Applications of Projection Pursuit in Functional Data Analysis: Goodness-of-fit, Forecasting, and Change-point Detection

by

Yijun Xie

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2021

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Jiguo Cao Professor, Department of Statistics and Actuarial Science Simon Fraser University
Supervisor(s):	Adam Kolkiewicz Associate Professor, Department of Statistics and Actuarial Science University of Waterloo Gregory Rice Assistant Professor, Department of Statistics and Actuarial Science University of Waterloo
Internal Member:	Shoja'eddin Chenouri Associate Professor, Department of Statistics and Actuarial Science University of Waterloo Joel Dubin Associate Professor, Department of Statistics and Actuarial Science University of Waterloo
Internal-External Member:	Stanko Dimitrov Associate Professor, Department of Management Sciences University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Dimension reduction methods for functional data have been avidly studied in recent years. However, existing methods are primarily based on summarizing the data by their projections into principal component subspaces, namely the functional principal component analysis (fPCA). While fPCA could be effective sometimes, in this thesis we show with both real and synthetic data examples some pitfalls of this approach, especially when the components of interest of the functional data are orthogonal to the leading principal components.

In multivariate data analysis, a possible alternative, the projection pursuit technique, was proposed by Kruskal (1972) and Friedman and Tukey (1974). In this thesis, we extend the idea of projection pursuit to functional data analysis. We develop several new computational tools needed to implement the high-dimensional projection pursuit. We apply this functional projection pursuit technique to three problems: (i) normality test for functional data; (ii) forecasting the functional time series; and (iii) change point detection for functional data. For each problem, a simulation study and several data analyses are provided to show the advantages of our proposed method to existing methods in the literature that mostly based on principal component analysis.

Acknowledgements

First and foremost, I want to thank my PhD supervisors, Professor Adam Kolkiewicz and Professor Gregory Rice, for their tutelage and support throughout my PhD study at University of Waterloo. Without their broad knowledge, encouragement, and guidance, I would never be able to finish this thesis. Beyond academics, they also provide me great role models. Professor Kolkiewicz is always patient, smiles with insightful ideas. Professor Rice has the greatest passion I have ever seen. I would never forget the Saturday afternoon we met in his office and he told me the secret of Bill Belichick's success is *execution*. This could be my motto in the rest of my professional career.

I also want to thank my PhD thesis committee, Professor Jiguo Cao, Professor Stanko Dimitrov, Professor Shoja'eddin Chenouri, and Professor Joel Dubin. They provide me many constructive and critical inputs that significantly improve the final version of my thesis.

My sincere thanks to Professor Natalia Nolde at University of British Columbia and Professor Fang Liu at University of Notre Dame. They encouraged me to pursue a PhD degree, and give me all the help they could. I want express my great gratitude to Ms Huijie Xu, my high school math teacher, at Changzhou Senior High School. I was such a headache to her at that time, but she still firmly believed my talent and encouraged me to study mathematics in college.

I would like to thank all my friends, especially Wenying Gu, Cheng Shi, and Jing Cai. You really helped me to get through those most difficult days, and I can never appreciate enough. I also want to thank Seine River, and special thanks to NJX. You have enlightened me at the darkest times.

During my PhD study, I also received enormous help and friendship from my fellow PhD colleagues. I would like to express my thanks to Dr. Lichen Chen, Zehao Xu, and Chi-Kuang Yeh. It always feels great to have buddies around you. I want to thank the whole Department of Statistics and Actuarial Science at University of Waterloo. There is such a warm and supportive environment, and I receive generous financial support from the department. Special thanks to our Graduate Studies Coordinator Mary Lou Dufton, she helped me with all the administrative work during my time at Waterloo.

Last but not least, I want to thank my family. Without your support, none of these will be possible. I want to thank my grandparents, Weichuan Xie and Hexiu Chen. They always tell me that I am their favourite grandchild, and I hope I have made them proud. I want to end my acknowledgements with the most special thanks to my parents, Zhaoliang Xie and Huaping Xie, for their unconditional love to me in my whole life.

Dedication

To my parents.

Table of Contents

List of Tables x

List of Figures xi

List of Algorithms xiii

1 Introduction 1

1.1 Functional Data Analysis 1

1.2 Problems Studied 7

1.2.1 Normality test for functional data 7

1.2.2 Forecasting functional time series 8

1.2.3 Change point detection for functional data 9

1.3 Contributions and Organization of the Thesis 9

2 Dimension Reduction for Functional Data 11

2.1 Multivariate Principal Component Analysis 11

2.2 Functional Principal Component Analysis 14

2.3 Multivariate Projection Pursuit 16

2.4 Functional Projection Pursuit 17

3 Projection pursuit based tests of normality with functional data **21**
3.1 Introduction . . . 21
3.2 Problem statement, definition of test statistics, and their asymptotic properties 23
3.2.1 Large sample properties . . . 25
3.3 Implementation and a Simulation Study . . . 27
3.3.1 Simulation study . . . 29
3.4 Data Analysis . . . 36
3.4.1 Fertility rate in Australia . . . 37
3.4.2 Conditional intra-day stock prices . . . 37
3.4.3 Yearly lower temperature profiles in Australia . . . 39

4 Functional Time Series Forecasting via Projection Pursuit **43**
4.1 Introduction . . . 43
4.2 Methodology . . . 46
4.2.1 Functional projection pursuit for forecasting . . . 46
4.2.2 Details of implementation . . . 49
4.3 Simulation Study . . . 51
4.4 Data Analysis . . . 59
4.4.1 PM10 concentration data . . . 59
4.4.2 French male mortality rate data . . . 63

5 Change-points Detection in Functional Data **68**
5.1 Introduction . . . 68
5.2 Methodology . . . 70
5.2.1 Functional projection pursuit for change point detection . . . 70
5.2.2 Approximate optimal direction . . . 72
5.2.3 Choice of $\boldsymbol{d}(\boldsymbol{r}, \boldsymbol{u})$ . . . 73
5.2.4 Statistical significance of estimated change point . . . 75

