REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

NOVEL APPROACHES TO REGRESSION MODELS FOR FUNCTIONAL DATA

Müge MUTİŞ

MASTER OF SCIENCE THESIS
Department of Statistics
Program of Statistics

Supervisor
Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK

Co-supervisor
Assoc. Prof. Dr. Ufuk BEYAZTAŞ

June, 2023

REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

NOVEL APPROACHES TO REGRESSION MODELS FOR FUNCTIONAL DATA

A thesis submitted by Müge MUTİŞ in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** is approved by the committee on 02.06.2023 in Department of Statistics, Program of Statistics.

Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK
Yildiz Technical University
Supervisor

Assoc. Prof. Dr. Ufuk BEYAZTAŞ
Marmara University
Co-supervisor

Approved By the Examining Committee

Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK, Supervisor
Yildiz Technical University _____

Assoc. Prof. Dr. Atıf Ahmet EVREN, Member
Yildiz Technical University _____

Assoc. Prof. Dr. Ozan KOCADAĞLI, Member
Mimar Sinan Fine Arts University _____

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of Novel Approaches to Regression Models for Functional Data supervised by my supervisor, Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Müge MUTİŞ

Signature

This study was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) Grant No: 120F270.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK, who has never withheld her interest and support since my undergraduate education, and to my co-supervisor Assoc. Prof. Dr. Ufuk BEYAZTAŞ for his guidance, support and most importantly patience throughout my graduate education. I would not be able to complete this study without his belief and encouragement.

I would like to thank the members of my thesis committee Assoc. Prof. Dr. Atıf Ahmet EVREN and Assoc. Prof. Dr. Ozan KOCADAĞLI for nice comments and opinions.

I would like to thank Prof. Dr. Han Lin SHANG and Assoc. Prof. Dr. Zaher Mundher YASEEN for their valuable contributions to the studies contained in this thesis. I would also like to thank Prof. Dr. Müjgan TEZ, whom I had the pleasure of getting to know during this process, for providing me with precious experiences.

Besides, I would like to thank TUBITAK for all its contributions to this study carried out within the scope of project 120F270.

Finally, I would like to express my special thanks to my family, who is always there for me and especially to my dear sister Müzeyyen MUTİŞ, whose presence and support I feel under all circumstances.

Müge MUTİŞ

TABLE OF CONTENTS

LIST OF ABBREVIATIONS vii

LIST OF FIGURES viii

ABSTRACT x

ÖZET xi

1 INTRODUCTION 1

1.1 Literature Review 1

1.2 Objective of the Thesis 4

1.3 Hypothesis 5

2 GENERAL FRAMEWORK FUNCTIONAL DATA ANALYSIS 6

2.1 Transforming Observed Data into Functional Form 7

2.1.1 Basis Expansion Functions 7

2.1.2 Curve Fitting 11

2.2 Functional Principal Components Analysis 13

2.3 Functional Partial Least Squares Regression 14

3 A ROBUST FUNCTIONAL LOGISTIC REGRESSION METHOD FOR CLASSIFICATION 15

3.1 Scalar on Function Logistic Regression 15

3.2 Proposed Method 17

3.2.1 Weighted Likelihood Estimation 17

3.3 Numerical Results 21

3.3.1 Monte-Carlo Experiments 21

3.3.2 Real-world Examples 27

4 FUNCTIONAL QUANTİLE AUTOREGRESSIVE REGRESSION 30

4.1 Functional Autoregressive Regression 30

4.2 Proposed Method 31

4.3 Numerical Results 32

4.3.1 Monte-Carlo Experiment 32

4.3.2 Real-world Example 34

5 RESULTS AND DISCUSSION **40**

REFERENCES **42**

PUBLICATIONS FROM THE THESIS **45**

LIST OF ABBREVIATIONS

ARIMA	Autoregressive Integrated Moving Average
AUC	Area Under the Receiver Operating Characteristic Curve
CCR	Correct Classification Ratio
FAR(1)	Functional Autoregressive Model of Order One
FDA	Functional Data Analysis
FPC	Functional Principal Components
FPCR	Functional Principal Components Regression
FLogR	Scalar-on-Function Logistic Regression
FPLS	Functional Partial Least Squares
FPLSR	Functional Partial Least Squares Regression
FQAR(1)	Functional Quantile Autoregressive Model of Order One
IMSE	Integrated Mean Squared Error
MAPE	Mean Absolute Prediction Error
MIA	Modified Index of Agreement
MLE	Maximum Likelihood Estimator
NSE	Nash-Sutcliffe Efficiency
RE	Relative Error
RFPLS	Robust Functional Partial Least Squares
RMESPE	Root Median Squared Prediction Error
RMSPE	Root Mean Squared Prediction Error
WLE	Weighted Likelihood Estimator

LIST OF FIGURES

Figure 2.1 Two example demonstrations of observed data at discrete sampling point (black circles) and their functional forms (colored lines). 6
Figure 2.2 Graphical display of the first 7 functions of the Fourier (left board) and B-spline (right board) basis function expansion (first line). Smoothing of functional data (black line) using 10 (blue line) and 70 (red line) Fourier (left board) and B-spline (right board) basis functions (second line). . . 10
Figure 3.1 Plots of the generated 25 sample curves for both $Y = 0$ and $Y = 1$ (left boards; scenario 1 (first line) and scenario 2 (second line)) and actual generated regression coefficient function (right boards; scenario 1 (first line) and scenario 2 (second line)). 22
Figure 3.2 Plots of the generated 50 sample curves for $Y = 0$ (left board) and $Y = 1$ (right board). The curves generated under scenario 1 are given in black while the curves generated under scenario 2 are given in gray.. 23
Figure 3.3 Boxplots of the computed CCR and AUC values under Case 1 for the FPC, FPLS, and RFPLS methods. 24
Figure 3.4 Boxplots of the computed IMSE values under Case 1 for the FPC, FPLS, and RFPLS methods. 25
Figure 3.5 Boxplots of the computed CCR and AUC values under Case 2 for the FPC, FPLS, and RFPLS methods. 25
Figure 3.6 Plots of the generated true parameter functions (blue lines), estimated parameter functions (gray lines), and mean functions of the estimated parameter functions (black lines). FPC (first stripe), FPLS (second stripe), and RFPLS (third stripe). The data are created under Case 1 when no outlier is present in the data (first row), [1%, 5%, 10%] of the generated data are contaminated by outliers (second row, third row, and fourth row, respectively.) 26
Figure 3.7 Graphical display growth curves of the boys (left board), girls (middle board), and detected outliers (right board). 27
Figure 3.8 Boxplot of the computed CCR and AUC values for the growth data. . . . 28

Figure 3.9 Graphical display of the mid-infrared spectra of strawberry fruit purees (left board), non-strawberry fruit purees (middle board), and detected outliers (right board). 28
Figure 3.10 Boxplot of the computed CCR and AUC values for the strawberry puree data. 29
Figure 4.1 Plots of the generated 50 sample curves (left board) and generated bivariate coefficient function $\beta(t,s)$ (right board). 33
Figure 4.2 Boxplots of the computed MAPE, RMSPE, RMESPE (first row), RE, NSE, and MIA values (second row) for the ARIMA, FPCR, FPLSR, and FQAR methods. Some extreme results are discarded from the plots to provide a better visualization. 35
Figure 4.3 Traditional (first row) and smoothed functional time series (second row) of the flow level for three different stations in the Australian region: Mary river (first column), Lockyer Valley (second column), and Albert river (third column). 36
Figure 4.4 Boxplots of the calculated MAPE, RMSPE, RMESPE, RE, NSE, and MIA values for the ARIMA, FPCR, FPLSR, and FQAR methods. Rows and columns represent performance metrics and stations, respectively. 37
Figure 4.5 An example demonstration of one-step-ahead point forecasts (blue lines) vs observed data point (black lines) and parameter functions at $0.25, 0.5, 0.75, 0.9$ quantile levels (columns, respectively) for all stations (in twos rows, respectively) obtained by the FQAR method. 38
Figure 4.6 Taylor diagrams of the one-step-ahead forecasts of flow level for three different stations in the Australian region with ARIMA, FPCR, FPLSR, and FQAR methods: Mary river (first column), Lockyer valley (second column), and Albert river (third column). 39
Figure 4.7 Box plots of computed interval score values for one-step-ahead prediction intervals generated by the FQAR method. 39

ABSTRACT

Novel Approaches to Regression Models for Functional Data

Müge MUTİŞ

Department of Statistics
Master of Science Thesis

Supervisor: Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK
Co-supervisor: Assoc. Prof. Dr. Ufuk BEYAZTAŞ

In recent years, in parallel with the development of technology, functional data analysis has been an active research area in statistics. In this thesis, two new approaches developed for functional data analysis are discussed. The first approach focuses on the functional logistic regression model where the predictors are functions to accurately estimate the binary response in the existence of outliers. The method based on the functional partial least squares components uses the weighted likelihood methodology for both the extraction of the components and the estimation of the functional regression parameter. The second approach considers the functional autoregressive model in which the past realizations of the functional time series are functional predictors. The functional principal components based approach uses quantile regression to characterize entire distribution of future realizations and to generates prediction intervals as well as the mean estimate of future realizations as an alternative to existing methods. Outcomes from simulation studies and real-world examples are also presented to demonstrate the performance of proposed new approaches. Our results are quite promising for both proposed approaches.

Keywords: Functional logistic regression, Functional autoregressive regression

ÖZET

Fonksiyonel Veriler için Regresyon Modellerine Yeni Yaklaşımlar

Müge MUTİŞ

İstatistik Anabilim Dalı
Yüksek Lisans Tezi

Danışman: Prof. Dr. Gülhayat GÖLBAŞI ŞİMŞEK
Eş-Danışman: Doç. Dr. Ufuk BEYAZTAŞ

Son yıllarda teknolojinin gelişmesine paralel olarak fonksiyonel veri analizi istatistikte aktif bir araştırma alanı olmuştur. Bu tezde, fonksiyonel veri analizi için geliştirilmiş iki yeni yaklaşım tartışılır. İlk yaklaşım, aykırı değerlerin varlığında ikili yanıtı doğru tahmin etmek için tahmin edicilerin fonksiyonlar olduğu fonksiyonel lojistik regresyon modeline odaklanır. Fonksiyonel kısmi en küçük kareler bileşenlerine dayalı yöntem hem bileşenlerin çıkarılmasında hem de fonksiyonel regresyon parametresinin tahmin edilmesinde ağırlıklı olabilirlik metodolojisini kullanır. İkinci yaklaşım, fonksiyonel zaman serisinin geçmiş gerçekleşmelerinin fonksiyonel tahmin ediciler olduğu fonksiyonel otoregresif modeli ele alır. Fonksiyonel temel bileşenlere dayalı yaklaşım, mevcut yöntemlere alternatif olarak gelecekteki gerçekleşmelerin ortalama tahmininin yanı sıra tüm dağılımını karakterize etmek için kantil regresyondan yararlanır ve tahmin aralıkları oluşturur. Önerilen yeni yaklaşımların performansını göstermek için simülasyon çalışmalarından ve gerçek dünyadan örneklerden sonuçlar da sunulur. Sonuçlarımız önerilen her iki yaklaşım için de oldukça umut vericidir.

Anahtar Kelimeler: Fonksiyonel lojistik regresyon, Fonksiyonel otoregresif regresyon

1
INTRODUCTION

1.1 Literature Review

In recent years, advancements in technology have led to an increase in data collection tools and database capacities. As a result, it has become possible to conduct more extensive observations or measurements based on time or specific conditions. When the studied area expands, or in another saying, when the number of observation points included in the sample increases, it is assumed that the data collected at these discrete points are sampled from an underlying real function. These observations are commonly referred to as functional data. The set of methods used to analyze such data is called functional data analysis. While the term functional data analysis (FDA) was introduced by Ramsay and Dalzell in 1991, the development of statistical techniques for analyzing functional data can be attributed to the work conducted in that earlier era. The roots of FDA can be traced back to the 1800s when Gauss and Legendre made efforts to model and predict the trajectory of a comet, which followed a curved path. In the FDA, which gained significant momentum especially with the contributions of Ramsay and Silverman, prediction models such as regression and classification have become widely used methodologies. These prediction models play a crucial role, particularly in applications related to chemometrics, medicine, food industry, economics, environmental and nature studies.

When it comes to regression models for functional data, the functional linear regression model takes precedence as the pioneering model that has received significant attention in the literature. It is undeniably the most popular and widely studied among the range of functional regression models available. The functional linear regression model was initially developed to handle scalar response variable. The model aims to capture the linear association among the a scalar response variable and the functional predictor variables. One notable extension of the functional regression models is the scalar-on-function logistic regression (FLogR). Introduced in the literature by [1], this framework is designed to handle binary outcomes as the response variable and random curves as predictors. FLogR has emerged as a versatile

approach for investigating the relation among the predictor curves and the binary response, as well as for classifying the binary outcome.

There are some examples in the literature dealing with the applications of the FLogR model in various fields. [1] used the FLogR model to estimate the probability of a high-risk birth outcome using periodically stimulated fetal heart rate monitoring. [2] created the FLogR model, which takes the change in temperature as an explanatory variable to predict the presence of drought risk, and also provided the explication of the parameter function in terms of odds ratios. [3] considered the FLogR model to classify gene expression curves according to known gene groups. [4] proposed a FLogR model to estimate the probability of a lupus flare based on stress level. More recently, in their case-control study of pancreatic cancer, [5] utilized the FLogR model for gene detection. Due to the high dimensionality and multicollinearity inherent in functional estimators, most of the studies mentioned included functional principal component analysis (FPCA). On the other side, a considerable body of research has concentrated on accurately estimating the regression coefficient function in the FLogR model, aiming to achieve precise estimation of the functional parameter. For example, [6] worked on the FLogR model, which takes the reduced principal components (PCs) as a covariate instead of the original design matrix in the FLogR model. Comparing different approaches for selecting optimum PCs, they concluded that including PCs in the given order with the stepwise method based on conditional likelihood ratio tests provides an accurate parameter function estimation with fewer PCs. Also, alternative approaches like penalized log likelihood [7] and LASSO [8] have been suggested as means to estimate the regression coefficient function.

Recently, [9] conducted a comparison among the aforementioned three methods. According to the findings of their, the LASSO yielded superior performance in terms of misclassification rate and parameter estimation compared to FPC regression and penalized log likelihood. However, the LASSO and penalized log-likelihood methods may produce unstable estimates under two situations: when a substantial number of basis expansion functions are employed to approximate the functional predictors or when numerous of functional predictors are used in the method [see, for example 10]. Therefore, using dimension reduction methods like FPC regression is more often than using other methods like log-likelihood and LASSO in scalar-on-function regression [11]. As an alternate to FPC regression, [12] proposed a functional partial least squares (FPLS) regression to estimate the FLogR model. FPLS regression uses both response and predictors when extracting the FPLS components, capturing most information with fewer terms. Thus, in scalar-on-function regression models, the FPLS regression is usually preferred to FPC regression [13–15].

Such as the logistic regression models, time series analysis constitutes a crucial area within the literature on functional regression. When the prediction of future developments in a process involves accounting for a continuous path of the past rather than a large number of past values within the statistical model, a functional regression model comes into play. Unlike traditional time series models, functional time series models do not encounter excessive correlation issues among repeated measurements due to their holistic perspective on the entire curve as a single observation. In simpler terms, functional models are a suitable choice for time series regression due to their ability to mitigate the problems caused by autocorrelation and non-stationarity, which are prevalent in time series data. Hence, functional time series modeling has received increasing attention in many scientific fields including hydrology and hydrometeorology. For example, [16] proposed a nonparametric functional time series model to forecast monthly sea surface temperatures. They focused on point forecast accuracy improvements using different dynamic updating methods. [17] employed functional models, which consider both a scalar response and a function, to examine the flow volume and the entire streamflow curve for a given period, utilizing precipitation curves as the main explanatory factors. They compared the functional model with the neural network approach, and comprehensively discussed the strengths and weaknesses associated with each model. In their research, [18] utilized a non-parametric functional time series model to make predictions on hourly air temperature, with a forecast horizon of up to one day, and subsequently compared its performance with the seasonal autoregressive moving average model. Their findings showed that the functional model gives superior results for the part other than the first six hours of the day. [19] proposed a functional time series model based on FPCs to forecast drought for different stations and compared it with the linear exponential smoothing and autoregressive integrated moving average methods. Their results demonstrated that their proposed method provides reliable forecasts. In addition to point forecasts, they also created pointwise interval prediction using non-parametric bootstrap technique. The impact of exogenous variables on river flow was examined by [20], who introduced a model incorporating past lags of functional time series along with other hydrometeorological variables. Additionally, they put forth a forward selection process as a means to ascertain the important functional lagged variables with statistical significance in their analysis. Comparing the model which they used FPLS regression to estimate the parameter function with many commonly used models, they demonstrated that their proposed approach delivers pretty competitive performance based on their findings.

