Some New Methods and Models in Functional Data Analysis

by

Tianyu Guan

M.Sc., Simon Fraser University, 2014
B.Sc., Jilin University, 2011

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2020

Approval

Name:	Tianyu Guan
Degree:	**Doctor of Philosophy (Statistics)**
Title:	**Some New Methods and Models in Functional Data Analysis**
Examining Committee:	**Chair:** Jinko Graham Professor
	Jiguo Cao Senior Supervisor Associate Professor
	Liangliang Wang Supervisor Associate Professor
	Lloyd Elliott Internal Examiner Assistant Professor Department of Statistics and Actuarial Science
	Yehua Li External Examiner Professor Department of Statistics University of California, Riverside
Date Defended:	**June 19, 2020**

Abstract

With new developments in modern technology, data are recorded continuously on a large scale over finer and finer grids. Such data push forward the development of functional data analysis (FDA), which analyzes information on curves or functions. Analyzing functional data is intrinsically an infinite-dimensional problem. Functional partial least squares method is a useful tool for dimension reduction. In this thesis, we propose a sparse version of the functional partial least squares method which is easy to interpret. Another problem of interest in FDA is the functional linear regression model, which extends the linear regression model to the functional context. We propose a new method to study the truncated functional linear regression model which assumes that the functional predictor does not influence the response when the time passes a certain cutoff point. Motivated by a recent study of the instantaneous in-game win probabilities for the National Rugby League, we develop novel FDA techniques to determine the distributions in a Bayesian model.

Keywords: Functional data analysis; Functional linear regression; Group bridge approach; Penalized B-splines; Functional partial least squares; Locally sparse; Bayesian analysis

Dedication

To my husband, Yang Zhou, who inspires me to be the best version of myself. Thank you for your love and encouragement which give me the best support. To my dear parents and grandparents, thanks for your care and unconditional love.

Acknowledgements

I would like to express my most sincere gratitude to my advisor Dr. Jiguo Cao for bringing me to the field of functional data analysis. His tremendous support, continuous encouragement, invaluable guidance, and immense knowledge helped me in all the time of research. I am truly fortunate to have him as my advisor. He saw the potential in me and encouraged me to pursue my dream of being a researcher.

I am deeply grateful to my thesis committee members for taking their time to read my thesis and giving invaluable comments for my thesis. I would like to express my sincere appreciation to Professor Liangliang Wang for being my mentor in both academia and real life. Many thanks to Professor Lloyd Elliott for serving on my committee. I would like to sincerely thank Professor Yehua Li from the University of California, Riverside for taking time from his busy schedule to serve as the external examiner, and Professor Jinko Graham for chairing my defence.

I gratefully acknowledge Professor Tim Swartz for supervising me on projects involving statistics in sports and for supporting me on my job searching. I would like to thank Professor Zhenhua Lin from the National University of Singapore for his wonderful guidance, kind help, and remarkable insights into my research. I have learned so much from the collaboration with him. I would also like to thank Professor Nancy Heckman from the University of British Columbia, who is very supportive for my research and career development.

I deeply thank all the staff and faculty members of the Department of Statistics and Actuarial Science for their dedication, especially Dr. Richard Lockhart, Dr. Gary Parker, Dr. Cary Tsai, Dr. Yi Lu, Dr. Joan Hu, Dr. Tom Loughin, Dr. Steve Thompson, Dr. Haolun Shi, Dr. Luke Bornn, and Dr. Dave Campbell whose courses greatly broadened my knowledge in statistics. I would also like to take the opportunity to thank Dr. Rachel Altman, Dr. Derek Binghan, Dr. Brad McNeney, and

Dr. Boxin Tang, Dr. Harsha Perera for their support and guidance. Many thanks to Sadika Jungic, Charlene Bradbury, Kelly Jay, Jay Young, and Jina Nam for always being there when I need help.

I would also like to thank my fellow graduate students, especially Luyao Lin, Chenlu Shi, Lu Wang, Lulu Guo, Yin Zhang, Peijun Sang, Yunlong Nie, Jianghu Dong, Shijia Wang, Shufei Ge, Yi Xiong, Dongmeng Liu, Zhiyang Zhou, Haoxuan Zhou, Trevor Thomson, Robert Nguyen, Peter Tea, Mengyun Li, Yang Bai, Dongdong Li, Yuping Yang, Haixu Wang, Chuyuan Lin, Sidi Wu, Boyi Hu, Yanjun Liu, Jingxue Feng, Botao Han, Barinder Thind, Joel Theirren, Jie Wang, Jiarui Zhang, Yifan Wu, Meng Sun, Hua Liu, Haitao Chen, Xiong Cai, Mengxiao Xu, Ying Yu, Sichen Liu, Haiyang Jiang, Anqi Chen, Sihan Cheng, Siyuan Chen, Michael Grosskopf, Jiying Wen, Wei-Hsiang Lin, Haoyao Ruan, Abdollah Safari, Jing Wang, Yueren Wang, Payman Nickchi, Pulindu Ratnasekera, William Ruth, Nate Sandholtz, Jacob Mortensen, Chirith Karunarathna. Thank you so much for the joyful discussions, the companionship, and the laughs we shared together during these years.

Table of Contents

Approval ii

Abstract iii

Dedication iv

Acknowledgements v

Table of Contents vii

List of Tables x

List of Figures xii

1 Introduction 1

2 Estimating Truncated Functional Linear Models with a Nested Group Bridge Approach 3

2.1 Introduction 3

2.2 Methodology 6

2.2.1 Nested Group Bridge Approach 6

2.2.2 Computational Method 8

2.3 Asymptotic Properties 10

2.4 Simulation Studies 13

2.5 Application: Particulate Matter Emissions Data 17

2.6 Conclusion and Discussion 18

2.7 Appendix 19

2.7.1 Penalized B-Splines Method 19
2.7.2 Effect of the Group Bridge Parameter γ in Section 2.4 19
2.7.3 Additional Simulation Results in Section 2.4 20
2.7.4 Proofs in in Section 2.3 23

3 Sparse Functional Partial Least Squares Regression 33

3.1 Introduction 33
3.2 Methodology 36
3.2.1 The First Sparse Functional Partial Least Squares Basis Function 36
3.2.2 Computational Method 39
3.2.3 An Algorithm for the Sparse Functional Partial Least Squares 39
3.3 Simulation Studies 41
3.4 Applications 46
3.4.1 OSB Furnish Data 46
3.4.2 Particulate Matter Emissions Data 48
3.5 Conclusion 49
3.6 Appendix 49
3.6.1 Functional Partial Least Squares Regression 49
3.6.2 Functional Principal Component Regression 51
3.6.3 Effect of the parameter κ in Section 3.3 52
3.6.4 Effect of the smoothing parameter γ in Section 3.3 52
3.6.5 Effect of the parameter δ in Section 3.3 53

4 In-game Win Probabilities for the National Rugby League 55

4.1 Introduction 55
4.2 Model Development 57
4.2.1 Available Data 57
4.2.2 Model Overview 59
4.2.3 Estimation of Model Components using FDA 59
4.3 Results 63

4.4 Model Validation 67

4.5 Conclusion 72

5 Summary and Future Work **74**

Bibliography **76**

List of Tables

Table 2.1 The mean of estimators for δ based on 200 simulation replications with the corresponding Monte Carlo standard deviation included in parentheses. . . . 15

Table 2.2 Mean integrated squared errors of estimators for $\beta(t)$ based on 200 simulation replications with the corresponding Monte Carlo standard deviation included in parentheses. 16

Table 2.3 Investigation of effect of the group bridge parameter γ on $\hat{\delta}$ and mean integrated squared errors (ISE) of estimators for $\beta(t)$. The results are obtained based on 200 simulation replications with the corresponding Monte Carlo standard deviations included in parentheses. 20

Table 2.4 The mean of estimators for δ based on 200 simulation replications with the corresponding Monte Carlo standard deviation included in parentheses. . . . 21

Table 2.5 Mean integrated squared errors of estimators for $\beta(t)$ based on 200 simulation replications with the corresponding Monte Carlo standard deviation included in parentheses. 22

Table 3.1 Mean integrated squared error, ISE_0, defined on the null region for $\hat{\beta}(t)$ based on 100 simulation replications with the corresponding Monte Carlo standard deviation included in parentheses. 43

Table 3.2 Mean integrated squared errors, ISE, defined on the entire region for $\hat{\beta}(t)$ based on 100 simulation replications with the corresponding Monte Carlo standard deviation included in parentheses. 44

Table 3.3 The prediction mean squared errors on test data based on 100 simulation replications with the corresponding Monte Carlo standard deviation included in parentheses. 45

Table 3.4 Investigation of effect of κ on ISE_0 and ISE for the estimators of $\beta(t)$, and the prediction mean squared errors (PMSE) on test data. The results are obtained based on 100 simulation replications with the corresponding Monte Carlo standard deviations included in parentheses. 52

Table 4.1 Potential choices of event data where all variables are measured with respect to the home team and larger values denote increasing superiority. 63

Table 4.2 Descriptive statistics of the scores corresponding to the 731 matches from the four regular seasons $(2016 - 2019)$ of the NRL. 65

Table 4.3 Descriptive statistics of the missed tackles corresponding to the 731 matches from the four regular seasons $(2016 - 2019)$ of the NRL. 66

List of Figures

Figure 2.1 (a) 108 smoothed engine acceleration curves. (b) Estimated slope function using the penalized B-splines approach (Cardot et al., 2003). The arrows indicate the direction of time. 4

Figure 2.2 The slope functions in three scenarios. The dashed vertical lines indicate the true values of δ. 13

Figure 2.3 Histograms of the estimated $\hat{\delta}$ in 200 simulation replications in the three scenarios. The results were obtained based on 200 Monte Carlo simulations with $n = 500$. The vertical lines indicate the true values of δ. 16

Figure 2.4 (a) 10 randomly selected smoothed acceleration curves. (b) Estimated $\hat{\beta}(t)$ using the penalized B-splines method (dashed line) and the proposed nested group bridge approach (solid line). 18

Figure 2.5 Estimated coefficient functions with rough functional covariates and $n = 500$ in one randomly selected simulation replicate for various methods (———, the proposed nested group bridge method; - - - - -, the penalized B-splines method; ⋯⋯⋯, the truncation method A; -·-·-·-, the truncation method B; — — —, the FLiRTI method; —·—·, the SLoS method; ———, the true $\beta(t)$). 22

Figure 2.6 Estimated coefficient functions with smooth functional covariates and $n = 500$ in one randomly selected simulation replicate for various methods (———, the proposed nested group bridge method; - - - - -, the penalized B-splines method; ⋯⋯⋯, the truncation method A; -·-·-·-, the truncation method B; — — —, the FLiRTI method; —·—·, the SLoS method; ———, the true $\beta(t)$). 23

Figure 3.1 (a) Spectra traces of 182 OSB fines samples. (b) Selected spectra traces of OSB fines samples with different proportions of sound wood (——, 100%; - - - - -, 80%; ··········, 40%; -·-·-·, 0%). 34

Figure 3.2 Estimated $\hat{\beta}(t)$ for various methods (——, the SpaFunPLS method; - - - - -, FunPLS method; -·-·-·-, FPLS_R method; — — —, the FPCR method; ——, the true $\beta(t)$) in a randomly selected simulation replicate with $n = 100$. . 46

Figure 3.3 Estimated slope function $\hat{\beta}(t)$ by the SpaFunPLS method (- - - - -) with $n = 100$ along with the true $\beta(t)$ (——). 46

Figure 3.4 (a) 10 randomly selected smoothed centered spectral curves. (b) Estimated $\hat{\beta}(t)$ using the conventional FunPLS method (- - - - -) and the proposed SpaFunPLS approach (——). 47

Figure 3.5 (a) 10 randomly selected smoothed centered acceleration curves. (b) Estimated $\hat{\beta}(t)$ using the conventional FunPLS method (- - - - -) and the proposed SpaFunPLS approach (——). 49

Figure 3.6 Estimated slope functions by SpaFunPLS method using different values of γ for Scenario II with $n = 500$ in one randomly selected simulation replicate. 53

Figure 3.7 Estimated slope functions by the SpaFunPLS method using different values of δ for Scenario II with $n = 500$ in one randomly selected simulation replicate. 54

Figure 4.1 Cubic B-spline basis functions defined on 9 equally spaced knots over the interval $[0, 80]$ minutes. 65

Figure 4.2 The score differential curves for the 731 matches. 66

Figure 4.3 The missed tackle differential curves for the 731 matches. 67

Figure 4.4 Upper left: estimated $\hat{\mu}_X(t|W, p_0)$. Upper right: estimated $\hat{\mu}_X(t|\overline{W}, p_0)$. Lower left: estimated $\hat{\mu}_D(t|W, p_0)$. Lower right: estimated $\hat{\mu}_D(t|\overline{W}, p_0)$. 68

Figure 4.5 The correct prediction rate for the 2019 NRL season. 69

Figure 4.6 Predicted instantaneous in-game win probabilities for four randomly selected matches from the 2019 season where the home team won. The dashed horizontal lines indicate the values of 0, 0.5, and 1. The dashed vertical lines indicate the times when the score changed. 70

Figure 4.7 Predicted instantaneous in-game win probabilities for four randomly selected matches from the 2019 season where the home team lost. The dashed horizontal lines indicate the values of 0, 0.5, and 1. The dashed vertical lines indicate the times when the score changed. 71

Figure 4.8 Predicted instantaneous in-game win probabilities for the match Storm versus Bulldogs on 6th April 2019 by Scenario I (———) and Scenario II (··········). The grey bars indicate the score differentials of the match. . . . 72

Figure 4.9 Predicted instantaneous in-game win probabilities for the match Storm versus Bulldogs on 6th April 2019 by Scenario I (———) and Scenario II (··········). The grey bars indicate the missed tackle differentials of the match. 73

Chapter 1

Introduction

Functional data analysis (FDA) deals with the analysis and theory of samples of curves, images, or other types of functions rather than scalars or vectors. Analyzing functional data is intrinsically an infinite-dimensional problem. It involves smoothing, dimension reduction, and regularization. In this thesis, we propose a sparse version of the functional partial least squares method to achieve dimension reduction. In FDA one of the most useful models is that of functional linear regression. We study a truncated functional linear regression model, which is a special case of the conventional functional linear regression model. Besides, we develop FDA methods to provide instantaneous in-game win probabilities for the National Rugby League.

The functional linear regression model extends the linear regression model to the functional context. In classical statistics, the linear regression model assumes that the response variable depends on the predictors in a linear form with random errors. The classic functional linear regression model relates a scalar response to a functional predictor. A scalar-on-function truncated linear regression model is a special case of the classic functional linear regression model, which assumes that the functional predictor does not influence the response when the time passes a certain cutoff point. In Chapter 2, we propose a new nested group bridge approach to estimate the slope function and the cutoff time point in the truncated functional linear model. Combined with the B-spline basis expansion and penalized least squares, the nested group bridge approach can identify the cutoff time and produce a smooth estimate of the slope function simultaneously. The proposed nested group bridge estimator is shown to be consistent, while its numerical performance is illustrated by simulation studies. The proposed nested group bridge method is demonstrated with an application of determining the effect of the past engine acceleration on the current particulate matter emission.

For infinite-dimensional functional data, dimension reduction is mandatory. Partial least squares is a very popular dimension reduction technique that has been successful in spectrometric prediction in chemometrics community. In Chapter 3 we propose a sparse version of the functional partial least squares method in the context of the functional linear regression model. The proposed method studies a functional linear regression model with a slope function that is zero on a substantial portion of its defining domain, which we call a locally sparse slope function. We expand the slope function with functional partial least squares basis functions. We aim at achieving locally sparse estimates for the functional partial least squares basis functions, and more importantly, the locally sparse estimate for the slope function. The new approach applies a functional regularization technique to each iteration step of the functional partial least squares and implements a computational method that identifies non-zero subregions on which the slope function is estimated. We illustrate the numerical performance of the proposed method via simulation studies and two real applications on the oriented strand board furnish data and the particulate matter emissions data.

In Chapter 4 we develop FDA methods for providing instantaneous in-game win probabilities for the National Rugby League (NRL). The NRL attempts to add a graphic that displays in-game win probabilities in a small corner of the screen, and be continually updated as the game circumstances change. We develop a Bayesian model where our main interest concerns the evaluation of the in-game posterior win probability. The underlying distributions in the Bayesian model are specified using novel FDA techniques.

Chapter 2

Estimating Truncated Functional Linear Models with a Nested Group Bridge Approach

2.1 Introduction

In this chapter we consider a scalar-on-function truncated linear regression model where the functional predictor $X_i(t), i = 1, \ldots, n$, is defined on a time interval $[0, T]$ but influences the scalar response Y_i only on $[0, \delta]$ for some unknown cutoff time $\delta \leq T$. Specifically, the model is written as

$$Y_i = \mu + \int_0^\delta X_i(t)\beta(t) \, \mathrm{d}\, t + \varepsilon_i, \tag{2.1}$$

where, without loss of generality, $X_i(\cdot)$ is assumed to be centered, i.e., $\mathbf{E}X_i(t) \equiv 0$, μ is then the mean of Y_i, $\beta(t)$ is the slope function (or coefficient function), and ε_i represents the noise that is independent of $X_i(\cdot)$.

An example of the scalar-on-function truncated linear regression is to determine the effects of the past engine acceleration on the current particulate matter emission. The response variable is the current particulate matter emission and the explanatory function is the smoothed engine acceleration curve for the past 60 seconds. Figure 2.1(a) displays 108 smoothed engine acceleration curves against the backward time, in which 0 means the current time, while Figure 2.1(b) shows the slope function estimated by the penalized B-splines method (Cardot et al., 2003). The penalized B-splines method is detailed in the appendix. We observe from Figure 2.1(b) that the acceleration over the past 20–60 seconds does not have apparent contribution to predicting the current particulate matter emission. Intuitively, the particulate matter emissions shall depend on the recent acceleration, but not the ancient one. Therefore, if a linear relation between the particulate matter emission and the acceleration curve is assumed, one might naturally use the truncated linear model (2.1) to analyze such data, where the task includes identifying the cutoff time beyond which the engine acceleration has no influence on the current particulate matter emission.

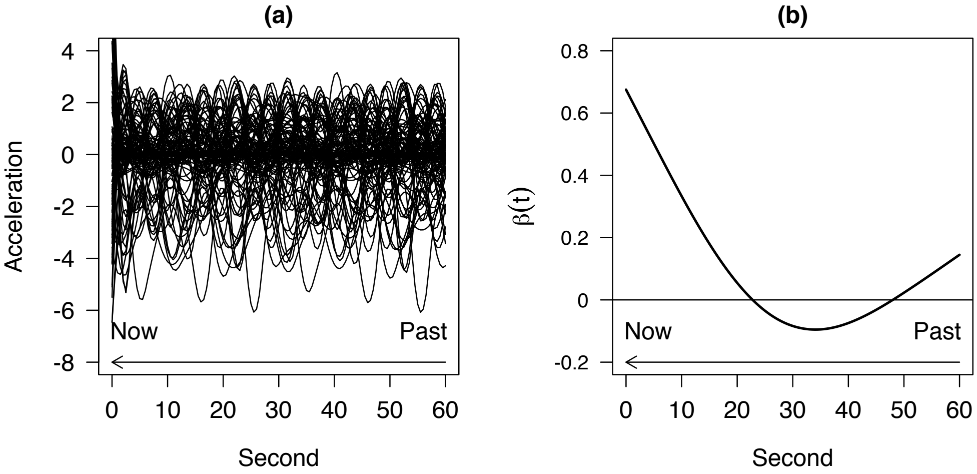

Figure 2.1: (a) 108 smoothed engine acceleration curves. (b) Estimated slope function using the penalized B-splines approach (Cardot et al., 2003). The arrows indicate the direction of time.

The degenerate case $\delta = T$ in model (2.1) corresponds to the classic functional linear regression that has been studied in vast literature. Hastie and Mallows (1993) pioneered the smooth estimation of $\beta(t)$ via penalized least squares and/or smooth basis expansion. Cardot et al. (2003) adopted B-spline basis expansion, while Li and Hsing (2007) utilized Fourier basis, both with a roughness penalty to control the smoothness of estimated slope functions. Data-driven bases such as eigenfunctions of the covariance function of the predictor process $X_i(t)$ were considered in Cardot et al. (2003), Cai and Hall (2006) and Hall and Horowitz (2007). Yuan and Cai (2010) took a reproducing kernel Hilbert space approach to estimate the slope function. The case of sparsely observed functional data was studied by Yao et al. (2005). These estimation procedures for classic functional linear regression do not apply to the truncated linear model where $\delta \leq T$ is often assumed. For models beyond linear regression and a comprehensive introduction to functional data analysis, readers are referred to the monographs by Ramsay and Silverman (2005), Ferraty and Vieu (2006), Hsing and

Eubank (2015) and Kokoszka and Reimherr (2017), as well as the review papers by Morris (2015) and Wang et al. (2016) and references therein.

Model (2.1) has been investigated by Hall and Hooker (2016) who proposed to estimate $\beta(t)$ and δ by penalized least squares with a penalty on δ^2. The resulting estimates for $\beta(t)$ are discontinuous at $t = \hat{\delta}$ where $\hat{\delta}$ stands for the estimator of δ. This feature might not be desirable when $\beta(t)$ is *a priori* assumed to be continuous. For example, it is more reasonable to assume the acceleration function influences the particulate matter emission in a continuous and smooth manner. Alternatively, we observe that model (2.1) is equivalent to a classic functional linear model with $\beta(t) = 0$ for all $t \in [\delta, T]$. Such a slope function $\beta(t)$ is a special case of locally sparse functions which by definition are functions being zero in a substantial portion of their defining domains. Locally sparse slope functions have been studied in Lin et al. (2017), as well as pioneering works of James et al. (2009) and Zhou et al. (2013). For example, in Lin et al. (2017), a general functional shrinkage regularization technique, called fSCAD, was proposed and demonstrated to be able to encourage the local sparseness. Although these endeavors are able to produce a smooth and locally sparse estimate, they do not specifically focus on the tail region $[\delta, T]$. Therefore, the estimated slope functions produced by such methods might not be zero in the region that is very close to the endpoint T, in particular when the boundary effect is not negligible.

In this chapter, we propose a new nested group bridge approach to estimate the slope function $\beta(t)$ and the cutoff time δ. Compared to the existing methods, the proposed nested group bridge approach has two features. First, it is based on the B-spline basis expansion and penalized least squares with a roughness penalty. Therefore, the resulting estimator of $\beta(t)$ is continuous and smooth over the entire domain $[0, T]$, contrasting the discontinuous estimator of Hall and Hooker (2016). Second, it employs a new nested group bridge shrinkage method proposed in Section 2.2 to specifically shrink the estimated function on the tail region $[\delta, T]$. Group bridge was proposed in Huang et al. (2009) for variable selection, and utilized by Wang and Kai (2015) for locally sparse estimation in the setting of nonparametric regression. In our approach, we creatively organize the coefficients of B-spline basis functions into a sequence of nested groups and apply the group bridge penalty to the groups. With the aid from B-spline basis expansion, such nested structure enables us to shrink the tail of the estimated slope function. This fixes the problem of the aforementioned generic locally sparse estimation procedures. An R package `ngr` has been developed for implementing the proposed method and is available at `https://github.com/caojiguo/TruFunLM`.

We structure the rest of the chapter as follows. In Section 2.2 we present the proposed nested group bridge estimation method for the slope function and the cutoff time, and also provide computational details. In Section 2.3 we investigate the asymptotic properties of the derived estimators. Simulation studies are discussed in Section 2.4, and an application to the particulate matter emissions data is given in Section 2.5. Conclusion and discussion are given in Section 2.6. In the appendix, we provide proofs and additional discussion.

2.2 Methodology

2.2.1 Nested Group Bridge Approach

Our estimation method utilizes B-spline basis functions that are detailed in de Boor (2001). Let $\boldsymbol{B}(t) = (B_1(t), \ldots, B_{M+d}(t))^{\mathrm{T}}$ be a vector that contains $M + d$ B-spline basis functions defined on $[0, T]$ with degree d and $M + 1$ equally spaced knots $0 = t_0 < t_1 < \cdots < t_M = T$. For $m \geq 0$, let $\boldsymbol{B}^{(m)}(t) = (B_1^{(m)}(t), \ldots, B_{M+d}^{(m)}(t))^{\mathrm{T}}$ denote the vector of the m-th derivatives of the B-spline basis functions. Each of these basis functions is a piecewise polynomial of degree d. B-spline basis functions are well known for their compact support property, i.e., each basis function is positive over at most $d + 1$ adjacent subintervals. Due to this compact support property, if we approximate $\beta(t)$ by a linear combination of B-spline basis functions, then such approximation is locally sparse if the coefficients are sparse in groups.

We shall further introduce some notations. Let $I_j = (t_{j-1}, t_M)$, and $A_j = \{j, j+1, \ldots, M+d\}$ for $j = 1, \ldots, M$. Intuitively, each group A_j represents the indices of B-spline basis functions that are nonzero on I_j. For a vector $\boldsymbol{b} = (b_1, \ldots, b_{M+d})^{\mathrm{T}}$ of scalars, we denote by $b_{A_j} = \{b_k : k \in A_j\}$ the subvector of elements whose indices are in the j-th group A_j. We shall use $\|\boldsymbol{a}\|_1 = |a_1| + \cdots + |a_q|$ to denote the L_1 norm of a generic q-dimensional vector $\boldsymbol{a}$, and use $\|x\|_2$ to denote the L_2 norm of a generic function $x(t)$. As our focus is on the estimation of $\beta(t)$ and δ, without loss of generality, we assume that $\mu = 0$ in model (2.1) in the sequel.

For a fixed $0 < \gamma < 1$, the historically sparse (zero on the tail region) and smooth estimators for β and δ are defined as

$$\hat{\beta}_n(t) = \hat{\boldsymbol{b}}_n^{\mathrm{T}} \boldsymbol{B}(t), \quad \hat{\delta}_n = t_{J_0 - 1}, \tag{2.2}$$

where $J_0 = \min\{M + 1, \min\{l : \hat{b}_{nk} = 0, \text{for all } k \geq l\}\}$ and $\hat{\boldsymbol{b}}_n = (\hat{b}_{n1}, \ldots, \hat{b}_{nM+d})^{\mathrm{T}}$ minimizes the penalized least squares

$$\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{k=1}^{M+d} b_k \int_0^T X_i(t) B_k(t) \, \mathrm{d}\, t \right)^2 + \kappa \left\| \boldsymbol{b}^{\mathrm{T}} \boldsymbol{B}^{(m)} \right\|_2^2 + \lambda \sum_{j=1}^{M} c_j \left\| b_{A_j} \right\|_1^{\gamma}, \tag{2.3}$$

with known weights c_j and nonnegative tuning parameters κ and λ. In the above criterion, the first term is the ordinary least squares error that encourages the fidelity of model fitting, while the second term is a roughness penalty that aims to enforce smoothness of the estimate $\hat{\beta}_n(t)$. In practice, $m = 2$ is a common choice, which corresponds to measuring the roughness of a function by its integrated curvature.

The last term in the objective function (2.3) is designed to shrink the estimated slope function toward zero specifically on the tail region. It originates from the group bridge penalty that was introduced by Huang et al. (2009) for simultaneous selection of variables at both the group and within-group individual levels. In (2.3), the groups have a special structure: $A_1 \supset \cdots \supset A_M$. In other words, the groups are nested as a sequence and hence we call the last term in (2.3) *nested group bridge*. Due to such nested nature, if $k > j$, then one can observe in (2.3) that (i) the coefficient b_k appears in all groups where the coefficient b_j also appears, and (ii) b_k appears in more groups than b_j. As a consequence, b_k is always penalized more heavily than b_j. These two features suggest that the nested group bridge penalty spends more effort on shrinking those coefficients of B-spline basis functions whose support is in a closer proximity to T. As B-spline basis functions enjoy the aforementioned compact support property and our estimate is represented by a linear combination of such basis functions as in (2.2), the progressive shrinkage of nested group bridge encourages the estimate of $\beta(t)$ to be locally sparse specifically on the tail part of the time domain. Such estimate is exactly what we are after in the scalar-on-function truncated linear model (2.1). The weights c_j are introduced to adjust the number of elements in the set A_j. A simple choice for c_j is $c_j \propto |A_j|^{1-\gamma}$, where $|A_j|$ denotes the cardinality of A_j (Huang et al., 2009). Borrowing the idea of the adaptive lasso (Zou, 2006), we practically choose $c_j = |A_j|^{1-\gamma}/\|b_{A_j}^{(0)}\|_2^\gamma$, where $\boldsymbol{b}^{(0)}$ can be obtained by the penalized B-splines method (Cardot et al., 2003). As Huang et al. (2009) pointed out, when $\gamma = 1$, the group bridge penalty is the lasso penalty and can only do individual variable selection. When $0 < \gamma < 1$, the group bridge penalty can be used for variable selection at the group and with-in group individual levels simultaneously. We also conduct a simulation study to compare the lasso and the nested group bridge penalty; see the appendix for details.