5.3 Simulation Study . . . 76
5.4 Data Examples . . . 81
5.4.1 Australian fertility data . . . 81
5.4.2 Daily low temperature profile in Gayndah . . . 84

6 Concluding Remarks and Future Works **87**
6.1 Functional Projection Pursuit Algorithm . . . 87
6.2 Functional Normality Test . . . 88
6.3 Forecasting Functional Time Series . . . 88
6.4 Change-points Detection in Functional Data . . . 88

References **90**

APPENDICES **105**

A Appendix for Chapter 1 **106**
A.1 Motivation . . . 106
A.2 PACE Algorithm . . . 107
A.3 Multivariate Functional Principal Component Analysis . . . 108

B Appendix for Chapter 3 **110**
B.1 Proof of Theorem 3.2.1 and Theorem 3.2.2 . . . 110
B.2 Selection of Parameters J and M . . . 118
B.3 Selection of Number of Basis Functions . . . 120
B.4 Non-smooth Curves: Simulation Study . . . 121

C Appendix for Chapter 4 **122**
C.1 Selection of the Stopping Criterion . . . 122
C.2 Selection of the Tuning Parameters . . . 123

List of Tables

3.1 Percentage of rejections under the fast decaying covariance matrix Σ_{fast}. . 34

3.2 Percentage of rejections under the slow decaying covariance matrix Σ_{slow}. . 35

3.3 Percentage of rejections under the random covariance matrix Σ_{ran}. 36

4.1 Average integrated squared error for each forecasting method and DGP considered. 56

4.2 Percentage of 1-step forward forecasting error ratios less than 1. 59

4.3 Comparison of integrated squared prediction errors for PM10 data. 61

4.4 Comparison of average integrated squared prediction errors for French male log mortality data. 64

5.1 Average locations of change point based on different detection methods. . . 79

5.2 Average adjusted Rand indices for estimated locations of change point based on different detection methods. 80

5.3 Percentage of rejections under the null, **MSF**, and **MW**. 81

B.1 Percentage of rejections under the slow decaying covariance matrix Σ_{slow}. . 121

List of Figures

1.1 NOx level (in mglm^3) in Barcelona Spain from February 23 2005 to June 26 2005. 2

3.1 Fertility rate by age in Australia from 1921 to 2006. 37
3.2 Daily curves of the transformed IBM prices from 06/15/2006 to 04/02/2007. 39
3.3 Daily lowest temperature at Gayndah Australia from 1894 to 2008. 40
3.4 Comparison between Gaussian and non-Gaussian components for Gayndah's temperature data. 41
3.5 The p-values of proposed normality test under different number basis functions. 42

4.1 A basic schematic for forecasting functional time series through dimension reduction. 44
4.2 Histograms of ratios of 1-step forward forecasting errors of 100 simulated samples. 58
4.3 PM10 concentration in Graz-Mitte, Austria. 60
4.4 Forecasted PM10 concentration level curves. 62
4.5 Log mortality rate for French Male between 0 and 100 years old from 1816 to 2006. 64
4.6 Forecasted French male log mortality rate curve in 1997. 65
4.7 Forecasted French male log mortality rate curve in 2006 65
4.8 Forecasted log mortality rate curve for French male between 18 and 45 years old in 2006. 66

4.9 Forecasted log mortality rate curve for French male between 60 and 90 years old in 2006. 67

5.1 Australian fertility rate from 1921 to 2006. 83

5.2 Estimated change functions and projection scores on estimated change direction for Australian Fertility Data. 83

5.3 The direction on which the change is most significant and the corresponding projection scores on estimated change direction for 1^{st} and 2^{nd} order differenced Australian Fertility Data. 84

5.4 Daily low temperature profile in Gayndah, Australia from 1894 to 2008. . . 85

5.5 Estimated change functions and the projection scores on estimated change direction for Daily Low Temperature Profile in Gayndah. 86

5.6 Estimated change functions and the projection scores on estimated change direction for detrended Daily Low Temperature Profile in Gayndah. 86

B.1 Estimated 95% quantiles of $\hat{S}_n$ (left panels) and $\hat{K}_n$ (right panels) with $n = 450$ and $k = 21$. 119

B.2 The test power of proposed normality test under different number B-spline basis functions. 120

C.1 The 1-step cross-validated forecasting errors for the PM10 data as a function of the number of dimensions included in the proposed projection pursuit method. 123

C.2 The 1-step and 10-step forecasting errors for different values of r range from 0.01 to 0.1. 124

List of Algorithms

2.4.1 Two-Step Approximation Algorithm for $\hat{Q}^{L,k}$ 20

3.3.1 Two-Step Approximation Algorithm for $\hat{S}_n^{L_k}$ 28

4.2.1 Two-Step optimization algorithm for finding most predictable direction. . . 48

5.2.1 Multi-layer optimization algorithm for functional change-point detection. . . 73

Chapter 1

Introduction

1.1 Functional Data Analysis

In elementary statistics courses, we studied scalar or multivariate data which usually have a relatively low dimension. In recent years, high dimensional data have drawn more and more attention from researchers and practitioners. However, all these data are considered as vectors in a finite dimensional Euclidean space. In functional data analysis, we instead focus on individual observations that can be thought as elements of a, perhaps infinite dimensional, function space. One assumes in this case that the data are intrinsically continuous functions $x_1, \cdots, x_n$, and for each function x_i we observe discrete observations $x_i(t_1), \cdots, x_i(t_{p_i})$, where $t_1, \cdots, t_{p_i}$ are in the domain of x_i. The range and domain of functional data do not need to be on the real line. For example, Ramsay (2000) and Ramsay and Silverman (2007) consider handwriting data for which the range is the spacial location on a plane. In this thesis, we mainly focus on real valued functions whose domain is a compact set on the real line.

The study done in Dauxois et al. (1982) is among the first to distinguish functional data from multivariate data. In Ramsay and Silverman (1997), the idea of functional data analysis is introduced comprehensively. Another textbook-level treatment is in Horváth and Kokoszka (2012a), where focus is put on dependent functional data, in particular functional time series.

