 100, w = 1$ in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 3.9: Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting III: $m = 100, w = 1$ in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 3.10: The estimated nonparametric functions averaged over 100 simulation runs with MLE and REML type of methods for simulation setting IV: $m = 100, w = 5$ in model (3.16). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.
Figure 3.11: Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting IV: \( m = 100, w = 5 \) in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 3.12: Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting IV: $m = 100, w = 5$ in model (3.16). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 3.13: The estimated nonparametric functions averaged over 100 simulation runs with MLE and REML type of methods for simulation setting V: $m = 100, w = 1, b_i \sim N(0, 0.2), \beta_0 = -5.5$. The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.
3.5 A Real Data Application

We apply the proposed method for model selection and estimation to the periodontal disease data set (Stoner, 2000). This is an observational study of patients with chronic periodontal disease. The data set includes 791 patients with 8 continuous variables and a binary response variable. Each patient was given an initial exam between 1988 and 1992 and followed a few years after that. One of the research goals is to select the informative covariates that are most associated with tooth loss and therefore to improve the treatment in periodontal disease. The response variable \( y_{ij} \) is 1 if patient \( i \) has at least one tooth extraction in the year \( j \), and \( y_{ij} = 0 \) otherwise. The short description of the attributes is shown as follows.

- Teeth: the number of teeth present at time of initial visit
- Sites: the number of diseased sites at time of initial visit
- Pddis: mean pocket depth (mm) in diseased sites at time of initial visit
- Pdall: mean pocket depth (mm) in all sites at time of initial visit
- Age: actual age (years) at each follow up measurement
- Nonsurg: the number of non-surgical periodontal procedures in a year
- Surg: the number of surgical periodontal procedures in a year
- Dent: the number of periodontal dental treatments in a year

Out of 5972 responses, there are 541 of \( y \)'s being as 1's, nearly 10%. We use \( p_0 = 0.97 \) in model selection and \( p_0 = 0.997 \) in estimation. We only consider the random intercept model to fit the data. Both the MLE and REML type of methods select the same four covariates: Teeth, Pddis, Dent, and Pdall. The previous simulation study on the rare event case indicates that the MLE type of method does not give a good estimation on selected functions. Thus, we just use the REML type of method to perform the estimation. Figure 3.14 presents the estimated functions with the 95% pointwise Bayesian and frequentist confidence intervals. The plot shows that the estimated nonparametric functions on Pddis, Dent, and Pdall have an overall increasing pattern. It implies that people would have a high probability to get tooth extracted if they have a large mean

86
pocket depth in diseased sites, get many periodontal dental treatments, or have a high mean pocket depth. The estimated function on the variable Teeth has an interesting pattern. The probability of having tooth extraction remains high for people with few teeth and drops as the number of teeth increases. The probability reaches the bottom when the number of teeth is around 27, and goes up slightly again.
Figure 3.14: The estimated functions corresponding to selected covariates in the Periodontal Disease data using the REML type of method. The red solid lines are estimated functions, the blue dot-dashed lines are the 95% frequentist confidence intervals and the black dotted lines are the 95% Bayesian confidence intervals of the selected functions.
3.6 Summary

In this chapter, we focus on model selection and estimation for generalized additive mixed models. We propose the adaptive LASSO in GAMMs with the eigenvalue-eigenvector decomposition approach to simplify the matrix calculation. We develop the approximate EM algorithms to obtain the estimates for variance components. If a nonparametric function has a significant variance component estimate, we need to select the corresponding variable in the model. After the variables are chosen in the model, we estimate the nonparametric functions by approximate BLUPs.