1.2 Objective of the Thesis

The objective of this thesis is to contribute to the function data analysis literature by suggesting new approaches. In this scope, we divide the thesis into five chapters, and we also review the existing literature in the Chapter 1 where we present our objective and hypothesis. In the Chapter 2, we create a general framework of FDA.

In the Chapter 3, we focus on FLogR. Because, in all the studies mentioned above, the parameter function in the FLogR model is estimated via the least-squares (LS) type estimator. But, the LS estimator is severely affected by the presence of outliers. In such a case, the LS estimator generates biased estimates, leading to an increased probability of misclassification. Tiny care has been paid to robustly estimating the parameter function of the FLogR model in the presence of outliers. We found only one study by [21] that estimates the parameter function of the FLogR model robustly. By using the minimum covariance determinant method, [21] computed the FPC scores. Then, the binary outcome and the computed scores are used to approximate the FLogR model, where the maximum likelihood estimator is used to estimate this model. However, this model may not have a full robust property since the maximum likelihood estimator is used in the approximate model. As far as our knowledge extends, there is currently no robust approach available for the computation of FPLS components and estimation of the regression coefficient function in the FLogR model. In this thesis, we propose a robust FPLS (RFPLS) method to robustly estimate the parameter function in the FLogR model and classify the binary response in the presence of outliers.

In the Chapter 4, we introduce quantile regression for functional time series. All the aforementioned papers are based on the conditional mean regression, i.e., they are useful only to find average relationship among the response and predictor variables. However, it may be crucial to model and forecast hydrological processes at non-central locations of the distribution of response variable. Most importantly, because the climatically distinct regions necessitate diverse risk managements, the decision makers may need to obtain forecasts of hydrological processes at the lower and/or upper tail of the distribution of the process to manage the decision-making processes properly [see e.g., 22, 23]. For example, a river flow dataset obtained from a dry region needs to be analyzed and its future realizations need to be forecasted at the lower tail of the distribution to predict possible droughts properly. In contrast, a river flow dataset obtained from a rainy region needs to be analyzed and its future realizations need to be forecasted at the higher part of the distribution to predict possible river floods properly. To the best of our knowledge, there has been no functional time series method integrated with the quantile regression to obtain the forecasts of time series datasets at the non-central distribution of the target variable.

In this thesis, we propose a functional quantile autoregressive of order one method to obtain future realizations of functional river flow datasets at the non-central locations. In addition to its capability to describe the complete conditional distribution of the response variable, the proposed method possesses all the attractive properties of the quantile regression that are not fulfilled by conditional mean regression-based functional time series methods. For example, 1) because the quantile regression belongs to a robust family, the proposed approach is more robust to unusual observations compared with the conditional mean regression-based methods; 2) the proposed approach does not require any distributional assumptions for the error terms; 3) the proposed approach is able to produce more efficient modeling results than the existing conditional mean regression-based methods when the error terms follow non-Gaussian and/or heavy-tailed distribution; 4) unlike the conditional mean regression-based methods, the proposed approach may produce better results in case of heteroscedasticity since it does not rely on the assumption of constant variance for the response variable; 5) the proposed approach can easily be used to construct forecast intervals by fitting the same method using different quantile levels.

In Chapter 5, we conclude the thesis by making a brief evaluation of our findings.

1.3 Hypothesis

In the Chapter 3, the proposed method for the FLogR model produces to its competitors significantly superior results in both estimations of parameter function and classification of the binary response variable in the presence of outliers. The proposed approach is robust to vertical outliers in the response variable and leverage points in the predictor variable. In addition, our proposed method provides competitive performance when there is no outlier in the data.

The functional autoregressive quantile regression model that we propose in Chapter 4 for functional time series of hydrological processes is an excellent alternative to existing mean regression methods at the 0.5 quantile level (median regression). Also, the proposed method offers a more thorough explanation of the connection among response and predictor curves at various quantile levels, providing a more extensive understanding of the relationship. Moreover, this feature of the proposed method allows for effortlessly generation of pointwise prediction intervals for functional response.

2
GENERAL FRAMEWORK FUNCTIONAL DATA ANALYSIS

As part of the statistical discipline, FDA focuses on analyzing data that provides knowledge into curves, surfaces, or other mathematical objects and heavily utilizes the underlying continuous properties of data. In conceptual terms, functional data are defined as continuous entities, although in practical applications, they are commonly observed or measured at discrete sampling points. Therefore, first and foremost of FDA is to convert the data observed at sampling values t_j, which may or may not be evenly spaced on the discrete grid $\{t_1, t_2, \ldots, t_J\} \in T$, into functional form. While sampling points t_j are mostly time, it need not always be a physical time, but rather any other continuum such as age, height, spatial location, frequency, etc. Figure 2.1 provides a prototype for the considered functional data. Afterwards, these obtained functions allow the application of functional equivalents of various multivariate analyzes as well as running a functional version of descriptive statistics that helps us perform further analysis by looking for noteworthy structures. In other words, in FDA, the initial step involves recovering the underlying functions from the observed data, followed by conducting various types of analyses on these functions.

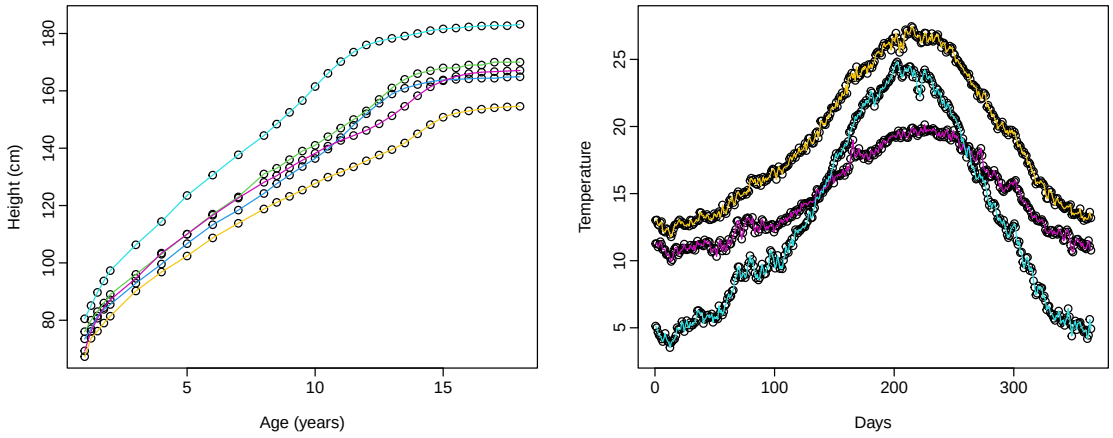


Figure 2.1 Two example demonstrations of observed data at discrete sampling point (black circles) and their functional forms (colored lines).

2.1 Transforming Observed Data into Functional Form

Let $X = \{X(t) : \forall t \in T\}$ be a stochastic process in a separable Hilbert space L_T^2 defined on the closed and bounded interval T, satisfying the condition $\int_T X^2(t)dt < \infty$ with the inner product $< X_1, X_2 >= \int_T X_1(t)X_2(t)dt, \quad \forall X_1, X_2 \in L_T^2$. Unlike multivariate analysis, FDA assumes that the sample functions are smooth curves of stochastic process $X = \{X(t) : \forall t \in T\}$. The property of smoothness which ensures that a pair of adjacent sample points are probably not very different from each other and are linked to a certain extent means having a certain number of derivatives. Also as stated in [24], if this smoothness property were not available, there would not be much advantage in evaluating the data functionally rather than multivariate. However, real-world observed data typically contain inherent observational errors or noise that can mask the underlying real function. Therefore, the sample functions might be far from being smooth. Let $Y = \{Y_1, Y_2, \ldots, Y_n\}$ be a n-dimensional random sample of functional data. A functional data defined as follows:

$$Y_i = X_i(t) + \varepsilon_i, \qquad 1 \leq i \leq n, \ \ \forall t \in T. \tag{2.1}$$

Y_i is i-th noisy observation of the stochastic process $X_i(t)$ with random error ε_i.

As stated earlier, the initial step of FDA is to reconstruct the functional structure of the discrete sampling points in finite-dimensional space. If the discrete observations are supposed to be free of errors, the process of converting them to functions is interpolation; however, if there is some measurement error and noise in the observations, then converting the discrete data to functions may require smoothing. Although there are various smoothing methods used for this purpose, a fixed basis extension function method is mostly adopted as a general and practical approach. Given that the basis expansion functions have the similar properties as the process that generates the data, they provide a good approximation to the functional data and minimize noise.

2.1.1 Basis Expansion Functions

Basis expansion functions are a collection of mathematically orthogonal curves, demonstrated by $\phi(t)$ and used to arbitrarily approximate any function $X_i(t)$ in L_T^2. In order to approximate to the function, a basis system consisting of K basis expansion functions is selected, then the functional observation is constructed from the linear combination of the basis functions as follows:

$$X_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t) \qquad 1 \leq i \leq n, \ \ \forall t \in T \tag{2.2}$$

where c parameters are the coefficients that corresponding to the basis expansion functions and determine the form of the curve. By letting $c_i = [c_{i1}, c_{i2}, \ldots, c_{iK}]^T$ indicate the vector of the coefficients for i-th observation and $\Phi = [\phi_1(t), \phi_2(t), \ldots, \phi_K(t)]^T$ as the vector of the basis expansion functions, we can also express (2.2) in matrix notation as follows:

$$X_i(t) = c_i^T \Phi = \Phi^T c_i \quad 1 \leq i \leq n. \tag{2.3}$$

K represents the number of basic expansion functions. If the K is too small, we run the risk of overlooking essential characteristics of the smooth curve $X_i(t)$ that we are attempting to estimate. On the other hand, larger K, larger K leads to a better fit to the data however, we run the risk of also fitting noise or variability that we aim to disregard [24]. Also, the process is interpolation for $K = t_J$. Considering all these, the choice of K is important to provide a compromise between fit and smoothness. Stepwise variable selection or variable pruning methods in the tremendous literature on multiple regression, and various algorithms such as generalized cross validation (GCV), generalized bayesian information criteria (GBIC), modified akaike information criterion (MAIC), cross validation (CV) are adopted to specify the number of basis expansion functions. Another significant issue regarding the basis expansion functions is the selection of the basis function system. The greater the resemblance between the properties of the basis functions and the data, in another saying, the better they represent the characteristics that are inherent in the data, the better the approximation to the underlying function. Some possible basis expansion functions are polynomial basis functions (which are constructed from the monomials $\phi_k(t) = t^{k-1}$), Bernstein polynomial basis functions (which are constructed from $1, 1-t, t, (1-t)^2, 2t(1-t), t^2, \ldots$), radial basis functions, wavelet basis functions, Fourier basis functions, B-spline basis functions which are a kind of spline system [see 24, 25, for more detailed information]. Compared to the others, the B-spline basis functions that we used in this study and Fourier basis functions are generally more preferred.

Fourier basis functions: Perhaps the most well-known basis expansion is provided by the Fourier series:

$$\phi_0(t) = \frac{1}{\sqrt{|T|}}, \quad \phi_{2r-1}(t) = \frac{\sin rwt}{\sqrt{\frac{|T|}{2}}}, \quad \phi_{2r}(t) = \frac{\cos rwt}{\sqrt{\frac{|T|}{2}}} \qquad \forall t \in T$$

where the constant w defines the period or the length of the interval $|T| = \frac{2\pi}{w}$. Assuming the series commonly comprises both elements of each pair, the number of basis functions corresponds to $K = 1 + 2r$. Note that the computation of

Fourier basis requires K to be an odd number.

In short, to define a Fourier basis system, simply two crucial details are needed:

1. the number of basis functions and
2. the period which is frequently equivalent to the length of the interval $|T|$.

In addition to this, if the values of t_j are equally spaced on T, then the basis is orthogonal. Fourier series are beneficial for expanding functions that lack distinct local features and exhibit a relatively constant curvature. Fourier series typically produce expansions which are uniformly smooth therefore they use for periodic or nearly periodic data. They are not suitable for data that contain discontinuities either in the function itself or in low order derivatives [26]. In [24], an interesting but pretty suitable simile is made: A Fourier series is similar to margarine: it is affordable and can be applied to nearly anything, but do not expect the outcome to be invariably thrilling.

B-spline basis functions: When it comes to non-periodic functional data or parameters, spline basis functions are the most prevalent utilized approximation system. Spline basis functions offer greater flexibility, making them more intricate compared to Fourier bases. However, the advantage of splines are that they are fast to compute corresponding c coefficients and they achieve a good approximation of the underlying function with a relatively small number of K [26]. Splines are piecewise polynomials in the bounded interval T which is divided by subintervals L separated by values of ς_l , $l = 0, \ldots, L-1$ called breakpoints or knots. Across any subinterval, the spline function takes the form of a polynomial with a fixed degree or order, although the characteristics of the polynomial transition as one moves to the subsequent subinterval. The highest power within a polynomial is referred to as the degree. Also, for a polynomial, the order is equal to the degree plus one. A prevalent choice is to utilize order four splines, which are constructed from cubic polynomial segments (degree three). In brief, spline basis systems exhibit the following defining attributes:

1. the break points defining subintervals,
2. the degree or order of the polynomial segments, and
3. the sequence of knots.

The number K of basis functions in a spline basis system is determined by order $+$ number of interior knots. The term interior knots denotes knots specifically located at break points within the domain of definition of function T, excluding the beginning and end points.

There are many diverse kinds of splines such as truncated power functions, B-splines, M-splines, I-splines and natural splines. The B-spline basis system developed by [27] is the most popular. The set of B-spline basis functions of order m for the knot sequence ς is derived using a recursion formula as follows:

$$\phi_{k,m}(t) = \frac{t-\varsigma_k}{\varsigma_{k+m-1}-\varsigma_k}\phi_{k,m-1}(t) + \frac{\varsigma_{k+m}-t}{\varsigma_{k+m}-\varsigma_{k+1}}\phi_{k+1,m-1}(t) \qquad m \geq 2, \ \ \forall t \in T.$$

Also as stated in [25], for $m = 1$ these functions are called Haar based functions and identified as follows:

$$\phi_{k,1}(t) = \begin{cases} 1, & t \in [\varsigma_k, \varsigma_{k+1}] \\ 0, & \text{otherwise.} \end{cases}$$

A graphical representation of the B-spline and Fourier basis expansions functions and the made smoothing with these functions is presented in Figure 2.2.

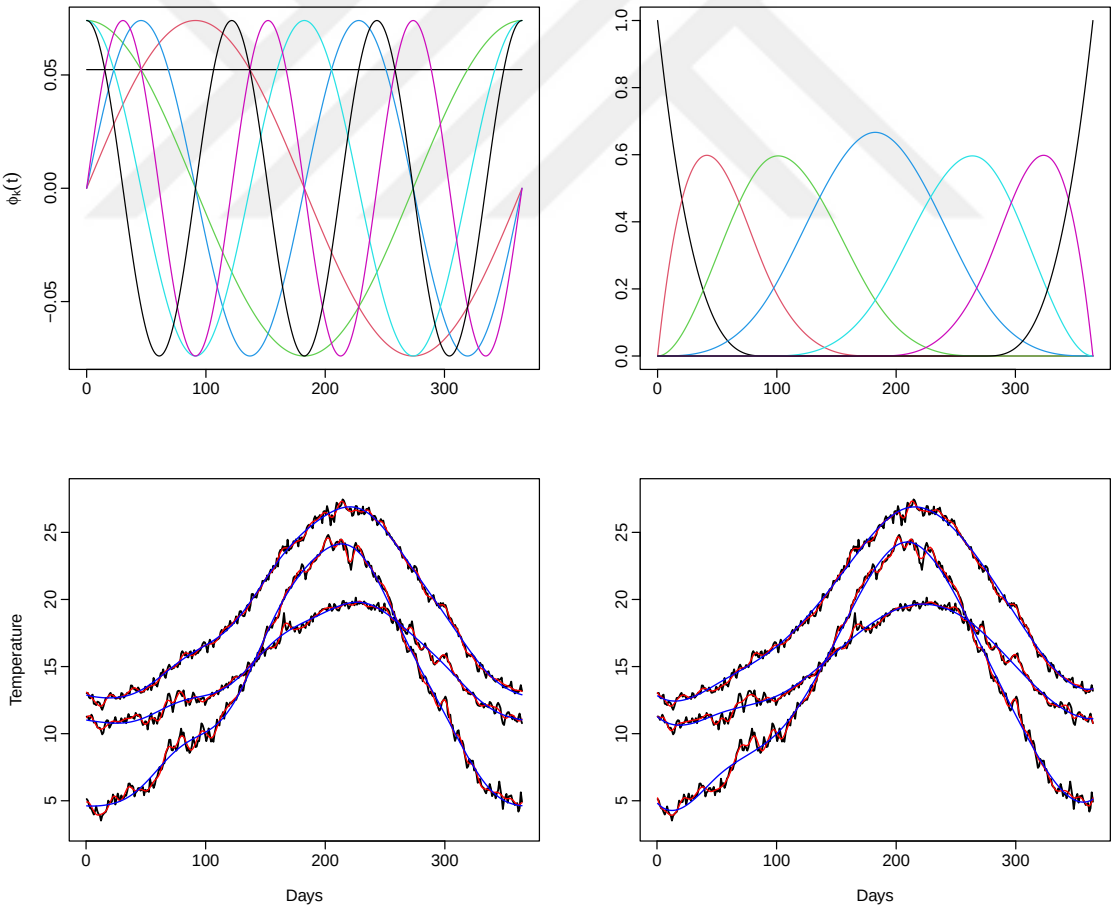


Figure 2.2 Graphical display of the first 7 functions of the Fourier (left board) and B-spline (right board) basis function expansion (first line). Smoothing of functional data (black line) using 10 (blue line) and 70 (red line) Fourier (left board) and B-spline (right board) basis functions (second line).