There are several reasons why we want to focus on functional data. Below we present some of these reasons using the NOx[1] pollution data available in the R package `fda.usc`

[1] NOx, or nitrogen oxides, is a very common source of air pollution that will lead to acid rain and smog.

(Febrero-Bande and Oviedo de la Fuente, 2012). This dataset contains the NOx levels measured in every hour in a control station in Barcelona, Spain. Each curve in Figure 1.1 contains 24 measurements of NOx concentrations recorded from 12 a.m. to 11 p.m. during one day. While the data are stored as vectors of length 24, one might wish to model the NOx pollution as functional data for the following reasons:

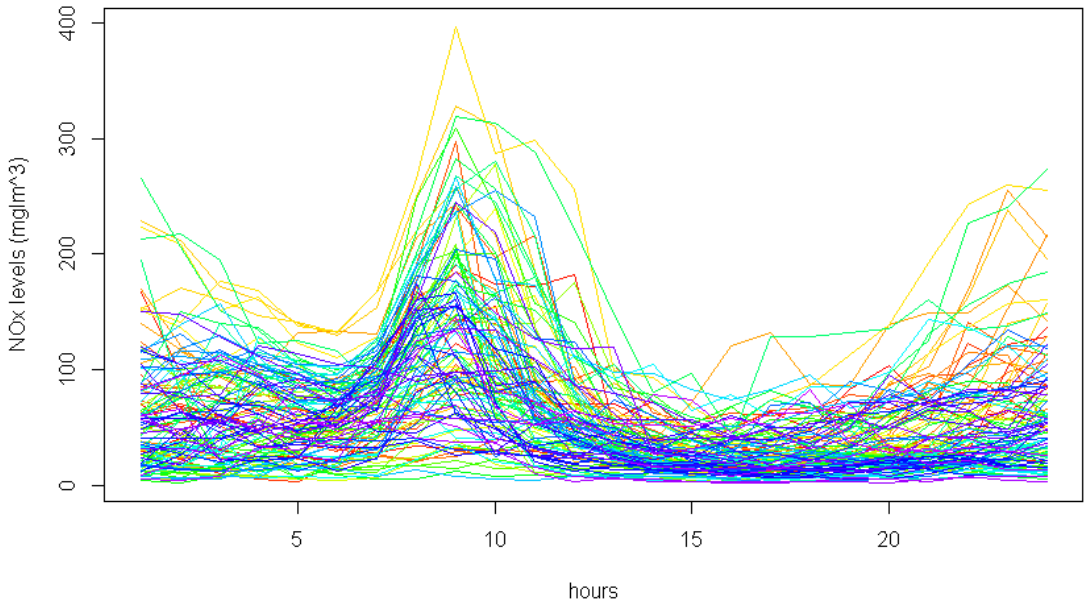

Figure 1.1: NOx level (in mglm^3) in Barcelona Spain from February 23 2005 to June 26 2005.

- First, a lot of data are continuous by nature. Since it is reasonable to assume that the level of pollution changes continuously in time, the NOx data should be treated as discrete observations of a continuous function at selected times. Also, the continuity of the unobserved data generating process will introduce a correlation structure of neighbouring observations which can be easily modeled by functional data analysis. For multivariate data corresponding to high-frequency observations, one would need a large covariance matrix to model the covariance structure, which could be quite challenging to work with. Moreover, for multivariate data one can always switch its coordinates while preserving the performance of most multivariate data analysis

tools (see, for example, the multivariate principal component analysis that will be introduced in Chapter 2). However, if the data come from a continuous function, then such operation may obscure the correlation structure of neighbouring observations.

- Second, when there are vectors with different lengths, it is not always clear how to proceed using multivariate analysis methods. Let us imagine a hypothetical situation when there is an automated weather station that can record NOx level in every 1 ms (i.e. 1000 measures per second). Suppose each day a few sudden power outages happen randomly, and the different power outages last for different lengths of time. Under this situation there would be a large number of random missing values for the daily records. We will encounter a similar example in Chapter 3 and Chapter 5, where the daily lower temperature data have many random missing values. Another example is the stock tick data, in which case the stock price is recorded every time the best bid or ask price is changed. Since the number of changes in each day could be different, the number of daily observations would also be different. The above examples would pose serious challenges for multivariate data analysis methods. One could align the data via curve estimation, and then discretize them over a dense grid to make the multivariate observations aligned. However, this approach might lead to biases, especially from the interpolation process. On the contrary, such situations can be easily handled in the functional data analysis framework. A functional data object can be constructed based on arbitrary evaluations of the function within its domain, and hence can deal with scenarios described above.

- Lastly, working with functional data can provide us with additional information from the data that is not available in the multivariate framework, such as information on derivatives or integrals. For example, the first order derivative describes how fast the data change and second order derivative describes the acceleration of the change. An interesting illustration of the use of derivatives can be found in Ramsay and Silverman (2007), where the researchers extract first and second order derivative information from the smoothed curve of human growth data to discover the growth pattern of adolescents. While there are methods from numerical analysis to estimate derivatives from discrete observations, it would be more natural to treat the data as a continuous function and calculate derivatives directly.

To construct functional data objects from discrete observations, a common practice is to follow the smoothing method suggested in Ramsay and Silverman (1997). To explain the main idea behind this method, let us assume that we observe discrete realizations $y_j, j = 1, \ldots, J$, of a function x such that $y_j = x(t_j) + \epsilon_j$, where $\{\epsilon_j\}$ is a white noise,

and $t_j, j = 1, \ldots, J$, are in the domain of x. Suppose further that the function x can be expanded using a set of basis functions ϕ_k as follows:

$$x = \sum_{k=1}^{K} \xi_k \phi_k = \xi' \phi,$$

where $\xi = [\xi_1, \ldots, \xi_K]$ is the coefficient vector of length K, and $\phi = [\phi_1, \cdots, \phi_K]'$ is a set of basis functions. Let Φ be a $J \times K$ matrix such that $\Phi_{kj} = \phi_k(t_j)$. Then a possible approach to the estimation of the coefficient vector ξ is by using the least square principle. This is equivalent to the minimization of the sum of squared errors

$$SSE(y|\xi) = \sum_{j=1}^{J} \left[y_j - \sum_{k=1}^{K} \xi_k \phi_k(t_j) \right]^2$$

with respect to the coefficient vector ξ. The standard results from linear regression show that the least square estimator for ξ is

$$\hat{\xi} = (\Phi'\Phi)^{-1}\Phi' y,$$

and the function is approximated as $\hat{x} = \hat{\xi}'\phi$.