We investigate the empirical performances of the proposed method by simulation studies. To compare the performances of model selection for different cases, we present the frequency table, the average model size, the average correct model size, and the average incorrect model size. We calculate the average model errors to evaluate the model estimation accuracy, and plot the true functions vs. the estimated ones. The pointwise frequentist and Bayesian confidence intervals are calculated and their coverage probabilities are plotted. The simulation results show that the proposed method performs well in model selection and estimation for GAMMs. In addition, we apply the proposed method in a real data case.
Chapter 4

Conclusion and Future Work

In this dissertation, we proposed a new approach to perform model selection and estimation in generalized additive models (GAMs) and generalized additive mixed models (GAMMs). With smoothing spline estimation for nonparametric functions, there are induced random effects with normal distribution of mean zero and variance as the inverse of the smoothing parameter multiplied by a known matrix. Therefore, the inverse of the smoothing parameters are treated as induced variance components. The importance of a nonparametric function is then controlled by its variance component.

Based on the generalized linear mixed model representation, we propose the adaptive LASSO to perform the model selection. The approach shrinks the induced variance components estimates and makes those corresponding to noninformative covariates to zero. Two types of EM algorithms are utilized to get the parameter estimates. But both EM algorithms are approximate procedures since the random effects given the observation and parameter estimate approximately follow a normal distribution. After the informative functions are identified, we proceed to estimate them with the quadratic smoothing splines. To speed up the matrix calculation, the eigenvalue-eigenvector decomposition approach is applied. In addition, pointwise confidence intervals are calculated for the selected functions.

The empirical performances of proposed methods are investigated through simulation studies. For GAMMs, both the random intercept only models and random intercept plus random slope models are investigated. The simulation results show that the proposed methods give a reasonably good model selection and estimation especially when the sample size gets larger. Besides, we apply the methods on two data sets, the South
African Heart Disease and the periodontal disease data.

In GAMMs, we assume the subject-specific random effects to have a normal distribution. In practice, the normality is a strong assumption, which is not always valid. One future research topic is to relax the normality assumption. We only assume that the subject-specific random effects have mean zero and some variance. In this framework, how to identify the informative nonparametric functions and estimate them becomes an interesting point. In addition, the simulation results indicate that the estimates of variance components in the subject-specific random effects are underestimated. Therefore, we need to propose a method to perform the bias correction.
REFERENCES


93


APPENDICES
Figure 1: The estimated nonparametric functions averaged over 100 simulation runs with the MLE and the REML type of methods for simulation setting VI: $m = 75$, $w = 1$ in model (3.18). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.
Figure 2: Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting VI: $m = 75, w = 1$ in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 3: Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting VI: $m = 75, w = 1$ in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 4: The estimated nonparametric functions averaged over 100 simulation runs with the MLE and the REML type of methods for simulation setting VII: $m = 75, w = 5$ in model (3.18). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.
Figure 5: Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting VII: $m = 75, w = 5$ in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 6: Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting VII: $m = 75, w = 5$ in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 7: The estimated nonparametric functions averaged over 100 simulation runs with the MLE and the REML type of methods for simulation setting VIII: $m = 100, w = 1$ in model (3.18). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.
Figure 8: Empirical coverage probabilities of 95\% pointwise confidence intervals with the MLE type of method for simulation setting VIII: $m = 100, w = 1$ in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 9: Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting VIII: \( m = 100, w = 1 \) in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 10: The estimated nonparametric functions averaged over 100 simulation runs with the MLE and the REML type of methods for simulation setting IX: $m = 100, w = 5$ in model (3.18). The blue solid lines are the true underlying functions. The black dot-dashed lines are estimates using the MLE type of method. The red dotted lines are estimates using the REML type of method.
Figure 11: Empirical coverage probabilities of 95% pointwise confidence intervals with the MLE type of method for simulation setting IX: $m = 100, w = 5$ in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.
Figure 12: Empirical coverage probabilities of 95% pointwise confidence intervals with the REML type of method for simulation setting IX: $m = 100, w = 5$ in model (3.18). The blue solid lines are coverage probability based on frequentist estimation and the red dotted lines are coverage probability based on Bayesian estimation.