2.1.2 Curve Fitting

The least squares (LS) method, the most familiar method among statisticians, can be used to fit the c coefficients corresponding the basis expansion functions from the model given by (2.1). The LS estimation problem based on minimization of the sum of squared error (SSE) define as follows:

$$\text{SSE} = \sum_{j=1}^{J} \left[Y_{ij} - X_i(t_j) \right]^2 = \sum_{j=1}^{J} \left[Y_{ij} - \sum_{k=1}^{K} c_{ik} \phi_k(t) \right]^2 \quad 1 \leq i \leq n, \ \forall t_j \in T. \tag{2.4}$$

Equation (2.4) can be rewritten in a matrix form as follow:

$$\text{SSE} = (Y - \Phi \boldsymbol{c})^T (Y - \Phi \boldsymbol{c}) \tag{2.5}$$

where Y represent the matrix of the discrete observations and $\boldsymbol{c} = [c_1, c_2, \ldots, c_n]^T$. When the derivative of the criterion SSE is taken with respect to c, the equation obtained is $2\Phi\Phi^T \boldsymbol{c} - 2\Phi^T Y = 0$. Then, by solving this equation for c, supplies the estimate $\hat{\boldsymbol{c}}$ which minimizes the LS solution

$$\hat{\boldsymbol{c}} = \left(\Phi^T \Phi \right)^{-1} \Phi^T Y.$$

This transaction is suitable when we make the assumption of the standard model for errors. If errors are not assumed independently distributed with zero mean and constant variance, we include weight W in the (2.5)

$$\text{SSE} = (Y - \Phi \boldsymbol{c})^T W (Y - \Phi \boldsymbol{c})$$

where W is a symmetric positive definite matrix which is defined to be the inverse of the variance-covariance matrix of errors, $\sum_{\varepsilon}^{-1}$ [see 24, for more detailed information]. Thus, the weighted LS estimate $\hat{\boldsymbol{c}}$ of the coefficient vector $\boldsymbol{c}$ is

$$\hat{\boldsymbol{c}} = \left(\Phi^T W \Phi \right)^{-1} \Phi^T W Y.$$

Note that an appropriate number of basis functions K should be chosen to accurately capture the functional form of discrete observations. As expressed in [28], the regression-based approach for data smoothing is effective only when the number of basis functions K is significantly smaller than the number of sampling points t_j. This can be better explained by the basic principle of the statistic as follows:

$$\text{MSE} = \text{Bias}^2 \left[X(t) \right] + \text{Var} \left[X(t) \right]$$

where $\text{Bias}[X(t)] = Y - \text{E}[X(t)]$ and $\text{Var}[X(t)] = \text{E}\left[(X(t) - \text{E}[X(t)])^2\right]$.

The bias in estimating $X(t)$ becomes insignificant as K increases in magnitude. However, in this case the sampling variance increases rapidly. Whilst an unbiased estimate is desirable, the significant variance interrelated with the curve will leads to a compromised mean squared error (MSE). Providing an improved MSE by reducing variance leads us to look for smaller values of K. However, caution must be exercised to ensure that the chosen values do not result in an unacceptable level of bias. As a trade-off, the penalized LS procedure aims to make smallest the SSE by finds a balance among the original data and the smoothed data with an most conveniently selected K. The fundamental concept of the approach is akin to LS estimation, but with the inclusion of a penalty term in equation (2.5) as follows:

$$\text{SSE} = (Y - \Phi \boldsymbol{c})^T (Y - \Phi \boldsymbol{c}) + \lambda \text{PEN}_2(X(t)) \tag{2.6}$$

where λ represents the smoothing parameter that governs the level of roughness. It specifies the emphasis on the second term penalizing curvature relative to goodness of fit quantified in the sum of squared errors in the first term [28]. $\text{PEN}_2(X(t))$ denotes a measure of the function's roughness. A natural measure of a function's roughness is the integrated squared second derivative or total curvature

$$\text{PEN}_2(X(t)) = \int_T \left[D^2 X(t)\right]^2 dt = \boldsymbol{c}^T \boldsymbol{R} \boldsymbol{c}$$

where $\boldsymbol{R}$ is called the penalty matrix and defined as the integral of outer products of 2-th derivatives of the basis functions $\boldsymbol{R} = \int_T D^2\phi(t) D^2\phi^T(t) dt$. From (2.6), the penalized LS estimate of $\boldsymbol{c}$ is obtained as follows:

$$\hat{\boldsymbol{c}} = \left(\Phi^T \Phi + \lambda \boldsymbol{R}\right)^{-1} \Phi^T Y.$$

Note that when $\lambda = 0$, penalized LS estimate reduces to the LS estimate. With the increase of λ, non-linear functions are subjected to a progressively pronounced roughness penalty due to the term $\text{PEN}_2(X(t))$. In addition, it can also be noticed that as $\lambda \to \infty$, the estimated curve approaches the standard linear regression curve. Consequently, several techniques, including GCV, MAIC and GBIC have been adopted to select the optimal smoothing parameter λ similar to the selection of K [see 24, 25, for more detailed information].

2.2 Functional Principal Components Analysis

Principal components analysis is a prevalent method used in multivariate analysis as a dimension reduction technique. In FDA as the data are curves that are assumed to be sampled from a continuous stochastic process $X = \{X_1, X_2, \ldots, X_n\}$, this dimension reduction technique is known as functional principal component analysis (FPCA). FPCA allows dimension reduction and leads to reduced complexity of data by transforming the variables to uncorrelated functional principal components (FPCs) explaining with maximum variance the structure of individual curves. Let the covariance function for random sample X is defined as follows:

$$\mathrm{Cov}_X(t,s) = \frac{1}{n-1}\sum_{i=1}^{n}\Big(\big(X_i(t)-\bar{X}(t)\big)\big(X_i(s)-\bar{X}(s)\big)\Big) \tag{2.7}$$

where, $\bar{X}(t)$ and $\bar{X}(s)$ indicate sample mean, e.g. $\bar{X}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t)$, also $\forall t,s \in T$. In the FPCA, the bivariate covariance function is decomposed into eigenfunctions and eigenvalues based on Mercer's theorem. Then (2.7) can be restated as follows:

$$\mathrm{Cov}_X(t,s) = \sum_{k\geq 1} \nu_k \psi_k(t)\psi_k(s)$$

where ψ_k and ν_k represent k-th eigenfunction and its associated positive eigenvalues, respectively. Eigenvalues of the bivariate covariance function are indicators of the importance of FPCs that maximize the curve variation. In addition eigenfunctions that describe the main variation components are also obtained sequentially by conforming to the orthogonality conditions with $\|\psi_k\|^2 = 1$ and the corresponding positive eigenvalues are in descending sequence as $\nu_k \geq \nu_{k+1}$. Note that usually the first few components capture most of the variation.

In terms of the mean integrated square error, FPCA yields the most accurate empirical basis expansion for functional data. Therefore, it can offer superior performance compared to other presentments of functional data using stable basis functions like Fourier series or B splines. As mentioned in the previous section, the drawback of stable basis functions is that a greater quantity of them may be required to catch a established sample of functions with reduced bias. Furthermore, the estimated coefficients of basis functions are correlated and may be less suitable for subsequent implementations for instance functional regression. Considering all these, as in the functıonal quantile autoregressıve regressıon proposed in Chapter 4, providing data based expansion by using FPCs as follows while reconstructing the functional structure

of the data in finite dimensional space has been a frequently used approach in FDA

$$X_i(t) = \sum_{k=1}^{K} \zeta_{ik}\psi_k(t) = \zeta_i^T \psi(t) \quad 1 \leq i \leq n$$

where ζ_{ik} denote the FPC scores and define as the projections of the curves onto the eigenfunction bases $\zeta_{ik} = \int_T X_i \psi_k(t) dt$. Thus, essentially infinite number of variables is reduced to a few orthogonal regressors.

2.3 Functional Partial Least Squares Regression

Functional partial least squares (FPLS) is a powerful alternative to FPCs as an efficient dimension reduction technique for dealing with infinite dimensional curves. Similar to FPCs, its main purpose is to produce an approximation of these infinite dimensional curves in finite dimensional space as follows:

$$X_i(t) = \sum_{k=1}^{K} \xi_{ik}\gamma_k(t) = \xi_i^T \gamma(t) \quad 1 \leq i \leq n$$

where ξ_{ik} and γ_k denote the FPLS scores and orthogonal components, respectively. FPLS approach is generally used with scalar-on-function regression models, where the response is a scaler, and the predictor is a random curve, therefore it is known as functional partial least squares regression (FPLSR). FPLSR orthogonal components are calculated based on iterative eigendecomposition by maximizing the covariance between response and predictors variables of the functional model. Thus, compared with FPCs regression, one potential disadvantage of FPCs regression is that the selection of regressed components is driven solely by their ability to explain predictor variance, without taking into account their effectiveness in explaining the responses. In contrast to alternative procedures, FPLSR focuses on identifying components that have the highest relevance in predicting the outcome. Note that there are a number of partial least squares algorithms that have been extended to functional form due to various differences, such as the way variables are initialized and intermediate coefficients are normalized, the way sequential steps are orthogonalized [see 29, 30, for more detailed information]. In this thesis, extraction of orthogonal FPLS components with the algorithm considered is given in detail in Chapter 3.

3
A ROBUST FUNCTIONAL LOGISTIC REGRESSION METHOD FOR CLASSIFICATION

In this chapter, we introduce a scalar-on function logistic regression model and propose a robust partial least squares method (RFLPS) to estimate the regression coefficient function based on the weighted likelihood method that reduces the effect of outliers in the response and predictor.

3.1 Scalar on Function Logistic Regression

Let us consider an independent and identically distributed (i.i.d.) random sample $\{Y_i,\ X_i(t):\ \ i=1,2,\ldots,n\}$ from a population $\{Y,X\}$, where $Y\in\{0,1\}$ is a binary outcome and $X=\{X(t)\}_{t\in T}$ is a stochastic process defined in a L_2 separable Hilbert space with a bounded and closed interval $t\in T$. The FLogR model used to explore the relationship between Y and $X(t)$ is defined as follows [1]:

$$Y_i=\pi_i+\varepsilon_i, i=1,\ldots,n, \tag{3.1}$$

where ε_i is the independent random error term with mean-zero and variance $\pi_i(1-\pi_i)$ and π_i is the probability that the response variable takes value one given an observation of the functional predictor

$$\pi_i=P\left[Y=1|X(t)=X_i(t)\right]=\frac{\exp\left\{\alpha+\int_T X_i(t)\beta(t)dt\right\}}{1+\exp\left\{\alpha+\int_T X_i(t)\beta(t)dt\right\}}, \tag{3.2}$$

where α is the intercept parameter and $\beta(t)$ denotes the regression coefficient function. From (3.2), the logit transformation yields

$$l_i=\ln\left(\frac{\pi_i}{1-\pi_i}\right)=\ln\frac{P\left[Y=1|X(t)=X_i(t)\right]}{P\left[Y=0|X(t)=X_i(t)\right]}.$$

In what follows, the FLogR model (3.1) can be expressed in terms of logit transformation as follows:

$$l_i = \alpha + \int_T X_i(t)\beta(t)dt. \tag{3.3}$$

By utilizing Equation (3.3) it becomes possible to interpret the regression coefficient curve so that the integral of $\beta(t)$ multiplied by a constant μ can be commentated as the multiplicative change in the odds of response $Y = 1$ obtained when a functional observation is incremented constantly in μ units along T [2, 12]. But, the estimating Model (3.3) puts forth a hard work due to the regression coefficient curve $\beta(t)$ residing in an infinite-dimensional space. Another problem related to the estimation of Model (3.3) is that while the functional predictor $X(t)$ belongs to an infinite-dimensional space, it is observed in a finite set of discrete points. This chapter contemplates the B-spline basis functions approach to reconstruction the functional form of the discretely measured predictor curves and estimate the infinite-dimensional regression coefficient curve.

With a adequately big quantity of basis expansion functions K, the functional predictor $X(t)$ can be approximated as a linear combination of the basis functions $\phi(t)$ and the corresponding basis expansion coefficients a as follows:

$$X(t) \approx \sum_{k=1}^{K} a_k \phi_k(t) = \boldsymbol{a}^\top \boldsymbol{\phi}(t). \tag{3.4}$$

Similarly, the regression coefficient function can be stated in the basis $\boldsymbol{\phi}(t)$ as follows:

$$\beta(t) \approx \sum_{k=1}^{K} \beta_k \phi_k(t) = \boldsymbol{\beta}^\top \boldsymbol{\phi}(t). \tag{3.5}$$

Substituting (3.4) and (3.5) into (3.3), the FLogR model in terms of logit transformation is expressed as follows:

$$l_i = \alpha + \sum_{k=1}^{K} a_{ik} \phi_k(t) \beta_k \phi_k(t).$$

Let us denote by $\boldsymbol{\Psi} = \int_T \boldsymbol{\phi}(t)\boldsymbol{\phi}^\top(t)dt$ the $K \times K$ inner product matrix of the basis functions. Then, the logit transformation expression of the FLogR model (3.1) has the following matrix form:

$$\boldsymbol{L} = \mathbf{1}\alpha + \boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{\beta},$$

where $\boldsymbol{L} = [l_1, \ldots, l_n]^\top$, $\mathbf{1} = [1, 1, \ldots, 1]^\top$, and $\boldsymbol{A}$ is the $n \times K$ matrix with row entries

$a_i = [a_{i1}, a_{i2}, \ldots, a_{iK}]$.

The results presented above demonstrate that the infinite-dimensional FLogR model of Y on $X(t)$ can be reduced to a simple finite-dimensional logistic regression (LogR) model of $\boldsymbol{L}$ on the random matrix $\boldsymbol{A\Psi}$. Let $\widehat{\beta}$ denote an estimate of β. Then, the regression coefficient function in (3.3) can be approximated using the basis functions and the estimated basis expansion coefficient as follows:

$$\widehat{\beta}(t) = \widehat{\boldsymbol{\beta}}^{\top} \boldsymbol{\phi}(t).$$

Due to the temperament of functional data, the among columns of obtained design matrix $\boldsymbol{A\Psi}$ become highly correlated (multicollinearity). In such a case, the estimate of $\widehat{\boldsymbol{\beta}}$ obtained via the LS method may have a large variance. While the FPLS method of [12] overcomes the multicollinearity problem, it may be significantly affected by outliers because the FPLS is based on the LS estimator. When outliers are present in the data, the effects of these points are included in the FPLS components approximated from the random matrix $\boldsymbol{A\Psi}$. The effects of outliers are also included in the estimated regression coefficient function $\widehat{\beta}(t)$ because its basis expansion coefficient $\boldsymbol{\beta}$ is estimated from the LogR model of $\boldsymbol{L}$ on the extracted PLS components. Thus, in the avaiable of outliers, the FPLS method based on the LS type estimator may produce biased parameter estimates and incorrect classification results. Therefore, we propose an RFPLS method to robustly estimate the regression coefficient function $\beta(t)$ and correctly classify the values of the binary response variable in the presence of outliers. While doing so, we first extend the PLS method of [31] to the FLogR model. Then, to approximate the FPLS components and regression parameter curve robustly, we consider the weighted likelihood estimator (WLE) of [32].

3.2 Proposed Method

Before introducing the proposed method, we start with a brief introduction of the weighted likelihood-based LogR model.