In Green and Silverman (1993), the authors further extend this approach by introducing a roughness penalty to fit a more smooth curve. The roughness is measured by the integrated squared second order derivatives:

$$\text{PEN}(x) = \int [x^{(2)}(s)]^2 ds,$$

where $x^{(2)}(s)$ denotes the second order derivative of x evaluated at time s. We define the penalized residual sum of squares as

$$\text{PENSSE}(x|y) = [y - x]'[y - x] + \lambda \text{PEN}(x),$$

where λ is a smoothing parameter. The authors show that in this case the estimated coefficient vector is

$$\hat{\xi} = (\Phi'\Phi + \lambda R)^{-1}\Phi' y,$$

where

$$R_{ij} = \int \phi_i^{(2)}(s)\phi_j^{(2)}(s)ds$$

is the penalty matrix.

Functional data analysis has been successfully applied to various fields. For example, in Cardot et al. (1999) and Ramsay and Silverman (1997), the authors first consider a functional linear model, which has been further extended to generalized functional linear models in Müller et al. (2005) and to functional mixed models in Scheipl et al. (2015). In Müller and Yao (2008) the authors propose a functional additive model. This model is applied to analyze the longitudinal data in Müller et al. (2008). See Malfait and Ramsay (2003) for a brief history, and Horváth and Kokoszka (2012a) for a comprehensive review of functional linear models.

Functional data analysis has also been applied to medical and neural science. In Shepstone et al. (1999) the authors analyze the shape of bones as a curve to help diagnose arthritis, and in Sidhu et al. (2012) the researchers use functional principal component analysis to study Attention-Deficit Hyperactivity Disorder (ADHD) from functional Magnetic Resonance Image (fMRI) data. Some other examples of applications of functional data analysis include Kargin and Onatski (2008), where the researchers propose functional time series models to forecast the Eurodollar futures and credit card transactions, and Ramsay and Bock (2002), in which the authors study the classic Berkeley Growth Study data in Tuddenham (1954) from the prospective of human growth speed and acceleration. For more examples and a thorough review of different applications of functional data analysis, see Ramsay (2005) and Ullah and Finch (2013).

In many applications of functional data analysis, certain degree of dimension reduction is necessary, as it might be difficult to work with the original curves directly. The arguably most prevalent dimension reduction technique for functional data is functional principal component analysis (fPCA), which is an extension of the principal component analysis (PCA) for multivariate data. Below we present two examples to illustrate the use of fPCA.

Example 1: Functional regression model

Here we consider a functional regression model and assume that the predictor is a square integrable random function $X(\cdot)$ defined on a domain $\mathcal{S}$ and the response is a random function $Y(\cdot)$ defined on domain $\mathcal{T}$. Let $\mu_Y = E[Y]$, $\mu_X = E[X]$, $Y^c = Y - \mu_Y$, and

$X^c = X - \mu_X$. Then the linear functional regression model can be defined as:

$$E[Y|X] = \mu_Y + \int_{\mathcal{S}} \beta(s,t) X^c(s) ds,$$

where $\beta(s,t)$ is the regression parameter function. In He et al. (2000) the authors suggest a fPCA based method to estimate $\beta(s,t)$. Suppose we have the following Karhunen-Loève expansion

$$Y^c = \sum_{k=1}^{\infty} \zeta_k \psi_k,$$
$$X^c = \sum_{j=1}^{\infty} \xi_j \phi_j,$$

where ψ_k and ϕ_j are functional principal components of Y^c and X^c respectively, and ζ_k and ξ_j are the corresponding functional principal component scores. Then the regression parameter function is obtained as

$$\beta(s,t) = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \frac{E(\xi_j \zeta_m)}{E(\xi_j^2)} \phi_j(s) \psi_k(t).$$

Example 2: Longitudinal biomarker data

This example is based on Jiang et al. (2020). The goal of this project is to develop a dynamic prediction method for the survival probability of patients with Alzheimer's Disease. While prediction methods that use base line covariates have been thoroughly studied, in this project we are trying to integrate longitudinal observations of biomarkers, especially the expression level of certain genes, into random survival forest. These biomarkers are observed at irregular time grids. For example, individual 1 has biomarker A measured at times $t = 0, 0.3, 0.4, 0.7$, and biomarker B measured at times $t = 0.1, 0.4, 0.5$. But individual 2 has biomarker A measured at times $t = 0, 0.2, 0.25, 0.8, 0.9$, and biomarker B measured at times $t = 0.2, 0.35, 0.4, 0.55$.

With a random forest algorithm we can input a large collection of variables, and during the training process the algorithm will not only fit the model but also estimate the variable importance simultaneously. Unfortunately, the traditional random forest algorithms cannot directly handle these irregularly observed data. However, one could instead treat these discretely observed values as realizations of some continuous functions. Then, similarly

to Example 1, we can represent these functions using the functional principal component analysis as a tool for dimension reduction. In the appendix to this chapter, we provide more details about the methods we have used to find the functional principal component scores from these irregularly observed data. We should notice that after proper dimension reduction the original data are now transformed into scalar scores of the same length, analysis of which is significantly easier. Hence, with the help of fPCA, we successfully link the multiple irregularly observed biomarker data with the well-studied random forest algorithm. We have achieved promising results for predicting Alzheimer's Disease patients' survival probabilities, and demonstrated that our proposed method could be used for understanding the cause and development of Alzheimer's Disease. The method has been implemented in the R package `funest` which is available for download on CRAN.