2.2.2 Computational Method

The objective function (2.3) is not convex and thus difficult to optimize. Huang et al. (2009) suggested the following formulation that was easier to work with. Based on Proposition 1 of Huang et al. (2009), for $0 < \gamma < 1$, if $\lambda = \tau^{1-\gamma}\gamma^{-\gamma}(1-\gamma)^{\gamma-1}$, then $\hat{\boldsymbol{b}}_n$ minimizes (2.3) if and only if $(\hat{\boldsymbol{b}}_n, \hat{\boldsymbol{\theta}})$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{k=1}^{M+d} b_k \int_0^T X_i(t)B_k(t)\,\mathrm{d}\,t\right)^2 + \kappa\left\|\boldsymbol{b}^{\mathrm{T}}\boldsymbol{B}^{(m)}\right\|_2^2 + \sum_{j=1}^{M}\theta_j^{1-1/\gamma}c_j^{1/\gamma}\|b_{A_j}\|_1 + \tau\sum_{j=1}^{M}\theta_j, \tag{2.4}$$

subject to $\theta_j \geq 0$ $(j = 1, \ldots, M)$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)^{\mathrm{T}}$ and $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_M)^{\mathrm{T}}$. Below we develop an algorithm following this idea.

Let $\boldsymbol{U}$ denote the $n \times (M+d)$ matrix with elements $u_{ij} = \int_0^T X_i(t)B_j(t)\,\mathrm{d}\,t$, and let $\boldsymbol{V}$ denote the $(M+d)\times(M+d)$ matrix with elements $v_{ij} = \int_0^T B_i^{(m)}(t)B_j^{(m)}(t)\,\mathrm{d}\,t$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, then the first term of (2.4) can be expressed as $1/n\,(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{b})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{b})$ and the second term of (2.4) yields $\kappa\boldsymbol{b}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{b}$. Since $\boldsymbol{V}$ is a positive semidefinite matrix, we write $\boldsymbol{V} = \boldsymbol{W}\boldsymbol{W}$, where $\boldsymbol{W}$ is symmetric. Define

$$\boldsymbol{U}_* = \begin{pmatrix} \boldsymbol{U} \\ \sqrt{n\kappa}\boldsymbol{W} \end{pmatrix} \quad \text{and} \quad \tilde{\boldsymbol{Y}} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{0}$ is the zero vector of length $M + d$. If we write $g_k = \sum_{j=1}^{\min\{k,M\}} \theta_j^{1-1/\gamma}c_j^{1/\gamma}$ for $k = 1, \ldots, M+d$, then (2.4) can be written in the form

$$\frac{1}{n}\left(\tilde{\boldsymbol{Y}} - \boldsymbol{U}_*\boldsymbol{b}\right)^{\mathrm{T}}\left(\tilde{\boldsymbol{Y}} - \boldsymbol{U}_*\boldsymbol{b}\right) + \sum_{k=1}^{M+d} g_k|b_k| + \tau\sum_{j=1}^{M}\theta_j. \tag{2.5}$$

Let $\boldsymbol{G}$ be the $(M+d)\times(M+d)$ diagonal matrix with the ith diagonal element $(ng_i)^{-1}$. With notation $\tilde{\boldsymbol{U}}=\boldsymbol{U}_*\boldsymbol{G}$ and $\tilde{\boldsymbol{b}}=\boldsymbol{G}^{-1}\boldsymbol{b}$, (2.5) can be expressed in a form of lasso problem (Tibshirani, 1996),

$$\frac{1}{n}\left\{\left(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{U}}\tilde{\boldsymbol{b}}\right)^{\mathrm{T}}\left(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{U}}\tilde{\boldsymbol{b}}\right)+\sum_{k=1}^{M+d}|\tilde{b}_k|\right\}+\tau\sum_{j=1}^{M}\theta_j,$$

where $\tilde{b}_k$ denote the kth element of vector $\tilde{\boldsymbol{b}}$. Now, we take the following iterative approach to compute $\hat{\boldsymbol{b}}_n$.

Step 1. Obtain an initial estimate $\boldsymbol{b}^{(0)}$.

Step 2. At iteration s, $s=1,2,\ldots$, compute

$$\begin{aligned}\theta_j^{(s)}=&c_j\left(\frac{1-\gamma}{\tau\gamma}\right)^{\gamma}\|b_{A_j}^{(s-1)}\|_1^{\gamma},\quad j=1,\ldots,M,\\ g_k^{(s)}=&\sum_{j=1}^{\min\{k,M\}}(\theta_j^{(s)})^{1-1/\gamma}c_j^{1/\gamma},\quad k=1,\ldots,M+d,\end{aligned}$$

$$\boldsymbol{G}^{(s)}=n^{-1}\mathrm{diag}\left(1/g_1^{(s)},\ldots,1/g_{M+d}^{(s)}\right),\quad \tilde{\boldsymbol{U}}^{(s)}=\boldsymbol{U}_*\boldsymbol{G}^{(s)}.$$

Step 3. At iteration s, compute

$$\boldsymbol{b}^{(s)}=\boldsymbol{G}^{(s)}\arg\min_{\tilde{\boldsymbol{b}}}\left(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{U}}^{(s)}\tilde{\boldsymbol{b}}\right)^{\mathrm{T}}\left(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{U}}^{(s)}\tilde{\boldsymbol{b}}\right)+\sum_{k=1}^{M+d}|\tilde{b}_k|. \tag{2.6}$$

Step 4. Repeat *Step 2* and *Step 3* until convergence is reached.

A choice for the initial estimate is $\boldsymbol{b}^{(0)}=(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}+n\kappa\boldsymbol{V})^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{Y}$, which is obtained by the penalized B-splines method (Cardot et al., 2003). Once $\hat{\boldsymbol{b}}_n$ is produced, the estimates for β and δ are given in (2.2). As the nested group bridge penalty is not convex, the above algorithm converges to a local minimizer. It is worth emphasizing that (2.6) is a lasso problem, which can be efficiently solved by the least angle regression algorithm (Efron et al., 2004).

In our fitting procedure, there are a few tuning parameters including the smoothing parameter κ, the shrinkage parameter λ, and the parameters for constructing the B-spline basis functions such as the degree d of the B-spline basis and the number of knots $M+1$. Following the schemes of

Marx and Eilers (1999), Cardot et al. (2003) and Lin et al. (2017), we choose M to be relatively large to capture the local features of $\beta(t)$. In addition, δ is estimated by the knot t_{J_0-1}, therefore a small M may lead to a large bias of the estimator $\hat{\delta}_n$. The effect of potential overfitting caused by a large number of knots can be offset by the roughness penalty. Compared to M, the degree d is of less importance, and therefore we fix it to a reasonable value, i.e., $d = 3$.

Once the number of B-spline basis functions is fixed, we can proceed to select the shrinkage parameter λ, as well as the smoothing parameter κ. In Hall and Hooker (2016) where the idea of penalized least squares is also employed, the shrinkage parameter is selected to minimize the mean-squared error of a parametric surrogate estimator of $\beta(t)$. In our case, for a given finite sample, the estimator in (2.2), which is represented by a finite number of B-spline basis functions, serves as such a surrogate. Therefore, we can adopt the same strategy to select λ. Instead of the mean-squared error, we employ the Bayesian information criterion (BIC) to encourage model sparsity, as follows.

Let $\hat{\boldsymbol{b}}_n = \hat{\boldsymbol{b}}_n(\kappa, \lambda)$ be the estimate based on a chosen pair of κ and λ. Let $\boldsymbol{U}_{\kappa,\lambda}$ denote the submatrix of $\boldsymbol{U}$ with columns corresponding to the nonzero $\hat{\boldsymbol{b}}_n(\kappa, \lambda)$, and $\boldsymbol{V}_{\kappa,\lambda}$ denote the submatrix of $\boldsymbol{V}$ with rows and columns corresponding to the nonzero $\hat{\boldsymbol{b}}_n(\kappa, \lambda)$. The approximated degree of freedom for κ and λ is

$$\mathrm{df}(\kappa, \lambda) = \mathrm{trace}\left(\boldsymbol{U}_{\kappa,\lambda}(\boldsymbol{U}_{\kappa,\lambda}^{\mathrm{T}}\boldsymbol{U}_{\kappa,\lambda} + n\kappa\boldsymbol{V}_{\kappa,\lambda})^{-1}\boldsymbol{U}_{\kappa,\lambda}^{\mathrm{T}}\right).$$

Then, Bayesian information criterion (BIC) can be approximated by

$$\mathrm{BIC}(\kappa, \lambda) = n\log(\|\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{b}}_n(\kappa, \lambda)\|_2^2/n) + \log(n)\mathrm{df}(\kappa, \lambda).$$

The optimal κ and λ are selected to minimize $\mathrm{BIC}(\kappa, \lambda)$.

2.3 Asymptotic Properties

Let δ_0 and $\beta_0(t)$ be the true values of the cutoff time δ and the slope function $\beta(t)$, respectively. We assume that realizations $X_1, \ldots, X_n$ are fully observed, while notice that the analysis can be extended to sufficiently densely observed data. Without loss of generality, we assume $T = 1$. If $\delta_0 = 0$, set $J_1 = 0$, and if $\delta_0 = 1$, let $J_1 = M$. Otherwise, let J_1 be an integer such that $\delta_0 \in [t_{J_1-1}, t_{J_1})$.

According to Theorem XII(6) of de Boor (2001), there exists some $\beta_s(t) = \sum_{j=1}^{M+d} b_{sj}B_j(t) = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{b}_s$ with $\boldsymbol{b}_s = (b_{s1}, \dots, b_{sM+d})^{\mathrm{T}}$ with $\inf_j |b_{sj}| \geq C_0' M^{-p_0}$, such that $\|\beta_s - \beta_0\|_\infty \leq C_0 M^{-p_0}$ for some positive constants C_0', C_0 and p_0. More specifically, if $\beta_0(t)$ satisfies condition *C.2*, then $p_0 = k + \nu$. Define $b_{0j} = b_{sj}I_{(j \leq J_1)}$, $j = 1, \dots, M + d$. Define Γ as the covariance operator of the random process X, and Γ_n as the empirical version of Γ, which is defined by

$$(\Gamma_n x)(v) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 X_i(v)X_i(u)x(u)\,\mathrm{d}\,u.$$

For two functions g and f defined on $[0, 1]$, we define the inner product in the Hilbert space L^2 as $\langle g, f\rangle = \int_0^1 g(t)f(t)\,\mathrm{d}\,t$. Let $\boldsymbol{H}$ be the $(M+d)\times(M+d)$ matrix with elements $h_{i,j} = \langle \Gamma_n B_i, B_j\rangle$. In order to establish our asymptotic properties, we assume that the following conditions are satisfied.

C.1 $E\|X\|_2^2 < \infty$.

C.2 The kth derivative $\beta^{(k)}(t)$ exists and satisfies the Hölder condition with exponent ν, that is $|\beta^{(k)}(t') - \beta^{(k)}(t)| \leq c|t' - t|^\nu$, for some constant $c > 0$, $\nu \in (0, 1]$. Define $p = k + \nu$. Assume $3/2 < p \leq d$.

C.3 $M = o(n^{1/2})$, $M = \omega(n^{\frac{1}{2p}})$ and $\kappa = o(n^{-1/2}M^{1/2-2m})$.

C.4 There are constants $C_{max} > C_{min} > 0$ such that

$$C_{min}M^{-1} \leq \rho_{min}(\boldsymbol{H}) \leq \rho_{max}(\boldsymbol{H}) \leq C_{max}M^{-1}$$

with probability tending to one as n goes to infinity, where ρ_{min} and ρ_{max} denote the smallest and largest eigenvalues of a matrix, respectively.

C.5 $\lambda = O(n^{-1/2}M^{-1/2}\eta^{-1})$, where $\eta = \big(\sum_{j=1}^{J_1} c_j^2\|b_{0A_j}\|_1^{2\gamma-2}|A_j|\big)^{1/2}$ with $c_j \propto |A_j|^{1-\gamma}$.

C.6 $\dfrac{\lambda}{M^{1-\gamma}n^{\gamma/2-1}} \to \infty$.

The condition *C.1* assures the existence of the covariance function of X. The second condition concerns the smoothness of the slope function β, which has been used by Cardot et al. (2003) and Lin et al. (2017). In condition *C.3* we set the growth rate for the smoothing tuning parameter

κ. Our analysis applies to $m = 0$, which is equivalent to Tikhonov regularization in Hall and Horowitz (2007) and simplifies our analysis. A similar result can be derived for $m > 0$. The last two conditions together pose certain constraints on the decay rate of λ. Similar conditions appear in Wang and Kai (2015). Here, η is a sequence of constants varying with M and determined by β_0 and γ. It can be shown that, when $\beta_0(t) \neq 0$ for some t, $C_1 M^{1/2} \leq \eta \leq C_2 M^{(2-\gamma)+(1-\gamma)p}$ for constants $C_1, C_2 > 0$, and otherwise $\eta \equiv 0$. These conditions can be realized, for example, by $\lambda \asymp n^{-1/2} M^{\gamma-(1-\gamma)p-5/2}$ and $M \asymp n^{(1-\gamma)/(8-4\gamma+2p-2p\gamma)}$.

Below we state the main results, and relegate their proofs to the appendix. Our first result provides the convergence rate of the estimator $\hat{\beta}_n$ defined in (2.2).

Theorem 1 (Convergence Rate)**.** *Suppose that conditions* C.1–C.6 *hold. Then,* $\|\hat{\beta}_n - \beta_0\|_2 = O_p(Mn^{-1/2} + M^{-p})$.

The convergence rate consists of two competing components, the variance term $Mn^{-1/2}$ and the bias term M^{-p}. With an increase of M, the approximation to $\beta(t)$ by B-spline basis functions is improved, however, at the cost of increased variance.

In addition, we observe that the smoothing parameter κ has negligible impact on the rate of the proposed estimator when its asymptotic rate is bounded by the threshold stated in the condition *C.3*. This is aligned with the classic results for penalized spline estimator (e.g., Claeskens et al., 2009, Theorem 1). Moreover, as the nested group bridge penalty has the effect of shrinkage, it also penalizes the roughness of the estimator. This partially explains why the κ shall be chosen smaller than the one in Claeskens et al. (2009). On the other hand, in practice, as the sample size is often limited, κ plays an important role in regulating the roughness/variability of the estimator, in particular when a large number of B-spline basis functions are required to reduce estimation bias. The next result shows that the null tail of $\beta(t)$, as well as the cutoff time δ, can be consistently estimated.

Theorem 2 (Consistency)**.** *Suppose that conditions* C.1–C.6 *hold.*

(i) For any $\zeta \in (0, 1 - \delta_0)$*,* $\hat{\beta}_n(t) = 0$ *for all* $t \in [\delta_0 + \zeta, 1]$ *with probability tending to* 1.

(ii) $\hat{\delta}_n$ *converges to* δ_0 *in probability.*

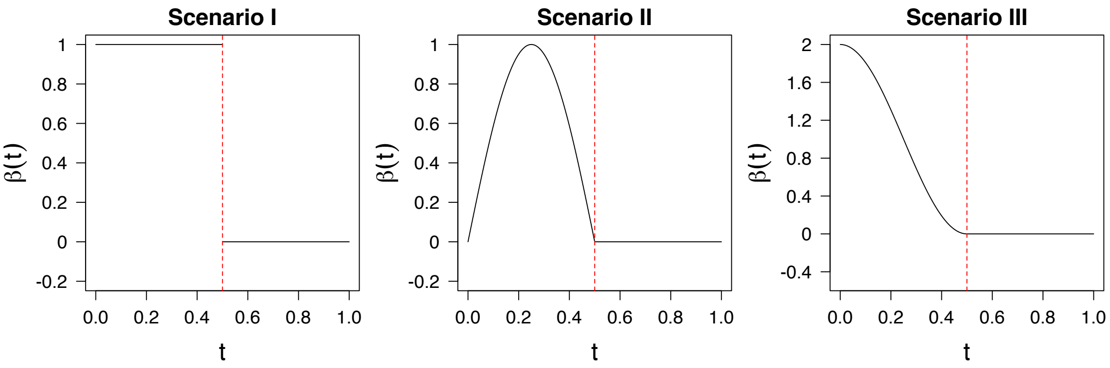

Figure 2.2: The slope functions in three scenarios. The dashed vertical lines indicate the true values of δ.

2.4 Simulation Studies

We conduct simulation studies to evaluate the numerical performance of the proposed nested group bridge method, and compare the results with the penalized B-splines approach (Cardot et al., 2003), the two truncation methods (Hall and Hooker, 2016), and two locally sparse modeling methods, the FLiRTI method (James et al., 2009) and the SLoS method (Lin et al., 2017). The truncation methods first expand the slope function with a sequence of principal component functions and then penalize δ by adding a penalty on δ^2 to the least squares. Two estimation procedures were suggested by Hall and Hooker (2016). The first one (called Method A) estimates δ and $\beta(t)$ simultaneously, while the second one (called Method B) estimates them in an iterative fashion. The FLiRTI method proposed by James et al. (2009) achieves local sparseness by applying variable selection to various derivatives at some discrete grid points. The SLoS method is based on fSCAD, a functional regularization technique.

In our studies, for the purpose of fair comparison, we consider the same scenarios for $\beta(t)$ in Hall and Hooker (2016), namely,

Scenario I. $\beta(t) = I_{(0 \le t < 0.5)}$,

Scenario II. $\beta(t) = \sin(2\pi t) I_{(0 \le t < 0.5)}$,

Scenario III. $\beta(t) = (\cos(2\pi t) + 1)\, I_{(0 \le t < 0.5)}$,

where $I_{(\cdot)}$ denotes the indicator function. For all cases the slope function $\beta(t) > 0$ on $(0, 0.5)$ and $\beta(t) = 0$ on $[0.5, 1]$. As illustrated in Figure 2.2, the slope function is discontinuous for Scenario I, and the first and second derivatives of the slope functions are discontinuous for Scenario II and III, respectively. The predictor functions $X_i(t)$ are generated by $X_i(t) = \sum a_{ij}B_j(t)$, where $B_j(t)$ are cubic B-spline basis functions defined on 64 (the number 64 is randomly selected between 50 and 100) equally spaced knots over $[0, 1]$, and the coefficients a_{ij} are generated independently from the standard normal distribution. The errors ε are normally distributed and sampled so that the signal-to-noise ratio equals to 2. We consider sample sizes $n = 100$ and $n = 500$. For each of the three scenarios and for each sample size, we replicate the simulation independently for 200 times. We also consider smooth functional covariates, which are generated in the same set up, except that the signal-to-noise ratio is 5 and $X_i(t)$ are generated as a linear combinations of 25 Fourier basis functions 1, $\sin(2\pi t)$, $\cos(2\pi t)$, $\ldots$, $\sin(2^{12}\pi t)$, $\cos(2^{12}\pi t)$ defined on $[0, 1]$, with the coefficients corresponding to the jth Fourier basis function generated independently from the normal distribution with mean 0 and variance $1/j^{1.2}$, $j = 1, \ldots, 25$. The results regarding the smooth functional covariates are provided in the appendix.

For the proposed nested group bridge method, the penalized B-splines approach and the SLoS method, we expand the slope function with cubic B-splines with 101 equally spaced knots. For the FLiRTI method, we use cubic B-splines with the number of knots selected according to the model selection method introduced in James et al. (2009). For the proposed nested group bridge method, we follow Huang et al. (2009) and set the group bridge parameter $\gamma = 0.5$ in all numerical studies. We discuss the effect of γ in the appendix. The tuning parameters of the proposed nested group bridge method are chosen by the procedure reported in Section 2.2.2. The smoothing parameter of the penalized B-splines approach is chosen by BIC. For the two truncation methods, the number of empirical principal components is chosen from $2 - 15$ by BIC. The FLiRTI method is implemented by the Dantzig selector (Candes and Tao, 2007). The two truncation methods and the FLiRTI and SLoS estimators are implemented and tuned according to Hall and Hooker (2016), James et al. (2009) and Lin et al. (2017), respectively.

Table 2.1 summarizes the Monte Carlo mean and standard deviation of $\hat{\delta}$. The results suggest that the proposed nested group bridge estimator is more accurate than the other methods in Scenario III when the second derivative of the slope function is discontinuous. In Scenario II when the first derivative of the slope function is discontinuous, the proposed nested group bridge method is comparable to the truncation methods. In Scenario I when the slope function is discontinuous, truncation method A is the most accurate. The FLiRTI and SLoS method do not focus on the tail region and therefore exhibit larger variability. The results for the smooth functional covariates reported in the appendix are similar. The histograms shown in Figure 2.3 provide more details of the performance of our method. They indicate that when $\beta(t)$ is not smooth, the proposed nested group bridge estimator is conservative, in the sense that the estimate $\hat{\delta} > \delta_0$.

To examine the quality of the estimation for $\beta(t)$, we report the mean integrated squared errors of the estimated $\hat{\beta}(t)$ in Table 2.2. It is observed that in general, the proposed nested group bridge estimator outperforms the other methods. The truncation methods do not regularize the roughness of the estimated slope function, which leads to a less favorable performance when the predictor function is relatively rough. The penalized B-splines method, the FLiRTI method and the SLoS method are comparable to the proposed nested group bridge method in terms of the estimation accuracy of $\beta(t)$, but the penalized B-splines method is unable to provide an estimate for δ. The results for the smooth functional covariates are reported in the appendix, which shows that the FLiRTI method does not perform as well as the other methods. To display the results more intuitively, we provide in the appendix the figures that compare the estimated coefficient functions for various methods.

Table 2.1: The mean of estimators for δ based on 200 simulation replications with the corresponding Monte Carlo standard deviation included in parentheses.

	NGR	TR (Method A)	TR (Method B)	FLiRTI	SLoS	True Value
Scenario I						
$n = 100$	0.66 (0.06)	0.48 (0.04)	0.35 (0.07)	0.81 (0.18)	0.69 (0.18)	0.50
$n = 500$	0.65 (0.05)	0.50 (0.02)	0.48 (0.05)	0.83 (0.17)	0.60 (0.09)	0.50
Scenario II						
$n = 100$	0.60 (0.07)	0.41 (0.04)	0.38 (0.06)	0.77 (0.21)	0.61 (0.18)	0.50
$n = 500$	0.59 (0.03)	0.45 (0.02)	0.45 (0.03)	0.71 (0.19)	0.55 (0.08)	0.50
Scenario III						
$n = 100$	0.50 (0.09)	0.31 (0.04)	0.30 (0.03)	0.73 (0.25)	0.55 (0.21)	0.50
$n = 500$	0.51 (0.04)	0.34 (0.03)	0.33 (0.04)	0.72 (0.23)	0.49 (0.08)	0.50

NGR, the proposed nested group bridge method; TR (Method A), the truncation method that estimates δ and $\beta(t)$ simultaneously proposed by Hall and Hooker (2016); TR (Method B), the truncation method that estimates δ and $\beta(t)$ iteratively (Hall and Hooker, 2016); FLiRTI, the FLiRTI method proposed by James et al. (2009); SLoS, the SLoS method proposed by Lin et al. (2017).

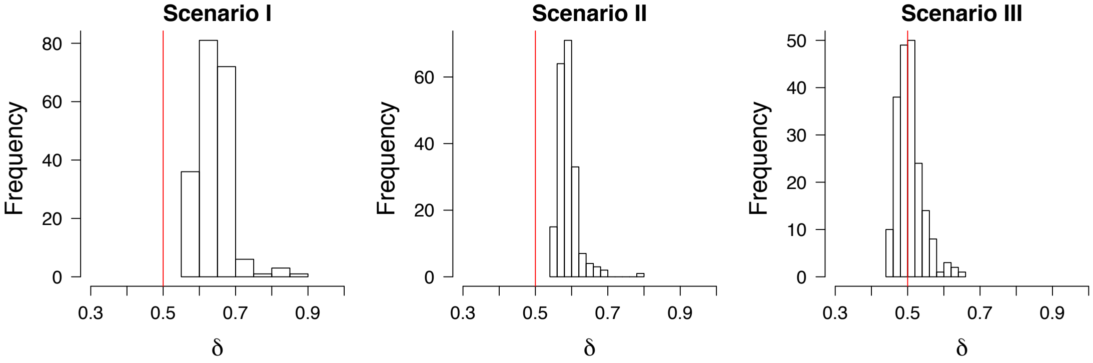

Figure 2.3: Histograms of the estimated $\hat{\delta}$ in 200 simulation replications in the three scenarios. The results were obtained based on 200 Monte Carlo simulations with $n = 500$. The vertical lines indicate the true values of δ.

Table 2.2: Mean integrated squared errors of estimators for $\beta(t)$ based on 200 simulation replications with the corresponding Monte Carlo standard deviation included in parentheses.

	NGR	PS	TR (Method A)	TR (Method B)	FLiRTI	SLoS
Scenario I $(\times 10^{-2})$						
n = 100	2.54 (0.93)	4.57 (1.70)	14.08 (5.13)	28.48 (8.54)	4.97 (2.22)	3.04 (1.21)
n = 500	1.42 (0.38)	1.89 (0.50)	3.34 (1.11)	9.65 (5.17)	1.88 (0.53)	1.38 (0.35)
Scenario II $(\times 10^{-2})$						
n = 100	0.64 (0.44)	1.44 (0.70)	5.69 (2.12)	10.33 (4.43)	1.40 (0.93)	0.95 (0.69)
n = 500	0.21 (0.11)	0.24 (0.15)	1.17 (0.41)	3.08 (1.33)	0.30 (0.16)	0.14 (0.10)
Scenario III $(\times 10^{-2})$						
n = 100	1.36 (1.05)	2.46 (1.50)	14.55 (6.99)	29.68 (13.91)	4.48 (3.16)	1.97 (1.67)
n = 500	0.34 (0.25)	0.64 (0.44)	4.25 (1.44)	11.68 (5.07)	0.87 (0.52)	0.46 (0.33)

NGR, the proposed nested group bridge method; PS, the penalized B-splines method; TR (Method A), the truncation method that estimates δ and $\beta(t)$ simultaneously proposed by Hall and Hooker (2016); TR (Method B), the truncation method that estimates δ and $\beta(t)$ iteratively (Hall and Hooker, 2016); FLiRTI, the FLiRTI method proposed by James et al. (2009); SLoS, the SLoS method proposed by Lin et al. (2017).

2.5 Application: Particulate Matter Emissions Data

In this subsection, we demonstrate the proposed nested group bridge approach to analyze the particulate matter emissions data which are taken from the Coordinating Research Councils E55/E59 research project (Clark et al., 2007). In this project, trucks were placed on the chassis dynamometer bed to mimic inertia and particulate matter was measured by an emission analyzer on standard test cycles. The engine acceleration of diesel trucks was also recorded. We are interested in determining the effects of the past engine acceleration on the current particulate matter emission, and in particular, identifying the cutoff time in the past that has a predicting power on the current particulate matter emission. The problem was originally addressed by Asencio et al. (2014) in their case study. As noted in Hall and Hooker (2016), we obtain observation every 10 second after the first 120 seconds to remove dependences in the data. Let Y_i be the logarithm of the particulate matter emission measured at the i-th 10 second after the first 120 seconds, and $X_i(t), t \in [0, 60]$, be the corresponding engine acceleration at the past time t. Both Y_i and $X_i(t)$ are centered such that $\mathbf{E}Y_i \equiv 0$ and $\mathbf{E}X_i(t) \equiv 0$. We estimate the functional linear model (2.1), where $\mu = 0$, the engine acceleration in the past 60 seconds $X_i(t)$ is the predictor curve, and $T = 60$. In total, we have 108 such samples. Figure 2.4(a) displays 10 randomly selected smoothed engine acceleration curves recorded on every second for 60 seconds.