3.2.1 Weighted Likelihood Estimation

The weighted likelihood estimator (WLE) is derived from the weighted likelihood estimation equations by modifying the weights, which are functions of properly defined residuals of the maximum likelihood estimation equations. The WLE of an unknown common parameter θ for an n-dimensional random sample $\{Y_1, \ldots, Y_n\}$ is

obtained as a solution of the following estimating equation:

$$\sum_{i=1}^{n} w(Y_i;\theta,f^*)\,u(Y_i,\theta)=0,$$

where $u(Y_i,\theta)$ is the maximum likelihood score function from the hypothesized model and $w(Y_i;\theta,f^*)$ denotes the weights. Weights are used to decrease the effects of outliers in the data set and on the score equations. The weights are defined as follows:

$$w(Y;\theta,f^*)=\max\left(\min\left(\frac{\mathscr{A}\left[\delta(Y;\theta,f^*)\right]+1}{\delta(Y;\theta,f^*)+1},1\right),0\right),$$

where $\delta(Y;\theta,f^*)$ and $\mathscr{A}\left[\delta(\cdot)\right]$ denote the Pearson residual function and residual adjustment function (RAF), respectively. The Pearson residual function, which indicates the harmony between the hypothesized model and the empirical distribution of observations is defined by $\delta(Y;\theta,f^*)=\frac{f^*(Y)}{m^*(Y;\theta)}-1$. For the LogR model, $f^*(Y)$ is the proportion of sample observations with value Y and $m^*(Y;\theta)$ is the corresponding probability under the hypothesized model. The RAF, on the other hand, is a strictly increasing and twice differentiable function defined in the range $[-1,\infty)$.

Let us consider n independent binary responses $Y=\{Y_1,,\ldots,Y_n\}$ and the corresponding $n\times K$ design matrix $\boldsymbol{H}=\boldsymbol{A\Psi}$. For the LogR model, the maximum likelihood estimator (MLE) is found by equating the score functions acquired by taking the derivative of the following log-likelihood function with respect to the parameters to zero:

$$\ln L(\boldsymbol{\beta})=\sum_{i=1}^{n} Y_i\ln[\pi(\boldsymbol{H}_i)]+(1-Y_i)\ln[1-\pi(\boldsymbol{H}_i)],$$

where $\pi(\boldsymbol{H}_i)=P(Y=1|\boldsymbol{H}=\boldsymbol{H}_i)=\frac{\exp\left(\boldsymbol{H}_i^\top\beta\right)}{1+\exp\left(\boldsymbol{H}_i^\top\beta\right)}$ is the conditional probability of the outcome where $\boldsymbol{H}_i=[1,\boldsymbol{H}_{i1},\ldots,\boldsymbol{H}_{iK}]^\top$ is the i-th row of the design matrix $\boldsymbol{H}$ and $\boldsymbol{\beta}=\left[\beta_0,\beta_1,\ldots,\beta_p\right]^\top$ denotes the parameter vector. The score equations are nonlinear in $\boldsymbol{\beta}$, and thus, the iterative reweighted least squares (IRLS) method is usually used to obtain the solutions For more information, see also [33]. In IRLS, the adjusted binary response variable vector $\boldsymbol{Z}$ with i-th element $\boldsymbol{Z}_i=\frac{Y_i-\pi(\boldsymbol{H}_i)}{\Omega_i}+\boldsymbol{H}_i^\top\boldsymbol{\beta}$ is regressed on the columns of design matrix $\boldsymbol{H}$ using the weight diagonal matrix $\boldsymbol{\Omega}$ with diagonal elements $\boldsymbol{\Omega}_i=\pi(\boldsymbol{H}_i)[1-\pi(\boldsymbol{H}_i)]$. Note that the weight values ($\boldsymbol{\Omega}_i$'s) are the functions of $\pi(\boldsymbol{H}_i)$. Therefore, the method is iterative and the $\boldsymbol{\Omega}_i$'s are renewed at each step to achieve the following solution

$$\widehat{\boldsymbol{\beta}}_{MLE}=\left(\boldsymbol{H}^\top\boldsymbol{\Omega H}\right)^{-1}\boldsymbol{H\Omega Z}.$$

The WLE of $\boldsymbol{\beta}$ is obtained by adding the weight diagonal matrix $\boldsymbol{W}$ based on the Pearson residuals to the maximum likelihood estimation equations and its two-step solution as follows:

$$\widehat{\boldsymbol{\beta}}_{WLE} = \left(\boldsymbol{H}^{\top}\boldsymbol{\Omega}\boldsymbol{W}\boldsymbol{H}\right)^{-1}\boldsymbol{H}\boldsymbol{\Omega}\boldsymbol{W}\boldsymbol{Z}.$$

In the first step of the solution, the weight diagonal matrix $\boldsymbol{W}$ is evaluated, and in the second step, the IRLS is used with a fixed $\boldsymbol{W}$. In each iteration of the IRLS method, $\boldsymbol{W}$ is updated and added to IRLS. The algorithm terminates when the values of $\boldsymbol{W}$s obtained in the consecutive steps do not change. Note that, the distribution of Pearson residuals tends to be skewed in generalized linear models with non Gaussian response distributions. Therefore, the Anscombe residuals whose asymptotic distributions are as close to Gaussian distribution as possible under the hypothesized model are usually used as an alternative to a more general solution. For the LogR model with a binomial response distribution, the Anscombe residual is defined as:

$$r_{A_i} = \sqrt{m_i}\left[B\left(Y_i, \frac{2}{3}, \frac{2}{3}\right) - B\left(\pi(\boldsymbol{H}_i), \frac{2}{3}, \frac{2}{3}\right)\right]\left[\pi(\boldsymbol{H}_i)(1-\pi(\boldsymbol{H}_i))\right]^{-\frac{1}{6}},$$

where m_i denotes the trial at i-th observation and $B(z,a,b) = \int_0^z t^{a-1}(1-t)^{b-1}\,dt$ denotes the beta function. The Anscombe residuals can easily be calculated using the R package "wle" [34]. For more information, see also [32].

The PLS method of [31] starts with centering and scaling the predictor matrix. However, we consider the L_1 median as defined by [35] and the median absolute deviation for robustly centering and scaling the random matrix $\boldsymbol{H}$, respectively. In addition, we consider a weighted version of the random matrix $\boldsymbol{H}$ to reduce the effects of outliers in the basis expansion coefficients. Let us denote by w_i the weight for the i-th observation as follows:

$$w_i = \underset{k}{\text{median}}\{w_{ik}\}, \qquad i = 1,\ldots,n, \quad k = 1,\ldots,K,$$

where w_{ik} is the weight for i-th residual corresponding to the k-th column. Then, similar to [36], we consider weighting the random matrix $\boldsymbol{H}$ by multiplying each row of it by the square root of weights. As proven by [37], the weights obtained from the weighted likelihood-based model hold $\sup|w_i - 1| \xrightarrow{p} 0$. This outcome demonstrates that the weighted and unweighted versions of the design matrix $\boldsymbol{H}$ will be similar when the absence of outliers in the data, and the weighted likelihood procedure tends to perform like to the maximum likelihood technique. On the other side, whensoever accounting for the presence of outliers in the data, the weights downweigh the effects of outliers in the design matrix. In this case, the weighted likelihood method is expected to perform better than the maximum likelihood technique.

Let $\widetilde{\boldsymbol{H}}$ denote the robustly centered and scaled as well as weighted version of $\boldsymbol{H}$. Let also $\widetilde{\boldsymbol{H}}_j$ for $j = 1, \ldots, K$ denote the columns of the design matrix. Then, our proposed RFPLS method for the FLogR model is presented as follows.

1) Calculation of the RFPLS components. For $l = 1, 2, \ldots$; repeat:

   (a) Fit the weighted likelihood based LogR model $Y/\widetilde{\boldsymbol{H}}_j$, $(j = 1, \ldots, K)$. Denote by $\widehat{\boldsymbol{\vartheta}}_l = \left[\widehat{\vartheta}_{l1}, \ldots, \widehat{\vartheta}_{lK}\right]^\top$ the estimated slope parameters and $\boldsymbol{V}_l = [v_{l1}, \ldots, v_{lK}]^\top$ the normalized vector of $\widehat{\boldsymbol{\vartheta}}_l$ so that $v_{lj} = \widehat{\vartheta}_{lj}/\|\widehat{\vartheta}_l\|$ where $\|\cdot\|$ denotes the Euclidean norm.

   (b) Set equal to zero the coefficients v_{lj} which are not significant according to the Wald test. More precisely, the elements of $\boldsymbol{V}_l$ satisfying $\left|\widehat{\vartheta}_{lj}/\mathrm{SE}\left(\widehat{\vartheta}_{lj}\right)\right| \leq z_{\alpha/2}$ is deleted. Here, the $\mathrm{SE}\left(\widehat{\vartheta}_{lj}\right)$ and $z_{\alpha/2}$ denote the the estimated standard deviation of $\widehat{\vartheta}_{lj}$ and $(\alpha/2)$-th quantile of the standard Gaussian distribution, respectively.

   (c) Compute the l-th robust PLS component as $\boldsymbol{T}_l = v_{l1}\widetilde{\boldsymbol{H}}_1 + \cdots + v_{lK}\widetilde{\boldsymbol{H}}_K$.

   (d) Adjust the robust PLS component $\boldsymbol{T}_l$ by replacing the columns of the design matrix $\widetilde{\boldsymbol{H}}$ by their residuals as follows:

$$\boldsymbol{H}_j - \mathrm{E}\left(\boldsymbol{H}_j|\boldsymbol{T}_l\right) = \mathrm{E}\left(\boldsymbol{H}_j\right) + \mathrm{Cov}\left(\boldsymbol{H}_j, \boldsymbol{T}_l\right)\mathrm{Var}\left(\boldsymbol{T}_l\right)^{-1}\left[\boldsymbol{T}_l - \mathrm{E}\left(\boldsymbol{T}_l\right)\right].$$

   The algorithm stops when all the elements of $\boldsymbol{T}_l$ are not significant because none of the estimated slope parameters $\widehat{\vartheta}_{lj}$ is not significant.

2) LogR fitting of Y on the retained robust PLS components.
   Let $\boldsymbol{\Gamma}$ specify the matrix of robust logit PLS components extracted from the design matrix $\widetilde{\boldsymbol{H}}$, i.e., $\boldsymbol{\Gamma} = \widetilde{\boldsymbol{H}}\boldsymbol{V}$, with $\boldsymbol{V}$ being the the matrix with column entries the coefficients of the robust logit PLS components in terms of original predictors. Then, the logit model in terms of the robust PLS components has the following expression:

$$\widehat{\boldsymbol{L}} = \mathbf{1}\widehat{\alpha} + \boldsymbol{\Gamma}\widehat{\boldsymbol{\vartheta}},$$

   where $\widehat{\boldsymbol{\vartheta}} = \left[\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_l\right]^\top$ is the vector of WLEs of the coefficients of the logit model in terms of the l robust PLS components obtained in the first step.

3) Robust estimation of the parameter function.
   The robust PLS regression model in terms of the original predictors can be expressed as follows:

$$\widehat{\boldsymbol{L}} = \mathbf{1}\widehat{\alpha} + \widetilde{\boldsymbol{H}}\boldsymbol{V}\widehat{\boldsymbol{\vartheta}}.$$

Accordingly, the basis expansion coefficient of the regression coefficient curve can be obtained as $\widehat{\boldsymbol{\beta}} = \boldsymbol{V}\widehat{\boldsymbol{\vartheta}}$. Finally, the robust estimation of the regression coefficient function is obtained as $\widehat{\beta}(t) = \widehat{\boldsymbol{\beta}}^{\top}\boldsymbol{\phi}(t)$.

In summary, in our proposed RFPLS approach, the effects of outliers in the functional predictors are downweighed by weighting the basis expansion coefficients (i.e., by weighting the design matrix $\boldsymbol{H}$) and computing the PLS components via the weighted likelihood method. In addition, in the last step of the proposed approach, the basis expansion coefficients of the regression coefficient function are robustly obtained via the weighted likelihood method. Therefore, our proposed approach is robust to both leverage points in the predictor variable and vertical outliers in the response variable.

3.3 Numerical Results

3.3.1 Monte-Carlo Experiments

All along the Monte Carlo experiments, we consider two different data generation processes as follows:

Case 1. In this case, we consider the data generation process of [9]. First, we produce the trajectories of the functional predictor $X(t)$ at 256 equally spaced point in the interval $T = [0, 10]$. While doing so, 13 cubic B-spline basis functions $\{\phi_k(t)\}_{k=1}^{13}$ are generated at nine equally spaced knots over the interval $[0, 10]$. In addition, the random basis expansion coefficients are generated as $\boldsymbol{C} = \boldsymbol{Z}U$, where $\boldsymbol{Z}$ is a 500×13 matrix of random values from the standard Gaussian distribution and $\boldsymbol{U}$ is a 13×13 random matrix consisting of values generated from the uniform distribution on $[0, 1]$. Then, the trajectories of the functional predictor is generated as $X_i(t_j) = \sum_{k=1}^{13} c_{ik}\phi_k(t_j)$ where $j \in [0, 10]$. Finally, the values of the binary response are generated as a linear functional of the functional predictor and parameter function as follows:

$$l_i = \int_0^{10} X_i(t)\beta(t)dt,$$

where $\beta(t) = \sin\left(\frac{t\pi}{3}\right)$ and $t \in [0, 10]$.

In this case, we consider two scenarios. In the initial scenario, the data set is generated from the smooth process as presented above. In this scenario, we aim to confirm the correctness of the proposed method so that the proposed RFPLS approach is expected to perform similarly to its traditional competitors. In the second scenario, the randomly selected $n \times [1\%,\ 5\%,\ 10\%]$ of the

generated data are contaminated by outliers. In doing so, the randomly selected $n \times [1\%,\ 5\%,\ 10\%]$ trajectories from each class of the binary response of the functional predictor are generated similar to the first scenario but using $\boldsymbol{U} \sim$ uniform on $[0,5]$ and $\beta(t) = 2\sin\left(\frac{4t\pi}{3}\right)$ where $t \in [0,10]$. In addition, the corresponding binary response values are replaced by opposite classes so that the response variable contains outliers. Thus, we research the performance of the proposed method against also vertical outliers besides that leverage points. In this case, the proposed RFPLS approach is expected to generate improved empirical performance over the classical methods. Figure 3.1 shows a graphical representation of the data generated in this case.

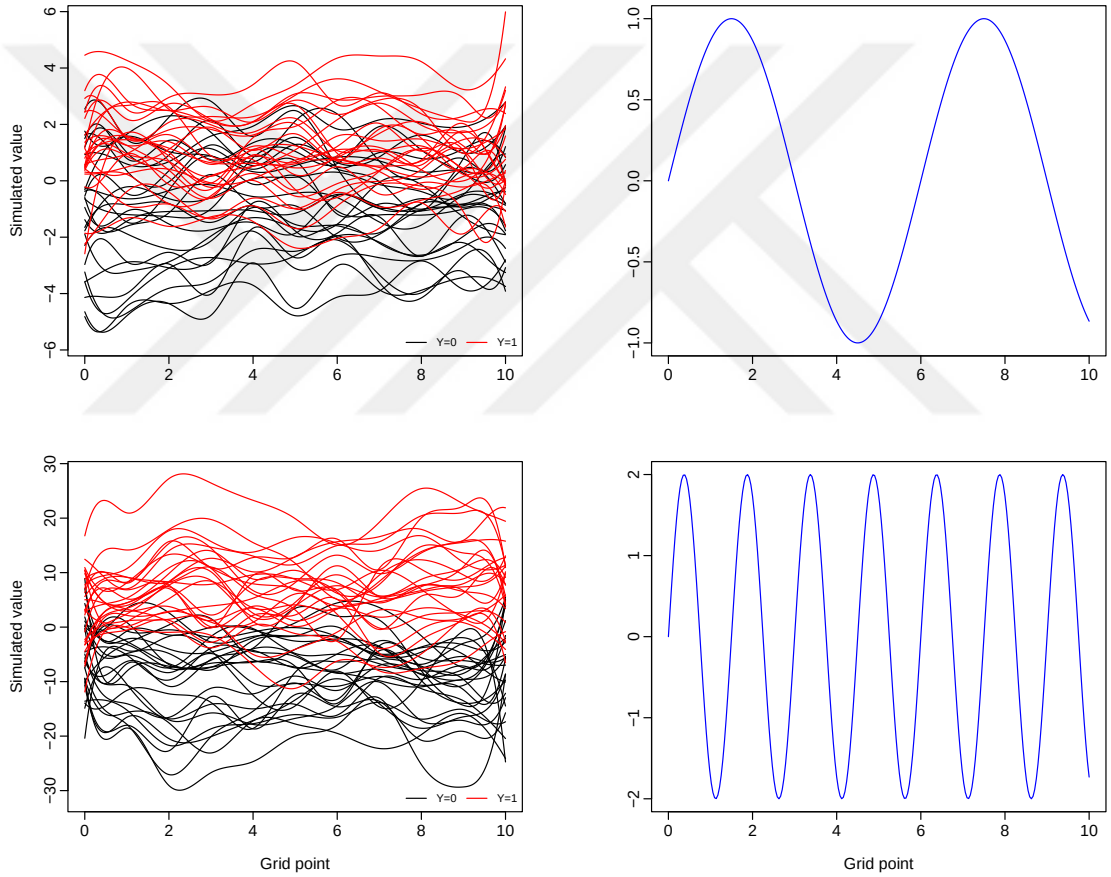


Figure 3.1 Plots of the generated 25 sample curves for both $Y = 0$ and $Y = 1$ (left boards; scenario 1 (first line) and scenario 2 (second line)) and actual generated regression coefficient function (right boards; scenario 1 (first line) and scenario 2 (second line)).