We discuss the functional principal component analysis in greater details in Chapter 2. However, we should notice that fPCA aims to minimize the L^2 loss between original observations and reconstructed data. In some applications, such a subspace constructed by functional principal components might not be optimal, especially when the features of the data one is interested in are not related to the second moment. Furthermore, fPCA implements the same dimension reduction scheme to all different problems, which is evidently not realistic.

1.2 Problems Studied

In this thesis, our goal is to introduce a flexible dimensional reduction framework, namely projection pursuit, for functional data as well as high dimension data. While this new framework aims to be adapted to any arbitrary case, in this thesis we focus on the following three problems.

1.2.1 Normality test for functional data

Statistical methods based on the assumption of normality of the observations and/or model errors are ubiquitous in classical statistics, and are also widely used in more modern settings when the data to be analyzed are high-dimensional or functional in nature. To give some recent examples, Panaretos et al. (2010) and Cuevas et al. (2004) assume normality in order to perform two sample and analysis of variance tests with functional data, and in Kowal et al. (2017), Kowal et al. (2019), normality of the data is used in performing Bayesian inference with complex functional data. Some further applications of normality in this

setting can be found in Gromenko et al. (2017b), Yao et al. (2005a), and Constantinou et al. (2017), although this list is far from being exhaustive. Given the usefulness of these procedures, it is important to have ways of measuring the validity of the assumption of normality for a given sample of functional data. At the least such a validation would lend further credibility to the conclusions of procedures in which normality is assumed, although evidence for normality of functional data may also be of independent interest.

Methods for validating the assumption of normality of functional data have been only lightly developed to date, with existing methods based primarily on the idea of summarizing the data by their projections onto random or principal component subspaces, and then applying multivariate normality tests to the vectors of scores defining these projections. While this is effective in some cases, there could be some pitfalls of this approach, including their sensitivity to the basis used to smooth the raw data.

In Chapter 3, we introduce a new normality test for functional data based on the projection pursuit technique that overcomes some of these challenges. We also furnish a way of decomposing functional data into its approximately Gaussian and non-Gaussian components, which is useful for the purpose of data visualization and subsequent analyses.

1.2.2 Forecasting functional time series

One of the most fundamental problems in time series analysis is forecasting future values. Suppose we observe a length n stretch of a time series $x_1, \ldots, x_n$. The problem of forecasting this series at horizon h can be framed as finding a function F_h with which we predict x_{n+h} as $\hat{x}_{n+h} = F_h(x_1, \ldots, x_n)$. One typically wishes to choose the function F_h optimally in the sense that some specified loss $L(x_{n+h}, \hat{x}_{n+h})$ is minimized.

The forecasting problem in both the univariate and the multivariate settings has been thoroughly studied, and one can find comprehensive discussions on this topic in Hyndman and Athanasopoulos (2018), Shumway and Stoffer (2017), Brockwell and Davis (2013), and Lütkepohl (2013). In recent years, methods for forecasting functional time series have also been actively studied. However, most existing methods either focus on forecasting within the framework of functional autoregressive models, or rely on dimension reduction using functional principal component analysis, and subsequent forecasting of the resulting multivariate series. While fPCA performs well for this purpose under certain conditions, principal component analysis as a general tool is not tailored for forecasting, and it can be sub-optimal or even misleading in the presence of non-stationarity or when the predictable components of a given functional time series do not coincide with the principal components. In Chapter 4, we propose a new forecasting method based on dimension reduction using

a functional projection pursuit technique that aims to optimize the dimension reduction step for forecasting. Emphasis is put on the cases where functional time series are non-stationary, or when the forecastable components are orthogonal to the leading principal components.

1.2.3 Change point detection for functional data

Change point detection aims at locating an index of a sequence of data such that the observations prior and post to this point have different characteristics. Most statistical methods assume homogeneity of the data, hence successfully identifying a change point is essential for many statistical applications in environmental science (Jarušková, 1997; Reeves et al., 2007), finance (Spokoiny et al., 2009), quality control (Lai, 1995), and health science (Muggeo and Adelfio, 2011), to name a few.

In the past few decades, a variety of change point detection methods have been developed for univariate and multivariate data. One can find a comprehensive review of classical methods for change point detection methods in Aue and Horváth (2013) and Horváth and Rice (2014). A recent change point detection method is discussed in Matteson and James (2014), where the authors propose an approach based on empirical characteristic functions to detect distributional changes for multivariate data.

The available functional change point detection methods are limited in their applicability, as most of them focus on detecting changes for a specific type of characteristics of the observed data. For example, method proposed by Berkes et al. (2009) can only be applied to detect a change in the mean level, while the method proposed by Aue et al. (2020) can only be applied to detect the change in second moment structures. In Chapter 5, we propose a new change point detection method for functional data that can work for arbitrary change point types, and can provide information about the mechanics behind the change point for more insightful analysis.

1.3 Contributions and Organization of the Thesis

The rest of this thesis is organized as follows. In Chapter 2 we first review the functional principal component analysis, and present a new dimensional reduction method based on projection pursuit. We apply the projection pursuit method to test normality of functional data in Chapter 3, to forecast functional time series in Chapter 4, and to detect a change point in a sequence of functional data in Chapter 5.

The major contribution of this thesis is an extension of the projection pursuit method to functional data. To this end, we propose a new framework to efficiently select and search a finite rank subspace of the potentially infinite dimensional function space. We also develop novel computational tools to overcome the burden of the high dimensional optimization problem in the implementation of projection pursuit. This part is presented in Chapter 2. We further make the following contributions. In Chapter 3, we propose a new normality test for functional data that can also decompose the data into Gaussian and non-Gaussian components. In Chapter 4, we propose a new forecasting framework for functional time series that emphasizes proper identification of the predictable components. In Chapter 5, we propose a new change point detection method for functional data that works for arbitrary types of a change point.