Figure 2.4(b) provides estimates for $\beta(t)$ obtained by the proposed nested group bridge approach with the group bridge parameter $\gamma = 0.5$ and the penalized B-splines method, respectively, both of which use cubic B-spline basis functions with 121 knots. We choose the number of knots to be equal to the number of time points of the observed acceleration, which is 121. With a sample size 108 and number of knots 121, the roughness penalty plays an important role of reducing the variability of the estimates. The proposed nested group bridge estimate $\hat{\beta}(t)$ is zero over $[20, 60]$ and the estimate for δ is 20s. It suggests that the engine acceleration influences particulate matter emission for no longer than 20 seconds. A similar trend can be observed for the penalized B-splines method which, however, does not give a clear cutoff time of the influence of acceleration on particulate matter emission. Hall and Hooker (2016) suggested that the point estimate for δ is 13s using Method A and 15s using Method B, both of which are more aggressive than our estimator.

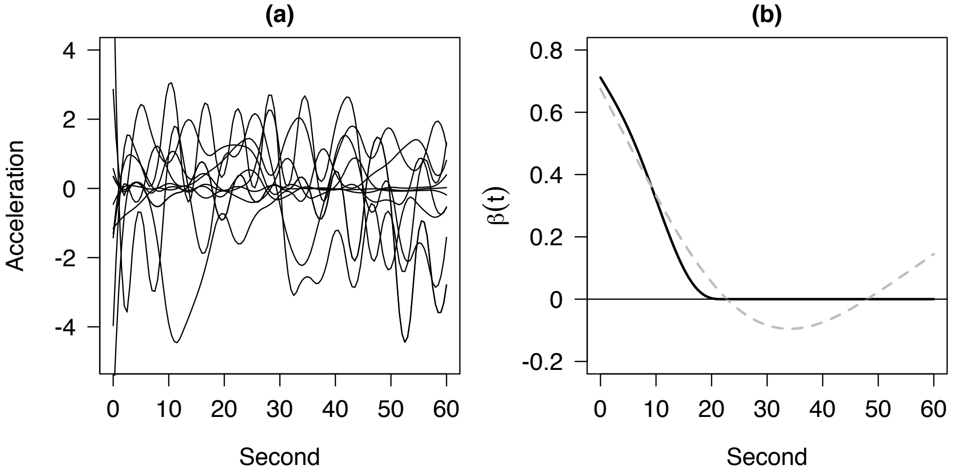

Figure 2.4: (a) 10 randomly selected smoothed acceleration curves. (b) Estimated $\hat{\beta}(t)$ using the penalized B-splines method (dashed line) and the proposed nested group bridge approach (solid line).

2.6 Conclusion and Discussion

In this chapter, we consider to study the relation between a scalar response and a functional predictor in a truncated functional linear model. We propose a nested group bridge approach to achieve the historical sparseness, which reduces the variability and enhances the interpretability. Compared with the truncation methods by Hall and Hooker (2016), the proposed nested group bridge approach is able to provide a smooth and continuous estimate for the coefficient function and performs much better when the coefficient function tends to zero more smoothly. The proposed nested group bridge estimator of the cutoff time enjoys the estimation consistency. We demonstrate in simulation studies and a real data application that the proposed nested group bridge approach performs well for predictor functions that are not very smooth. We also show that even when the signal to noise ratio is low, the proposed nested group bridge approach can still accommodate the situation very well.

The question then arises as to in practice whether to use the proposed nested group bridge method or the truncation methods. We believe it depends on how smoothly the coefficient function tends to zero and how smooth the functional covariates are. Based on our simulation studies, we know that when the coefficient function is discontinuous at the cutoff time, the truncation methods perform better than the proposed nested group bridge method in terms of estimating the cutoff time.

However, for relatively rough functional covariates, the truncation methods estimate the coefficient function less accurately than the proposed method. When the coefficient function goes to zero more smoothly, the proposed nested group bridge method outperforms the truncation methods in both estimating the cutoff time and the coefficient function. In practice, we can first obtain an estimate of $\beta(t)$ using penalized B-splines method. If the estimated $\hat{\beta}(t)$ does not have a steep slope at the tail region, the proposed nested group bridge method is recommended. When the estimated $\hat{\beta}(t)$ goes steeply to the tail region, for more accurate estimate of the cutoff time, the truncation methods should be applied. However, if the functional covariates are relatively rough, the proposed nested group bridge method provides more accurate estimate for the coefficient function.

2.7 Appendix

2.7.1 Penalized B-Splines Method

Let $\mathcal{S}_{dM}$ be the linear space spanned by the B-spline basis functions $\{B_k(t) : k = 1, \ldots, M+d\}$ with degree d and $M + 1$ equally spaced knots defined on $[0, T]$. The penalized B-splines estimator of $\beta(t)$ proposed by Cardot et al. (2003) is the one in $\mathcal{S}_{dM}$ which is defined as

$$\hat{\beta}_{PS}(t) = \sum_{k=1}^{M+d} \hat{b}_k B_k(t) = \hat{\boldsymbol{b}}^{\mathrm{T}} \boldsymbol{B}(t)$$

where $\hat{\boldsymbol{b}}$ minimizes the penalized least squares

$$\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{k=1}^{M+d} b_k \int_0^T X_i(t) B_k(t) \,\mathrm{d}\, t \right)^2 + \kappa \left\| \boldsymbol{b}^{\mathrm{T}} \boldsymbol{B}^{(m)} \right\|_2^2,$$

with smoothing parameter $\kappa > 0$. The tuning parameter κ can be chosen by cross validation, AIC or BIC.

2.7.2 Effect of the Group Bridge Parameter γ in Section 2.4

We conduct a simulation study to numerically investigate the effect of the group bridge parameter γ. The setting is the same as Scenario III with the functional covariates generated by a linear combination of B-spline basis functions, the signal-to-noise ratio 2 and the sample size $n = 100$.

We can observe that the results are similar when $0 < \gamma < 1$, and they are better than the results based on $\gamma = 1$.

Table 2.3: Investigation of effect of the group bridge parameter γ on $\hat{\delta}$ and mean integrated squared errors (ISE) of estimators for $\beta(t)$. The results are obtained based on 200 simulation replications with the corresponding Monte Carlo standard deviations included in parentheses.

γ	$\hat{\delta}$	True δ	ISE ($\times 10^{-2}$)
0.2	0.48 (0.11)	0.50	1.50 (1.32)
0.5	0.50 (0.09)	0.50	1.36 (1.05)
0.8	0.51 (0.06)	0.50	1.38 (1.03)
1	0.66 (0.15)	0.50	1.54 (1.21)

2.7.3 Additional Simulation Results in Section 2.4

In Table 2.4 and 2.5, we display the simulation results for the smooth functional covariates discussed in Section 2.4. We compare the estimated coefficient curves for various methods with rough functional covariates in Figure 2.5 and smooth functional covariates in Figure 2.6.

Table 2.4: The mean of estimators for δ based on 200 simulation replications with the corresponding Monte Carlo standard deviation included in parentheses.

	NGR	TR (Method A)	TR (Method B)	FLiRTI	SLoS	True Value
Scenario I						
$n = 100$	0.64 (0.07)	0.46 (0.06)	0.50 (0.09)	0.59 (0.13)	0.59 (0.16)	0.50
$n = 500$	0.63 (0.04)	0.49 (0.03)	0.52 (0.05)	0.69 (0.19)	0.61 (0.08)	0.50
Scenario II						
$n = 100$	0.56 (0.06)	0.41 (0.05)	0.42 (0.06)	0.56 (0.16)	0.53 (0.16)	0.50
$n = 500$	0.55 (0.03)	0.43 (0.02)	0.45 (0.04)	0.56 (0.14)	0.55 (0.06)	0.50
Scenario III						
$n = 100$	0.49 (0.07)	0.31 (0.03)	0.35 (0.09)	0.55 (0.20)	0.48 (0.11)	0.50
$n = 500$	0.49 (0.03)	0.30 (0.01)	0.39 (0.07)	0.58 (0.18)	0.50 (0.08)	0.50

NGR, our proposed nested group bridge method; TR (Method A), the truncation method that estimates δ and $\beta(t)$ simultaneously proposed by Hall and Hooker (2016); TR (Method B), the truncation method that estimates δ and $\beta(t)$ iteratively (Hall and Hooker, 2016); FLiRTI, the FLiRTI method proposed by James et al. (2009); SLoS, the SLoS method proposed by Lin et al. (2017).

Table 2.5: Mean integrated squared errors of estimators for $\beta(t)$ based on 200 simulation replications with the corresponding Monte Carlo standard deviation included in parentheses.

	NGR	PS	TR (Method A)	TR (Method B)	FLiRTI	SLoS
Scenario I						
n = 100	0.06 (0.06)	0.08 (0.03)	0.08 (0.15)	0.07 (0.05)	0.50 (0.30)	0.20 (0.27)
n = 500	0.03 (0.01)	0.04 (0.01)	0.02 (0.02)	0.04 (0.01)	0.20 (0.19)	0.03 (0.02)
Scenario II						
n = 100	0.02 (0.04)	0.05 (0.02)	0.04 (0.03)	0.03 (0.02)	0.13 (0.11)	0.07 (0.10)
n = 500	0.01 (0.00)	0.02 (0.02)	0.03 (0.00)	0.01 (0.01)	0.02 (0.04)	0.00 (0.00)
Scenario III						
n = 100	0.03 (0.05)	0.04 (0.02)	0.10 (0.02)	0.08 (0.05)	0.50 (0.50)	0.03 (0.05)
n = 500	0.01 (0.01)	0.01 (0.01)	0.09 (0.01)	0.04 (0.01)	0.15 (0.18)	0.01 (0.01)

NGR, our proposed nested group bridge method; PS, the penalized B-splines method; TR (Method A), the truncation method that estimates δ and $\beta(t)$ simultaneously proposed by Hall and Hooker (2016); TR (Method B), the truncation method that estimates δ and $\beta(t)$ iteratively (Hall and Hooker, 2016); FLiRTI, the FLiRTI method proposed by James et al. (2009); SLoS, the SLoS method proposed by Lin et al. (2017).

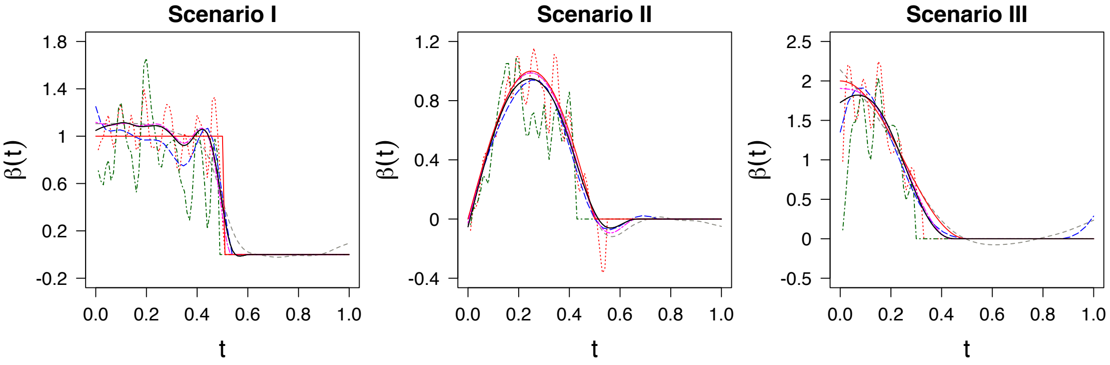

Figure 2.5: Estimated coefficient functions with rough functional covariates and $n = 500$ in one randomly selected simulation replicate for various methods (———, the proposed nested group bridge method; - - - - -, the penalized B-splines method; ··········, the truncation method A; -·-·-·-, the truncation method B; — — —, the FLiRTI method; –·–·–, the SLoS method; ———, the true $\beta(t)$).

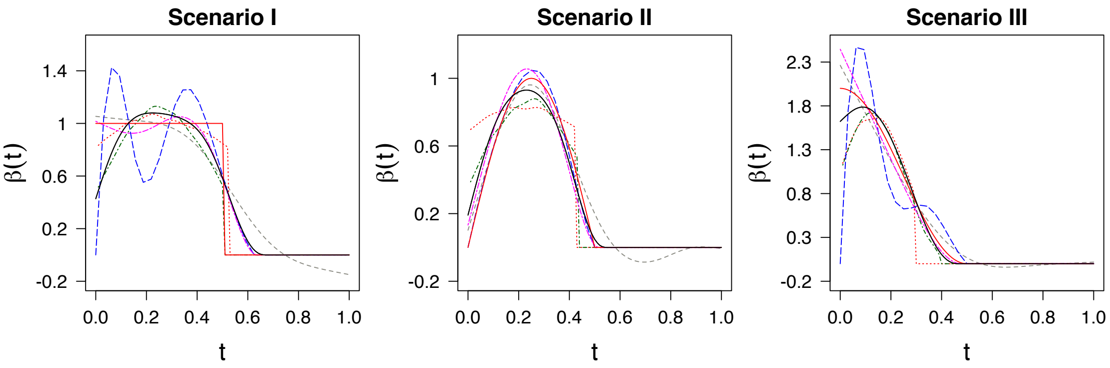

Figure 2.6: Estimated coefficient functions with smooth functional covariates and $n = 500$ in one randomly selected simulation replicate for various methods (———, the proposed nested group bridge method; - - - - -, the penalized B-splines method; ··········, the truncation method A; -·-·-·-, the truncation method B; — — —, the FLiRTI method; -·—·-, the SLoS method; ———, the true $\beta(t)$).

2.7.4 Proofs in in Section 2.3

Without of loss of generality, we assume that $T = 1$. We will first collect some remarks on notations that are used in the sequel. Boldface symbol is used to denote matrix or vector. If $0 < q < \infty$, L^q is defined as the space of functions $f(t)$ over the interval $[0, 1]$ such that $\int_0^1 |f(t)|^q \,\mathrm{d}\,t < \infty$. Two functions $g(t)$ and $f(t)$ are identified as the same if $g(t) = f(t)$ almost everywhere over $[0, 1]$ with respect to the usual Lebesgue measure. With this convention, L^q is treated as a Banach space with the norm $\|f\|_q = (\int_0^1 |f(t)|^q \,\mathrm{d}\,t)^{1/q}$. When $q = 2$, we get the Hilbert space L^2 with the inner product $\langle g, f\rangle = \int_0^1 g(t)f(t)\,\mathrm{d}\,t$ and the L^2 norm $\|\cdot\|_2$. Since $\mathbb{R}^m$ for a positive integer m is also a Hilbert space, we use the same notation $\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \boldsymbol{u}^{'}\boldsymbol{v}$ and $\|\boldsymbol{u}\|_2 = (\boldsymbol{u}^{'}\boldsymbol{u})^{1/2}$ to denote the inner product and the norm of vector $\boldsymbol{u}$ and $\boldsymbol{v}$. Here, $\boldsymbol{u}^{'}$ is used to denote the transpose of $\boldsymbol{u}$. To reduce notational burden and make our presentation concise, we use $\langle f, \boldsymbol{B}\rangle$ to denote the vector $(\langle f, B_1\rangle, \langle f, B_2\rangle, \ldots, \langle f, B_M\rangle)$. The supremum norm of a function $f(t)$ is conventionally denoted by $\|f\|_\infty$ and defined as $\|f\|_\infty = \sup\{|f(t)| : t \in [0, 1]\}$. Similarly, the supremum norm of a vector $\boldsymbol{u} = (u_1, u_2, \ldots, u_m) \in \mathbb{R}^m$ is also denoted by $\|\boldsymbol{u}\|_\infty$ and defined as $\|\boldsymbol{u}\|_\infty = \max\{|u_i| : i = 1, 2, \ldots, m\}$. The operator norm of a linear operator Λ on a Hilbert space $\mathcal{H}$, is traditionally denoted by $\|\Lambda\|$ and defined as $\|\Lambda\| = \sup\{\|\Lambda f\|_2 : f \in \mathcal{H}, \|f\|_2 = 1\}$. Here, $\mathcal{H}$ could be L^2 or $\mathbb{R}^m$.

As our estimator is based on B-spline basis, before we dive into proofs of the theorems, we briefly discuss some basic properties of B-spline basis that are used in our proofs. A detailed treatment of B-spline can be found in de Boor (2001). The B-spline basis has a local support property, which means each B-spline basis function is nonzero over no more than $d + 1$ adjacent subintervals. Also, each B-spline basis function is non-negative and they form a partition of unity, that is, $\sum_{j=1}^{M+d} B_j(t) = 1$ for all $t \in [0, 1]$. Assume $\beta_0(t)$ satisfies condition *C.2*, according to Theorem XII(6) of de Boor (2001), there exists some $\beta_s(t) = \sum\limits_{j=1}^{M+d} b_{sj}B_j(t) = \boldsymbol{B}^{'}\boldsymbol{b}_s$ with $\boldsymbol{b}_s = (b_{s1}, \dots, b_{s,M+d})^{'}$, such that $\|\beta_s - \beta_0\|_\infty \le C_0M^{-p}$ for some positive constant C_0 and p. Define $b_{0j} = b_{sj}I_{(j\le J_1)}$ and put $\beta_{0s}(t) = \sum\limits_{j=1}^{M+d} b_{0j}B_j(t) = \boldsymbol{B}^{'}\boldsymbol{b}_0$, where $\boldsymbol{b}_0 = (b_{01}, \dots, b_{0,M+d})^{'}$. It is easy to see that $b_{0j} = 0$ if the support of $B_j(t)$ is contained in $[\delta_0, 1]$. It is obvious that $\|\beta_{0s} - \beta_0\|_\infty \le C_1M^{-p}$ for some positive constant C_1.

The following lemmas are established to prove the theorems in Section 2.3.

Lemma 1. *If* C.1 *and* C.3 *hold, then for some positive constants C_{ρ_1} and C_{ρ_2},*

$$P(C_{\rho_1}\kappa n/M < \rho_{\min}(\boldsymbol{U}^{'}\boldsymbol{U} + n\kappa\boldsymbol{V}) \le \rho_{\max}(\boldsymbol{U}^{'}\boldsymbol{U} + n\kappa\boldsymbol{V}) < C_{\rho_2}n/M) \to 1, \tag{2.7}$$

where $\rho_{\min}$ and $\rho_{\max}$ denote the smallest and largest eigenvalues of a matrix, respectively.

Proof. This is a consequence of Lemma 6.1 and 6.2 of Cardot et al. (2003). □

Lemma 2. $\sup_{i,j} |v_{ij}| = O(M^{2m-1})$.

Proof. Let B_{jd} denote the jth normalized B-spline defined on $[0, 1]$ with degree d and $M + 1$ equispaced knots $0 = t_0 < t_1 < ... < t_M = 1$, $j = 1, ..., M + d$. The knots divide $[0, 1]$ into M subintervals with equal length $\Delta = 1/M$. Now consider $B_{d+1,d}$, $B_{d+2,d}$, ... , and $B_{M,d}$ that are positive on $d + 1$ such subintervals. Let $B^{'}_{jd}$ and $B^{''}_{jd}$ denote the first and second derivatives of B_{jd} respectively. Then it follows from X(8) of de Boor (2001) that

$$\begin{aligned} B^{'}_{jd}(t) &= \frac{1}{\Delta}(B_{j-1,d-1}(t) - B_{j,d-1}(t)) \\ &= M(B_{j-1,d-1}(t) - B_{j,d-1}(t)), \quad j = d + 1, ..., M, \end{aligned} \tag{2.8}$$

Since $0 \leq B_j(t) \leq 1$, $|B_{jd}^{'}| \leq M$. By taking derivative of (2.8), we have

$$\begin{aligned} B_{jd}^{''}(t) &= \frac{1}{\Delta}(B_{j-1,d-1}^{'}(t) - B_{j,d-1}^{'}(t)) \\ &= \frac{1}{\Delta^2}(B_{j-2,d-2}(t) - 2B_{j-1,d-2}(t) + B_{j,d-2}(t)) \\ &= M^2(B_{j-2,d-2}(t) - 2B_{j-1,d-2}(t) + B_{j,d-2}(t)). \end{aligned}$$

and hence $|B_{jd}^{''}| \leq 2M^2$. Then we can deduce that $|B_{jd}^{(m)}| \leq C_m M^m$, where C_m is some constant depending on m. Since $|B_{jd}^{(m)}| \geq 0$ on at most $d+1$ subintervals, $\left\| B_{jd}^{(m)} \right\|_2 \leq 2C_m(d+1)^{1/2}\Delta^{1/2}M^m$. This further implies that

$$\sup_{i,j} |v_{ij}| = \sup_{i,j} |\langle B_{id}^{(m)}, B_{jd}^{(m)} \rangle| \leq \sup_{i,j} \left\| B_{id}^{(m)} \right\|_2 \left\| B_{jd}^{(m)} \right\|_2 \leq 4C_m^2(d+1)M^{2m-1}, \tag{2.9}$$

which yields the conclusion of the lemma. □

Let $\ell(\boldsymbol{b}) = n^{-1}(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{b})^{'}(\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{b}) + \kappa \boldsymbol{b}^{'}\boldsymbol{V}\boldsymbol{b}$. We can write $\ell(\boldsymbol{b})$ as

$$\begin{aligned} \ell(\boldsymbol{b}) &= \frac{1}{n}\sum_{i=1}^{n} \Big(\langle \beta, X_i \rangle - \langle \boldsymbol{B}^{'}\boldsymbol{b}, X_i \rangle + \varepsilon_i \Big)^2 + \kappa \boldsymbol{b}^{'}\boldsymbol{V}\boldsymbol{b} \\ &= \frac{1}{n}\sum_{i=1}^{n} \Big(\langle \beta - \boldsymbol{B}^{'}\boldsymbol{b}, X_i \rangle^2 + 2\varepsilon_i \langle \beta - \boldsymbol{B}^{'}\boldsymbol{b}, X_i \rangle + \varepsilon_i^2 \Big) + \kappa \boldsymbol{b}^{'}\boldsymbol{V}\boldsymbol{b} \\ &= \langle \Gamma_n(\beta - \boldsymbol{B}^{'}\boldsymbol{b}), \beta - \boldsymbol{B}^{'}\boldsymbol{b} \rangle + \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \langle \beta - \boldsymbol{B}^{'}\boldsymbol{b}, X_i \rangle + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 + \kappa \boldsymbol{b}^{'}\boldsymbol{V}\boldsymbol{b}, \end{aligned}$$

where Γ_n is the empirical version of the covariance operator Γ of X, and is defined by

$$(\Gamma_n x)(v) = \frac{1}{n}\sum_{i=1}^{n} \int_0^1 X_i(v)X_i(u)x(u)\,\mathrm{d}\,u.$$

Let $\boldsymbol{H}$ be the $(M+d) \times (M+d)$ matrix with element $h_{i,j} = \langle \Gamma_n B_i, B_j \rangle$. Then the gradient of ℓ with respect to $\boldsymbol{b}$ is

$$\nabla \ell(\boldsymbol{b}) = 2\boldsymbol{H}\boldsymbol{b} - 2\langle \Gamma_n \beta, \boldsymbol{B}\rangle - \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \langle X_i, \boldsymbol{B}\rangle + 2\kappa \boldsymbol{V}\boldsymbol{b}$$

and the Hessian is

$$\nabla^2 \ell(\boldsymbol{b}) = 2\boldsymbol{H} + 2\kappa \boldsymbol{V}.$$

At the point $\boldsymbol{b} = \boldsymbol{b}_0$, the gradient of ℓ can be written as

$$\begin{aligned}
\nabla \ell(\boldsymbol{b}_0) &= 2\boldsymbol{H}\boldsymbol{b}_0 - 2\langle \Gamma_n \beta, \boldsymbol{B}\rangle - \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \langle X_i, \boldsymbol{B}\rangle + 2\kappa \boldsymbol{V}\boldsymbol{b}_0 \\
&= 2\langle \Gamma_n \boldsymbol{B}, \beta_{0s}\rangle - 2\langle \Gamma_n \beta, \boldsymbol{B}\rangle - \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \langle X_i, \boldsymbol{B}\rangle + 2\kappa \boldsymbol{V}\boldsymbol{b}_0 \\
&= 2\langle \Gamma_n \beta_{0s}, \boldsymbol{B}\rangle - 2\langle \Gamma_n \beta, \boldsymbol{B}\rangle - \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \langle X_i, \boldsymbol{B}\rangle + 2\kappa \boldsymbol{V}\boldsymbol{b}_0 \\
&= 2\langle \Gamma_n (\beta_{0s} - \beta), \boldsymbol{B}\rangle - 2\langle \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i X_i, \boldsymbol{B}\rangle + 2\kappa \boldsymbol{V}\boldsymbol{b}_0.
\end{aligned}$$

In other words, for each $j = 1, 2, \ldots, M + d$,

$$\frac{\partial \ell(\boldsymbol{b}_0)}{\partial b_{0j}} = 2\langle \Gamma_n (\beta_{0s} - \beta), B_j\rangle - 2\langle \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i X_i, B_j\rangle + 2\kappa V_j \boldsymbol{b}_0, \tag{2.10}$$

where V_j represents the jth row of $\boldsymbol{V}$.

Below we first provide bounds of $\frac{\partial \ell(\boldsymbol{b}_0)}{\partial b_{0j}}$ and $\nabla^2 \ell(\boldsymbol{b})$ in Lemma 3 and 4, respectively.

Lemma 3. *For any $\epsilon > 0$, there exists a constant C_2 such that*

$$P\left(\sup_j \left|\frac{\partial \ell(\boldsymbol{b}_0)}{\partial b_{0j}}\right| \le C_2 M^{-1/2} n^{-1/2}\right) > 1 - \epsilon \tag{2.11}$$

holds for all sufficiently large M and n.

Proof. Below we will develop bounds for each term in (2.10). In the first term, Γ_n converges to Γ almost surely according to Proposition 1 in Dauxois et al. (1982). Thus, $\|\Gamma_n\|$ converges to $\|\Gamma\|$ almost surely, since the operator norm is continuous. Recall that, the function β_{0s} is chosen to satisfy $\|\beta_{0s} - \beta\|_\infty \le C_1 M^{-p}$, where C_1 is a positive constant. This implies that $\|\beta_{0s} - \beta\|_2 =$

$\sqrt{\int_0^1 (\beta_{0s}(t) - \beta(t))^2 \,\mathrm{d}\,t} \leq \sqrt{\int_0^1 (C_1 M^{-p})^2 \,\mathrm{d}\,t} = C_1 M^{-p}$. We know that each B-spline basis function is nonzero over no more than $d+1$ adjacent subintervals. Also, each B-spline basis function is non-negative and the basis functions form a partition of unity. The two properties together imply that

$$\|B_j\|_2^2 = \int_0^1 B_j^2(t) \,\mathrm{d}\,t \leq (d+1)M^{-1}. \tag{2.12}$$

Applying Cauchy-Schwarz inequality and (2.12) yields

$$\sup_j |\langle \Gamma_n(\beta_{0s} - \beta), B_j \rangle| \leq \sup_j \|\Gamma_n\| \, \|\beta_{0s} - \beta\|_2 \, \|B_j\|_2 \leq C_1 (d+1)^{1/2} M^{-p-1/2} \|\Gamma_n\|.$$

Since $\|\Gamma_n\|$ converges to $\|\Gamma\|$ almost surely and hence in probability, we conclude that, for any given $\epsilon > 0$, there is a constant $\rho_1(\epsilon)$ depending on ϵ, such that for all sufficiently large n and M,

$$P\left(\sup_j |\langle \Gamma_n(\beta_{0s} - \beta), B_j \rangle| \leq \rho_1(\epsilon) M^{-p-1/2} \right) > 1 - \epsilon. \tag{2.13}$$

For the second term, by Condition *C.1* and CLT (Aldous, 1976), $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i\right)$ converges to a Gaussian random element in $L^2([0,1])$ in distribution, whose mean is 0. This implies that, for any given $\epsilon > 0$, there is a constant $\rho_2(\epsilon)$ which only depends on ϵ, such that for sufficiently large n,

$$P\left(\sqrt{n} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|_2 < \rho_2(\epsilon) \right) > 1 - \epsilon.$$

Each X_i is a mapping from sample space Ω to the space $L^2([0,1])$. Specifically, we let $X_i^\omega \in L^2([0,1])$ be the image of the sample $\omega \in \Omega$ under the mapping X_i. We then denote $\Omega_\epsilon \subset \Omega$ the set of ω that makes $\left\| \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i^\omega \right\|_2 < \rho_2(\epsilon) n^{-1/2}$ hold. Thus, $P(\Omega_\epsilon) > 1 - \epsilon$. Then by Cauchy-Schwarz inequality, on Ω_ϵ, it holds

$$\sup_j \left| \langle \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i^\omega, B_j \rangle \right| \leq \sup_j \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i^\omega \right\|_2 \|B_j\|_2 \leq \rho_2(\epsilon)(d+1)^{1/2} n^{-1/2} M^{-1/2}. \tag{2.14}$$

For the third term, according to lemma 2, we first have $\sup_{ij} |v_{ij}| \leq 4C_m^2(d+1)M^{2m-1}$ and the number of non-zero elements in each row of $\boldsymbol{V}$ is at most $2d+1$. For $\boldsymbol{b}_0$, we have $\|\boldsymbol{b}_0\|_\infty \leq C_3$ for some positive constant C_3 that does not depend on M. This conclusion can be derived from the fact that $\beta_{0s}(t) = \boldsymbol{B}^{'}(t)\boldsymbol{b}_0$, continuity of $\beta(t)$ and the discussion on page 145-149 of de Boor (2001). Given these bounds, we have $\sup_j |V_j \boldsymbol{b}_0| \leq 4C_3 C_m^2 (d+1)(2d+1)M^{2m-1}$. Combining this result with (2.13) and (2.14), and using Condition *C.3*, we deduce (2.11). □

We write $u = \Theta_P(v)$ if u/v is bounded away from 0 and ∞ with probability tending to one. The next result concerns the order of $\nabla^2 \ell(\boldsymbol{b})$.