Case 2. In the second case, we focus on the curve discrimination accuracy of the FLogR model as in [12]. In this case, two different functional predictors are produced at 101 equally spaced point in the interval $T \in [1,21]$ based on the classes of the binary response, i.e., $Y = 0$ or $Y = 1$. For the first class ($Y = 0$),

the trajectories of the functional predictor are generated as follows:

$$X(t) = uh_1(t) + (1-u)h_2(t) + \varepsilon(t),$$

where u is a random variable generated from the uniform distribution on $[0,1]$, $\varepsilon(t)$ is a random error generated from the standard Gaussian distribution, $h_1(t) = \max\{6-|t-11|,0\}$, and $h_2(t) = h_1(t-4)$. For the second class ($Y=1$), on the other side, the trajectories of the functional predictor are generated as follows:

$$X(t) = uh_1(t) + (1-u)h_3(t) + \varepsilon(t),$$

where $h_3(t) = h_1(t+4)$.

Similar to **Case 1**, we also consider two scenarios in this case. In the first scenario, the data set is generated as presented above. In the second scenario, the randomly chosen $n \times [1\%,\ 5\%,\ 10\%]$ trajectories of the functional predictor are generated as follows:

$$X(t) = uh_4(t) + (1-u)h_2(t) + \varepsilon^*(t), \qquad \text{(when } Y=0\text{)},$$
$$X(t) = uh_4(t) + (1-u)h_3(t) + \varepsilon^*(t), \qquad \text{(when } Y=1\text{)},$$

where $\varepsilon^*(t) \sim N(5,1)$ and $h_4(t) = h_1(t+50)$. In addition, the randomly chosen $n \times [1\%,\ 5\%,\ 10\%]$ of the binary response values are replaced by opposite class. Figure 3.2 displays a graphical representation of the data generated under this scenario.

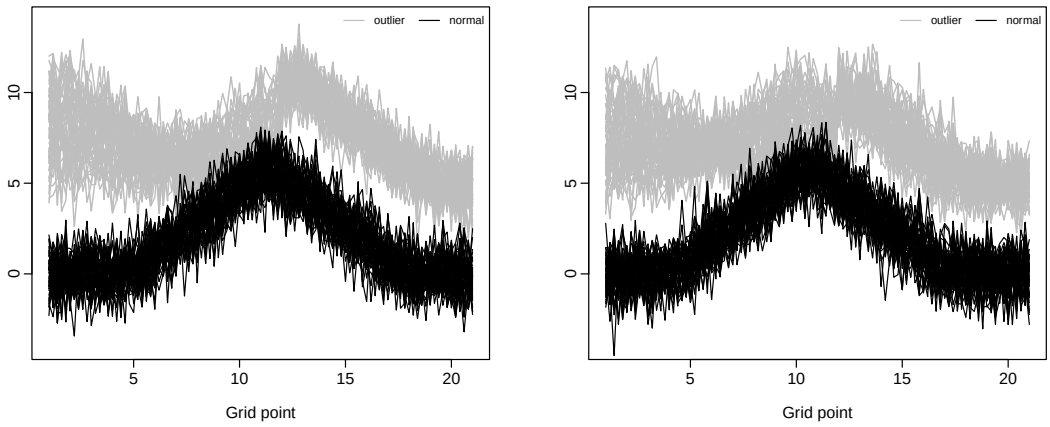


Figure 3.2 Plots of the generated 50 sample curves for $Y=0$ (left board) and $Y=1$ (right board). The curves generated under scenario 1 are given in black while the curves generated under scenario 2 are given in gray..

For each case and scenario, $n = 1000$ trajectories for the functional predictor are generated. The following procedure is repeated 500 times to compare the empirical

performance of the proposed RFPLS and traditional FPLS and FPC methods. First, the generated data set is randomly divided into training that include outliers and test samples with sizes 700 and 300, respectively. In each replication, the correct classification ratio (CCR) by taking 0.5 as a cut-point and the area under the receiver operating characteristic curve by taking into account all the possible cut-points (AUC) are computed to evaluate the classification performance of the methods. In addition, the integrated mean squared error (IMSE) is calculated under **Case 1** to compare the estimation performance of the methods as follows:

$$\text{IMSE} = \int_T \left[\beta(t) - \widehat{\beta}(t)\right]^2 dt.$$

We note that in the Monte Carlo experiments, the data generated without outliers are first divided into the training and test samples. Then, only the randomly selected observations in the training samples are replaced by outliers. We also note that $K = 15$ basis expansion functions are used in the estimation phase of the methods.

Our results are presented in Figures 3.3-3.6. From Figures 3.3 and 3.5, if there are no outliers present in the data, the proposed approach yields similar or slightly smaller CCR and AUC values compared with the traditional FPLS and FPC procedures. On the other side, while outliers are available in the data, our proposed method importantly improves CCR and AUC values than FPLS and FPC. Especially, the difference between the classification performance of the proposed approach and its traditional competitors becomes more prominent as the contamination level increases. Compared with FPC, the FPLS is generally more affected by outliers and produces smaller CCR and AUC values.

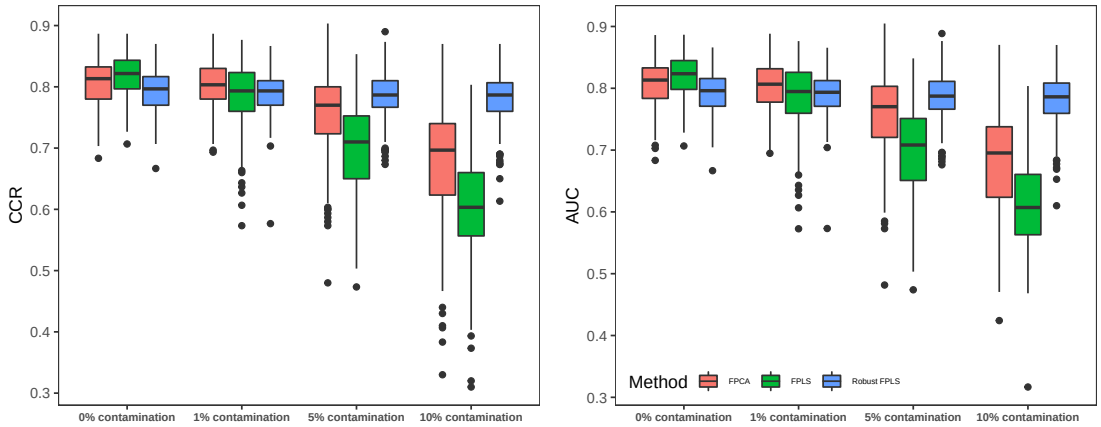


Figure 3.3 Boxplots of the computed CCR and AUC values under Case 1 for the FPC, FPLS, and RFPLS methods.

The results for the computed IMSE under Case 1 are presented in Figure 3.4.

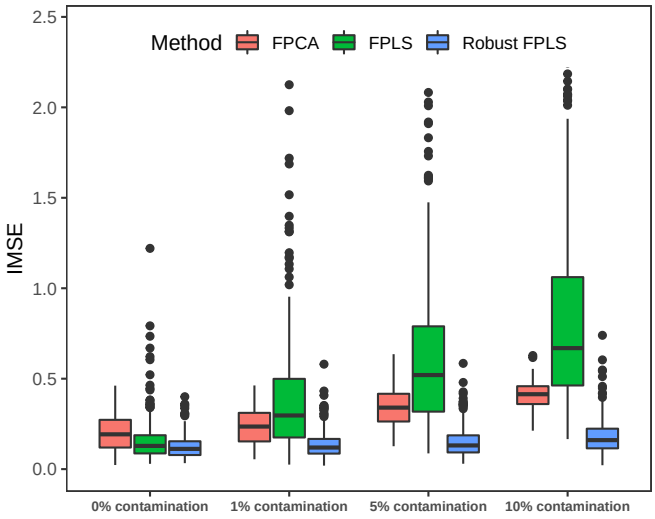


Figure 3.4 Boxplots of the computed IMSE values under Case 1 for the FPC, FPLS, and RFPLS methods.

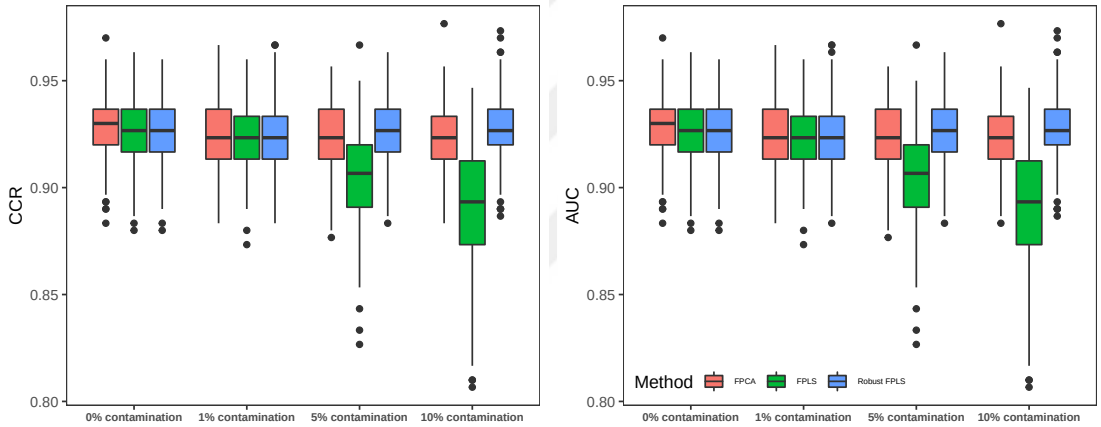


Figure 3.5 Boxplots of the computed CCR and AUC values under Case 2 for the FPC, FPLS, and RFPLS methods.

This figure shows that our proposed approach generates better parameter estimates than the FPC and FPLS for all scenarios. Note that our proposed method is expected to produce similar IMSE values when no outlier is present in the data. However, it generates improved IMSE values even in this case. This result is because the considered data generation process may generate some outliers with small or moderate magnitudes. The proposed method downweighs such observations and produces smaller IMSE values than other methods. In addition, the plots of the generated true regression coefficient function and its estimates by the methods are presented in Figure 3.6. This figure also supports our findings presented in Figure 3.4 and further confirm that our proposed method produces improved parameter estimates in all scenarios compared with the FPC and FPLS.

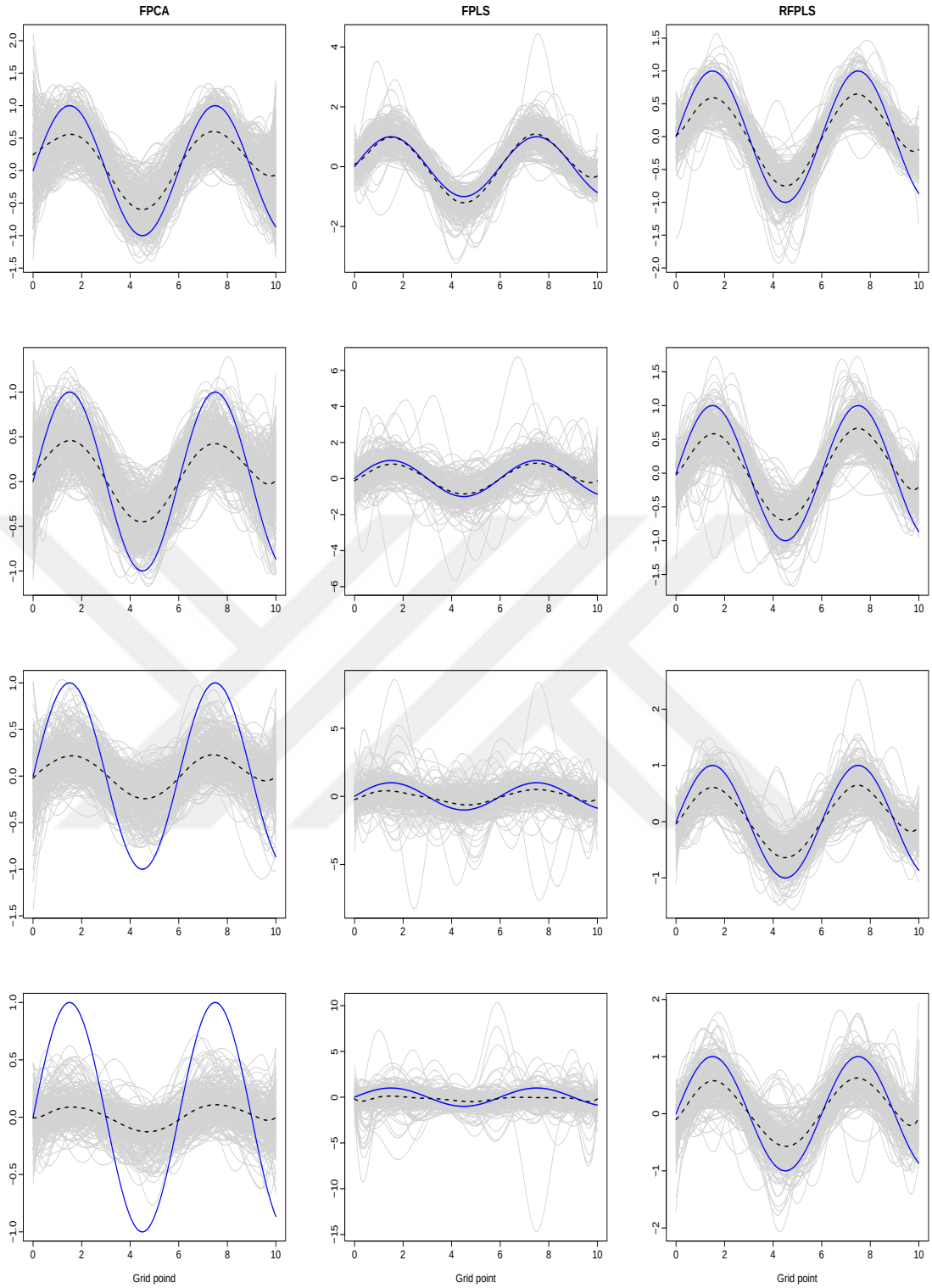


Figure 3.6 Plots of the generated true parameter functions (blue lines), estimated parameter functions (gray lines), and mean functions of the estimated parameter functions (black lines). FPC (first stripe), FPLS (second stripe), and RFPLS (third stripe). The data are created under Case 1 when no outlier is present in the data (first row), [1%, 5%, 10%] of the generated data are contaminated by outliers (second row, third row, and fourth row, respectively.)

3.3.2 Real-world Examples

Growth data: The growth data available in the R package "fda" [38] include the heights of 54 girls and 39 boys between the ages of 1 and 18, measured at 31 discrete time points. For this dataset, the presence of outliers is investigated via the functional boxplot technique proposed by [39] (existing in the R package "fdaoutlier" [40]) and 4(4.3%) outliers are detected. However, as presented in Figure 3.7, the magnitudes of the detected outliers are very small, and thus, the proposed method is expected to generate similar results with the FPLS and FPC.

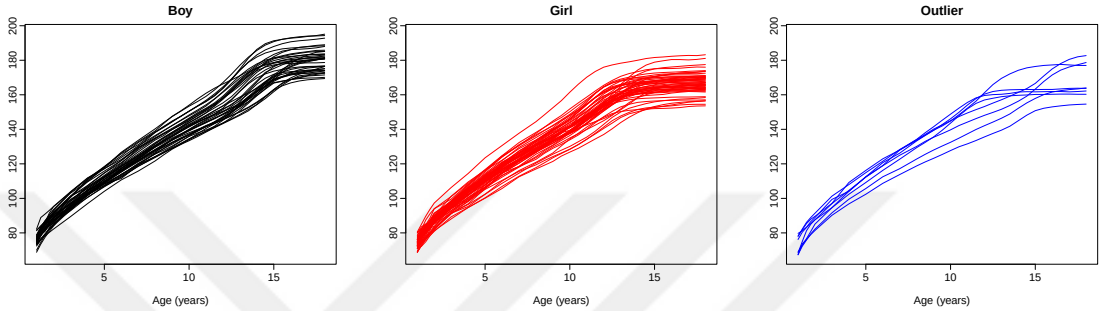


Figure 3.7 Graphical display growth curves of the boys (left board), girls (middle board), and detected outliers (right board).