Chapter 2

Dimension Reduction for Functional Data

While it is exciting to work with functional data, their infinite dimensional nature makes them difficult to analyze directly. Therefore, proper dimension reduction for functional data is desired. In this chapter we first introduce the arguably most commonly used dimension reduction method for functional data, namely the functional principal component analysis (fPCA). We then introduce a new dimension reduction framework for functional data based on general projection pursuit technique.

2.1 Multivariate Principal Component Analysis

Before discussing the functional principal component analysis, we first introduce the principal component analysis in the multivariate setting. Let $X = [X_1, \cdots, X_d]'$ be a d-dimensional random vector with zero mean. Suppose $v_m = [v_{m1}, \cdots, v_{md}]'$ is a vector of length d and $||v_m|| = v_m' v_m = \sqrt{\sum_{j=1}^{d} v_{mj}^2} = 1$. By Y_m we denote the inner product of v_m and X, or the projection score of X onto v_m. That is,

$$Y_m = \langle v_m, X \rangle = v_m' X = \sum_{j=1}^{d} v_{mj} X_j. \tag{2.1.1}$$

Let

$$v_1 = \underset{v_m \in \mathbb{R}^d, ||v_m||=1}{\operatorname{argmax}} \operatorname{Var}(Y_m),$$

and the first principal component y_1 is defined as

$$y_1 = \langle v_1, X \rangle.$$

Intuitively, we are looking for a unit length vector in the d-dimensional Euclidean space such that the projection scores of the data onto this unit length vector has the maximum variance among projection scores of the data onto all unit length vectors. Such procedure will decompose the data into two parts: the projections that are in the same direction as v_1, and residuals that are orthogonal to v_1. Each subsequent v_j is defined as the direction that will maximize the sample variance of the residuals after the $(j-1)^{th}$ projection, i.e.

$$v_j = \underset{\substack{v_m \in \mathbb{R}^d, ||v_m||=1, \\ \langle v_m, v_q \rangle = 0 \text{ for } q<k}}{\operatorname{argmax}} \operatorname{Var}(Y_m).$$

Then the j^{th} principal component y_j is defined as

$$y_j = \langle v_j, X \rangle,$$

and the PCA decomposition of X has the form

$$X = \sum_{j=1}^{d} \langle v_j, X \rangle v_j.$$

Let Σ be the covariance matrix of X, and we further assume that the eigenvalues of Σ satisfy $\lambda_1 > \lambda_2 > \cdots > \lambda_d > 0$. Then $\operatorname{Var}(Y_m) = \operatorname{Var}(v_m' X) = v_m' \Sigma v_m$. To find v_1, we would like to maximize $v_1' \Sigma v_1$ subject to $v_1' v = 1$. One possible approach is to use the technique of Lagrange multipliers. That is, we would like to maximize

$$v_1' \Sigma v_1 + \lambda(v_1' v_1 - 1), \tag{2.1.2}$$

where λ is a constant. Differentiate (2.1.2) with respect to v_1 gives us

$$\Sigma v_1 - \lambda v_1 = 0,$$

which is equivalent to

$$(\Sigma - \lambda I)v_1 = 0, \tag{2.1.3}$$

where I is a $d \times d$ identity matrix. The form of (2.1.3) suggests that λ is an eigenvalue of Σ and v_1 is the corresponding eigenvector. Therefore,

$$v_1'\Sigma v_1 = v_1'\lambda v_1 = \lambda v_1'v_1 = \lambda.$$

Hence, the maximum of $v_1'\Sigma v_1$ is achieved when $\lambda = \lambda_1$, the largest eigenvalue of Σ, and v_1 coincides with the eigenvector corresponding to the largest eigenvalue λ_1. Similarly, one can show that v_j is the eigenvector corresponding to the j^{th} largest eigenvalue λ_j. More details could be found in Jolliffe (2011). Another possible approach is based on the spectrum decomposition of the covariance matrix Σ. Principal axis theorem suggests that the full rank covariance matrix can be decomposed as

$$\Sigma = U'\Lambda U,$$

where $U = [u_1, \cdots, u_d]$ is an orthonormal matrix whose columns are eigenvectors of Σ, and $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_d)$. To maximize $v_1'\Sigma v_1$ is equivalent to maximize $w_1'\Lambda w_1$ where $w_1 = U'v_1 = [w_{11}, \cdots, w_{1d}]'$. Since U is orthonormal, $||w_1|| = ||v_1|| = 1$, and hence we are looking for a unit length vector w_1 such that $w_1'\Lambda w_1$ is maximized. Since Λ is a diagonal matrix,

$$w_1'\Lambda w_1 = \sum_{j=1}^{d} w_{1j}^2\lambda_j, \tag{2.1.4}$$

which is maximized when $w_1 = [1, 0, \cdots, 0]$. Therefore, $v_1 = Uw_1 = u_1$, the eigenvector corresponds to λ_1. One can further show that v_j is the eigenvector corresponds to λ_j. See Horváth and Kokoszka (2012a) for more discussion about this approach.

Directions $v_1, \cdots, v_d$ can be estimated through eigenvectors of the sample covariance matrix of observed data. Suppose the sample covariance matrix of d-dimensional observations $x_i, i = 1, \cdots, n$, is $\hat{C}$. Assume its eigenvalues satisfy $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_d$, and the corresponding eigenvectors are $\hat{v}_1, \cdots, \hat{v}_d$. Then the data can be approximated using the first p eigenvectors as

$$x_i \approx \sum_{j=1}^{p} \langle \hat{v}_j, x_i \rangle \hat{v}_j, i = 1, \cdots, n,$$

where p could be determined by, for example, Akaike information criterion (AIC) or the total variance explained (TVE) (Wang et al., 2016). The purpose of this truncated summation of the first p terms is to use a lower dimensional subspace (as $p < d$) to summarize as much information in terms of the variation of the original data as possible.

We should notice that in much of the literature in multivariate data analysis, principal components refer to $\langle v_j, X \rangle$, the projection scores of the data onto the principal component directions. See, for example, Jolliffe (2011), Johnson et al. (2002), and Vidal et al. (2005). However, in functional data analysis literature, the principal component directions are commonly referred to as the principal components. See pp. 40 of Horváth and Kokoszka (2012a) for an example. While it should not cause much confusion in the context, we will call the principal component directions as the principal components.