Lemma 4. *$\nabla^2 \ell(\boldsymbol{b})$ is positive-definite and $\|\nabla^2 \ell(\boldsymbol{b})\| = \Theta_P(M^{-1})$. Furthermore, $\sup_{ij} |\frac{\partial^2 \ell(\boldsymbol{b})}{\partial b_i \partial b_j}| = O_P(M^{-1})$.*

Proof. By Condition *C.4*, $\rho_{\min}(\boldsymbol{H}) = \Theta_P(M^{-1})$ and $\rho_{\max}(\boldsymbol{H}) = \Theta_P(M^{-1})$. Since $\rho_{\max}(\kappa \boldsymbol{V}) = O(\kappa M^{-1})$ and $\kappa = o(1)$ according to the condition *C.3*, we then have $\|\nabla^2 \ell(\boldsymbol{b})\| = \Theta_P(M^{-1})$.

By Cauchy-Schwarz inequality and (2.12), we have

$$\sup_{ij} |h_{ij}| = \sup_{ij} |\langle \Gamma_n B_i, B_j \rangle| \leq \|\Gamma_n\| \|B_i\|_2 \|B_j\|_2 = O_P(M^{-1}).$$

Combining this result with (2.9), and using the Condition *C.3*, we conclude that $\sup_{ij} |\frac{\partial^2 \ell(\boldsymbol{b})}{\partial b_i \partial b_j}| = O_P(M^{-1})$. □

Lemma 5. *Suppose* C.1 - C.5 *hold. Then $\|\hat{\boldsymbol{b}}_n - \boldsymbol{b}_0\|_2 = O_p(Mn^{-1/2})$.*

Proof. Let $\hat{\boldsymbol{b}}_n = \boldsymbol{b}_0 + \delta_n \boldsymbol{u}$ with $\|\boldsymbol{u}\| = 1$. Therefore, it is sufficient to show that $\delta_n = O_p(Mn^{-1/2})$.

Denote $D(\boldsymbol{u}) = Q(\boldsymbol{b}_0 + \delta_n \boldsymbol{u}) - Q(\boldsymbol{b}_0)$, where Q is the objective function (2.3). Then $D(\boldsymbol{u})$ is the sum of $D_1 \equiv \ell(\boldsymbol{b}_0 + \delta_n \boldsymbol{u}) - \ell(\boldsymbol{u})$ and $D_2 \equiv \lambda \sum_{j=1}^M c_j \|\hat{b}_{nA_j}\|_1^\gamma - \lambda \sum_{j=1}^M c_j \|b_{0A_j}\|_1^\gamma$. For D_1, according to Lemma 3 and Lemma 4, also noting that it is a quadratic function of $\boldsymbol{b}$, by Taylor expansion, we can show that

$$\begin{aligned} D_1 &= \ell(\boldsymbol{b}_0 + \delta_n \boldsymbol{u}) - \ell(\boldsymbol{b_0}) \\ &= \delta_n \nabla^{'} \ell(\boldsymbol{b}_0)\boldsymbol{u} + \frac{1}{2}\delta_n^2 \boldsymbol{u}^{'} \nabla^2 \ell(\boldsymbol{b}_0)\boldsymbol{u} \\ &\geq \delta_n O_P(M^{-1/2}n^{-1/2}) + C_4 \delta_n^2 M^{-1} \end{aligned}$$

for some constant $C_4 > 0$ with probability tending to one. Since $b^\gamma - a^\gamma \le 2(b-a)b^{\gamma-1}$ for $0 \le a \le b$, we have

$$\begin{aligned}
-D_2 &\le 2\lambda \sum_{j=1}^{M} c_j \|b_{0A_j}\|_1^{\gamma-1} \left(\|b_{0A_j}\|_1 - \|\hat{b}_{nA_j}\|_1\right) \\
&\le 2\lambda \sum_{j=1}^{M} c_j \|b_{0A_j}\|_1^{\gamma-1} \left(|A_j| \|b_{0A_j} - \hat{b}_{nA_j}\|_2^2\right) \\
&\le 2\lambda\eta \left(\sum_{j=1}^{M} \|b_{0A_j} - \hat{b}_{nA_j}\|_2^2\right)^{1/2} \\
&\le 2\lambda\eta\delta_n M^{1/2}.
\end{aligned}$$

Since $\hat{\boldsymbol{b}}_n$ minimizes $Q(\boldsymbol{b})$, we have $D_1 + D_2 \le 0$. According to condition *C.5*, $\lambda\eta = O(n^{-1/2}M^{-1/2})$. Then

$$\delta_n O_P(M^{-1/2}n^{-1/2}) + C_4\delta_n^2 M^{-1} - O(\delta_n n^{-1/2}) \le 0$$

with probability tending to one, which implies that $\delta_n = O_p(Mn^{-1/2})$. Thus the conclusion of the lemma follows. □

Lemma 6. *Suppose conditions* C.1 - C.6 *hold. Then* $P\left(\hat{b}_{nA_j} = 0 \text{ for } j > J_1\right) \to 1$.

Proof. Define $\tilde{\boldsymbol{b}}_n = (\tilde{b}_{n1}, ..., \tilde{b}_{n,M+d})'$ by $\tilde{b}_{nk} = \hat{b}_{nk} I_{(k \le J_1)}$, $k = 1, ..., M+d$. We have $\hat{\theta}_j^{1-1/\gamma} = \lambda\gamma c_j^{1-1/\gamma} \|\hat{b}_{nA_j}\|_1^{\gamma-1}$. The Karush-Kuhn-Tucker condition for (2.6) implies

$$2(\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{b}}_n)' U_k - 2n\kappa\hat{\boldsymbol{b}}_n' V_k = \sum_{j=1}^{\min\{k,M\}} n\hat{\theta}_j^{1-1/\gamma} c_j^{1/\gamma} \operatorname{sgn}(\hat{b}_{nk}), \quad \hat{b}_{nk} \ne 0,$$

where U_k is the kth column of $\boldsymbol{U}$ and V_k is the kth column of $\boldsymbol{V}$. Observe that $\|\hat{b}_{nA_j}\|_1 - \|\tilde{b}_{nA_j}\|_1 = \sum_{k=\max\{j,J_1+1\}}^{M+d} |\hat{b}_{nk}|$ and $(\hat{b}_{nk} - \tilde{b}_{nk})\operatorname{sgn}(\hat{b}_{nk}) = |\hat{b}_{nk}| I_{(k \ge J_1+1)}$. Thus

$$\begin{aligned}
2(\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{b}}_n)' \boldsymbol{U}(\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n) &= 2n\kappa \sum_{k=1}^{M+d} (\hat{b}_{nk} - \tilde{b}_{nk}) \hat{\boldsymbol{b}}_n' V_k + n\lambda\gamma \sum_{j=1}^{M} \sum_{k=\max\{j,J_1+1\}}^{M+d} c_j \|\hat{b}_{nA_j}\|_1^{\gamma-1} |\hat{b}_{nk}| \\
&= 2n\kappa \sum_{k=1}^{M+d} (\hat{b}_{nk} - \tilde{b}_{nk}) \hat{\boldsymbol{b}}_n' V_k + n\lambda\gamma \sum_{j=1}^{M} c_j \|\hat{b}_{nA_j}\|_1^{\gamma-1} (\|\hat{b}_{nA_j}\|_1 - \|\tilde{b}_{nA_j}\|_1).
\end{aligned}$$

Since $\gamma b^{\gamma-1}(b-a) \le b^{\gamma} - a^{\gamma}$ for $0 \le a \le b$, for $j \le J_1$, we have

$$\gamma\|\hat{b}_{nA_j}\|_1^{\gamma-1}(\|\hat{b}_{nA_j}\|_1 - \|\tilde{b}_{nA_j}\|_1) \le \|\hat{b}_{nA_j}\|_1^{\gamma} - \|\tilde{b}_{nA_j}\|_1^{\gamma},$$

Observe that $\|\tilde{b}_{nA_j}\|_1 = 0$ for $j > J_1$. Thus

$$\begin{aligned}
&2(\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{b}}_n)'\boldsymbol{U}(\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n) \\
&\le 2n\kappa \sum_{k=1}^{M+d} (\hat{b}_{nk} - \tilde{b}_{nk})\hat{\boldsymbol{b}}_n' V_k + n\lambda \sum_{j=1}^{J_1} c_j(\|\hat{b}_{nA_j}\|_1^{\gamma} - \|\tilde{b}_{nA_j}\|_1^{\gamma}) + n\lambda\gamma \sum_{j=J_1+1}^{M} c_j\|\hat{b}_{nA_j}\|_1^{\gamma}.
\end{aligned} \tag{2.15}$$

By the optimality of $\hat{\boldsymbol{b}}_n$, we have

$$\begin{aligned}
&\|\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{b}}_n\|_2^2 + n\kappa\hat{\boldsymbol{b}}_n'\boldsymbol{V}\hat{\boldsymbol{b}}_n + n\lambda\sum_{j=1}^{M} c_j\|\hat{b}_{nA_j}\|_1^{\gamma} \\
\le&\|\boldsymbol{Y} - \boldsymbol{U}\tilde{\boldsymbol{b}}_n\|_2^2 + n\kappa\tilde{\boldsymbol{b}}_n'\boldsymbol{V}\tilde{\boldsymbol{b}}_n + n\lambda\sum_{j=1}^{M} c_j\|\tilde{b}_{nA_j}\|_1^{\gamma}.
\end{aligned} \tag{2.16}$$

It follows from (2.15) and (2.16) that

$$\begin{aligned}
&2(\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{b}}_n)'\boldsymbol{U}(\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n) + (1-\gamma)n\lambda \sum_{j=J_1+1}^{M} c_j\|\hat{b}_{nA_j}\|_1^{\gamma} \\
\le& n\lambda\sum_{j=1}^{M} c_j\|\hat{b}_{nA_j}\|_1^{\gamma} - n\lambda\sum_{j=1}^{M} c_j\|\tilde{b}_{nA_j}\|_1^{\gamma} + 2n\kappa \sum_{k=J_1+1}^{M+d} \hat{\boldsymbol{b}}_n' V_k \hat{b}_{nk} \\
\le&\|\boldsymbol{Y} - \boldsymbol{U}\tilde{\boldsymbol{b}}_n\|_2^2 - \|\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{b}}_n\|_2^2 + n\kappa\tilde{\boldsymbol{b}}_n'\boldsymbol{V}\tilde{\boldsymbol{b}}_n - n\kappa\hat{\boldsymbol{b}}_n'\boldsymbol{V}\hat{\boldsymbol{b}}_n + 2n\kappa \sum_{k=1}^{M+d} (\hat{b}_{nk} - \tilde{b}_{nk})\hat{\boldsymbol{b}}_n' V_k \\
=&\|\boldsymbol{U}(\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n)\|_2^2 + 2(\boldsymbol{Y} - \boldsymbol{U}\hat{\boldsymbol{b}}_n)'\boldsymbol{U}(\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n) + n\kappa(\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n)'\boldsymbol{V}(\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n).
\end{aligned}$$

Consequently,

$$(1-\gamma)n\lambda \sum_{j=J_1+1}^{M} c_j\|\hat{b}_{nA_j}\|_1^{\gamma} \le (\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n)'(\boldsymbol{U}'\boldsymbol{U} + n\kappa\boldsymbol{V})(\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n).$$

By (2.7) and condition *C.3*,

$$(1-\gamma)n\lambda \sum_{j=J_1+1}^{M} c_j \|\hat{b}_{nA_j}\|_1^{\gamma} \le O_p(nM^{-1})\|\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n\|_2^2. \tag{2.17}$$

Given $|A_j| = M+d-j+1$ and $\boldsymbol{b}^{(0)}$, which can be obtained by the penalized B-splines method (Cardot et al., 2003), the constants $c_j = |A_j|^{1-\gamma}/\|b_{A_j}^{(0)}\|_2^{\gamma}$ can be scaled so that $\min_{j\le J} c_j \ge 1$ and

$$\sum_{j=J_1+1}^{M} c_j \|\hat{b}_{nA_j}\|_1^{\gamma} \ge \left(\sum_{j=J_1+1}^{M} \|\hat{b}_{nA_j}\|_1\right)^{\gamma} \ge \|\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n\|_1^{\gamma} \ge \|\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n\|_2^{\gamma}. \tag{2.18}$$

If $\|\hat{b}_{nA_{J_1+1}}\|_2 > 0$ which is equivalent to $\|\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n\|_2 > 0$, combination of (2.17) and (2.18) yields

$$(1-\gamma)n\lambda \le O_p(nM^{-1})\|\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n\|_2^{2-\gamma}.$$

Together with Lemma 5 and the fact that $\|\hat{\boldsymbol{b}}_n - \tilde{\boldsymbol{b}}_n\|_2 \le \|\hat{\boldsymbol{b}}_n - \boldsymbol{b}_0\|_2$, this implies that $(1-\gamma)n\lambda \le O_p(M^{1-\gamma}n^{\gamma/2})$. Now, by condition *C.6*,

$$P(\|\hat{b}_{nA_{J_1+1}}\|_2 > 0) \le P\left(\frac{\lambda}{M^{1-\gamma}n^{\gamma/2-1}} \le \; O_p(1)\right) \to 0.$$

Then the conclusion of the lemma follows. □

Proof of Theorem 1. By Lemma 5, $\|\hat{\beta}_n - \beta_{0s}\|_2^2 \le \|\hat{\boldsymbol{b}}_n - \boldsymbol{b}_0\|_2^2 \sum_{j=1}^{M+d}\int_0^1 B_k^2(t)dt = O_p(M^2 n^{-1})$. Since $\|\beta_0 - \beta_{0s}\|_\infty = O(M^{-p})$, $\|\beta_{0s} - \beta_0\|_2 \le O(M^{-p})$. Thus $\|\hat{\beta}_n - \beta_0\|_2 \le \|\hat{\beta}_n - \beta_{0s} + \beta_{0s} - \beta_0\|_2 \le \|\hat{\beta}_n - \beta_{0s}\|_2 + \|\beta_{0s} - \beta_0\|_2 = O_p(Mn^{-1/2} + M^{-p})$. □

Proof of Theorem 2. (i) We know that $\delta_0 \in [t_{J_1-1}, t_{J_1})$. By the compact support property of B-spline basis functions, for all $t \in [t_{J_1}, 1]$, $\hat{\beta}_n(t) = \sum\limits_{j=1}^{M+d} \hat{b}_{nj}B_j(t) = \sum\limits_{j=J_1+1}^{M+d} \hat{b}_{nj}B_j(t)$. If $\hat{b}_{nA_{J_1+1}} = 0$, then $\hat{\beta}_n(t) = 0$ on $[t_{J_1}, 1]$. Thus by Lemma 6, $P\left(\hat{\beta}_n(t) = 0 \text{ on } [t_{J_1}, 1]\right) \ge P\left(\|\hat{b}_{nA_{J_1+1}}\|_2 = 0\right) \to 1$. Therefore, given $0 < \zeta_1 < 1 - \delta_0$, for M sufficiently large, $\delta_0 + \zeta_1 > t_{J_1}$. Then

$$P\left(\hat{\beta}_n(t) = 0 \text{ on } [\delta_0 + \zeta_1, 1]\right) \ge P\left(\hat{\beta}_n(t) = 0 \text{ on } [t_{J_1}, 1]\right) \ge P\left(\|\hat{b}_{nA_{J_1+1}}\|_2 = 0\right) \to 1$$

.

(ii) We first argue that $P(\hat{b}_{nA_{J_1-d}} = 0) \to 0$. To see that, for some fixed $\zeta_2 > 0$, $\beta_0(t) \neq 0$ for some $t \in (\delta_0 - \zeta_2, \delta_0)$. Since $\|\beta_s - \beta_0\|_\infty = O(M^{-p})$, for sufficiently large M, there is some K such that $t_K \geq \delta_0 - \zeta_2$. We also have $|\hat{b}_{nK}| \neq 0$ with probability tending to one, which further implies that $P(\delta_0 - \zeta_2 \leq t_K \leq \hat{\delta}_n) \to 1$. On the other hand, from Lemma 6 we deduce that $P(\hat{\delta}_n \leq \delta_0 + \zeta_2) \to 1$. Therefore, together we obtain the claim that $\hat{\delta}_n$ converges to δ_0 in probability, by noting that $\zeta_2 > 0$ is arbitrary. □

Chapter 3

Sparse Functional Partial Least Squares Regression

3.1 Introduction

In this chapter, we develop a sparse version of the functional partial least squares in the context of the functional linear regression model. Our proposed method provides locally sparse estimators for the functional partial least squares basis functions. More importantly, the proposed method is able to produce a sparse estimate for the slope function in the functional linear regression model.

This work is motivated by oriented strand board (OSB) furnish research conducted by FPInnovations. In this study, a novel laboratory spectroscopy technique was developed for determining species identification of OSB strands and the relative proportions of sound wood, rot and bark in OSB fines samples. The log section samples were first retrieved from Canadian OSB mills and then separated into sound wood, rot and bark groups. A laboratory disk strander was used to prepare strands and a laboratory grinder ground the samples into coarse powder. Vis-NIR (visible and near infrared) spectroscopy measurement techniques were applied to acquire the spectra traces of the samples with 2150 individual wavelengths ranging from 350-2500nm.

Figure 3.1(a) displays spectra traces of 182 OSB fines samples. Currently mills do not monitor the key raw material constituents (sound wood, rot and bark) that play a major role in production operating efficiency and final product attributes. Periodically monitoring raw material can help mills identify problems associated with rot in logs, debarking inefficiency and species variability. Measurements can also provide data to assist in process adjustments and production planning and budgeting. Figure 3.1(b) illustrates spectra traces of OSB fines samples with different proportions of compositions of sound wood. It can be observed from Figure 3.1(b) that samples with different proportions of sound wood have distinct spectra traces and thereby we focus on predicting the proportions of sound wood in OSB fines samples from their spectra traces.

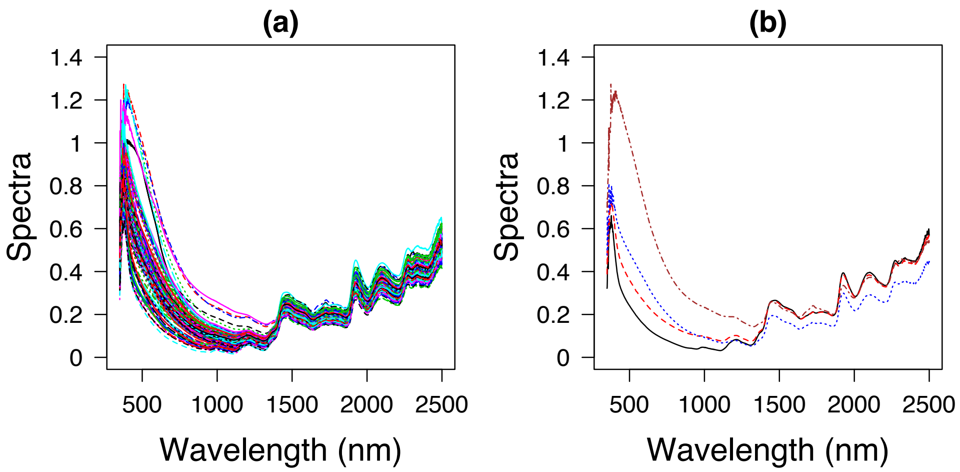

Figure 3.1: (a) Spectra traces of 182 OSB fines samples. (b) Selected spectra traces of OSB fines samples with different proportions of sound wood (———, 100%; - - - - -, 80%; ··········, 40%; -·-·-·- 0%).

Partial least squares (PLS) regression searches for components from the perspective of prediction and takes the response into account, and is often used in spectrometric prediction in chemometrics. Cook and Forzani (2019) pointed out that the best asymptotic behavior of PLS is reached in abundant regressions, which occurs in chemometrics applications. For instance, in the OSB furnish study, the spectra traces were acquired at 2150 individual wavelengths from 350-2500nm. On the one hand, we can treat the spectra as 2150 separate variables and apply the PLS method. However, this approach ignores the continuity, smoothness and order of the 2150 measurements. For this reason, instead of considering them as separate variables, we treat the spectra traces from 2150 individual wavelengths as realizations of a stochastic process. Therefore we aim to develop a functional partial least squares method. At the same time, we are also interested in detecting the subranges of the wavelengths that have no effect on the proportions of constituents.

The literature on PLS is abundant; see Wold (1975), Helland (1990) and Garthwaite (1994), for example. The PLS first gained popularity in chemometrics (e.g., Frank and Friedman, 1993; Martens and Næs, 1992), and more recently has been applied to other fields, such as bioinformatics (Boulesteix and Strimmer, 2006) and image detection (Schwartz et al., 2009). The PLS method was also extended to process functional data by Preda et al. (2007) and Delaigle and Hall (2012a) for classification. Escabias et al. (2007) adapted PLS to the functional logit model. In addition, PLS was employed in functional linear models by Preda and Saporta (2005) to determine the slope function. Reiss and Ogden (2007) proposed a smooth estimator of the slope function by combining the smoothing method and the PLS in the functional linear regression model. Delaigle and Hall (2012b) developed an alternative formulation of PLS in the functional setting that led to detailed theoretical results. Although these endeavors can produce satisfactory estimates for the slope function, they do not specifically focus on the locally sparse estimation based on the functional partial least squares method.

To address the aforementioned problem, we develop a new method, which we call *sparse functional partial least squares method* (SpaFunPLS), that simultaneously produces locally sparse estimates for the functional partial least squares (FunPLS) basis functions and the slope function in the functional linear regression model. Achieving locally sparse estimates for the FunPLS basis functions is not difficult. However, the locally sparse estimates for the FunPLS basis functions do not directly lead to a locally sparse estimate for the slope function, because the estimated sparse subregions of each basis function may not overlap. Our proposed SpaFunPLS method combines the ideas of the dimension reduction via PLS, the B-spline expansion, and the fSCAD (Lin et al., 2017) penalty. Similar to PLS, the SpaFunPLS method is an iterative procedure. In each iteration, we obtain a locally sparse FunPLS basis by employing the fSCAD penalty and identify the non-zero subregions, called active regions. On the subspace of the active regions, we update all the FunPLS basis functions and estimate the slope function. This is motivated by Chun and Keleş (2010), who proposed a sparse partial least squares method for simultaneous dimension reduction and variable selection.

The functional linear regression has been studied in vast literature. For example, Hastie and Mallows (1993) developed the smooth estimation of $\beta(t)$ via penalized least squares and/or smooth basis expansion, Cardot et al. (2003) and Li and Hsing (2007) used B-spline basis expansion and Fourier basis, respectively, both with a roughness penalty to control the smoothness of estimated slope functions, and Cardot et al. (2003), Cai and Hall (2006) and Hall and Horowitz (2007) considered to use the eigenfunctions of the covariance function of the predictor process as the bases.

Yao et al. (2005) extended the scope of the problem to study sparse longitudinal data. James et al. (2009) pioneered the locally sparse estimation of the slope functions. Lin et al. (2017) and Zhou et al. (2013) studied the locally sparse slope functions. Hall and Hooker (2016) and Guan et al. (2020) investigated the truncated functional linear regression models. Comparing to the existing methods, our proposed method makes multiple contributions. First, we propose a new method to provide locally sparse estimates for the functional partial least squares bases. Second, we produce a locally sparse estimate for the slope function via the locally sparse functional partial least squares bases. Third, we develop an efficient algorithm to implement the proposed method.

We structure the rest of the chapter as follows. In Section 3.2, we introduce the proposed SpaFunPLS method and provide its computational details. In Section 3.3, simulation results are presented to evaluate the performance of the proposed method. Applications of the SpaFunPLS approach to the OSB furnish data and the particulate matter emissions data are given in Section 3.4. Section 3.5 concludes the chapter. Additional discussions are provided in the appendix.

3.2 Methodology

3.2.1 The First Sparse Functional Partial Least Squares Basis Function

We consider the first functional partial least squares (FunPLS) basis function here and discuss estimation of the other basis functions in a separate subsection. Suppose we observe data pairs $(X_1, Y_1), ..., (X_n, Y_n)$, which are independently and identically distributed as (X, Y), where X is a random function defined on an interval $[0, T]$ and Y is a random scalar generated by $Y = \mu + \int_0^T X(t)\beta(t)\,\mathrm{d}\,t + \varepsilon$, with μ being the intercept, $\beta(t)$ the slope function, and ε representing the noise that is independent of X. The first FunPLS basis function might be found by

$$\begin{aligned} \max_{w} \quad & \mathrm{Cov}^2\left(Y, \int_0^T X(t)w(t)\,\mathrm{d}\,t\right), \\ \text{subject to} \quad & \|w\|_2^2 = 1. \end{aligned} \tag{3.1}$$

To estimate $w(t)$, we utilize B-spline basis functions that are detailed in de Boor (2001). Let $\boldsymbol{B}(t) = (B_1(t), \ldots, B_{M+d}(t))^{\mathrm{T}}$ be a vector that contains $M+d$ B-spline basis functions. Each basis function is defined on $[0, T]$ with degree d and $M+1$ equally spaced knots $0 = t_0 < t_1 < \cdots <$

$t_M = T$, which is a piecewise polynomial of degree d. B-spline basis functions are well known for their compact support property, i.e., each basis function is positive over at most $d+1$ adjacent subintervals. We approximate $w(t)$ by $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{B}(t)$ with $\boldsymbol{b} = (b_1, \ldots, b_{M+d})^{\mathrm{T}}$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ and denote by $\boldsymbol{U}$ the $n \times (M+d)$ matrix with elements $u_{ij} = \int_0^T X_i(t)B_j(t)\,\mathrm{d}\,t$. If we denote $\boldsymbol{V}_0$ the $(M+d) \times (M+d)$ matrix with the elements $v_{0ij} = \int_0^T B_i(t)B_j(t)\,\mathrm{d}\,t$, then $\boldsymbol{b}$ can be estimated by

$$\begin{aligned} \max_{\boldsymbol{b}} \quad & \boldsymbol{b}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{b} \\ \text{subject to} \quad & \boldsymbol{b}^{\mathrm{T}}\boldsymbol{V}_0\boldsymbol{b} = 1, \end{aligned}$$

where $\boldsymbol{H} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{U}$.