With this dataset, our aim is to classify kids (boy or girl) using their growth curves. To compare the classification performance of the methods, the following procedure is repeated for 250 times. The complete dataset is split into a training set of size $n_{train} = 53$ and a test set of size $n_{test} = 40$. Models that using the training sample are constructed to classify the observations in the test sample based on the CCR and AUC. Note that to estimate the constructed models, $K = 15$ basis expansion functions are used for all the methods. Our outcomes are displayed in Figure 3.8, which indicates that both traditional techniques produce slightly better CCR and AUC values than the proposed method. This is not a suprising result since the growth data include a small percentage of outliers with very small magnitudes.

Strawberry puree data: The strawberry purees data available at `https://csr.quadram.ac.uk` include a total of 983 mid-infrared spectrum collections, where 351 of which are pure strawberries (authentic samples) and 632 of which are non-pure strawberries (adulterated strawberries and other fruits). The fruit purees were collected using the Fourier Transform Infrared (FTIR) spectroscopy with attenuated total reflectivity sampling on the integrated spectral area over the range 899 $cm^{-1} - 1802\ cm^{-1}$ (235 points in total) [see 41, for a more detailed description of the dataset]. With this dataset, our aim is to classify the fruit purees (strawberry or non-strawberry) based on the mid-infrared spectra

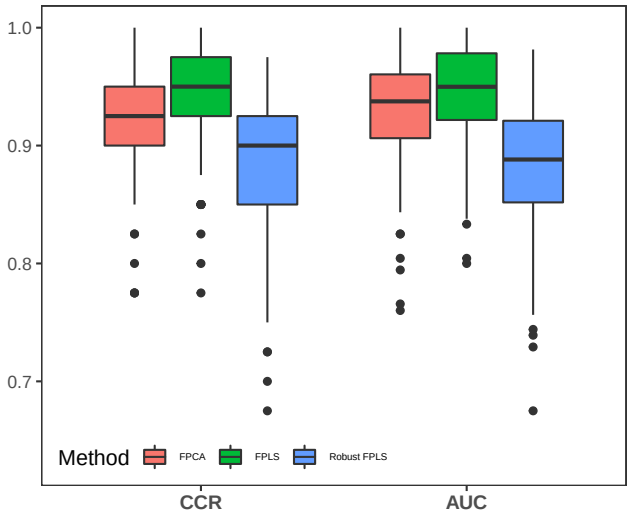


Figure 3.8 Boxplot of the computed CCR and AUC values for the growth data.

of fruit purees. According to the functional boxplot method, the functional predictor includes 122 (12.4%) outliers (please see Figure 3.9) with small and large magnitudes, which may lead to poor classification performance for the non-robust methods. The presence of a large amount of outliers in this dataset excite us to apply our proposed approach to robustly classify the fruit purees.

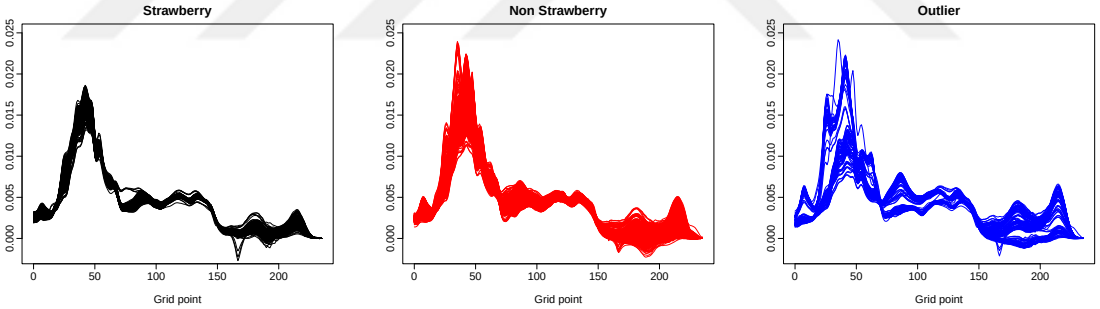


Figure 3.9 Graphical display of the mid-infrared spectra of strawberry fruit purees (left board), non-strawberry fruit purees (middle board), and detected outliers (right board).

The following procedure is repeated 250 times to compare the classification performance of the techniques. The complete data set is split into a training sample with size $n_{train} = 583$ and a test sample with size $n_{test} = 400$. The models are built with the training sample to classify the fruit purees in the test sample based on the CCR and AUC. Note that $K = 15$ basis expansion functions are used to estimate the constructed methods. The outcomes of our study are illustrated in Figure 3.10. From this figure, it is evident that the proposed approach importantly outperforms the FPLS and FPC so that it produces improved CCR and AUC values compare with the FPLS and FPC. In other words, the non-robust methods (FPLS and FPC) are affected by the

outliers and produce mis-classification results for the binary response variable. On the other hand, our proposed method improves classification accuracy for the response variable by downweighing the effects of the outliers.

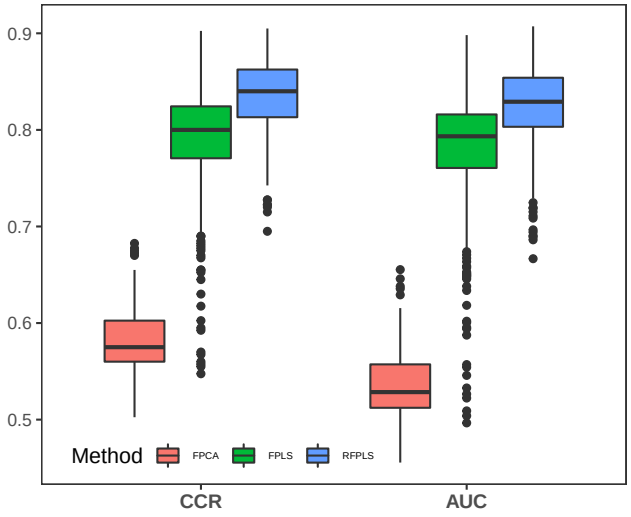


Figure 3.10 Boxplot of the computed CCR and AUC values for the strawberry puree data.

4
FUNCTIONAL QUANTİLE AUTOREGRESSIVE REGRESSION

In this chapter, we introduce the functional autoregressive regression model of order one and consider to integrate the quantile regression proposed by [42] into this model and approximate the regression coefficient function using the FPC decomposition of the functional response and prediction variable.

4.1 Functional Autoregressive Regression

Let $Y = \{Y_i(t) : i = 1, 2, \ldots, N\}$ denote a random sample of a functional time series, where each element $Y_i(t)$ of the sample is an observation of square integrable function belonging to a separable Hilbert space H with a bounded and closed interval T, $t \in T$. Suppose that $Z = \{Z_j(t) = Y_i(t) : i = 2, \ldots, N\}$ and $X = \{X_j(t) = Y_i(t) : i = 1, \ldots, N-1\}$, for $j = 1, \ldots, J$, are the functional response and a functional predictor (which is one lag of Z), respectively. [43] described the functional autoregressive model of order one FAR(1) process as follows:

$$\mathrm{E}\big[Z_j(t)|X_j(s)\big] = \beta_0(t) + \int_T X_j(s)\beta(s,t)ds + \varepsilon_j(t), \qquad \forall s, t \in T, \tag{4.1}$$

where $\mathrm{E}\big[Z_j(t)|X_j(s)\big]$ is the conditional mean of functional response variable given functional lagged predictor variable $X_j(s)$, $\beta_0(t)$ is the intercept function, $\beta(s,t)$ is the bivariate parameter function, and $\varepsilon_j(t)$ is the random error function which is assumed to be mean-zero process.

4.2 Proposed Method

The proposed FQAR(1) model is defined as follows:

$$Q_\tau\left[Z_j(t)|X_j(t)\right] = \beta_{0\tau}(t) + \int_T X_j(s)\beta_\tau(s,t)ds, \qquad \forall s,t \in T, \quad \tau \in (0,1). \tag{4.2}$$

where $Q_\tau\left[Z_j(t)|X_j(t)\right]$ denotes the conditional quantile of the response variable $Z_j(t)$ given the functional lagged variable $X_j(t)$ and $\beta_{0\tau}(t)$ and $\beta_\tau(s,t)ds$ are the intercept function and regression parameter function for a given quantile level $\tau \in (0,1)$, respectively. Let $\varepsilon_j(t) = Z_j(t) - Q_\tau\left[Z_j(t)|X_j(s)\right]$ denote the error function. Then, the parameters of the proposed FQAR(1) model are estimated by minimizing the sum of the weighted absolute deviations as follows:

$$\text{argmin}_{\beta_{0\tau}(t),\beta_\tau(s,t)} \sum_{\varepsilon_j(t)\geq 0} \tau\left|\varepsilon_j(t)\right| + \sum_{\varepsilon_j(t)<0} (1-\tau)\left|\varepsilon_j(t)\right| \tag{4.3}$$

It is clear from (4.3) that the error functions have positive and negative values based on their signs. The estimates of the parameter functions for each quantile regression model are obtained based on the entire dataset, not the observations in the given quantile level. Therefore, the estimation phase of the quantile parameters are computationally intensive and requires the use of linear programming algorithms such as simplex and Frisch-Newton interior point [see 44, 45, for more detailed information]. The solution to the minimization problem given in (4.3) is not possible with the existing algorithms because the functional objects belong to an infinite-dimensional space. To overcome this problem, the functional principal component (FPC) analysis method, where the basis functions are eigenfunctions extracted from data, is considered.

Based on the FPC decomposition, the random curves can be expressed as

$$Z_j(t) = \sum_{k=1}^{K} \delta_{jk}\phi_k(t) = \delta_j^T \phi(t), \quad X_j(s) = \sum_{k=1}^{K} \vartheta_{jk}\theta_k(s) = \vartheta_j^T \theta(s),$$

where δ_j and ϑ_j denote the FPC scores corresponding to the response and predictor variables, respectively. Similarly, the intercept and bivariate parameter functions can be expressed in the same bases as follows:

$$\beta_{0\tau}(t) = B_{0\tau}\phi(t), \quad \beta_\tau(s,t) = \theta(s)^T B_\tau \phi(t).$$

Then, the proposed FQAR(1) model in (4.2) can be rewritten as follows:

$$\begin{aligned}
\delta_j^T \phi(t) &= B_{0\tau}\phi(t) + \int_T \vartheta_j^T \theta(s)\,\theta(s)^T B_\tau \phi(t), \\
\delta_j^T \phi(t) &= B_{0\tau}\phi(t) + \vartheta_j^T B_\tau \phi(t), \\
\delta_j^T \int_T \phi(t)\,\phi(t)^T &= B_{0\tau}\int_T \phi(t)\,\phi(t)^T + \vartheta_j^T B_\tau \int_T \phi(t)\,\phi(t)^T, \\
\delta_j^T &= B_{0\tau} + \vartheta_j^T B_\tau, \qquad (4.4)
\end{aligned}$$

where $B_{0\tau}$ and B_τ denote the intercept and regression coefficient for the regression model constructed using FPC scores, respectively. The above result indicate that the infinite-dimensional FQAR(1) model in (4.2) can be converged to a finite-dimensional quantile regression problem (4.4) using the FPC basis functions and scores.

Let $e_j = \delta_j - Q_\tau\left[\delta_j|\vartheta_j\right]$ denote the residual obtained from the finite-dimensional quantile regression model (4.4). Then, the estimates of the parameters $\hat{B}_{0\tau}$, $\hat{B}_\tau$ are obtained as the solution to the following minimization problem

$$\text{argmin}_{B_{0\tau},B_\tau} \quad \sum_{e_j\geq 0}\tau\left|e_j\right| + \sum_{e_j<0}(1-\tau)\left|e_j\right|,$$

which can easily be done using the available R package "quantreg" [46]. Let $\hat{B}_{0\tau}$ and $\hat{B}_\tau$ denote the quantile estimates of $B_{0\tau}$ and B_τ, respectively. Then, the estimates of the regression parameter functions are obtained as follows:

$$\begin{aligned}
\hat{\beta}_{0\tau}(t) &= \hat{B}_{0\tau}\phi(t), \\
\hat{\beta}_\tau(s,t) &= \theta(s)^T \hat{B}_\tau \phi(t).
\end{aligned}$$

Finally, the one-step-ahead quantile forecast of the response variable conditionally on the available time N can be obtained as follows:

$$\hat{Q}_\tau\left[Y_{N+1}(t)|Z_J(s)\right] = \hat{\beta}_{0\tau}(t) + \int_T Z_J(s)\hat{\beta}_\tau(s,t)ds.$$

4.3 Numerical Results

4.3.1 Monte-Carlo Experiment

A Monte Carlo experiment is performed to evaluate the finite sample performance of the method and its performance is compared with three existing functional time series methods: 1) FAR(1) model integrated with the functional partial least squares regression (FPLSR), 2) functional principal component regression (FPCR), and

the autoregressive integrated moving average (ARIMA)-based functional time series method, which are the available in the R package "ftsa" [47]. An example R code for all the methods are available at (we will provide link here).

Throughout the experiments, 251 random curves at 101 equally spaced time points in the interval $T = [1, 10]$ are generated using the following process:

$$Y_i(t) = 0.3Y_{i-1}(t) + \int_1^{10} V_i(s)\beta(t,s)\,ds + \varepsilon_i(t),$$

where $\varepsilon_i(t)$ is the random error trajectory simulated from the zero-mean Ornstein Uhlenbeck process and $\beta(t,s) = e^{\frac{1}{5}t}\sin(\frac{\pi s}{2})$. In addition, the following process is used to generate $V_i(s)$:

$$V_i(s) = \int_1^{10} \Psi(s,l)\,V_{i-1}(l) + B_i(s),$$

where $B_i(s)$ is a realization of an independently and identically distributed standard Brownian motion stochastic process and $\Psi(s,l) = 0.34\exp^{\frac{1}{2}s^2+l^2}$. A graphical display of the generated random curves and the bivariate regression parameter function is presented in Figure 4.1.

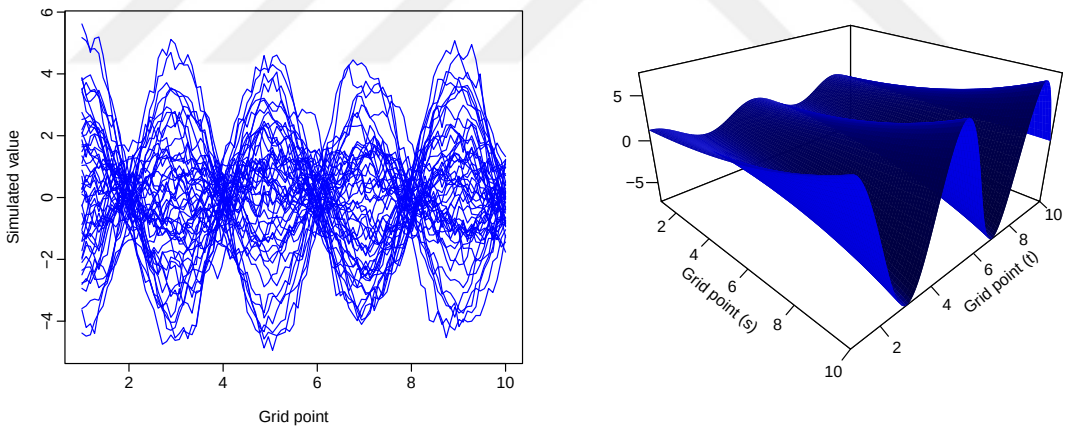


Figure 4.1 Plots of the generated 50 sample curves (left board) and generated bivariate coefficient function $\beta(t,s)$ (right board).

Throughout the experiments, $K = 5$ FPCs with 15 B-spline basis functions, that capture at least 98% variation in the data, are used to approximate the functional variables. In addition, because all the benchmark methods are based on the conditional mean regression, only $\tau = 0.5$ quantile level is considered for the proposed FQAR(1) method. With these choices, the aim is to show the correctness of the proposed method. Because the generated data are based on conditional mean regression, the proposed method is expected to produce similar forecasting accuracy to those of existing methods.

In order to evaluate the one-step-ahead forecasting performance of the methods, the first 250 curves are used to build models and the last curve is forecasted based on the constructed models. This process is repeated for MC = 1000 times. In each replication, various performance metrics, such as mean absolute prediction error (MAPE), root mean squared prediction error (RMSPE), root median squared prediction error (RMESPE), relative error (RE), Nash-Sutcliffe efficiency (NSE), and modified index of agreement (MIA) are computed as follows:

$$\text{MAPE} = \frac{1}{101}\sum_{m=1}^{101}\left|\frac{\hat{Y}(t_m) - Y(t_m)}{Y(t_m)}\right| \qquad \text{MIA} = 1 - \frac{\sum_{m=1}^{101}\left|Y(t_m) - \hat{Y}(t_m)\right|}{\sum_{m=1}^{101}\left(\left|Y(t_m) - \bar{Y}(t)\right| + \left|\hat{Y}(t_m) - \bar{Y}(t)\right|\right)}$$

$$\text{RMSPE} = \sqrt{\frac{1}{101}\sum_{m=1}^{101}\left(\frac{\hat{Y}(t_m) - Y(t_m)}{Y(t_m)}\right)^2} \qquad \text{RE} = \left(\frac{1}{101}\sum_{m=1}^{101}\frac{\hat{Y}(t_m) - Y(t_m)}{Y(t_m)}\right) \times 100$$

$$\text{RMESPE} = \sqrt{\mathscr{M}\left(\frac{\hat{Y}(t_m) - Y(t_m)}{Y(t_m)}\right)^2} \qquad \text{NSE} = 1 - \frac{\sum_{m=1}^{101}\left(Y(t_m) - \hat{Y}(t_m)\right)^2}{\sum_{m=1}^{101}\left(Y(t_m) - \bar{Y}(t)\right)^2}$$

where $Y(t_m)$ and $\hat{Y}(t_m)$ respectively are the observed and forecasted values at the m-th time point, $\bar{Y}(t)$ and $\bar{\hat{Y}}(t)$ denote the means of the observed and forecasted functions, respectively, and $\mathscr{M}$ denotes the median operator.