2.2 Functional Principal Component Analysis

Functional principal component analysis is very similar to its multivariate counterpart. We first introduce some notation used in the rest of this thesis. We assume, without loss of generality, that the domain of the observed functions is $[0, 1]$. We let $L^2([0, 1], \mathbb{R})$ denote the space of real valued functions with finite squared integral, which is a Hilbert space when equipped with the inner product defined for $x, y \in L^2([0, 1], \mathbb{R})$ by $\langle x, y \rangle = \int_0^1 x(t)y(t)dt$. The corresponding norm is defined by $\| \cdot \|^2 = \langle \cdot, \cdot \rangle$.

We let X be a random object defined in $L^2([0, 1], \mathbb{R})$ with zero mean. Let $C(t, s) = \text{cov}(X(t), X(s))$, then C defines a Hilbert-Schmidt integral operator of the form

$$c(f)(t) = \int_0^1 C(t, s)f(s)ds.$$

Let $v_i, i = 1, 2, \cdots$, be the eigenfunctions of c with the corresponding eigenvalues λ_i satisfying

$$\lambda_i c(v_i)(t) = \lambda_i v_i(t), \quad \lambda_1 \geq \lambda_2 \geq \cdots . \tag{2.2.1}$$

These eigenfunctions and eigenvalues can be estimated through estimates of C. The fPCA decomposition can be seen as a special case of the Karhunen-Loève decomposition,

which is of the form

$$X = \sum_{j=1}^{\infty} \langle v_j, X \rangle v_j.$$

In the context of functional data, we deal with an infinite-dimensional space, hence there will be infinitely many eigenfunctions. In practice, typically we need to estimate the covariance operator using the available data and then truncate the KL-decomposition to the first p eigenfunctions, where p could also be determined by AIC or TVE mentioned above. Suppose we observe a set of functions $x_1, \cdots, x_n$ defined in $L^2([0,1], \mathbb{R})$. Let $\bar{x}(t) = (1/n) \sum_{i=1}^{n} x_i(t), t \in [0,1]$. Then the natural estimator of C is

$$\hat{C}(t,s) = \frac{1}{n} \sum_{i=1}^{n} (x_i(t) - \bar{x}(t))(x_i(s) - \bar{x}(s)),$$

which leads to the following empirical version of the operator c

$$\hat{c}(f)(t) = \int_0^1 \hat{C}(t,s) f(s) ds.$$

The corresponding eigenfunctions and eigenvalues $\hat{v}_i$ and $\hat{\lambda}_i$ of v_i and λ_i then satisfy

$$\hat{\lambda}_i \hat{c}(\hat{v}_i)(t) = \hat{\lambda}_i \hat{v}_i(t), \quad i = 1, ..., n, \tag{2.2.2}$$

and the functional object may be approximated as

$$x_i \approx \sum_{j=1}^{p} \langle \hat{v}_j, x_i \rangle \hat{v}_j, t \in [0,1], i = 1, \cdots, n.$$

The fPCA decomposition represents the data in directions where the variance of the projection scores is maximized. While it is a convenient approach to summarize the data using eigenfunctions and eigenvalues, there are situations where such a decomposition, and the corresponding approximation, will not capture the desired features of the data. For example, variance is not necessarily related to the skewness of the data, the predictability of a time series, or the location of the change point under certain scenarios that will be discussed in greater details in later chapters.

2.3 Multivariate Projection Pursuit

The classic projection pursuit technique for multivariate data is first discussed in Kruskal (1972) and Friedman and Tukey (1974). In the multivariate setting, we define the projection index $Q(v)$ as a measure of "interestingness" of the scores of the data projected onto a vector v, where the score of projection is defined as (2.1.1). That is,

$$Q(v) = Q(\langle x_1, v\rangle, \cdots, \langle x_n, v\rangle)$$

for some function $Q : \mathbb{R}^n \to \mathbb{R}$. Similarly to the principal component analysis in the multivariate setting, our goal is to find the set of d-dimensional vectors $v_j = [v_{j1}, \cdots, v_{jd}]', j = 1, \cdots, d$, such that

$$\begin{aligned} v_1 &= \underset{\|v\|=1}{\operatorname{argmax}} Q(v) \text{ and} \\ v_j &= \underset{\substack{\|v\|=1, \\ v_j' v_m = 0 \text{ for } m<j}}{\operatorname{argmax}} Q(v) \text{ for } j = 2, 3, \cdots, d. \end{aligned}$$

Principal component analysis can be viewed as a special case of the projection pursuit, where we choose the projection index to be the variance measure. In this case the optimal directions coincide with the eigenvectors of the sample covariance matrix. Some applications of the projection pursuit method include finding robust principal components, like in Li and Chen (1985) and Bali et al. (2011). However, the projection pursuit in general leads to an orthogonal approximation of the set of optimal d-dimensional vectors in the sense of "interestingness" rather than in the sense of L^2 loss.

Usually the projection pursuit method has an intensive requirement for computational power. In most implementations the dimension p is limited to a smaller integer (see Croux and Ruiz-Gazen (1996) for an example where $p = 10$). In Croux and Ruiz-Gazen (2005) and Croux et al. (2007), the authors develop an efficient algorithm for searching the projection pursuit directions. However, this algorithm is based on coordinate descent optimization, which works well when the objection functions is smooth and without many local maximums. In Croux et al. (2007), the authors implement their algorithm to estimate robust principal components, where some robust variance measure is used as the projection index. Such a projection index is smooth in the space, and therefore the results seem promising. In applications where the objection function is not smooth or is unknown, one may have concerns about this algorithm's effectiveness.