To derive the first sparse FunPLS basis function, we introduce a surrogate $r(t)$ of $w(t)$ and impose the fSCAD penalty on the surrogate $r(t)$. We approximate the surrogate function $r(t)$ by $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{B}(t)$, where $\boldsymbol{c} = (c_1, \ldots, c_{M+d})^{\mathrm{T}}$. For $m > 0$, let $\boldsymbol{B}^{(m)}(t) = \left(B_1^{(m)}(t), \ldots, B_{M+d}^{(m)}(t)\right)^{\mathrm{T}}$ denote the vector of the m-th derivatives of the B-spline basis functions. The first sparse functional partial least squares basis function is then obtained by

$$\hat{w}(t) = \frac{\hat{\boldsymbol{c}}^{\mathrm{T}}}{\|\hat{\boldsymbol{c}}\|_2}\boldsymbol{B}(t), \tag{3.2}$$

where $\hat{\boldsymbol{c}} = (\hat{c}_1, \ldots, \hat{c}_{M+d})^{\mathrm{T}}$ solves

$$\begin{aligned} \min_{\boldsymbol{b},\boldsymbol{c}} \quad & -\kappa\boldsymbol{b}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{b} + (1-\kappa)(\boldsymbol{b}-\boldsymbol{c})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{b}-\boldsymbol{c}) + \gamma\left\|\boldsymbol{c}^{\mathrm{T}}\boldsymbol{B}^{(m)}\right\|_2^2 + \frac{M}{T}\int_0^T p_\lambda\left(|\boldsymbol{c}^{\mathrm{T}}\boldsymbol{B}(t)|\right)\mathrm{d}\,t + \delta\left\|\boldsymbol{c}^{\mathrm{T}}\boldsymbol{B}\right\|_2^2, \\ \text{subject to} \quad & \boldsymbol{b}^{\mathrm{T}}\boldsymbol{V}_0\boldsymbol{b} = 1, \end{aligned} \tag{3.3}$$

with non-negative parameters κ, γ, λ and δ. In this formulation, $p_\lambda(\cdot)$ is a SCAD penalty function proposed by Fan and Li (2001), which is defined on $[0, \infty]$ as

$$p_\lambda(u) = \begin{cases} \lambda u & \text{if } 0 \leq u \leq \lambda \\ -\frac{u^2 - 2a\lambda u + \lambda^2}{2(a-1)} & \text{if } \lambda < u < a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } u \geq a\lambda, \end{cases}$$

where a is a constant suggested to be 3.7. In (3.2), we use the rescaled $\hat{\boldsymbol{c}}$ as the B-spline coefficients, which has a norm 1.

In the objective function, the first term measures the interaction between the response and the latent component, while the second term enforces the surrogate function to be close to the functional partial least squares basis function in the same context. The third term is a roughness penalty that controls the smoothness of the estimated surrogate function. In practice, we choose $m = 2$, which corresponds to measuring the roughness of a function by its integrated curvature. The fourth term is the fSCAD penalty which is designed to regularize the sparseness of the estimated surrogate function. The last term in the objective function aims to stabilize the estimation procedure. The objective function is concave due to the first term, therefore we shall use a small κ to alleviate the concaveness.

Let $\boldsymbol{V}$ denote the $(M+d) \times (M+d)$ matrix with the elements $v_{ij} = \int_0^T B_i^{(m)}(t) B_j^{(m)}(t) \, \mathrm{d}\, t$. The roughness penalty term yields $\gamma \boldsymbol{c}^{\mathrm{T}} \boldsymbol{V} \boldsymbol{c}$. We follow Lin et al. (2017) to approximate the fSCAD penalty. Let $\boldsymbol{V}_{0j} = \int_{t_{j-1}}^{t_j} \boldsymbol{B}(t) \boldsymbol{B}^{\mathrm{T}}(t) \, \mathrm{d}\, t$ and define

$$G_\lambda\left(r^{(0)}\right) = \sum_{l=1}^{M} p_\lambda\left(\frac{\|r_{[l]}^{(0)}\|_2}{\sqrt{T/M}}\right) - \frac{1}{2}\sum_{l=1}^{M} p'_\lambda\left(\frac{\|r_{[l]}^{(0)}\|_2}{\sqrt{T/M}}\right)\frac{\|r_{[l]}^{(0)}\|_2}{\sqrt{T/M}},$$

$$\boldsymbol{Q}_\lambda^{(0)} = \frac{1}{2}\sum_{l=1}^{M}\left(\frac{p'_\lambda\left(\|r_{[l]}^{(0)}\|_2\sqrt{M/T}\right)}{\|r_{[l]}^{(0)}\|_2\sqrt{T/M}}\boldsymbol{V}_{0l}\right),$$

where $\|r_{[l]}^{(0)}\|_2 = \left(\int_{t_{l-1}}^{t_l} \left|r^{(0)}(t)\right|^2 \mathrm{d}\, t\right)^{1/2}$. Then

$$\frac{M}{T}\int_0^T p_\lambda\left(|\boldsymbol{c}^{\mathrm{T}}\boldsymbol{B}|\right) \mathrm{d}\, t \approx \boldsymbol{c}^{\mathrm{T}}\boldsymbol{Q}_\lambda^{(0)}\boldsymbol{c} + G_\lambda\left(r^{(0)}\right).$$

A choice for $r^{(0)}$ is the first FunPLS basis function obtained by solving (3.1). We may write the last term in the objective function as $\delta \boldsymbol{c}^{\mathrm{T}} \boldsymbol{V}_0 \boldsymbol{c}$. Therefore the objective function in (3.3) can be expressed as

$$-\kappa \boldsymbol{b}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{b} + (1-\kappa)(\boldsymbol{b}-\boldsymbol{c})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{b}-\boldsymbol{c}) + \gamma \boldsymbol{c}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{c} + \boldsymbol{c}^{\mathrm{T}}\boldsymbol{Q}_\lambda^{(0)}\boldsymbol{c} + G_\lambda\left(r^{(0)}\right) + \delta \boldsymbol{c}^{\mathrm{T}}\boldsymbol{V}_0\boldsymbol{c}. \quad (3.4)$$

3.2.2 Computational Method

To solve (3.3), we iterate between solving for $\boldsymbol{b}$ for fixed $\boldsymbol{c}$ and solving for $\boldsymbol{c}$ for fixed $\boldsymbol{b}$. Estimating $\boldsymbol{b}$ given $\boldsymbol{c}$ becomes the following problem

$$\min_{\boldsymbol{b}} \; -\kappa \boldsymbol{b}^{\mathrm{T}} \boldsymbol{H} \boldsymbol{b} + (1-\kappa)(\boldsymbol{b}-\boldsymbol{c})^{\mathrm{T}} \boldsymbol{H} (\boldsymbol{b}-\boldsymbol{c}),$$
$$\text{subject to } \boldsymbol{b}^{\mathrm{T}} \boldsymbol{V}_0 \boldsymbol{b} = 1.$$

Since $\boldsymbol{V}_0$ is symmetric and positive definite, we can write $\boldsymbol{V}_0 = \boldsymbol{W}_0 \boldsymbol{W}_0$, where $\boldsymbol{W}_0$ is symmetric. Let $\boldsymbol{b}_* = \boldsymbol{W}_0 \boldsymbol{b}$ and $\boldsymbol{c}_* = \boldsymbol{W}_0 \boldsymbol{c}$. Define $\boldsymbol{Z} = \boldsymbol{W}_0^{-1} \boldsymbol{U}^{\mathrm{T}} \boldsymbol{Y}$. For $0 < \kappa < 1/2$, $\boldsymbol{b}$ can be obtained by

$$\boldsymbol{W}_0^{-1} \min_{\boldsymbol{b}_*} \left(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{b}_* - \kappa' \boldsymbol{Z}^{\mathrm{T}} \boldsymbol{c}_*\right)^{\mathrm{T}} \left(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{b}_* - \kappa' \boldsymbol{Z}^{\mathrm{T}} \boldsymbol{c}_*\right),$$
$$\text{subject to } \boldsymbol{b}_*^{\mathrm{T}} \boldsymbol{b}_* = 1, \tag{3.5}$$

where $\kappa' = (1-\kappa)/(1-2\kappa)$. Now if we fix $\boldsymbol{b}$ in (3.4), $\boldsymbol{c}$ can be estimated by solving

$$\min_{\boldsymbol{c}} (1-\kappa)(\boldsymbol{b}-\boldsymbol{c})^{\mathrm{T}} \boldsymbol{H} (\boldsymbol{b}-\boldsymbol{c}) + \gamma \boldsymbol{c}^{\mathrm{T}} \boldsymbol{V} \boldsymbol{c} + \boldsymbol{c}^{\mathrm{T}} \boldsymbol{Q}_{\lambda}^{(0)} \boldsymbol{c} + \delta \boldsymbol{c}^{\mathrm{T}} \boldsymbol{V}_0 \boldsymbol{c}.$$

Based on Theorem 3 of Chun and Keleş (2010), for $0 < \kappa \leq 1/2$ and a fixed $\boldsymbol{c}$, the solution of (3.5) is $\hat{\boldsymbol{b}} = \boldsymbol{W}_0^{-1} \boldsymbol{Z} / \|\boldsymbol{Z}\|_2$, which does not depend on $\boldsymbol{c}$. Therefore the solution to (3.3) is

$$\hat{\boldsymbol{c}} = (1-\kappa) \left((1-\kappa)\boldsymbol{H} + \gamma \boldsymbol{V} + \boldsymbol{Q}_{\lambda}^{(0)} + \delta \boldsymbol{V}_0\right)^{-1} \boldsymbol{H} \hat{\boldsymbol{b}}.$$

Once $\hat{\boldsymbol{c}}$ is produced, the estimate for the first empirical sparse FunPLS basis function is given in (3.2).

3.2.3 An Algorithm for the Sparse Functional Partial Least Squares

Incorporating the formulation of the first sparse FunPLS basis into the iterative conventional FunPLS algorithm (see the appendix) enables us to obtain the subsequent sparse FunPLS basis functions. However, the sparse FunPLS basis functions derived this way are unlikely to have the same zero subregions and therefore the estimated slope function might not be sparse. To see this, without loss of generality, we assume $X(\cdot)$ and Y are centered, i.e., $\mathbf{E}X(t) \equiv 0$ and $\mathbf{E}Y \equiv 0$, so that μ is zero. Assume $w_1(t), \ldots, w_K(t)$ are the first K sparse FunPLS basis functions. We express the slope function in terms of the sparse FunPLS bases, which can be written as $\beta(t) = \sum_{k=1}^{K} \alpha_k w_k(t)$. Let $\boldsymbol{S}$ be the $n \times K$ matrix with elements $s_{ij} = \int_0^T X_i(t) w_j(t) \,\mathrm{d}\, t$. The columns of $\boldsymbol{S}$ represent the scores. The coefficient vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$ can be estimated using ordinary least squares by $\hat{\boldsymbol{\alpha}} = (\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S})^{-1}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{Y}$. It is obvious that if the basis functions $w_1(t), \ldots, w_K(t)$ do not have overlapping zero subregions, then $\hat{\beta}(t) = \sum_{k=1}^{K} \hat{\alpha}_k w_k(t)$ is not a locally sparse estimator.

To address the above problem, we propose the following algorithm. At each iteration, we first produce a sparse FunPLS basis function by solving (3.2) and (3.3), based on which we define an active region $\mathcal{A}$, and then we update all the FunPLS basis functions on $\mathcal{A}$. The slope function in the functional linear regression model is expanded by the FunPLS basis functions and is not zero only on the active region $\mathcal{A}$. Define $X_{\mathcal{A}}(t) = X(t) I_{(t \in \mathcal{A})}$, where $I_{(\cdot)}$ denotes the indicator function, and let $\boldsymbol{X}_{\mathcal{A}} = \{X_{1\mathcal{A}}, \ldots, X_{n\mathcal{A}}\}$. We first set the initial estimate $\hat{\beta}^{(0)}(t) = 0$ for all $0 \le t \le T$ and set $\mathcal{A}^{(0)} = \emptyset$. Set $\boldsymbol{Y}^{(1)} = \boldsymbol{Y}$. The algorithm is outlined below. For $k = 1, \ldots, K$,

1. Find $\hat{w}(t)$ by solving (3.2) and (3.3) with $\boldsymbol{H} = \left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{Y}^{(k)}\right)\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{Y}^{(k)}\right)^{\mathrm{T}}$.

2. Update $\mathcal{A}^{(k)}$ as $\{t : \hat{w}(t) \neq 0\} \cup \mathcal{A}^{(k-1)}$.

3. Update the first k FunPLS basis functions with $\boldsymbol{Y}$ and $\boldsymbol{X}_{\mathcal{A}^{(k)}}$ using the conventional FunPLS method detailed in the appendix.

4. Update $\hat{\beta}^{(k)}(t)$ by using the estimated FunPLS basis functions obtained in the step 3.

5. Update $\boldsymbol{Y}^{(k)} = \left(Y_1^{(k)}, \ldots, Y_n^{(k)}\right)^{\mathrm{T}}$ with $Y_i^{(k)} = Y_i - \int_0^T X_i(t)\hat{\beta}^{(k)}(t) \,\mathrm{d}\, t$.

The non-zero active region is updated at each iteration. For $k = K$, $\mathcal{A}^{(K)}$ is a combination of the non-zero regions for the first K sparse FunPLS basis functions. After completion of steps 1 to 5 for $k = 1, \ldots, K$, the estimated sparse FunPLS basis functions are obtained in step 3 when $k = K$, the estimated slope function $\hat{\beta}^{\mathrm{SpaFunPLS}}(t) = \hat{\beta}^{(K)}(t)$, and the estimated non-zero region for the slope function is $\mathcal{A}^{(K)}$.

In our fitting procedure, there are a few tuning parameters including κ, γ, λ, δ, K, and the parameters for constructing the B-spline basis functions such as the degree d of the B-spline basis and the number of knots $M+1$. The crucial tuning parameters are the shrinkage parameter λ and the number of latent components K and the smoothing parameter γ. Cardot et al. (2003) suggested that the choice of M is not important due to the roughness penalty and suggested a relatively large value for M to capture the local features of the functions; see also Marx and Eilers (1999) and Lin et al. (2017). The degree d, which is also of less importance, is fixed to a reasonable value, i.e., $d = 3$. As discussed in Section 3.2.1, we choose κ to be small to reduce the effect of the concaveness of the objective function. We set δ to a sensible value to stabilize the estimation procedure. The roughness penalty in (3.3) regularizes the smoothness of the surrogate function, the zero subregions of which determine the active region $\mathcal{A}$. In other words, if the roughness parameter γ is very small, the resulting active region will be very irregular and scattered; see the appendix for more details. The above algorithm implies that the sparse estimation of the slope function depends highly on the shrinkage parameter λ and the number K of components. Observe that $\mathcal{A}^{(1)} \subseteq \cdots \subseteq \mathcal{A}^{(K)} \subseteq [0, T]$, a large number of components may not lead to a sparse estimation of $\beta(t)$. To encourage model sparsity, the parameters γ, λ and K are tuned by minimizing the Bayesian information criterion (BIC). We follow Krämer and Sugiyama (2011) to calculate the degree of freedom of the model and BIC. Readers are referred to the appendix for a discussion on the effects of κ, γ and δ.

3.3 Simulation Studies

In this subsection we numerically illustrate the performance of the proposed method via simulations. We also compare the method with the conventional FunPLS method described in the appendix, the regularized version of the functional partial least squares (FPLS_R) proposed by Reiss and Ogden (2007), and the functional principal component regression (FPCR) which is also detailed in the appendix. The FPLS_R method adopts B-spline basis expansion and incorporates a roughness penalty in the regression. The roughness penalty parameter of FPLS_R is selected by fitting a linear mixed model through restricted maximum likelihood (REML) estimation (Reiss and Ogden, 2007).

We consider the following three scenarios for $\beta(t)$:

Scenario I: $\beta(t) = I_{(0 \leq t < 0.5)}$. $\beta(t)$ is a discontinuous function with a flat non-zero part on $[0, 0.5)$. The zero subregion is $[0.5, 1]$.

Scenario II: $\beta(t) = \sin(2\pi t) I_{(0\leq t<0.5)}$. $\beta(t)$ is a continuous function with a trigonometric non-zero part on $(0, 0.5)$. The zero subregion is $[0.5, 1]$.

Scenario III: $\beta(t) = 3t + e^{t^2}\cos(3\pi t) + 1$. $\beta(t)$ is a continuous function without zero subregions.

The function $\beta(t)$ in these scenarios are plotted in Figure 3.2 and 3.3. The functional predictors $X_i(t)$ are generated by $X_i(t) = \sum a_{ij} B_j(t)$, where $B_j(t)$ are cubic B-spline basis functions defined on 50 equally spaced knots over $[0, 1]$, and the coefficients a_{ij} are sampled from the standard normal distribution. The errors ε_i are independently generated from normal distributions so that the signal-to-noise ratio equals 5. In each study we run the simulation independently for 100 times with sample sizes $n = 100$ and $n = 500$. For each simulation replicate we also generate a separate test dataset with a sample size of 5000.

For the SpaFunPLS method and the FPLS_R method, we expand the slope function with cubic B-splines with 101 equally spaced knots. The FunPLS method adopts cubic B-splines with the number of knots and the number of components selected by Cross-Validation (CV) and BIC. The number of knots selected by the FunPLS method will be used in step 4 of the proposed SpaFunPLS algorithm introduced in Section 3.2.3 to reduce the computational complexity. We select the tuning parameters of the FPCR by CV and BIC. For the FunPLS method, the results based on BIC are better than CV and for the FPCR method the results based on CV are slightly better than BIC. Therefore, we provide results for FunPLS and FPCR method based on BIC and CV, respectively. For the FPLS_R method, the number of components is chosen by fivefold CV, while the roughness parameter is selected by REML. For the SpaFunPLS method, we set $\kappa = 0.1$ and $\delta = 0.1$ in all numerical studies. In the appendix, we discuss the effects of κ, γ and δ. The smoothing parameter γ, the sparse parameter λ and the number of components K of the SpaFunPLS method are chosen by the approach presented in Section 3.2.3.

The performance of the estimated $\hat{\beta}(t)$ is evaluated by the integrated squared errors on the zero subregions and the entire domain, which are, respectively, defined as

$$\text{ISE}_0 = \frac{1}{l_0}\int_{\mathcal{I}_0}\left(\hat{\beta}(t) - \beta(t)\right)^2 \mathrm{d}\, t \quad \text{and} \quad \text{ISE} = \frac{1}{l}\int_{\mathcal{I}}\left(\hat{\beta}(t) - \beta(t)\right)^2 \mathrm{d}\, t,$$

where $\mathcal{I}_0$ denotes the zero subregions of $\beta(t)$, $\mathcal{I}$ denotes the entire domain of $\beta(t)$, and l_0 and l are the lengths of the zero subregions and the entire region, respectively. In all scenarios, we have $\mathcal{I} = [0, 1]$ and $l = 1$, and for Scenario I and II, $\mathcal{I}_0 = [0.5, 1]$ and $l_0 = 0.5$.

Table 3.1 summarizes the ISE_0 of the estimators. The results in Table 3.1 indicate that the proposed SpaFunPLS method outperforms all the other methods on the zero subregions, thanks to the sparse penalty. In addition, the results in Table 3.2 indicate that the proposed SpaFunPLS method has the best performance in estimating the slope functions. In Scenario III when there is no zero subregion, the SpaFunPLS method has nearly identical results as the FunPLS method, as expected. It is also observed that the FPLS_R approach exhibits relatively larger errors. In terms of prediction, Table 3.3 suggests that, in general, the proposed SpaFunPLS method has better performance than the other methods. The FunPLS and the FPCR methods are comparable to the proposed method.

Table 3.1: Mean integrated squared error, ISE_0, defined on the null region for $\hat{\beta}(t)$ based on 100 simulation replications with the corresponding Monte Carlo standard deviation included in parentheses.

	SpaFunPLS	FunPLS	FPLS_R	FPCR
Scenario I $(\times 10^{-2})$				
n = 100	0.39 (0.57)	2.76 (1.06)	3.89 (1.28)	2.31 (1.33)
n = 500	0.25 (0.25)	0.95 (0.26)	0.93 (0.33)	0.85 (0.38)
Scenario II $(\times 10^{-3})$				
n = 100	0.43 (1.67)	6.85 (4.95)	17.97 (5.63)	6.91 (7.36)
n = 500	0.12 (0.35)	1.45 (0.83)	3.45 (1.28)	1.86 (1.38)

SpaFunPLS, the proposed sparse functional partial least squares method; FunPLS, the functional partial least squares method; FPLS_R, the regularized-regression version of the functional partial least squares proposed by Reiss and Ogden (2007); FPCR, the functional principal component regression method.

Table 3.2: Mean integrated squared errors, ISE, defined on the entire region for $\hat{\beta}(t)$ based on 100 simulation replications with the corresponding Monte Carlo standard deviation included in parentheses.

	SpaFunPLS	FunPLS	FPLS_R	FPCR
Scenario I $(\times 10^{-2})$				
n = 100	2.37 (1.42)	3.28 (1.04)	6.55 (1.63)	2.94 (1.33)
n = 500	1.37 (0.47)	1.49 (0.26)	2.67 (0.66)	1.34 (0.28)
Scenario II $(\times 10^{-3})$				
n = 100	4.78 (4.27)	7.76 (4.83)	27.12 (8.89)	7.38 (8.29)
n = 500	1.13 (0.82)	1.46 (0.66)	8.56 (4.47)	1.87 (1.54)
Scenario III $(\times 10^{-1})$				
n = 100	1.53 (1.28)	1.54 (1.29)	7.85 (2.20)	1.64 (1.74)
n = 500	0.34 (0.17)	0.34 (0.17)	2.55 (1.08)	0.40 (0.28)

SpaFunPLS, the proposed sparse functional partial least squares method; FunPLS, the functional partial least squares method; FPLS_R, the regularized-regression version of the functional partial least squares proposed by Reiss and Ogden (2007); FPCR, the functional principal component regression method.

Table 3.3: The prediction mean squared errors on test data based on 100 simulation replications with the corresponding Monte Carlo standard deviation included in parentheses.

	SpaFunPLS	FunPLS	FPLS_R	FPCR
Scenario I $(\times 10^{-3})$				
n = 100	2.14 (0.16)	2.31 (0.16)	2.68 (0.29)	2.26 (0.17)
n = 500	1.99 (0.06)	2.02 (0.06)	2.13 (0.11)	2.01 (0.06)
Scenario II $(\times 10^{-3})$				
n = 100	1.08 (0.07)	1.14 (0.08)	1.42 (0.17)	1.14 (0.11)
n = 500	1.03 (0.03)	1.04 (0.03)	1.16 (0.09)	1.04 (0.03)
Scenario III $(\times 10^{-2})$				
n = 100	3.29 (0.23)	3.29 (0.23)	4.17 (0.44)	3.31 (0.25)
n = 500	3.09 (0.09)	3.09 (0.09)	3.42 (0.23)	3.10 (0.09)

SpaFunPLS, the proposed sparse functional partial least squares method; FunPLS, the functional partial least squares method; FPLS_R, the regularized-regression version of the functional partial least squares proposed by Reiss and Ogden (2007); FPCR, the functional principal component regression method.

To visualize the performance of the proposed SpaFunPLS method, Figure 3.2 compares the estimated $\hat{\beta}(t)$ for various methods. We can see that the SpaFunPLS, FunPLS and FPCR methods provide smooth estimates for the slope function. However, for Scenario I and Scenario II, only the SpaFunPLS estimate is capable of correctly identifying a major portion of the zero subregions. For Scenario III where there is no zero subregions, the SpaFunPLS and FunPLS estimates are identical. It also shows that FPLS_R method might not have sufficient roughness regularization of the estimated slope function, which leads to less favorable performance. In Figure 3.3, we present the SpaFunPLS estimates for all simulation replicates with $n = 100$. The plots suggest that in general, the SpaFunPLS method locates the zero subregions with considerable accuracy and does a good job of estimating the slope function.

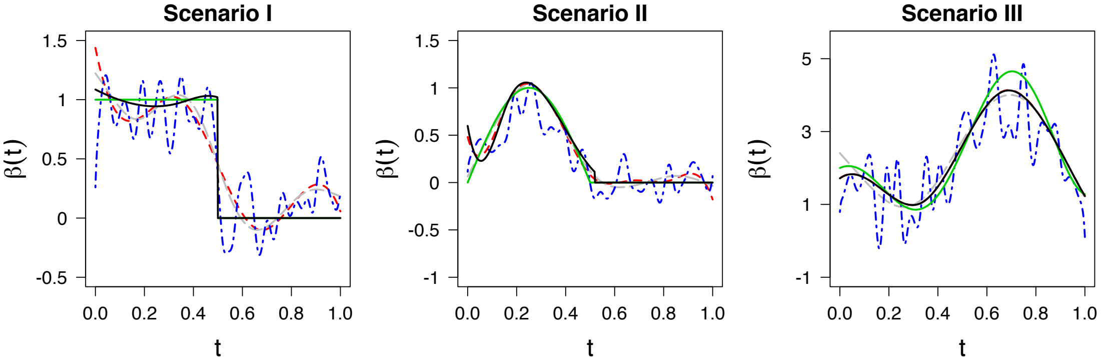

Figure 3.2: Estimated $\hat{\beta}(t)$ for various methods (———, the SpaFunPLS method; - - - - -, FunPLS method; -·-·-·-, FPLS_R method; — — —, the FPCR method; ———, the true $\beta(t)$) in a randomly selected simulation replicate with $n = 100$.

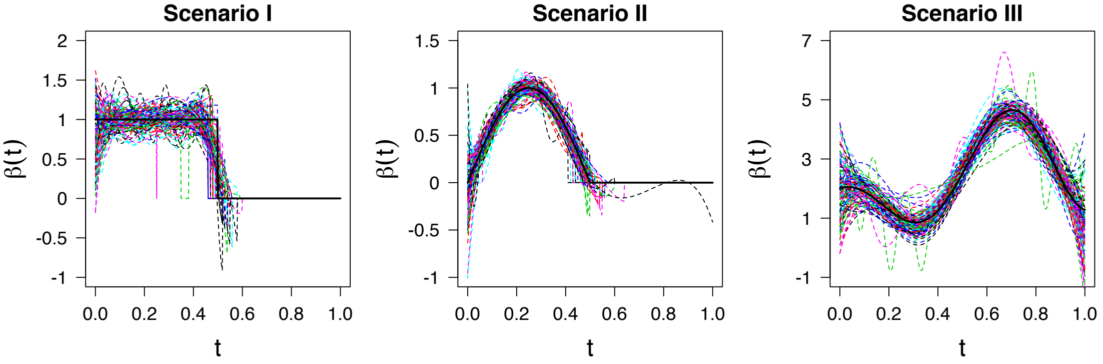

Figure 3.3: Estimated slope function $\hat{\beta}(t)$ by the SpaFunPLS method (- - - - -) with $n = 100$ along with the true $\beta(t)$ (———).

3.4 Applications

3.4.1 OSB Furnish Data

We apply our method to a novel dataset provided by FPInnovations which collected wood material from Canadian OSB mills. In total, 60 cookies from 60 different logs were collected with 10 from each of the following types of logs: aspen with core rot, sound aspen with bark, balsam poplar with core rot, balsam poplar with bark, birch with core rot and birch with bark. Then the bark and rot were removed from all samples and segregated into separate sample bags. Sound wood, rot and bark material from each species were ground into powder. Each sample was then oven dried to minimize the measurement error due to moisture variation and stored in a sealed sample bag. In the end, 182 mixtures with different proportions of sound wood were prepared.

The powder samples were scanned by a benchtop Vis/NIR (visible and near infrared) spectrometer ASD 5000 from Analytical Spectral Devices for acquiring spectra. Vis/NIR spectroscopy is an industrial proven measurement technique for classifying and quantifying the composition and properties of organic materials including wood. In the visible range from $350 - 700$nm the colour composition of the measured sample was recorded. The records in the NIR range from $700-2500$nm reveal special interactions between light at specific wavelengths and target sample molecular structures, from which specific sample constituents can be chemically identified and quantified. Specifically, for each sample, a spectral file was generated comprised of log inverse reflectance versus wavelength for 2150 individual wavelengths ranging from 350nm to 2500nm. For each of the 182 samples, two spectra replicates were acquired.

The objective of this study is to examine the relation between the spectral trajectories and the proportions of the sound wood in OSB fines samples. In the functional linear regression model, for the ith fine mixture sample, the response Y_i is the proportion of sound wood content and $X_i(t)$, $t \in [350, 2500]$ is its spectral curve. Both Y_i and $X_i(t)$ are centered such that $\mathbf{E}Y_i \equiv 0$ and $\mathbf{E}X_i(t) \equiv 0$. Figure 3.4(a) displays 10 randomly selected smoothed centered spectral curves.