The computed performance metrics are presented in Figure 4.2. From this figure, it is evident that all the methods produce similar results. In other words, the proposed method with quantile level $\tau = 0.5$ (i.e., median regression) provides consistent forecasts to those obtained by the existing conditional mean-based functional time series methods.

4.3.2 Real-world Example

The proposed method is applied to three river flow datasets obtained from three different stations in Australia (Mary River, Lockyer Valley, and Albert River). For each river, the dataset consist of hourly river flow measurements obtained from January 2009 to December 2014 (6 years in total). For each dataset, the observations are assumed to be functions of hours, i.e., there are 2190 functional observations $Y_i(t)(1 \le t \le 24,\ i = 1, \ldots, 2190)$ in total. A graphical display of the classical time series and functional time series plots of the river flow data is presented in Figure 4.3. Note that in the functional time series plots, the red colored curves denote the river

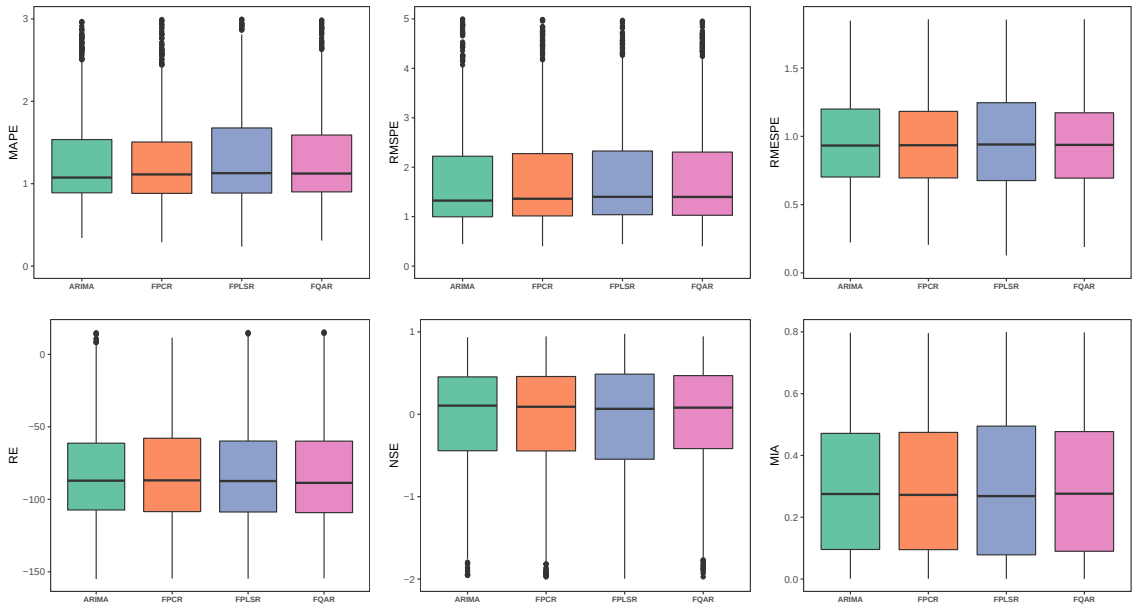


Figure 4.2 Boxplots of the computed MAPE, RMSPE, RMESPE (first row), RE, NSE, and MIA values (second row) for the ARIMA, FPCR, FPLSR, and FQAR methods. Some extreme results are discarded from the plots to provide a better visualization.

flow measurements for distant-past years and the river flow measurements for more recent years are given by violet. From this figure, compared with Lockyer Valley, Mary and Albert rivers include more heteroscedastic measurements. Because the proposed method is more robust to heteroscadasticity compared with conditional mean-based methods, it is expected to produce improved results for these stations.

For each dataset, an expanding-window approach is considered to investigate the forecasting accuracy of the methods. In doing so, the entire data are split into two parts: a training sample consisting of days 1/1/2009 to 1/10/2014 (2100 days in total) and a test sample comprising days from 2/10/2014 to 31/12/2014 (90 days in total). For each dataset, the entire training sample is used to forecast river flow measurement on 2/10/2014. Then the forecast measures for 3/10/2014 are obtained by increasing the training samples one day. This procedure is repeated until the training samples cover the entire datasets. For each dataset, $K = 5$ FPCs, that capture at least 99% of the total variation in the functional time series, are used to decompose the river flow curve time series. Similar to the Monte Carlo experiments, the quantile level is set to $\tau = 0.5$ to compare the forecasting accuracy of the proposed method with the existing ones. In each replication, the performance metrics are computed and the results are presented in Figure 4.4. From Figure 4.4, the proposed method generates smaller error values than all the benchmark models for all datasets. The superior performance of the proposed FQAR model is more prominent in Mary River because this river has considerably higher heteroscedastic riverflow measurements

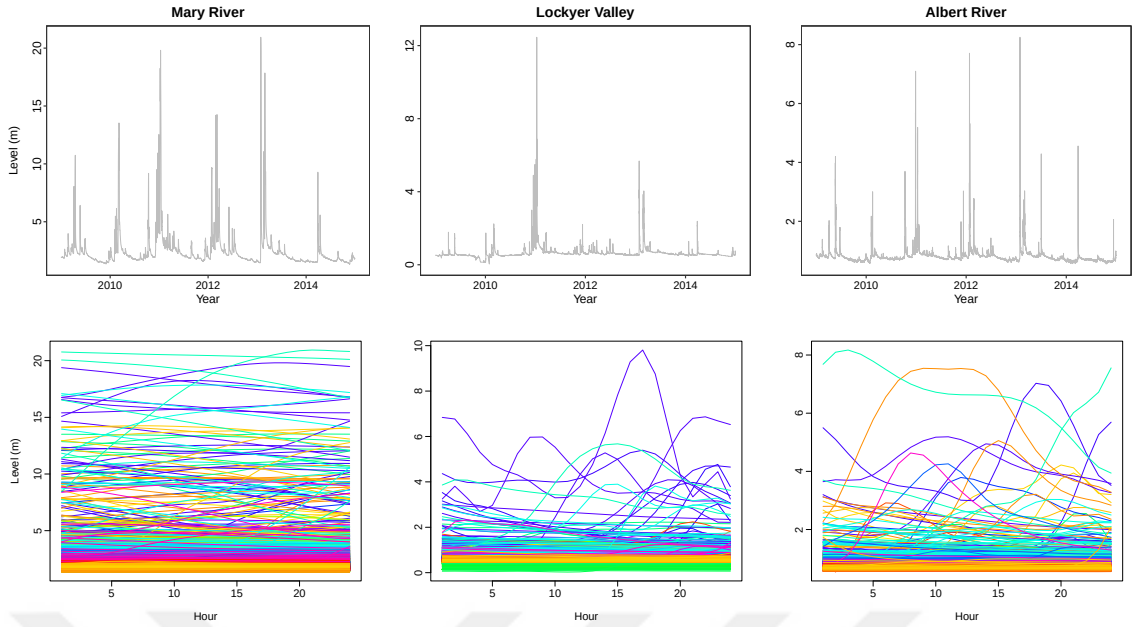


Figure 4.3 Traditional (first row) and smoothed functional time series (second row) of the flow level for three different stations in the Australian region: Mary river (first column), Lockyer Valley (second column), and Albert river (third column).

than those of Lockyer Valley and Albert River. For the existing functional time series methods, while the ARIMA-based method produce improved forecast accuracy than the FPCR and FPLSR-based methods for Lockyer Valley and Alvert River, it produce the worst performance metrics for the Mary River. Note that all these results are based on the conditional mean regression-based (existing methods) and median regression (proposed method with $\tau = 0.5$)-based methods. The proposed method is capable of characterizing the all conditional distribution of the response variable, i.e., it can be used to obtain forecasts at non-central locations. To demonstrate this capability of the proposed method, the forecasts of the last observations in the datasets obtained using distinct quantile levels ($\tau = 0.25, 0.50, 0.75, 0.90$) and the corresponding estimated bivariate regression parameter functions are presented in Figure 4.5. The results presented in this figure can be considered as possible scenarios that can occur for each river, i.e., the forecasts obtained with $\tau = 0.25$ can be used to prevent drought while the results obtained with $\tau = 0.9$ can be used to prevent river flooding. Moreover, the Taylor diagrams obtained using the observed and forecasted river flow measurements are presented in Figure 4.5 to further research the forecasting accuracy of the methods.

The proposed approach can be used to generate pointwise prediction interval for the future realizations of the river flow measurements by applying different quantile levels. The prediction interval may provide better inference than the point forecast because it takes into account the uncertainty of each point forecasts. The prediction interval can be obtained as $\left[\hat{Q}_{\alpha/2}, \hat{Q}_{(1-\alpha/2)}\right]$ for $100(1-\alpha)\%$ nominal level, where

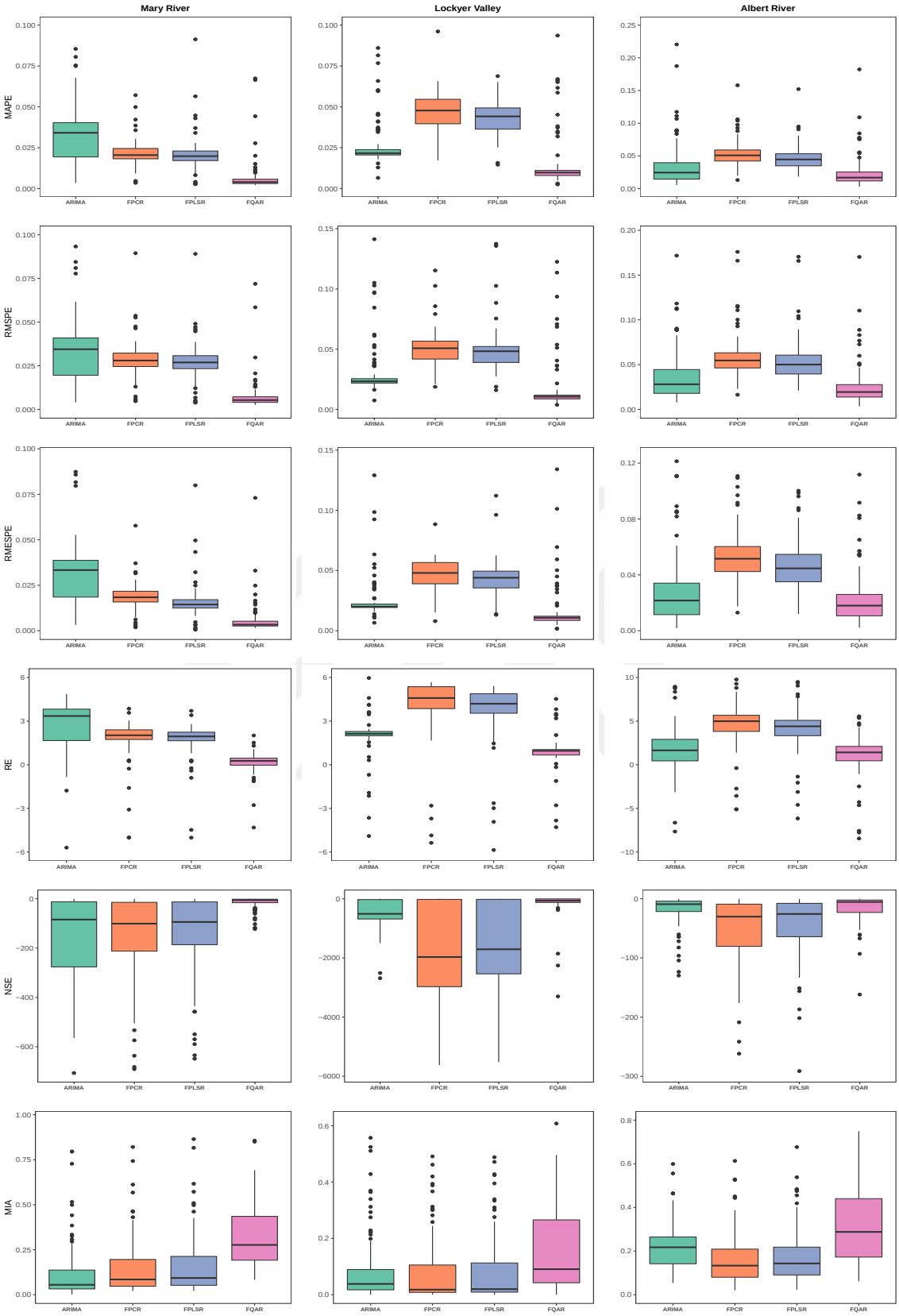


Figure 4.4 Boxplots of the calculated MAPE, RMSPE, RMESPE, RE, NSE, and MIA values for the ARIMA, FPCR, FPLSR, and FQAR methods. Rows and columns represent performance metrics and stations, respectively.

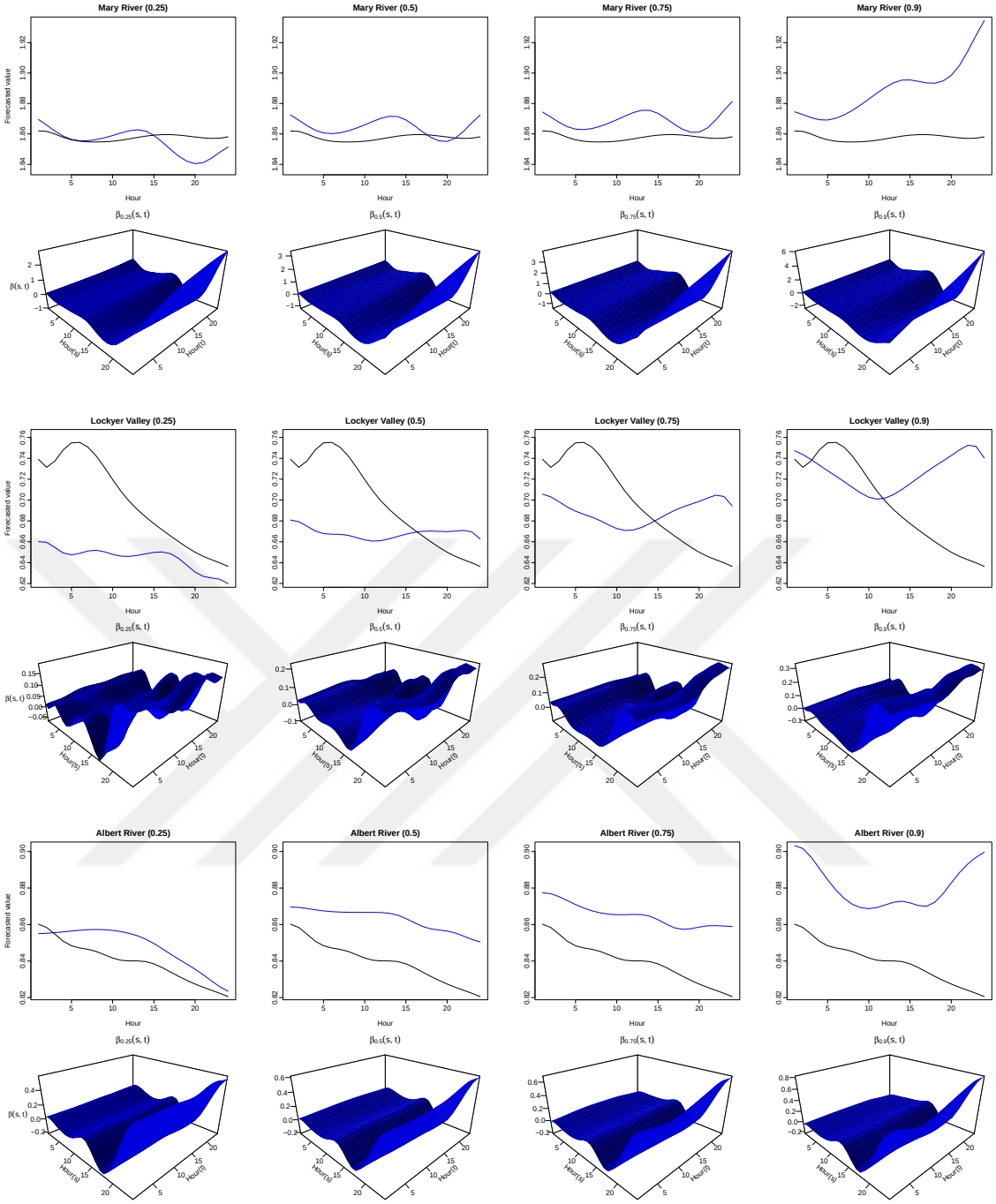


Figure 4.5 An example demonstration of one-step-ahead point forecasts (blue lines) vs observed data point (black lines) and parameter functions at $0.25, 0.5, 0.75, 0.9$ quantile levels (columns, respectively) for all stations (in twos rows, respectively) obtained by the FQAR method.