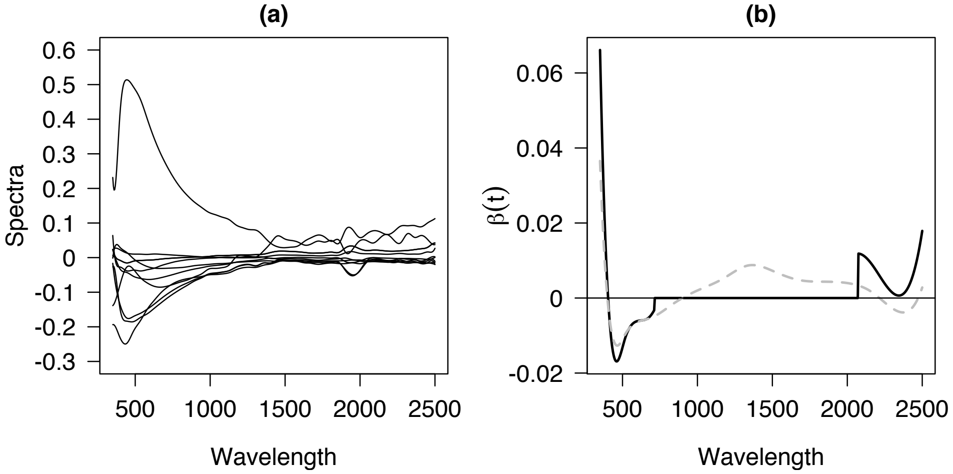

Figure 3.4: (a) 10 randomly selected smoothed centered spectral curves. (b) Estimated $\hat{\beta}(t)$ using the conventional FunPLS method (- - - - -) and the proposed SpaFunPLS approach (———).

Figure 3.4(b) depicts estimates for $\beta(t)$ obtained by the proposed SpaFunPLS approach and the conventional FunPLS method. The SpaFunPLS method uses cubic B-spline basis functions with 101 equally spaced knots. We set $\kappa = 0.1$, $\delta = 0.1$, and select γ,λ and the number of components

K by BIC. The SpaFunPLS estimate for the slope function is zero over wavelengths $715 - 2070$nm, which suggests the relationship between the sound wood proportions and spectra in the low and high ends of the wavelengths. The FunPLS method provides a similar estimate without giving clear subregions on which the spectra are not related to the sound wood proportions.

3.4.2 Particulate Matter Emissions Data

We further illustrate our proposed procedure by analyzing the particulate matter emissions (PM) data, which was studied in Asencio et al. (2014), Hall and Hooker (2016) and Guan et al. (2020). The data are detailed in the Coordinating Research Councils E55/E59 research project (Clark et al., 2007). In the experiment, drivers drove the trucks through a pre-set series of driving cycles on a chassis dynamometer test bed which was employed to simulate inertia, wind drag and tire rolling resistance. The particulate matter emissions were measured every second by an emission analyzer which was attached to the truck exhaust pipe. The engine acceleration of diesel trucks was also collected. Our interest here is to estimate the effects of the past engine acceleration on the current particulate matter emission. Intuitively, we expect earlier engine acceleration to have a smaller impact on the current particulate matter emission.

To remove dependences in the data, we follow Hall and Hooker (2016) to use PM observation every 10 seconds after the first 120 seconds. The response is the logarithm of the PM measured every 10 seconds after the first 120 seconds and the functional covariates are the engine acceleration for the past 60 seconds. We center both the PM and engine acceleration. Figure 3.5(a) illustrates 10 randomly selected smoothed centered engine acceleration curves recorded on every second for 60 seconds.

The proposed SpaFunPLS approach uses cubic B-spline basis functions with 101 equally spaced knots with $\kappa = 0.1$ and $\delta = 0.1$. The smoothing parameter γ, the sparse parameter λ and the number of components K are chosen by BIC. Figure 3.5(b) presents the SpaFunPLS and FunPLS estimates for $\beta(t)$. The FunPLS estimate suggests that the engine acceleration greatly influences the PM for the past 20 seconds. Compared to the FunPLS method, the SpaFunPLS method provides a more insightful and interpretable result. The SpaFunPLS estimate indicates a positive relationship that tapers to zero from second 0 to 16. This suggests that the engine acceleration has contribution to predicting PM for no longer than 16 seconds.

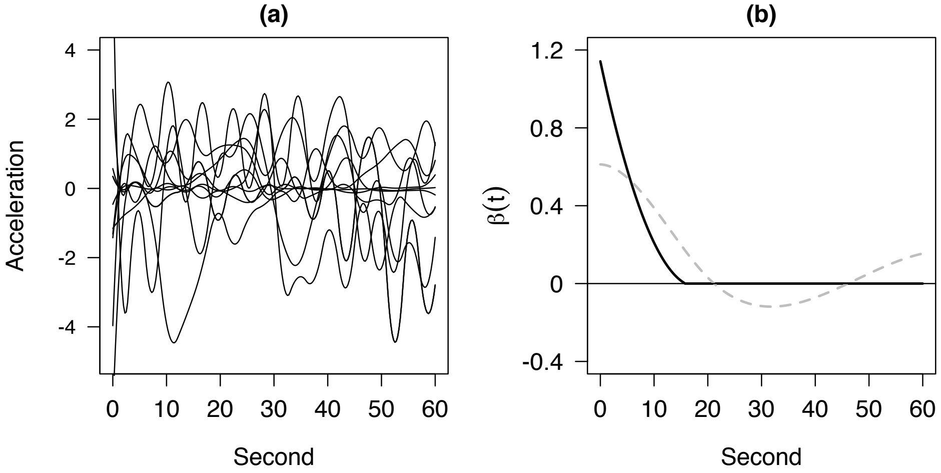

Figure 3.5: (a) 10 randomly selected smoothed centered acceleration curves. (b) Estimated $\hat{\beta}(t)$ using the conventional FunPLS method (- - - - -) and the proposed SpaFunPLS approach (———).

3.5 Conclusion

We proposed a sparse functional partial least squares regression to achieve a locally sparse estimate for the slope function in a functional linear regression model. The proposed method is effective in identifying nonactive subregion in the regression model. In practice, although there are a few tuning parameters to select, most of them can be fixed to a reasonable value. The important parameters are the smoothing parameter, the shrinkage parameter and the number of components. From our simulation studies we found that the proposed SpaFunPLS method improves the conventional FunPLS method and performs well for continuous/discontinuous and locally sparse/nonsparse functions. The real data applications show that the proposed method provides more interpretable estimates of the slope function.

3.6 Appendix

3.6.1 Functional Partial Least Squares Regression

In this subsection we introduce the functional partial least squares method (FunPLS) for estimating the slope function in the functional linear regression model. Without loss of generality, we assume that $\mu = 0$ in the functional linear regression model $Y_i = \mu + \int_0^T X_i(t)\beta(t)\,\mathrm{d}\,t + \varepsilon_i$ in the sequel. Given $w_1, \ldots, w_{k-1}$, the kth FunPLS basis function w_k is obtained by

$$
\begin{aligned}
&\max_{w} \quad \mathrm{Cov}^2\left(Y, \int_0^T X(t)w(t)\,\mathrm{d}\,t\right),\\
&\text{subject to} \quad \mathrm{Cov}\left(\int_0^T X(t)w_j(t)\,\mathrm{d}\,t, \int_0^T X(t)w(t)\,\mathrm{d}\,t\right) = 0 \quad \text{for } j = 1, \ldots, k-1 \quad \text{and}\\
&\|w\|_2^2 = 1.
\end{aligned} \tag{3.6}
$$

To estimate the FunPLS basis functions, we utilize B-spline basis functions that are detailed in de Boor (2001). Let $\boldsymbol{B}(t) = (B_1(t), \ldots, B_{M+d}(t))^{\mathrm{T}}$ be a vector that contains $M + d$ B-spline basis functions. Each basis function is defined on $[0, T]$ with degree d and $M + 1$ equally spaced knots $0 = t_0 < t_1 < \cdots < t_M = T$, which is a piecewise polynomial of degree d. Let $\mathcal{S}_{dM}$ denote the linear space spanned by the B-spline basis functions $\{B_j(t) : j = 1, \ldots, M + d\}$. In (3.6), we can approximate $w(t)$ and $w_j(t)$ by $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{B}(t)$ and $\boldsymbol{b}_j^{\mathrm{T}}\boldsymbol{B}(t)$ with $\boldsymbol{b} = (b_1, \ldots, b_{M+d})^{\mathrm{T}}$ and $\boldsymbol{b}_j = (b_{j1}, \ldots, b_{jM+d})^{\mathrm{T}}$, which are elements in $\mathcal{S}_{dM}$. Let $U_j = \int_0^T X(t)B_j(t)\,\mathrm{d}\,t$ and $U = (U_1, \ldots, U_{M+d})$, then $\mathrm{Cov}\left(\int_0^T X(t)w_j(t)\,\mathrm{d}\,t, \int_0^T X(t)w(t)\,\mathrm{d}\,t\right)$ yields $\mathrm{Cov}\,(U\boldsymbol{b}_j, U\boldsymbol{b})$ and $\|w\|_2^2$ is estimated by $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{V}_0\boldsymbol{b}$, where $\boldsymbol{V}_0$ denote the $(M + d) \times (M + d)$ matrix with the elements $v_{0ij} = \int_0^T B_i(t)B_j(t)\,\mathrm{d}\,t$. Since $\boldsymbol{V}_0$ is positive definite, we can write $\boldsymbol{V}_0 = \boldsymbol{W}_0\boldsymbol{W}_0$, where $\boldsymbol{W}_0$ is symmetric. Let $\boldsymbol{b}_* = \boldsymbol{W}_0\boldsymbol{b}$ and $\boldsymbol{b}_{*j} = \boldsymbol{W}_0\boldsymbol{b}_j$, then $\boldsymbol{b}_*$ is obtained by maximizing

$$
\begin{aligned}
&\mathrm{Cov}^2\left(Y, U\boldsymbol{W}_0^{-1}\boldsymbol{b}_*\right),\\
&\text{subject to} \quad \mathrm{Cov}\left(U\boldsymbol{W}_0^{-1}\boldsymbol{b}_{*j}, U\boldsymbol{W}_0^{-1}\boldsymbol{b}_*\right) = 0, \quad \text{for } j = 1, \ldots, k-1 \quad \text{and}\\
&\boldsymbol{b}_*^{\mathrm{T}}\boldsymbol{b}_* = 1.
\end{aligned} \tag{3.7}
$$

(3.7) is the criterion to construct the kth PLS weight vector of response Y and covariates $U\boldsymbol{W}_0^{-1}$.

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ and denote by $\boldsymbol{U}$ the $n \times (M + d)$ matrix with elements $u_{ij} = \int_0^T X_i(t)B_j(t)\,\mathrm{d}\,t$. The empirical version of (3.7) based on $\boldsymbol{Y}$ and $\boldsymbol{U}\boldsymbol{W}_0^{-1}$ can be obtained efficiently using SIMPLS (de Jong, 1993) or NIPALS (Wold, 1966) algorithms. Let $\hat{\boldsymbol{b}}_{*1}, \ldots, \hat{\boldsymbol{b}}_{*K}$ denote the first K empirical weight vectors and put $\boldsymbol{R}_K = (\hat{\boldsymbol{b}}_{*1}, \ldots, \hat{\boldsymbol{b}}_{*K})$. Then the first K empirical FunPLS scores $\boldsymbol{T} \in \mathbb{R}^{n \times K}$ is $\boldsymbol{T} = \boldsymbol{U}\boldsymbol{W}_0^{-1}\boldsymbol{R}_K$. Consider a linear regression model with response $\boldsymbol{Y}$ and covariates matrix $\boldsymbol{T}$. Let $\alpha = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$ be the corresponding coefficients. The slope function in the functional linear regression model can be estimated by

$$\hat{\beta}^{FunPLS}(t) = \hat{\alpha}^{\mathrm{T}} \boldsymbol{R}_K^{\mathrm{T}} \boldsymbol{W}_0^{-1} \boldsymbol{B}(t),$$

where $\hat{\alpha}$ is an estimate of α (e.g., ordinary least squares estimator).

To obtain a smooth estimate for the slope function, we need to tune the number of components K and the parameters of the B-spline basis functions d and M, which are used to expand the FunPLS basis functions. The degree d, which is of less importance, is fixed it to a reasonable value, i.e., $d = 3$, whereas K and M can be chosen by Cross-Validation (CV) or the information criteria, such as the Bayesian information criterion (BIC).

3.6.2 Functional Principal Component Regression

Let $G(s,t) = \mathrm{Cov}\{X(s), X(t)\}$ be the covariance function of the random process X. Assume that $\int_0^T E(X^2) < \infty$. Then we can write an orthogonal expansion of G as

$$G(s,t) = \sum_{k=1}^{\infty} \theta_k \phi_k(s) \phi_k(t),$$

where $\phi_1, \phi_2, \ldots$ are the eigenfunctions of the linear operator with kernel G and the $\theta_1 \geq \theta_2 \geq \cdots \geq 0$ are the corresponding eigenvalues. We can estimate G and expand it by

$$\begin{aligned}\hat{G}(s,t) &= \frac{1}{n} \sum_{i=1}^{n} \{(X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t))\} \\ &= \sum_{k=1}^{\infty} \hat{\theta}_k \hat{\phi}_k(s) \hat{\phi}_k(t),\end{aligned}$$

where $\bar{X}(t) = n^{-1} \sum_{i=1}^{n} X_i(t)$. The estimator for the slope function $\beta(t)$ is defined as

$$\hat{\beta}(t) = \sum_{k=1}^{p} \hat{b}_k \hat{\phi}_k(t) = \hat{\boldsymbol{b}}^{\mathrm{T}} \hat{\boldsymbol{\phi}}(t),$$

where $\hat{\boldsymbol{\phi}}(t) = (\hat{\phi}_1(t), \ldots, \hat{\phi}_p(t))^{\mathrm{T}}$, $p \le n$ is a positive integer, and $\hat{\boldsymbol{b}} = (\hat{b}_1, \ldots, \hat{b}_p)^{\mathrm{T}}$ minimizes the least squares

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{k=1}^{p} b_k \int_0^T X_i(t)\hat{\phi}_k(t)\,\mathrm{d}\,t\right)^2 .$$

Let $\boldsymbol{U}_\phi$ denote the $n \times p$ matrix with elements $u_{ij\phi} = \int_0^T X_i(t)\hat{\phi}_j(t)\,\mathrm{d}\,t$. The estimated $\hat{\boldsymbol{b}} = (\boldsymbol{U}_\phi^{\mathrm{T}}\boldsymbol{U}_\phi)^{-1}\boldsymbol{U}_\phi^{\mathrm{T}}\boldsymbol{Y}$. In practice, we penalize the roughness in principal components; see Chapter 9 in Ramsay and Silverman (2005) for more details regarding smoothing the functional principal components. The truncation number p and the smoothing parameter can be chosen by the Akaike information criterion (AIC), the Bayesian information criterion (BIC) or CV.

3.6.3 Effect of the parameter κ in Section 3.3

We conduct a simulation study to numerically investigate the effect of κ. The setting is the same as Scenario I with the functional covariates generated by a linear combination of B-spline basis functions, the signal-to-noise ratio 5 and the sample size $n = 500$. We can observe that the results based on different κ are very similar.

Table 3.4: Investigation of effect of κ on ISE$_0$ and ISE for the estimators of $\beta(t)$, and the prediction mean squared errors (PMSE) on test data. The results are obtained based on 100 simulation replications with the corresponding Monte Carlo standard deviations included in parentheses.

κ	ISE$_0$ ($\times 10^{-2}$)	ISE ($\times 10^{-2}$)	PMSE ($\times 10^{-3}$)
0.01	0.26 (0.25)	1.34 (0.34)	1.99 (0.06)
0.1	0.25 (0.25)	1.37 (0.47)	1.99 (0.06)
0.2	0.24 (0.25)	1.37 (0.47)	1.99 (0.06)
0.3	0.24 (0.24)	1.37 (0.49)	1.99 (0.06)
0.4	0.24 (0.24)	1.37 (0.50)	1.99 (0.06)
0.5	0.25 (0.24)	1.38 (0.52)	1.99 (0.06)

3.6.4 Effect of the smoothing parameter γ in Section 3.3

We conduct a simulation study to numerically investigate the effect of γ. The setting is the same as Scenario II with the functional covariates generated by a linear combination of B-spline basis functions, the signal-to-noise ratio 5 and the sample size $n = 500$. In Figure 3.6, we present the estimated slope functions by the SpaFunPLS method using different values of γ in one randomly selected simulation replicate. We can observe that small values for γ (when $\gamma = 10^{-30}$ and $\gamma = 10^{-10}$) lead to discontinuous, irregular and scattered zero subregions for estimated $\hat{\beta}(t)$, whereas relatively large values for γ (when $\gamma = 1$ and $\gamma = 10^5$) result in estimates that are not locally sparse. For γ within a certain range (e.g., $\gamma = 10^{-5}$ and $\gamma = 10^{-3}$), the proposed method has a favorable performance.

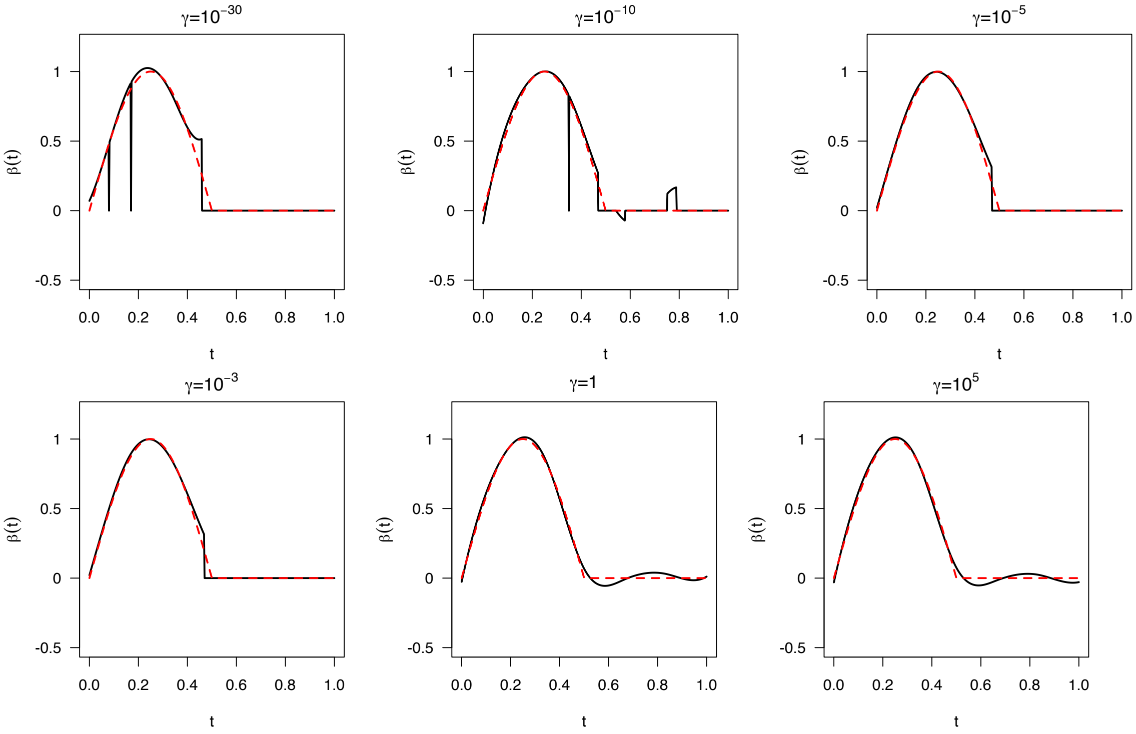

Figure 3.6: Estimated slope functions by SpaFunPLS method using different values of γ for Scenario II with $n = 500$ in one randomly selected simulation replicate.

3.6.5 Effect of the parameter δ in Section 3.3

In this subsection, we conduct a simulation study to numerically investigate the effect of δ. The setting is the same as Scenario II with the functional covariates generated by a linear combination of B-spline basis functions, the signal-to-noise ratio 5 and the sample size $n = 500$. The figure illustrates that the estimated slope functions are similar with different values for δ.

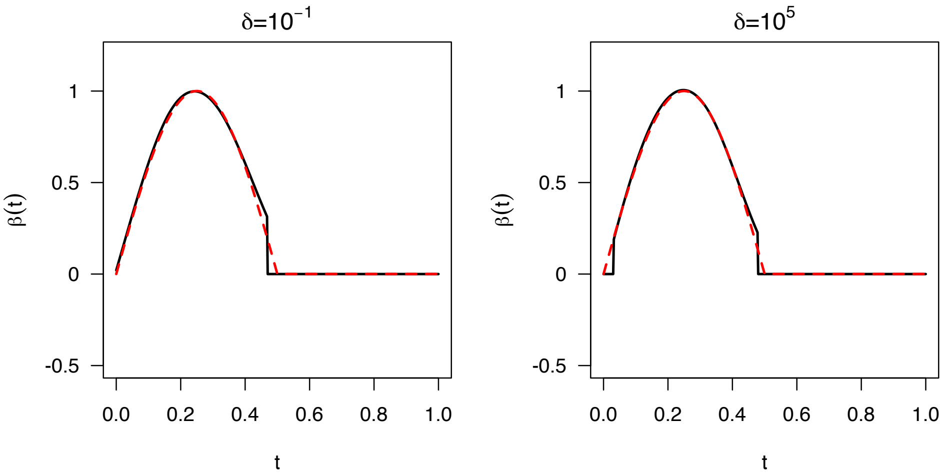

Figure 3.7: Estimated slope functions by the SpaFunPLS method using different values of δ for Scenario II with $n = 500$ in one randomly selected simulation replicate.

Chapter 4

In-game Win Probabilities for the National Rugby League

4.1 Introduction

In recent years, analytics have made a profound impact in sport where great investments have been made in the "big" professional sports of basketball (the National Basketball Association), football (the National Football league), soccer (major European leagues), hockey (the National Hockey League) and baseball (Major League Baseball). Many teams now have their own analytics staff where decisions are scrutinized in many areas of the sporting operation including strategy, drafting, salaries, player evaluation, and marketing. For a survey of some of the work that has been done in sports analytics, see Albert et al. (2017).

Whereas the National Rugby League (NRL) may be considered big sport (it has the greatest television viewership of any sport in Australia), the NRL is underrepresented in the sports analytics literature. For example, in a search of the archives of the *Journal of Quantitative Analysis in Sports* (founded in 2005), the authors were unable to find a single article devoted to rugby league. However, there have been many papers written on rugby league from the sports science perspective, and a small sample of these include Glassbrook et al. (2019), Booth and Orr (2017), Seitz et al. (2014), King et al. (2009) and Gabbett (2005).

In an attempt to grow the game, the NRL is adding an analytics focus to the sport (see `www.nrl.com.stats`). In particular, to provide additional excitement to the television viewing experience, the NRL would like to include in-game win probabilities. The idea is that such a graphic may be presented in a small corner of the screen, and be continually updated as the game circumstances change. The graphic would be appealing to the NRL fan base and also to punters. The continual update precludes highly computational techniques, and of course, the predictions of the in-game win probabilities would need to be accurate.

The NRL has provided us with four seasons of detailed event data ($2016 - 2019$) which we use to inform the in-game win probabilities. Our approach is Bayesian where our main interest concerns the evaluation of the in-game posterior win probability. The challenge is the development of an accurate model for which the posterior probability can be evaluated in real-time. The accuracy provided by the model relies on domain knowledge of the sport; hence we search for data and covariates that have high predictive capability.

The distributions that are specified in our Bayesian model are determined via functional data analysis (FDA). FDA is a relatively new branch of statistics where regression methods are extended to the study of functions ((Ramsay and Silverman, 2005); (Ferraty and Vieu, 2006)). For a practical introduction to FDA, see Ramsay et al. (2009). In sport, FDA techniques were used by Chen and Fan (2018) who investigated score differentials in basketball. FDA also has many applications in other areas. For instance, Ainsworth et al. (2011) applied FDA for ecosystem research, in which they studied the relationship between river flow and salmon abundance. Luo et al. (2013) estimated the intensity of ward admission and investigated its effect on emergency department access block in public hospitals by using some FDA methods.

The functional aspect of our problem is that a rugby league match is 80 minutes in duration and that circumstances change over the duration of the match. Therefore, the in-game win probability is a function of the time of the match. In FDA, a typical application involves the analysis of a sample of realizations from independent and identically distributed (iid) functions. A novelty in our work is that the matches are not iid, because each match is conditional on a unique kickoff win probability of the home team. We propose a weighted least squares method to estimate the functional parameters of each match by borrowing the information from matches with similar kickoff probabilities.

A key feature of our work is that the general approach for estimating in-game win probabilities may be used in any sport that has event data. Event data consists of a chronological record of well-defined events that occur during a match which are relevant to the match and are recorded with a time stamp. The necessary modifications would involve the determination of the relevant event data which is predictive and sport specific.

In Section 4.2, we begin with a discussion of the data that is at our disposal. We then outline the Bayesian model from which we obtain the in-game posterior win probability. In the Bayesian model, there are distributions that are specified via FDA methods. The FDA methodology is explained in detail. In particular, we develop novel estimation techniques to address the complexity of the problem. In Section 4.3, we consider the utilization of the event data to provide good predictions. There are many potential insights from a game that are relevant. We use the domain knowledge from the rugby league for the specification. We then demonstrate that our estimated in-game win probabilities change during a match in expected ways. In Section 4.4, we demonstrate that our estimated win probabilities are reliable. We conclude with a short discussion in Section 4.5.

4.2 Model Development

4.2.1 Available Data

The NRL consists of 16 teams. Each team plays 24 games during the regular season. The NRL has gratiously given us access to event data for the resultant 769 regular-season matches which have taken place during the four seasons $2016 - 2019$. Event data is detailed match data that goes well beyond box score data. With event data, every time an event occurs during a match (e.g. field goal, try, tackle, etc), characteristics of the event are recorded (e.g. location on the pitch, players involved, time of the match, etc). In the NRL, 2.1 events are recorded on average per second. The events and characteristics are obtained through cameras and optical recognition software that carry out the data collection process in real-time. In total, we have $8,144,905$ events obtained over the four seasons.

An important component of our work is the determination of relevant event data to inform in-game win probabilities. In our development and without loss of generality, win probabilities and data will refer to the home team. For the time being, for a particular match, we will refer to $X(t)$ as data arising from the game's event log at time $t = 1, \ldots, 80$. Note that $X(t)$ may be multivariate. For example, it is obvious that the average field position by the home team is a measure of dominance and it may be a good predictor of the home team's chance of winning the match.

Another important predictor of the in-game win probability of the home team is the current score differential. We will refer to $D(t)$ as the number of points by which the home team is defeating the road team at time t. Note that $D(t) < 0$ indicates that the road team is winning by $|D(t)|$ points at time t.

Finally, another important predictor of the in-game win probability of the home team is a measure of its strength relative to the road team. This is not something immediately available from the event data, and therefore we sourced an additional dataset. The website `http://www.aussportsbetting.com/data/historical-nrl-results-and-odds-data/` gives closing betting odds of NRL matches immediately prior to kickoff. A nice feature of the betting odds is that they take into account everything that is relevant to a match including home team advantage, injuries, travel, etc. Betting odds are also known to be efficient; otherwise sportsbooks would not exist. Therefore, we can rely on the betting odds as providing reliable information concerning the win-probability of the home team at the time of kickoff.

Betting odds arise in various formats, and we will refer to odds provided in the European format. Odds o_h on the home team indicate that a winning bet of \$1 on the home team will result in a payout of $\$o_h$. Clearly, $o_h \geq 1$. Similarly, odds o_r on the road team indicate that a winning bet of \$1 on the road team will result in a payout of $\$o_r$. We ignore the rare event that a match can end in a draw as this does not affect the subsequent calculations. Draws occur roughly 4.94% of the time in the NRL. Now, some simple probability calculations involving expectations yield that the probability of the home team winning is $p_h = 1/o_h$ and the probability of the road team winning is $p_r = 1/o_r$. However, these calculations do not take into account the vigorish (i.e. the expected profit) by the sportsbook, and therefore $p_h + p_r > 1$. We therefore remove the vigorish and set the kickoff probability that the home team wins the match as $p_0 = p_h/(p_h + p_r)$.

Therefore, to review, the inputs to our Bayesian model which we use to estimate in-game win probabilities for the home team are given by:

$$\begin{aligned}
X(t) \quad &\equiv \text{event data relative to the home team at time } t = 1, \ldots, 80, \\
D(t) \quad &\equiv \text{score differential in favour of the home team at time } t = 1, \ldots, 80, \qquad (4.1) \\
p_0 \quad &\equiv \text{kickoff probability of home team winning based on sportsbook odds.}
\end{aligned}$$

4.2.2 Model Overview

In this subsection, we present a Bayesian model based on the inputs given by (4.1). We let W denote the event that the home team wins the match, and it is the posterior probability of W which is our quantity of interest. Further, we use the notation $[A \mid B]$ to denote the generic conditional density of A given B. We therefore obtain the expression

$$\begin{aligned}\text{Prob}(W \mid X(t), D(t), p_0) &= \frac{[X(t),D(t)|W,p_0]\ \text{Prob}(W|p_0)}{[X(t),D(t)|W,p_0]\ \text{Prob}(W|p_0)\ +\ [X(t),D(t)|\overline{W},p_0]\ \text{Prob}(\overline{W}|p_0)} \qquad (4.2)\\ &= \frac{[X(t),D(t)|W,p_0]\ p_0}{[X(t),D(t)|W,p_0]\ p_0\ +\ [X(t),D(t)|\overline{W},p_0]\ (1-p_0)}\ .\end{aligned}$$

We observe that (4.2) is a simple expression. However, for the application to television broadcasts, we emphasize that it is necessary that the component distributions in (4.2) need to be evaluated instantaneously.

4.2.3 Estimation of Model Components using FDA

This is the most technical portion of this chapter where a nonstandard FDA structure is introduced and novel estimation techniques are developed to determine the probability distributions $[X(t), D(t) \mid W, p_0]$ and $[X(t), D(t) \mid \overline{W}, p_0]$ in (4.2). We illustrate the methodology with univariate $X(t)$ although the methods can be extended to multivariate $X(t)$. This subsection may be skimmed while still retaining the overall intent of the chapter.

We begin by focusing on the $[X(t), D(t) \mid W, p_0]$ term where $[X(t), D(t) \mid \overline{W}, p_0]$ is handled in a similar fashion. We assume that

$$\begin{aligned} X(t|W, p_0) &= \mu_X(t|W, p_0) + \epsilon_X(t),\\ D(t|W, p_0) &= \mu_D(t|W, p_0) + \epsilon_D(t),\end{aligned}$$

where $\mu_X(t|W, p_0)$ is the expected value of the event data $X(t)$ conditional on the home team winning and having a kickoff win probability of p_0. Similarly, $\mu_D(t|W, p_0)$ is the expected value of the score differential $D(t)$ conditional on the home team winning and having a kickoff win probability of p_0.

In Section 4.3, we consider various choices for $X(t)$ that affect the variance assumption and the resultant estimation procedure. Suppose for now that $\epsilon_X(t)$ is a random variable which consists of independent incremental contributions up to time t. Therefore, we assume that $\epsilon_X(t)$ has mean 0 and variance $t\sigma_X^2$. However, we note that the following theory may be modified to accommodate other variance assumptions such as a constant variance. With respect to $\epsilon_D(t)$, we also assume that it is based on a white noise process where we recognize that the score differential consists of incremental contributions during the match up to time t. Therefore, assuming that these contributions are independent and identically distributed, it is appropriate that $\epsilon_D(t)$ have mean 0 and variance $t\sigma_D^2$. Therefore, at time t, the noises are distributed as

$$\begin{bmatrix} \epsilon_X(t) \\ \epsilon_D(t) \end{bmatrix} \sim \text{Normal}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, t\begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_D \\ \rho\sigma_X\sigma_D & \sigma_D^2 \end{pmatrix}\right]. \tag{4.3}$$

For different time points t and t', we assume that $\epsilon_X(t)$ and $\epsilon_D(t')$ are independent.

We further assume that $\mu_X(t|W, p_0)$ and $\mu_D(t|W, p_0)$ are continuous smooth functions, and we approximate these functions as linear combinations of basis functions as follows

$$\begin{aligned} \mu_X(t|W, p_0) &= \textstyle\sum_{k=1}^K a_k(W, p_0) b_k(t), \\ \mu_D(t|W, p_0) &= \textstyle\sum_{k=1}^K c_k(W, p_0) b_k(t), \end{aligned} \tag{4.4}$$

where the $b_k(t)$ are predetermined basis functions. Up until this point, except for the variance assumptions associated with the noise terms, this is a standard setup in FDA applications (see, Chapter 3 in Ramsay and Silverman (2005), for example).

With our initial concentration on the specification of $[X(t), D(t) \mid W, p_0]$, we restrict our data to matches where the home team has won (i.e. W is observed). For a particular match i, we therefore have functional data $\{x_i(t_{ij}), d_i(t_{ij}) : i = 1, \ldots, N; j = 1, \ldots, n_i\}$ where $x_i(t_{ij})$ is the event data recorded at time t_{ij} (i.e. the realization of $X_i(t_{ij})$), and $d_i(t_{ij})$ is the score differential at time t_{ij} (i.e. the realization of $D_i(t_{ij})$). We also have the kickoff win probability p_{i0} associated with the ith match.

An aspect of our problem that makes it different from a typical FDA application is that the functional data are not iid. Specifically, the functional distribution of the ith match is conditional on p_{i0} (the kickoff win probability of the home team in the ith match). Therefore, to address the estimation of the a's and c's in (4.4), we minimize the functions

$$\begin{aligned} H_a(\mathbf{a}) &= \textstyle\sum_{i=1}^{N}\sum_{j=1}^{n_i} \dfrac{1}{t_{ij}} \left[\, x_i(t_{ij}) - \sum_{k=1}^{K} a_k(W,p_0) b_k(t)\right]^2 \, \exp\left\{\frac{-(p_0-p_{i0})^2}{\gamma}\right\} \\ H_c(\mathbf{c}) &= \textstyle\sum_{i=1}^{N}\sum_{j=1}^{n_i} \dfrac{1}{t_{ij}} \left[\, d_i(t_{ij}) - \sum_{k=1}^{K} c_k(W,p_0) b_k(t)\right]^2 \, \exp\left\{\frac{-(p_0-p_{i0})^2}{\gamma}\right\} \end{aligned} \tag{4.5}$$

where $\mathbf{a} = (a_1(W,p_0),\ldots,a_K(W,p_0))^T$, $\mathbf{c} = (c_1(W,p_0),\ldots,c_K(W,p_0))^T$, and $\gamma > 0$ is a tuning parameter. The term $\exp\{-(p_0-p_{i0})^2/\gamma\}$ assigns more weight to matches that have similar kickoff win probabilities to the generic value p_0.

The proposed estimation procedure based on the minimization of the functions H_a and H_c in (4.5) is nonstandard. However, it is motivated by least squares and maximum likelihood considerations (since the X's and D's are normally distributed). What makes the equations in (4.5) unusual is that $\mathrm{E}(X_i(t_{ij}))$ and $\mathrm{E}(D_i(t_{ij}))$ do not equal the specified parametric expressions. Equality would only exist if the x_i and d_i were observed under the generic value p_0, where again, we emphasize that the functional data are not iid. This provides the motivation for the exponential terms; we assign more weight to observations for which the generic p_0 is closer to the observed p_{i0}.

With a little bit of work, it can be shown that for fixed γ, the minimization of H_a and H_c yields the analytic expressions

$$\begin{aligned} \hat{\mathbf{a}}(p_0) &= \left(\textstyle\sum_{i=1}^{N} v_i \mathbf{B}_i^T \mathbf{G}_i \mathbf{B}_i\right)^{-1} \left(\textstyle\sum_{i=1}^{N} v_i \mathbf{B}_i^T \mathbf{G}_i \mathbf{x}_i\right), \\ \hat{\mathbf{c}}(p_0) &= \left(\textstyle\sum_{i=1}^{N} v_i \mathbf{B}_i^T \mathbf{H}_i \mathbf{B}_i\right)^{-1} \left(\textstyle\sum_{i=1}^{N} v_i \mathbf{B}_i^T \mathbf{H}_i \mathbf{d}_i\right), \end{aligned} \tag{4.6}$$

where $\hat{\mathbf{a}}(p_0) = (\hat{a}_1(W,p_0),\ldots,\hat{a}_K(W,p_0))^T$, $\hat{\mathbf{c}}(p_0) = (\hat{c}_1(W,p_0),\ldots,\hat{c}_K(W,p_0))^T$, $v_i = v_i(p_0) = \exp\{-(p_0-p_{i0})^2/\gamma\}$, $\mathbf{B}_i$ is the $n_i \times K$ matrix with (j,k)th element $b_k(t_{ij})$, $\mathbf{G}_i$ and $\mathbf{H}_i$ are $n_i \times n_i$ diagonal matrices with the jth diagonal element $1/t_{ij}$, $\mathbf{x}_i = (x_i(t_{i1}),\ldots,x_i(t_{in_i}))^T$ and $\mathbf{d}_i = (d_i(t_{i1}),\ldots,d_i(t_{in_i}))^T$.

With estimated vectors $\hat{\mathbf{a}}$ and $\hat{\mathbf{c}}$, we now turn to more traditional estimation procedures. Let $\hat{a}_{ik} = \hat{a}_k(W, p_{i0})$ and $\hat{c}_{ik} = \hat{c}_k(W, p_{i0})$. Based on the data and the modelling assumption (4.3), the resulting profile likelihood is given by

$$
\begin{aligned}
&L(\sigma_X, \sigma_D, \rho | W, p_0) \\
&= \prod_{i=1}^{N} \prod_{j=1}^{n_i} \frac{1}{2\pi\, t_{ij}\sigma_X\sigma_D\sqrt{1-\rho^2}} \exp\bigg\{ -\frac{1}{2(1-\rho^2)} \bigg[\frac{\Big(x_i(t_{ij}) - \sum_{k=1}^{K} \hat{a}_{ik} b_k(t_{ij})\Big)^2}{t_{ij}\sigma_X^2} - \\
&\frac{2\rho \Big(x_i(t_{ij}) - \sum_{k=1}^{K} \hat{a}_{ik} b_k(t_{ij})\Big) \Big(d_i(t_{ij}) - \sum_{k=1}^{K} \hat{c}_{ik} b_k(t_{ij})\Big)}{t_{ij}\sigma_X\sigma_D} + \frac{\Big(d_i(t_{ij}) - \sum_{k=1}^{K} \hat{c}_{ik} b_k(t_{ij})\Big)^2}{t_{ij}\sigma_D^2} \bigg] \bigg\}.
\end{aligned}
$$

The profile likelihood can then be maximized to provide estimates

$$
\begin{aligned}
\hat{\sigma}_X^2 &= D_{xx} \,/\, v_0, \\
\hat{\sigma}_D^2 &= D_{dd} \,/\, v_0, \\
\hat{\rho} &= D_{xd} \,/\, \sqrt{D_{xx} D_{dd}}\ ,
\end{aligned}
$$

where

$$
\begin{aligned}
v_0 &= \sum_{i=1}^{N} n_i, \\
D_{xx} &= \sum_{i=1}^{N} \sum_{i=1}^{n_i} \Big(x_i(t_{ij}) - \sum_{k=1}^{K} \hat{a}_{ik} b_k(t_{ij})\Big)^2 / t_{ij} \\
&= \sum_{i=1}^{N} \Big(\mathbf{x}_i - \mathbf{B}_i \hat{\mathbf{a}}_i\Big)^T \mathbf{G}_i \Big(\mathbf{x}_i - \mathbf{B}_i \hat{\mathbf{a}}_i\Big), && (4.7) \\
D_{xd} &= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left(x_i(t_{ij}) - \sum_{k=1}^{K} \hat{a}_{ik} b_k(t_{ij})\right) \left(d_i(t_{ij}) - \sum_{k=1}^{K} \hat{c}_{ik} b_k(t_{ij})\right) / t_{ij} \\
&= \sum_{i=1}^{N} \Big(\mathbf{x}_i - \mathbf{B}_i \hat{\mathbf{a}}_i\Big)^T \sqrt{\mathbf{H}_i} \sqrt{\mathbf{G}_i} \Big(\mathbf{d}_i - \mathbf{B}_i \hat{\mathbf{c}}_i\Big), && (4.8) \\
D_{dd} &= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left(d_i(t_{ij}) - \sum_{k=1}^{K} \hat{c}_{ik} b_k(t_{ij})\right)^2 / t_{ij} \\
&= \sum_{i=1}^{N} \Big(\mathbf{d}_i - \mathbf{B}_i \hat{\mathbf{c}}_i\Big)^T \mathbf{H}_i \Big(\mathbf{d}_i - \mathbf{B}_i \hat{\mathbf{c}}_i\Big),
\end{aligned}
$$

with $\hat{\mathbf{a}}_i = (\hat{a}_{i1}, \ldots, \hat{a}_{iK})^T$ and $\hat{\mathbf{c}}_i = (\hat{c}_{i1}, \ldots, \hat{c}_{iK})^T$. Finally, the parameter γ is tuned as described in Section 4.3.

Putting this all together, suppose that there is a new match l with kickoff probability p_{l0}, and we observe event data $x_l(t)$ and score differential $d_l(t)$ at time t. Then

$$[x_l(t), d_l(t)|W, p_{l0}] = \frac{1}{2\pi \; t\hat{\sigma}_X \hat{\sigma}_D \sqrt{1-\hat{\rho}^2}} \exp\Bigg\{ -\frac{1}{2(1-\hat{\rho}^2)} \Bigg[\frac{\Big(x_l(t) - \sum_{k=1}^K \hat{a}_k(W, p_{l0}) b_k(t)\Big)^2}{t\hat{\sigma}_X^2} - \frac{2\hat{\rho}\Big(x_l(t) - \sum_{k=1}^K \hat{a}_k(W, p_{l0}) b_k(t)\Big)\Big(d_l(t) - \sum_{k=1}^K \hat{c}_k(W, p_{l0}) b_k(t)\Big)}{t\hat{\sigma}_X \hat{\sigma}_D} + \frac{\Big(d_l(t) - \sum_{k=1}^K \hat{c}_k(W, p_{l0}) b_k(t)\Big)^2}{t\hat{\sigma}_D^2} \Bigg] \Bigg\}.$$

Similarly, we can obtain $[x_l(t), d_l(t)|\overline{W}, p_{l0}]$. Then using (4.2), we can simply estimate the posterior in-game win probability at time t for match l.

4.3 Results

We begin by considering appropriate choices for the event data $X(t)$. When a game is being viewed, there are often indications that one of the teams is gaining an upper hand in the match. The variable $X(t)$ is chosen to quantitatively reflect this sort of dominance as a predictor of winning the match. In Table 4.1, we propose several choices that are intended to reflect dominance by the home team. All of the variables presented in Table 4.1 are recorded with respect to the home team.

Table 4.1: Potential choices of event data where all variables are measured with respect to the home team and larger values denote increasing superiority.

Event Data	Description
$X_1(t)$	tackle differential up to time t
$X_2(t)$	tackle differential during the most recent 10 minutes
$X_3(t)$	missed tackle differential up to time t
$X_4(t)$	missed tackle differential during the most recent 10 minutes

Now, we are not suggesting that the variables proposed in Table 4.1 are the best choices. For clarity, a missed tackle is one where a player on the team of interest may have been tackled, but the tackle was unsuccessful. Therefore, the missed tack differential with respect to the home team is favourable to the home team if the variable is positive. For example, Parmar et al. (2017) investigate key performance indicators in professional rugby league. However, the variables in Table 4.1 are easy to calculate based on live match data. We imagine that experts with detailed domain knowledge of the rugby league may be able to propose improved variables from the point of view of prediction. However, to illustrate the proposed methods, we will hereafter use the variable $X_3(t)$ in Table 4.1 as the event data of interest. For ease of notation, we denote $X_3(t)$ as $X(t)$.

We also emphasize that the choice of the event data impacts the modelling distribution (4.3) and the estimation equations given by (4.5), (4.6), (4.7), and (4.8). For example, the noise terms $\epsilon_X(t)$ associated with X_2 and X_4 in Table 4.1 have constant variances. In this case, $\mathbf{G}_i = I_{n_i}$, where I_{n_i} is an $n_i \times n_i$ identity matrix. On the other hand, X_1 and X_3 lead to noise variances that are proportional to t, where $\mathbf{G}_i$ is the $n_i \times n_i$ matrix with the jth diagonal element $1/t_{ij}$.

The basis functions $b_k(t)$ introduced in (4.4) are cubic B-splines. For details on B-spline approximation, see de Boor (2001). Specifically, we choose 9 equally spaced knots over the interval $[0, 80]$ minutes and this results in $K = 11$ cubic basis functions as depicted in Figure 4.1. This selection of knots and splines leads to flexible shapes that can be used to express $\mu_X(t|W, p_0)$ and $\mu_D(t|W, p_0)$ in (4.4).

Before proceeding to estimation, it is good to have a sense of the data. We exclude 38 matches that ended in draws from the 769 regular-season matches which have taken place during the four seasons $2016 - 2019$. In Table 4.2 and 4.3, we provide descriptive statistics of data collected from the remaining 731 NRL regular season matches from $2016-2019$. We observe that there is indeed a home-field advantage as the average score differential in favour of the home team is 1.8 points. We also observe that the average missed tackle differential is positive which is also evidence of the home team advantage. The score differential curves and the missed tackle differential curves for the 731 matches are plotted in Figure 4.2 and 4.3 respectively. On average, it seems that both the differential and missed tackle differential are linear with respect to the time of the match. This is consistent with a process whereby the better team separates itself from the weaker team in a consistent manner over the course of a match.

Having specified the basis functions, the procedure in Section 4.2.3 first requires the estimation of the parameters σ_X, σ_D and ρ as specified in the multivariate normal distribution (4.3). We first

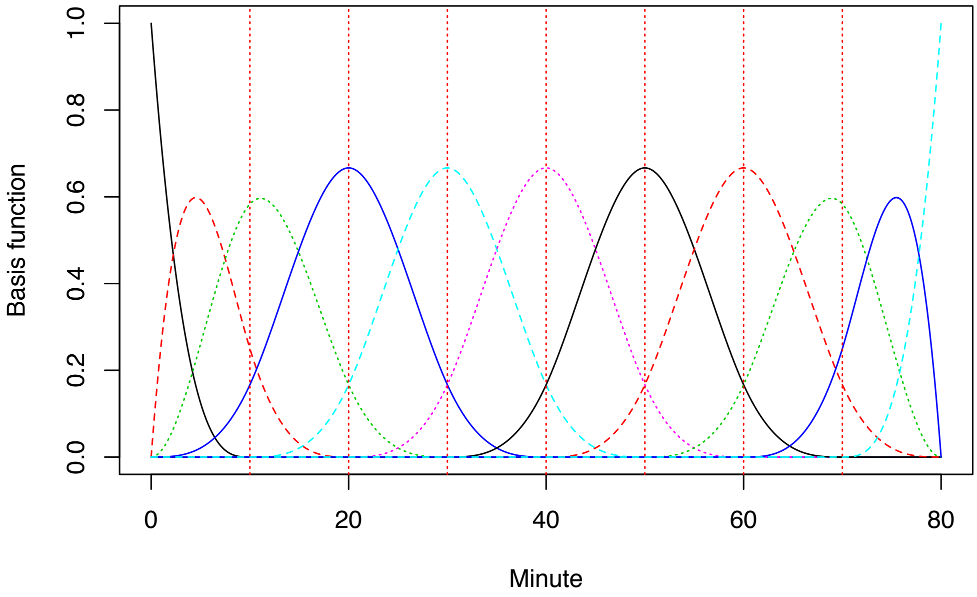

Figure 4.1: Cubic B-spline basis functions defined on 9 equally spaced knots over the interval $[0, 80]$ minutes.

Table 4.2: Descriptive statistics of the scores corresponding to the 731 matches from the four regular seasons $(2016 - 2019)$ of the NRL.

Variable	Min Value	Max Value	Average	Std Dev
Home Team Score	0	64	21.1	10.8
Road Team Score	0	62	19.2	10.1
Score Differential wrt Home Team	-62	58	1.8	16.9

restrict estimation to data where the home team has won (i.e. W) and we note that there are 311 matches corresponding to the training data $(2016 - 2018$ seasons) that fit this criterion. Based on the specification of the tuning parameter $\gamma = 0.01$, the chosen basis functions and the determination of the a_k and c_k terms, we obtain

$$\hat{\sigma}_X = 0.90,$$

$$\hat{\sigma}_D = 1.39,$$

$$\hat{\rho} = 0.40\ .$$

Table 4.3: Descriptive statistics of the missed tackles corresponding to the 731 matches from the four regular seasons ($2016-2019$) of the NRL.

Variable	Min Value	Max Value	Average	Std Dev
Home Team Missed Tackle	8	48	23.5	6.7
Road Team Missed Tackle	8	44	22.5	6.2
Missed Tackle Differential wrt Home Team	-31	34	1.0	9.6

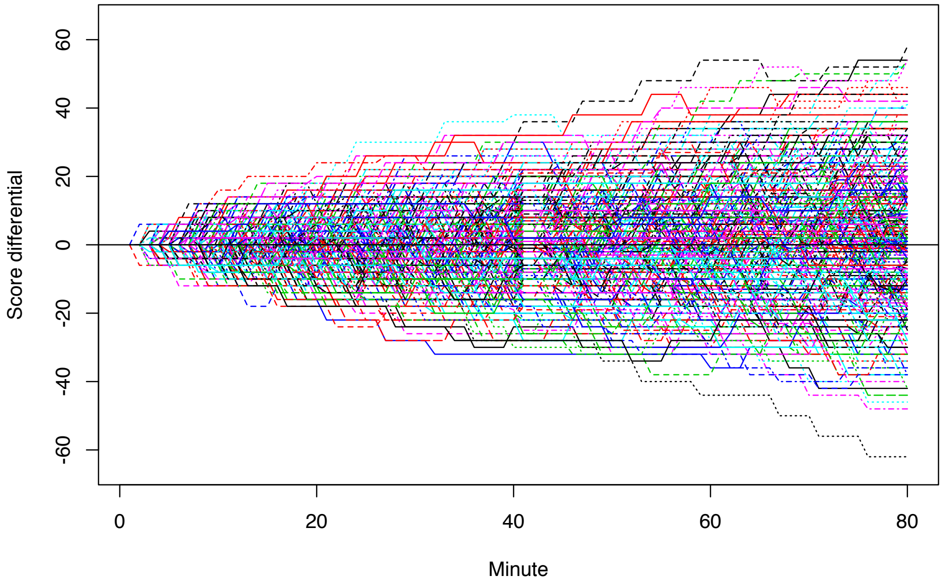

Figure 4.2: The score differential curves for the 731 matches.

These estimates appear to be sensible in terms of the descriptive statistics provided in Table 4.2 and 4.3. In particular, we note a positive correlation $\hat{\rho}$ which suggests that $X(t)$ and $D(t)$ tend to work in tandem.

Using the training data ($2016-2018$ seasons) where the home team has not won (i.e. $\overline{W}$), there are 241 matches, and we similarly obtain

$$\hat{\sigma}_X = 0.84,$$

$$\hat{\sigma}_D = 1.30,$$

$$\hat{\rho} = 0.40\ .$$

Our estimation procedure involves a tuning parameter γ. We select the tuning parameter by fivefold Cross-Validation. Specifically, we randomly split the $2016-2018$ season matches into

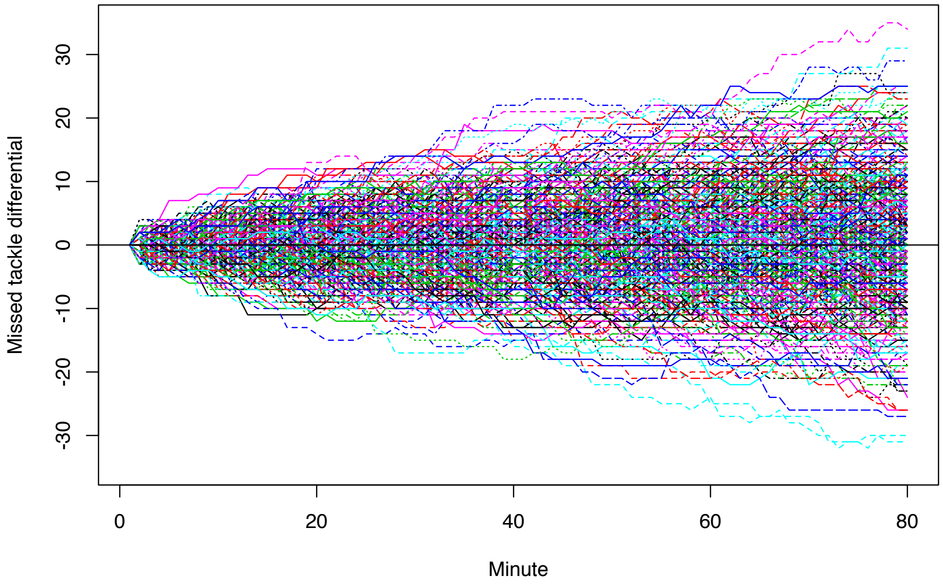

Figure 4.3: The missed tackle differential curves for the 731 matches.

five groups. For each unique group, we take it as a hold out data set and fit the model parameters $(a, c, \sigma_X, \sigma_D, \rho)$ as described above for a particular γ on the remaining groups. We then estimated home team win probabilities for the hold out data set at time t. If, for a given match at time t, the estimated winning probability is larger (smaller) than 0.5 and the home team eventually won (lost) the match, then the prediction was considered to be correct. We repeated this procedure over all matches in the test set and all times t to give the overall rate of correct predictions. The choice $\gamma = 0.01$ yielded the highest average overall rate of correct predictions over all the five hold out groups. In Figure 4.4, we show the estimated mean functions of the X and D processes based on $\gamma = 0.01$ with various p_0. We observe that the plots exhibit the expected behaviors. For example, in matches where the home team wins, mean differentials in both X and D increase as the game progresses. When a curve is wiggly, we attribute this to lack of data. For example in the upper right plot where $p_0 = 0.8$, there are not many matches where the home team is heavily favored and they lose.