$\hat{Q}_{\alpha/2}$ denotes $(\alpha/2)^{\text{th}}$ quantile. For each dataset, the 95% (i.e., $\alpha = 0.05$) pointwise prediction intervals for the river flow measurements are computed by using quantile levels $\tau = 0.025(\alpha/2)$ and $\tau = 0.975(1-(\alpha/2))$ quantile levels, respectively. To investigate the interval forecast accuracy of the proposed approach, the interval score which evaluates the coverage probability and the width of the prediction interval

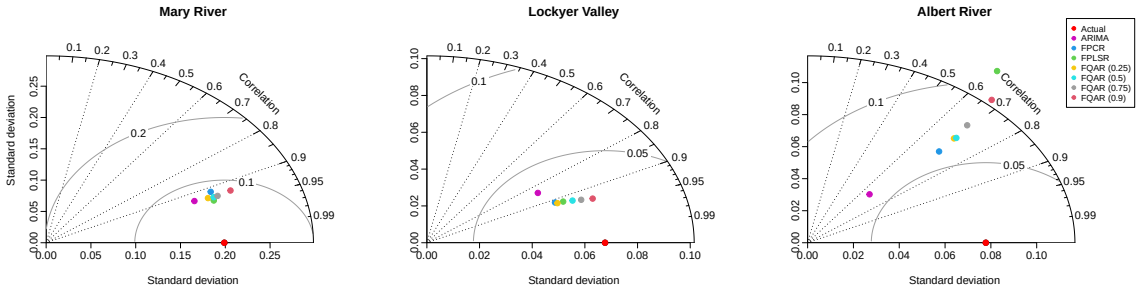


Figure 4.6 Taylor diagrams of the one-step-ahead forecasts of flow level for three different stations in the Australian region with ARIMA, FPCR, FPLSR, and FQAR methods: Mary river (first column), Lockyer valley (second column), and Albert river (third column).

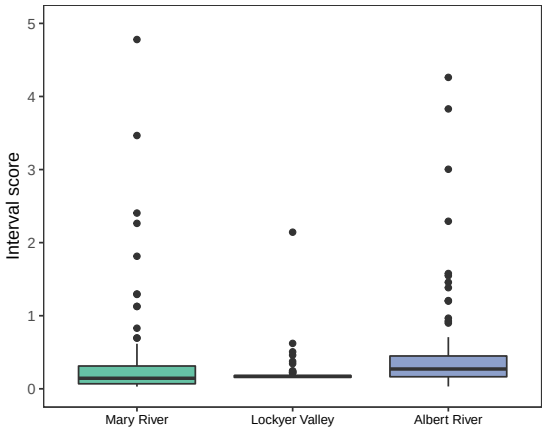


Figure 4.7 Box plots of computed interval score values for one-step-ahead prediction intervals generated by the FQAR method.

together are computed as follows:

$$\begin{aligned}\text{Interval score} = \frac{1}{24}\sum_{t=1}^{24}\left(\hat{Q}_{(1-\alpha/2)}(t) - \hat{Q}_{\alpha/2}(t)\right) + \\ \frac{2}{\alpha}\left(\hat{Q}_{\alpha/2}(t) - Y(t)\right)1\left\{Y(t) < \hat{Q}_{\alpha/2}(t)\right\} + \\ \frac{2}{\alpha}\left(Y(t) - \hat{Q}_{(1-\alpha/2)}(t)\right)1\left\{Y(t) > \hat{Q}_{(1-\alpha/2)}(t)\right\}\end{aligned}$$

where $1\{\cdot\}$ represents the indicator function. The results are presented in Figure 4.7. This figure demonstrates that the pointwise prediction intervals obtained by the proposed FQAR method produces intervals score values close to zero (except for some outlying observations). In other words, the future realizations of the river flow measurements are well covered by the prediction intervals constructed by using the proposed method. The computed interval score values close to zero is also an indication that the prediction intervals constructed by the proposed method are relatively narrow.

5
RESULTS AND DISCUSSION

This thesis introduces two novel approaches for analyzing functional data. The first approach is called RFPLS, which is a robust method for estimating the regression coefficient function of the FLogR model in the presence of outliers. The FLogR model is a popular method for analyzing functional logistics regression model, which involves a functional predictor and a binary response variable. However, the presence of outliers in the functional predictor and binary response can significantly affect the performance of the model. To address this issue, we propose a robust FPLS (RFPLS) approach to robustly estimate the parameter function in the FLogR model and classify the binary response robustly. In doing so, we not only use weighted likelihood estimators that are robust to vertical outliers, but also weight the design matrix so that our proposed approach is robust to horizontal outliers. Thus we ensure robustness against both horizontal and vertical outliers.

To gauge the performance of the proposed approach, different scenarios of two different Monte Carlo trials and two empirical data analyses are conducted, studying growth and strawberry purees. The results exhibit that the proposed approach generates improved estimates for the regression coefficient function of the FLogR model in the presence of outliers, compared with the existing non robust methods. Furthermore, it produces improved classification results for the binary response variable in the presence of outliers. These findings suggest that the proposed approach can be a valuable tool for classifying functional predictors in the presence of outliers.

In addition, our outcomes demonstrate that the proposed approach performs competitively with its traditional competitors even when the absence of outliers in the data. In other words, by evaluating the prediction and classification performance of the proposed approach in the absence of outliers, it is shown that the it can be used as an alternative to existing methods in the absence of outlier contamination.

The second approach proposed in this thesis is a FQAR model of order one, which is used to predict the overall distribution of future realizations of a functional time series. The FQAR model allows us to model the conditional quantiles of the response variable as a function of the past values of it. In this approach, we use the FPC decomposition, which is a common approach, to reconstruct the functional structure of the response and predictor variables in a finite-dimensional space.

To gauge the performance of the proposed FQAR approach, we conduct a Monte Carlo experiment and an empirical dataset analysis which focuses on the prediction of river flows for different stations. We use various performance measures such as root median squared prediction error, modified index of agreement, relative error and Nash Sutcliffe efficiency as benchmarks to evaluate our proposed approach against existing conditional mean regression. Note that we use the 0.5 quantile level (i.e., the median regression) for point forecasts to provide a logical comparison of the proposed method with the existing conditional mean regression. Our numerical results exhibit that the FQAR method produces competitive and even better estimation results compared to the comparison methods.

Moreover, we generate pointwise prediction intervals without using bootstrap methods for the hydrological process studied with our proposed method, and report the affirmative coverage performance of this generated interval. This finding suggests that the proposed FQAR method is a reliable and exciting tool for predicting the overall distribution of future realizations of a functional time series.

REFERENCES

[1] S. J. Ratcliffe, G. Z. Heller, L. R. Leader, "Functional data analysis with application to periodically stimulated foetal heart rate data. II: functional logistic regression," *Statistics in Medicine*, vol. 21, no. 8, pp. 1115–1127, 2002.

[2] M. Escabias, A. M. Aguilera, M. J. Valderrama, "Modeling environmental data by functional principal component logistic regression," *Environmetrics*, vol. 16, no. 1, pp. 95–107, 2005.

[3] X. Leng, H.-G. Müller, "Classification using functional data analysis for temporal gene expression data," *Bioinformatics*, vol. 22, no. 1, pp. 68–76, 2006.

[4] A. M. Aguilera, M. Escabias, M. J. Valderrama, "Discussion of different logistic models with functional data. Application to systemic lupus erythematosus," *Computational Statistics & Data Analysis*, vol. 53, no. 1, pp. 151–163, 2008.

[5] P. Wei, H. Tang, D. Li, "Functional logistic regression approach to detecting gene by longitudinal environmental exposure interaction in a case-control study," *Genetic Epidemiology*, vol. 38, no. 7, pp. 638–651, 2014.

[6] M. Escabias, A. Aguilera, M. Valderrama, "Principal component estimation of functional logistic regression: Discussion of two different approaches," *Journal of Nonparametric Statistics*, vol. 16, no. 3-4, pp. 365–384, 2004.

[7] J. Goldsmith, J. Bobb, C. M. Crainiceanu, B. Caffo, D. Reich, "Penalized functional regression," *Journal of Computational and Graphical Statistics*, vol. 20, no. 4, pp. 830–851, 2011.

[8] S. N. Mousavi, H. Sørensen, "Multinomial functional regression with wavelets and LASSO penalization," *Econometrics and Statistics*, vol. 1, pp. 150–166, 2017.

[9] S. N. Mousavi, H. Sørensen, "Functional logistic regression: A comparison of three methods," *Journal of Statistical Computation and Simulation*, vol. 88, no. 2, pp. 250–268, 2018.

[10] U. Beyaztas, H. L. Shang, "On function-on-function regression: Partial least squares approach," *Environmental and Ecological Statistics*, vol. 27, no. 1, pp. 95–114, 2020.

[11] P. T. Reiss, J. F. Goldsmith, H. L. Shan, R. T. Odgen, "Methods for scalar-on-function regression," *International Statistical Review*, vol. 85, no. 2, pp. 228–249, 2017.

[12] M. Escabias, A. M. Aguilera, M. J. Valderrama, "Functional PLS logit regression model," *Computational Statistics & Data Analysis*, vol. 51, no. 10, pp. 4891–4902, 2007.

[13] P. T. Reiss, R. T. Odgen, "Functional principal component regression and functional partial least squares," *Journal of the American Statistical Association: Theory and Methods*, vol. 102, no. 479, pp. 984–996, 2007.

[14] A. Delaigle, P. Hall, "Methodology and theory for partial least squares applied to functional data," *The Annals of Statistics*, vol. 40, no. 1, pp. 322–352, 2012.

[15] D. Yu, L. Kong, I. Mizera, "Partial functional linear quantile regression for neuroimaging data analysis," *Neurocomputing*, vol. 195, no. C, pp. 74–87, 2016.

[16] H. L. Shang, R. J. Hyndman, "Nonparametric time series forecasting with dynamic updating," *Mathematics and Computers in Simulation*, vol. 81, no. 7, pp. 1310–1324, 2011.

[17] P. Masselot, S. Dabo-Niang, F. Chebana, T. B. M. J. Ouarda, "Streamflow forecasting using functional regression," *Journal of Hydrology*, vol. 538, pp. 754–766, 2016.

[18] S. Curceac, C. Ternynck, T. B. M. J. Ouarda, F. Chebana, S. Dabo-Niang, "Short-term air temperature forecasting using nonparametric functional data analysis and SARMA models," *Environmental Modelling & Software*, vol. 111, pp. 394–408, 2019.

[19] U. Beyaztas, Z. M. Yaseen, "Drought interval simulation using functional data analysis," *Journal of Hydrology*, vol. 579, pp. 124–141, 2019.

[20] U. Beyaztas, H. L. Shang, Z. M. Yaseen, "A functional autoregressive model based on exogenous hydrometeorological variables for river flow prediction," *Journal of Hydrology*, vol. 598, p. 126 380, 2021.

[21] M. Denhere, N. Billor, "Robust principal component functional logistic regression," *Communications in Statistics - Simulation and Computation*, vol. 45, no. 1, pp. 264–281, 2016.

[22] S. A. Abbas, Y. Xuan, X. Song, "Quantile regression based methods for investigating rainfall trends associated with flooding and drought conditions," *Water Resources Management*, vol. 33, no. 12, pp. 4249–4264, 2019.

[23] U. Beyaztas, H. L. Shang, A. Alin, "Function-on-function partial quantile regression," *Journal of Agricultural, Biological and Environmental Statistics*, vol. in press, 2021.

[24] J. O. Ramsay, B. W. Silverman, *Functional Data Analysis*, 2nd. New York: Springer Science & Business Media, 2005.

[25] R. F. Essomba, "An investigation into functional linear regression modeling," M.S. thesis, University of Cape Town, 2015.

[26] L. Horváth, P. Kokoszka, *Inference for functional data with applications*. New York: Springer Science & Business Media, 2012, vol. 200.

[27] C. de Boor, *A Practical Guide to Splines*, Revised Edition. New York: Springer, 2001.

[28] J. O. Ramsay, G. Hooker, G. Spencer, *Functional Data Analysis with R and MATLAB*. New York: Springer Science & Business Media, 2009.

[29] P. T. Reiss, R. T. Ogden, "Functional principal component regression and functional partial least squares," *Journal of the American Statistical Association*, vol. 102, no. 479, pp. 984–996, 2007.

[30] C. Preda, G. Saporta, "Pls regression on a stochastic process," *Computational Statistics & Data Analysis*, vol. 48, no. 1, pp. 149–158, 2005.

[31] Y. Dodge, J. Whittaker, "Partial quantile regression," *Metrika*, vol. 70, no. 1, pp. 35–57, 2009.

[32] F. Alqallaf, C. Agostinelli, "Robust inference in generalized linear models," *Communications in Statistics - Simulation and Computation*, vol. 45, no. 9, pp. 3053–3073, 2016.

[33] J. Fox, *Applied Regression Analysis and Generalized Linear Models*, 3rd. Thousand Oaks, California: Sage, 2016.

[34] C. Agostinelli, S. C. M. Library, *Wle: Weighted likelihood estimation*, R package version 0.9-91, 2015. [Online]. Available: `https://CRAN.R-project.org/package=wle`.

[35] S. Serrneels, C. Croux, P. Filzmoser, P. J. V. Espen, "Partial robust M-regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 79, no. 1-2, pp. 55–64, 2005.

[36] A. Alin, C. Agostinelli, "Robust iteratively reweighted SIMPLS," *Journal of Chemometrics*, vol. 31, no. 3, e2881, 2017.

[37] C. Agostinelli, "Robust stepwise regression," *Journal of Applied Statistics*, vol. 29, no. 6, pp. 825–840, 2002.

[38] J. O. Ramsay, S. Graves, G. Hooker, *Fda: Functional data analysis*, R package version 5.1.9, 2020. [Online]. Available: `https://CRAN.R-project.org/package=fda`.

[39] Y. Sun, M. G. Genton, "Functional boxplots," *Journal of Computational and Graphical Statistics*, vol. 20, no. 2, pp. 316–334, 2011.

[40] O. T. Ojo, R. E. Lillo, A. F. Anta, *Fdaoutlier: Outlier detection tools for functional data analysis*, R package version 0.2.0, 2021. [Online]. Available: `https://CRAN.R-project.org/package=fdaoutlier`.

[41] J. K. Holland, E. K. Kemsley, R. H. Wilson, "Use of Fourier transform infrared spectroscopy and partial least squares regression for the detection of adulteration of strawberry purees," *Journal of the Science of Food and Agriculture*, vol. 76, no. 2, pp. 263–269, 1998.

[42] R. Koenker, G. Bassett-Jr., "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.

[43] D. Bosq, *Linear Processes in Function Spaces*. New York: Springer, 2000.

[44] R. Koenker, *Quantile Regression*. Cambridge: Cambridge University Press, 2005.

[45] L. Hao, D. Q. Naiman, *Quantile Regression*. California: Sage Publications, 2007.

[46] R. Koenker, *Quantreg: Quantile regression*, R package version 5.86, 2021. [Online]. Available: `https://CRAN.R-project.org/package=quantreg`.

[47] R. Hyndman, H. L. Shang, *Ftsa: Functional time series analysis*, R package version 6.1, 2021. [Online]. Available: `https://CRAN.R-project.org/package=ftsa`.

PUBLICATIONS FROM THE THESIS

Papers

1. M. Mutis, U. Beyaztas, G. G. Simsek, H. L. Shang, "A robust scalar-on-function logistic regression for classification", Communications in Statistics - Theory and Methods, pp.1–17, Taylor & Francis, 2022.

Conference Papers

1. M. Mutis, U. Beyaztas, G. G. Simsek, H. L. Shang, "A robust functional logistic regression method for classification", International Conference on Trends and Perspectives in Linear Statistical Inference, Bedlewo, Poland, 30 August - 03 September 2021, pp.83

Projects

1. M. Mutis (Scholarship recipient), U. Beyaztas (Project executive), "Novel Prediction, Estimation And Classification Methods Based On Partial Least Squares And Machine Learning In Functional Data Analysis", 120F270, TUBITAK Project, 2021-2023.