4.4 Model Validation

Obviously, there is a random component to sport and this is part of its appeal. If matches were perfectly predictable, then there would be no point in holding sporting competitions. Therefore, our

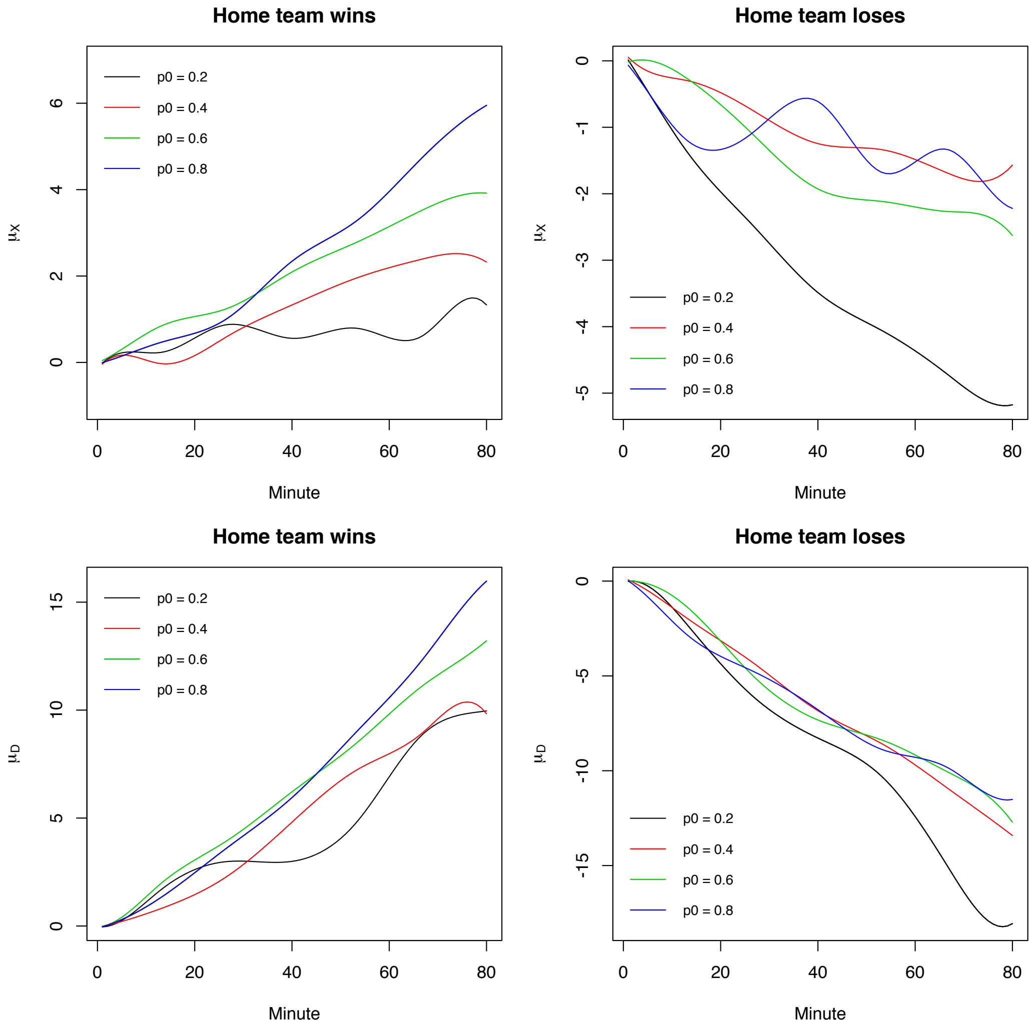

Figure 4.4: Upper left: estimated $\hat{\mu}_X(t|W, p_0)$. Upper right: estimated $\hat{\mu}_X(t|\overline{W}, p_0)$. Lower left: estimated $\hat{\mu}_D(t|W, p_0)$. Lower right: estimated $\hat{\mu}_D(t|\overline{W}, p_0)$.

investigation in this section involves an assessment of whether our predictions are reasonable - they cannot and should not be perfect predictions.

We should not use the same data to both fit models and carry out the model assessment. We therefore fit our model using the first three seasons $2016 - 2018$ of the event data, and we use the fitted model to predict outcomes for the 2019 season for which there are 179 matches. We then compare the actual 2019 match outcomes with the predicted outcomes.

In Figure 4.5, we investigate the predictive capability of our method. We consider the estimated probability that the home team wins at times $t = 1, \ldots, 75$ for the 2019 data. It is sensible to only consider predictions up to the 75th minute as many sportsbooks terminate in-match betting towards the end of matches. A reason for this is that possession of the ball near the end of a close match is critical and becomes more important than both X and D in the determination of fair betting odds. Punters could exploit this situation. If a probability exceeds 0.5, then this indicates a prediction in favour of the home team. At time t, we compare the 2019 match predictions with the actual match results, and obtain the correct prediction rate. As one would expect, Figure 4.5 demonstrates that the correct prediction rate improves as matches progress in time. We observe that our method yields good results exceeding 80% accuracy by the 55th minute.

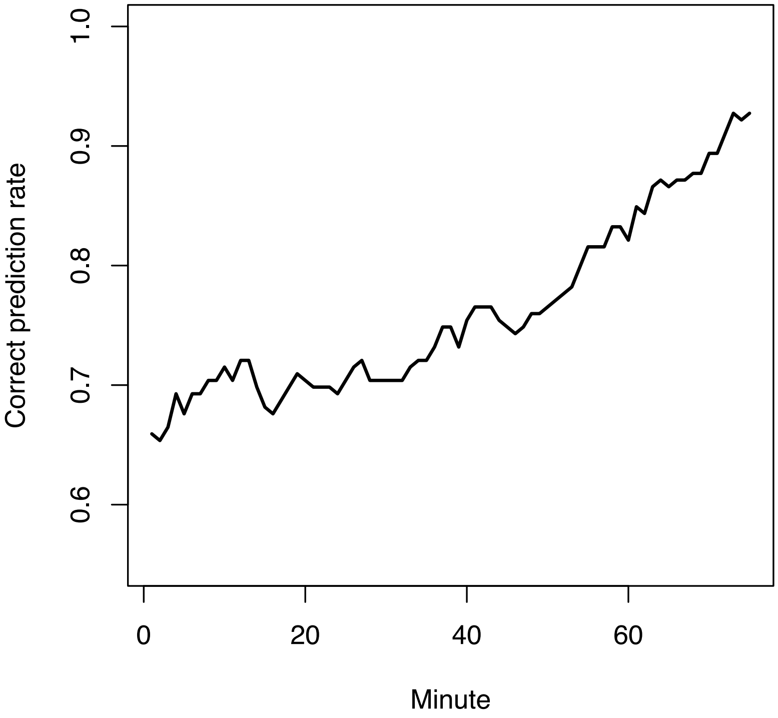

Figure 4.5: The correct prediction rate for the 2019 NRL season.

To investigate whether our estimated in-game win probabilities are reliable, we randomly selected four matches from the 2019 season where the home team won. In Figure 4.6, the solid curves are the predicted in-game win probabilities annotated with scoring events (dashed vertical lines). Sensibly, we observe that the predicted win probabilities are impacted by scoring (discontinuous jumps) and that match outcomes have more extreme probabilities near the 80th minute. Similarly, Figure 4.7 shows the predictions of the in-game win probabilities of four randomly selected matches from the 2019 season where the home team lost. Figure 4.7 also shows patterns that correspond to common sense.

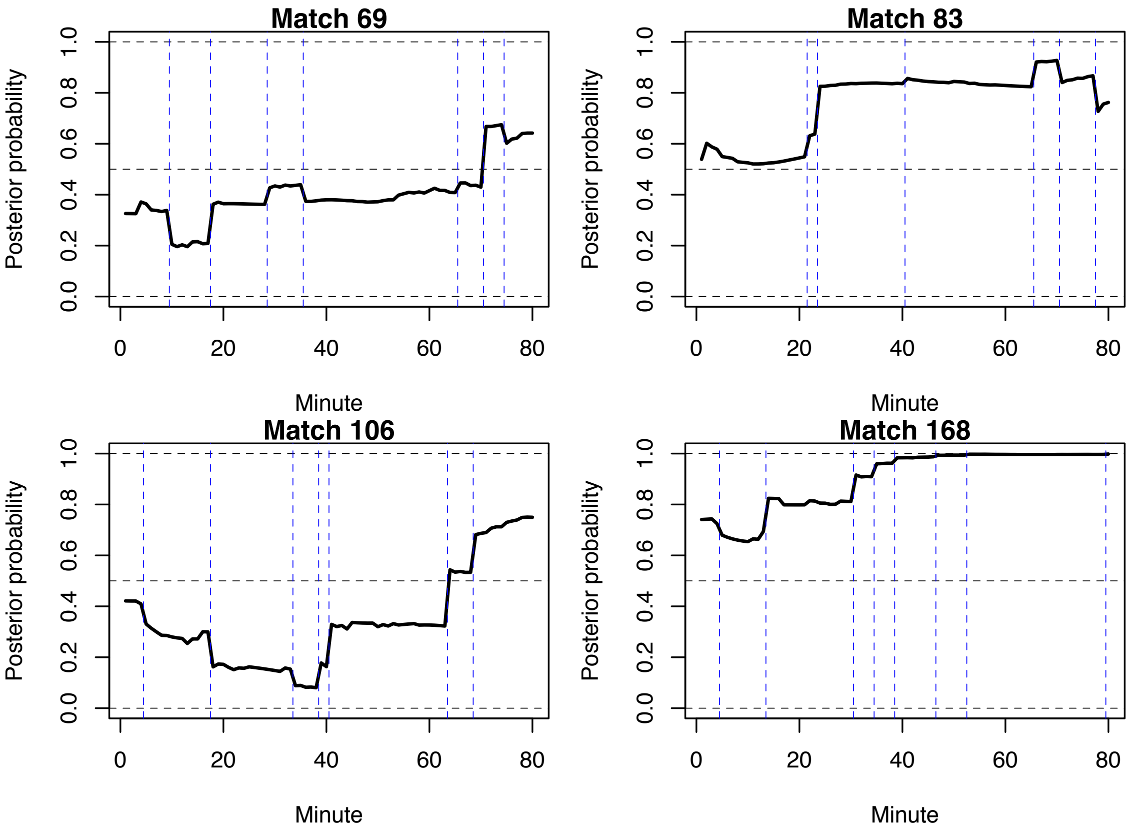

Figure 4.6: Predicted instantaneous in-game win probabilities for four randomly selected matches from the 2019 season where the home team won. The dashed horizontal lines indicate the values of 0, 0.5, and 1. The dashed vertical lines indicate the times when the score changed.

To see how $X(t)$ impacts the estimation procedure, we consider two scenarios. Scenario I predicts the in-game win probabilities using both the event data X and the score differential D. Scenario II predicts the in-game win probabilities using only the score differential D. We select a match played on 6th April 2019 between the Melbourne Storm (home) and the Canterbury-Bankstown Bulldogs. The half time score is 6 (Storm) - 12 (Bulldogs) and the full time score is 18 (Storm) - 16 (Bulldogs). More details about the match can be found at `https://www.nrl.com/draw/nrl-premiership/2019/round-4/storm-v-bulldogs/`.

In Figure 4.8, we present the predicted instantaneous in-game win probabilities for the match under Scenario I and Scenario II together with the score differentials. In Figure 4.9, the in-game win probabilities predicted by Scenario I and Scenario II are displayed together with the missed

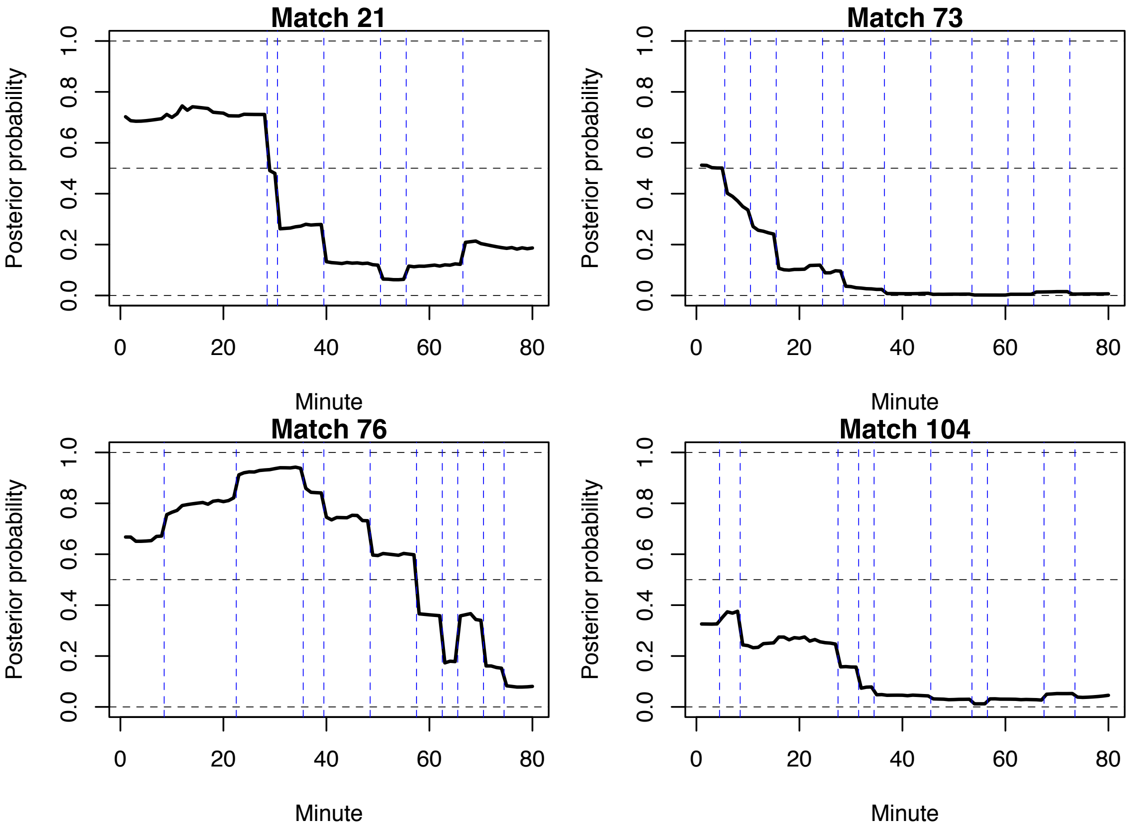

Figure 4.7: Predicted instantaneous in-game win probabilities for four randomly selected matches from the 2019 season where the home team lost. The dashed horizontal lines indicate the values of 0, 0.5, and 1. The dashed vertical lines indicate the times when the score changed.

tackle differential. In Figures 4.8 and 4.9, the solid curves represent the predictions based on both X and D, and the dotted curves represent the predictions based on D only. The kickoff probability $p_0 = 0.85$ indicates that the Storm were heavily favored. We can see from Figure 4.8 that the road team scored on the 6th minute of the match, and after that, the predicted in-game probabilities based only on D quickly decreased to around 0.6. In contrast, Figure 4.9 shows that the missed tackle differences keep positive for most of the time in the first half of the match. This indicates that even though the Storm were trailing, there was reason to be hopeful that they would turn the match around. We observe that the predicted in-game probabilities based on Scenario I are greater than those based on Scenario II for the entire game except for the short time interval between the 24th and 32nd minute. Clearly, the example demonstrates the added value in the event data $X(t)$ through the superiority of Scenario I over Scenario II.

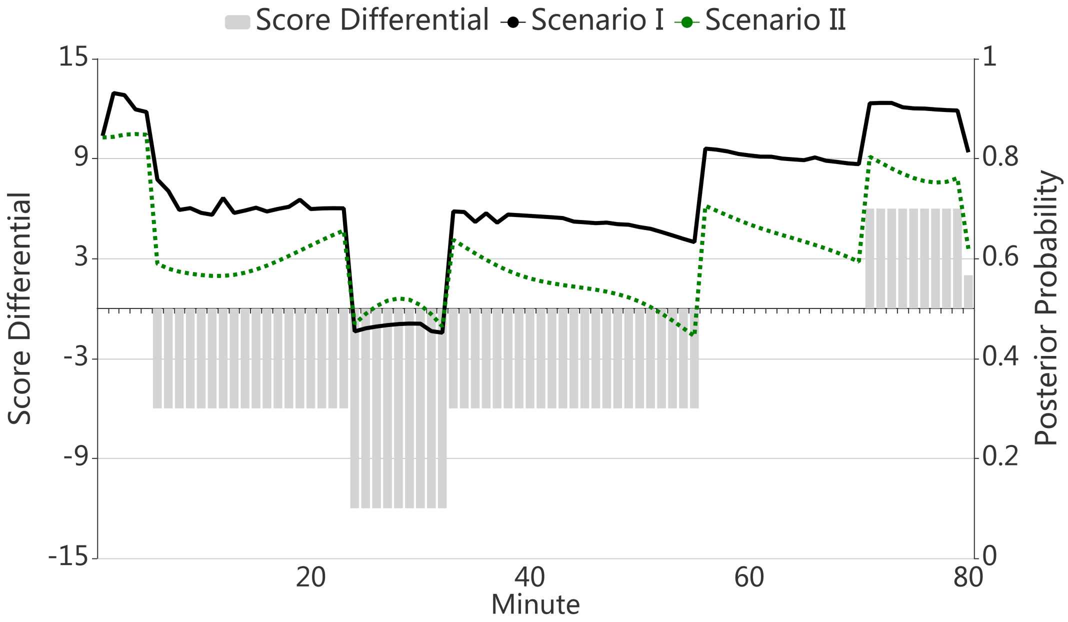

Figure 4.8: Predicted instantaneous in-game win probabilities for the match Storm versus Bulldogs on 6th April 2019 by Scenario I (———) and Scenario II (··········). The grey bars indicate the score differentials of the match.

4.5 Conclusion

We have developed a Bayesian model that provides instantaneous in-game win probabilities for the National Rugby League. The model has distributional components that are informed by novel FDA estimation techniques.

There are various future research directions associated with our work. First, the approach is general and is applicable to other sports whenever suitable event data $X(t)$ are available. Second, there are obvious gambling questions that may be explored with respect to our predictions. Finally, the choice of event data $X(t)$ impacts our estimation procedure, and we have focused on missed tackle differential. We believe that experts with detailed domain knowledge of rugby league may be able to propose better predictive choices for $X(t)$. Although we illustrate the use of univariate $X(t)$, our methods can be extended to multivariate settings.

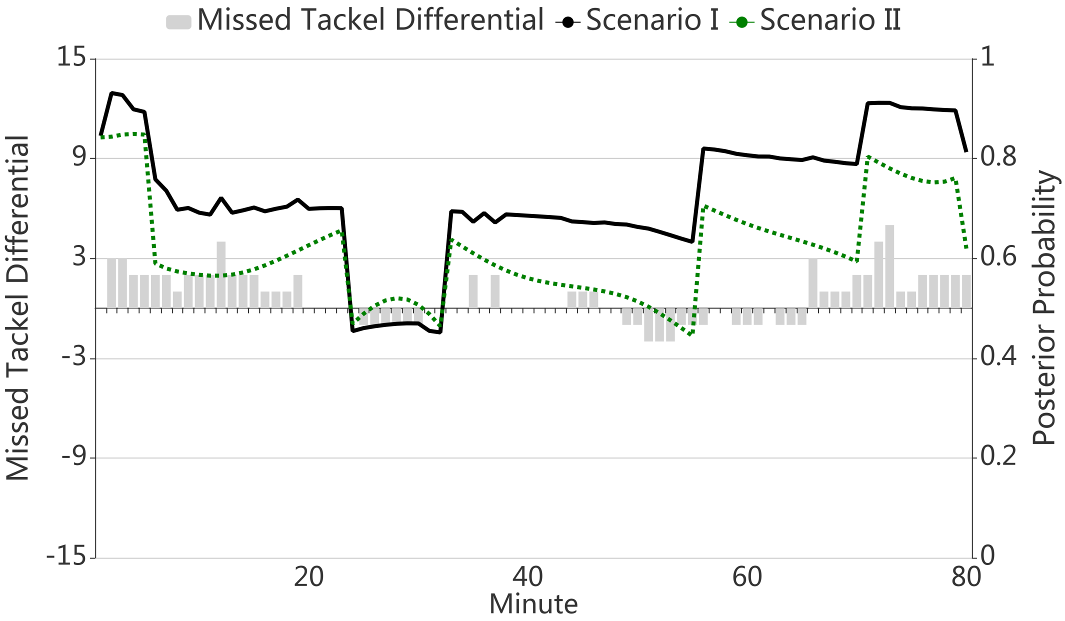

Figure 4.9: Predicted instantaneous in-game win probabilities for the match Storm versus Bulldogs on 6th April 2019 by Scenario I (———) and Scenario II (··········). The grey bars indicate the missed tackle differentials of the match.

Chapter 5

Summary and Future Work

With the development in modern technology and advances in data collection methods, FDA has become increasingly important in statistical research. In this thesis, we mainly focus on developing some new methods and models in functional data analysis (FDA).

A new nested group bridge approach to estimate a scalar-on-function truncated linear regression model was proposed in Chapter 2. Based on the B-spline basis expansion and penalized least squares with a roughness penalty, we proposed a new nested group bridge method to specifically shrink the tail region of the estimated function. The proposed nested group bridge estimator of the cutoff time is consistent. Moreover, compared with the truncation methods by Hall and Hooker (2016), we found that the proposed nested group bridge estimator of the slope function is smooth and continuous. The nested group bridge method can be generalized to other statistical contexts. For example, motivated by a Hong Kong horse racing dataset, we are currently studying a conditional logit regression with functional covariates via a nested group bridge method. The study aims to correctly predict the winner of a race and identify the cutoff time beyond which the longitudinal trajectory of the past rank of the horse has no prediction power.

Although the partial least squares method has been applied in the functional data context, little has been done on the locally sparse modelling of the functional partial least squares. In Chapter 3, we proposed a sparse version of the functional partial least squares regression. We estimated a locally sparse slope function in the functional linear regression model based on sparse functional partial least squares bases. The proposed method imposes sparsity in the dimension reduction stage, which simultaneously performs dimension reduction and locally sparse modelling. We applied the proposed method to quickly determine the relative proportions of rot, bark and sound wood in oriented strand board (OSB) fines samples from their Vis/NIR (visible and near-infrared) spectra, which offers an opportunity for commercial mills to identify problems associated with rot in logs, debarking inefficiency, and species variability.

Chapter 4 addresses the modelling of the in-game win probabilities for the National Rugby League. We developed a Bayesian model with the distributions determined based on FDA. To deal with the non-iid matches, a weighted least squares method was proposed to estimate the functional parameters of each match by borrowing the information from matches with similar kickoff probabilities. Our proposed predictions can be done in real-time and the proposed methods are applicable to other sports. In this work, we used univariate $X(t)$. A future direction is to generalize our method to the multivariate settings.

Bibliography

Ainsworth, L., R. Routledge, and J. Cao (2011). Functional data analysis in ecosystem research: the decline of Oweekeno Lake sockeye salmon and Wannock River flow. *Journal of Agricultural, Biological, and Environmental Etatistics 16*(2), 282–300.

Albert, J., M. E. Glickman, T. B. Swartz, and R. H. Koning (2017). *Handbook of Statistical Methods and Analyses in Sports*. Boca Raton: Chapman and Hall/CRC.

Aldous, D. J. (1976). A characterisation of hilbert space using the central limit theorem. *Journal London Mathematical Society 14*(2), 376–380.

Asencio, M., G. Hooker, and H. O. Gao (2014). Functional convolution models. *Statistical Modelling 14*(4), 315–335.

Booth, M. and R. Orr (2017). Time-loss injuries in sub-elite and emerging rugby league players. *Journal of Sports Science and Medicine 16*(2), 295–301.

Boulesteix, A.-L. and K. Strimmer (2006). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics 8*(1), 32–44.

Cai, T. T. and P. Hall (2006). Prediction in functional linear regression. *The Annals of Statistics 34*(5), 2159–2179.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics 35*(6), 2313–2351.

Cardot, H., F. Ferraty, and P. Sarda (2003). Spline estimators for the functional linear model. *Statistica Sinica 13*, 571–591.

Chen, T. and Q. Fan (2018). A functional data approach to model score difference process in professional basketball games. *Journal of Applied Statistics 45*(1), 112–127.

Chun, H. and S. Keleş (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(1), 3–25.

Claeskens, G., T. Krivobokova, and J. D. Opsomer (2009). Asymptotic properties of penalized spline estimators. *Biometrika 96*(3), 529–544.

Clark, N., M. Gautam, W. Wayne, D. Lyons, G. Thompson, and B. Zielinska (2007). Heavy-duty vehicle chassis dynamometer testing for emissions inventory, air quality modeling, source apportionment and air toxics emissions inventory: E55/59 all phases. Technical report, Coordinating Research Council, Alpharetta.

Cook, R. D. and L. Forzani (2019). Partial least squares prediction in high-dimensional regression. *The Annals of Statistics 47*(2), 884–908.

Dauxois, J., A. Pousse, and Y. Romain (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis 12*(1), 136–154.

de Boor, C. (2001). *A practical Guide to Splines*. New York: Springer-Verlag.

de Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems 18*(3), 251–263.

Delaigle, A. and P. Hall (2012a). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(2), 267–286.

Delaigle, A. and P. Hall (2012b). Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics 40*(1), 322–352.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*(2), 407–499.

Escabias, M., A. M. Aguilera, and M. J. Valderrama (2007). Functional PLS logit regression model. *Computational Statistics and Data Analysis 51*(10), 4891–4902.

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

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer-Verlag.

Frank, L. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35*(2), 109–135.

Gabbett, T. J. (2005). Science of rugby league football: A review. *Journal of Sports Sciences 23*(9), 961–976.

Garthwaite, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association 89*(425), 122–127.

Glassbrook, D. J., T. L. Doyle, J. A. Alderson, and J. T. Fuller (2019). The demands of professional rugby league match-play: a meta-analysis. *Sports Medicine-Open 5*(24).

Guan, T., Z. Lin, and J. Cao (2020). Estimating truncated functional linear models with a nested group bridge approach. *Journal of Computational and Graphical Statistics*, DOI: 10.1080/10618600.2020.1713797. URL: https://doi.org/10.1080/10618600.2020.1713797.

Hall, P. and G. Hooker (2016). Truncated linear models for functional data. *Journal of Royal Statistical Society, Series B (Statistical Methodology) 78*(3), 637–653.

Hall, P. and J. L. Horowitz (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics 35*(1), 70–91.

Hastie, T. and C. Mallows (1993, May). A discussion of "A statistical view of some chemometrics regression tools". *Technometrics 35*(2), 140–143.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scandinavian Journal of Statistics 17*(2), 97–114.

Hsing, T. and R. Eubank (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Chichester: Wiley.

Huang, J., S. Ma, H. Xie, and C. H. Zhang (2009). A group bridge approach for variable selection. *Biometrika 96*(2), 339–355.

James, G. M., J. Wang, and J. Zhu (2009). Functional linear regression that's interpretable. *The Annals of Statistics 37*(5A), 2083–2108.

King, T., D. Jenkins, and T. Gabbett (2009). A time–motion analysis of professional rugby league match-play. *Journal of Sports Sciences 27*(3), 213–219.

Kokoszka, P. and M. Reimherr (2017). *Introduction to Functional Data Analysis*. Boca Raton: Chapman and Hall/CRC.

Krämer, N. and M. Sugiyama (2011). The degrees of freedom of partial least squares regression. *Journal of the American Statistical Association 106*(494), 697–705.

Li, Y. and T. Hsing (2007). On rates of convergence in functional linear regression. *Journal of Multivariate Analysis 98*, 1782–1804.

Lin, Z., J. Cao, and L. W. . H. Wang (2017). Locally sparse estimator for functional linear regression models. *Journal of Computational and Graphical Statistics 26*(2), 306–318.

Luo, W., J. Cao, M. Gallagher, and J. Wiles (2013). Estimating the intensity of ward admission and its effect on emergency department access block. *Statistics in Medicine 32*(15), 2681–2694.

Martens, H. and T. Næs (1992). *Multivariate calibration*. New York: John Wiley & Sons.

Marx, B. D. and P. H. C. Eilers (1999). Generalized linear regression on sampled signals and curves: A *P*-spline approach. *Technometrics 41*(1), 1–13.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application 2*(1), 321–359.

Parmar, N., N. James, M. Hughes, H. Jones, and G. Hearne (2017). Team performance indicators that predict match outcome and points difference in professional rugby league. *International Journal of Performance Analysis in Sport 17*(6), 1044–1056.

Preda, C. and G. Saporta (2005). PLS regression on a stochastic process. *Computational Statistics and Data Analysis 48*, 149–158.

Preda, C., G. Saporta, and C. Lévéder (2007). PLS classification of functional data. *Computational Statistics 22*(2), 223–235.

Ramsay, J. O., G. Hooker, and S. Graves. (2009). *Functional Data Analysis with R and MATLAB*. New York: Springer.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. New York: Springer.

Reiss, P. T. and R. T. Ogden (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association 102*(479), 984–996.

Schwartz, W. R., A. Kembhavi, D. Harwood, and L. S. Davis (2009). Human detection using partial least squares analysis. *2009 IEEE 12th International Conference on Computer Vision*, 24–31.

Seitz, L. B., M. Rivière, E. S. De Villarreal, and G. G. Haff (2014). The athletic performance of elite rugby league players is improved after an 8-week small-sided game training intervention. *Journal of Strength and Conditioning Research 28*(4), 971–975.

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

Wang, H. and B. Kai (2015). Functional sparsity: global vesus local. *Statistica Sinica 25*, 1337–1354.

Wang, J.-L., J.-M. Chiou, and H.-G. Müller (2016). Review of functional data analysis. *Annual Review of Statistics and Its Application 3*, 257–295.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, 391–420.

Wold, H. (1975). Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. *Journal of Applied Probability 12*(S1), 117–142.

Yao, F., H.-G. Müller, and J.-L. Wang (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics 33*(6), 2873–2903.

Yuan, M. and T. T. Cai (2010). A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics 38*(6), 3412–3444.

Zhou, J., N.-Y. Wang, and N. Wang (2013). Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica 23*, 25–50.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.