2.4 Functional Projection Pursuit

In this section, we propose a projection pursuit algorithm for functional data that would be more robust in searching for the global maximum or minimum in the functional space. Let $U^\infty = \{v \in L^2([0,1],\mathbb{R}) \ : \ \|v\| = 1\}$ denote the unit sphere in $L^2([0,1],\mathbb{R})$, and $x_1, \cdots, x_n$ be a set of functions observed on $L^2([0,1],\mathbb{R})$. The projection index is defined as

$$Q(v) = Q(\langle x_1, v\rangle, \cdots, \langle x_n, v\rangle),$$

where $\langle \cdot, \cdot \rangle$ is the inner product defined in a Hilbert space. We want to find a set of functions $v_j, j = 1, 2, \cdots$ on the unit sphere on U^∞ such that

$$\begin{aligned} v_1 &= \underset{v \in U^\infty}{\operatorname{argsup}}\, Q(v) \text{ and} \\ v_j &= \underset{\substack{v \in U^\infty, \\ \langle v_j, v_m\rangle = 0 \text{ for } m<j}}{\operatorname{argsup}} Q(v) \text{ for } j = 2, 3, \cdots. \end{aligned}$$

An issue that presents itself here, in contrast with the multivariate setting, is that the maximum of $Q(v)$ is generally not well defined, owing to the fact that the unit sphere in $L^2([0,1],\mathbb{R})$ is not compact. An obvious way to fix this is to restrict the search for projections of the data to compact subsets of U^∞, which, as a result of Riesz's lemma (see e.g. Riesz and Sz.-Nagy (1990)), must be finite dimensional. Such a finite dimensional subset must be spanned by a finite collection of orthonormal basis functions, and hence a natural way to explore compact subsets of U^∞ is then to consider those that intersect a k dimensional linear subspace of the form $L_k = \text{span}(\phi_1, \ldots, \phi_k)$, for some orthonormal basis elements $\phi_1, ..., \phi_k$ chosen by the practitioner. For a chosen subspace L_k, we then instead consider maximizing

$$Q^{L,k}(v) = \sup_{v \in U^\infty \cap L_k} Q(v). \tag{2.4.1}$$

This supremum is well defined if the function $Q(v)$ is (almost surely) continuous over $U^\infty \cap L_k$, which holds in many cases under quite mild conditions in addition to $v \in U^\infty \cap L_k$, often basically entailing that L_k is not orthogonal to the data. The consequence of our restriction to the searching space is that the set of optimal directions are constructed as linear combinations of the basis functions $\phi_1, ..., \phi_k$ that span L_k. One might choose the subspace L_k and its dimension based on a number of considerations. If the observed functional data have been obtained by smoothing over a particular basis, such as the Fourier basis or a spline basis, then that basis and the dimension used for smoothing is

a natural choice for the subspace L_k. If departures from certain features are sought or expected in a particular way, then this information can also be used to select the basis. For instance, if it is believed that the functional data exhibits the feature of interest on a subset of its domain, then a Haar basis could be used.

The practical evaluation of the supremum defined in (2.4.1) requires maximizing the objective function $Q^{L,k}(v)$ over a high-dimensional unit sphere, which presents a difficult optimization problem. While there are algorithms proposed for projection pursuit in multivariate setting, due to the potentially high dimension of the unit sphere, as well as the complexity of the objective function that may arise in the present application, traditional methods might be ineffective. One reason for the poor performance of such methods is the fact that they may not search the high dimensional space thoroughly, and hence miss the global maximum. The method that we propose here to address this issue borrows from recent advances in the generation of low discrepancy sequences developed in the context of quasi Monte Carlo integration.

To explain the main idea behind these low discrepancy sequences, consider a p-dimensional unit hypercube $[0,1)^p$, and a sequence of points in the cube $Z = \{\mathbf{x}_j \in [0,1)^p,\ j = 0,1,2,\ldots\}$. Further, let $[a,b) = \{x \in [0,1)^p : a_i \leq x_i < b_i, i = 1,\ldots,p\}$ denote a sub-rectangular prism, and $A([\mathbf{a},\mathbf{b}),N)$ the number of the first N points from Z that lie in $[a,b)$. A desirable property of the sequence Z is that

$$\lim_{N\to\infty} \frac{A([\mathbf{a},\mathbf{b}),N)}{N} = \lambda_p([\mathbf{a},\mathbf{b}))$$

for any selection of the rectangular $[a,b)$, where λ_p denotes the p-dimensional Lebesgue measure. In order to quantify the rate at which the fraction $A([\mathbf{a},\mathbf{b}),N)/N$ converges to the limit, different measures of discrepancy have been proposed in the literature. Among them, the so-called star discrepancy has received a lot of attention:

$$D_N^*(S) = \sup_{\mathbf{b}\in[0,1]^p} \left| \frac{A([\mathbf{0},\mathbf{b}),N)}{N} - \lambda_p([\mathbf{0},\mathbf{b})) \right|.$$

In the context of numerical integration methods, the importance of star discrepancy stems from the Koksma-Hlawka inequality, which provides an upper bound for the error estimate for quasi Monte Carlo rules (see, for example, Niederreiter (1992) or Leobacher and Pillichshammer (2014)). This bound depends on the underlying integration nodes only through the star discrepancy, and this explains why sequences with low discrepancy are desirable.

As demonstrated by numerous authors, sequences with low discrepancy can also improve efficiency of some global optimization methods (for example, Kimura and Matsumura (2007), Pant et al. (2008), Georgieva and Jordanov (2009), and Monica et al. (2011)). In our problem, the goal is to generate a low discrepancy sequence on the unit sphere that could help us explore all regions of the unit sphere $U^\infty \cap L_k$. The recent work by Brauchart et al. (2015) provides an algorithm for generating such a sequence, which we use to propose a two-step optimization method to estimate the maximum of the projection index Q.

First, each function on $U^\infty \cap L_k$ can be expanded by the chosen basis functions $\phi_1, \cdots, \phi_k$, and the coefficients form a unit sphere in a k-dimensional Euclidean space. We generate a low discrepancy sequence of length J on this k-dimensional unit sphere as described in Brauchart et al. (2015). Denote these points by $\boldsymbol{\xi}_j = (\xi_{j,1}, \ldots, \xi_{j,k})$ for $j = 1, 2, \ldots, J$. For each $\boldsymbol{\xi}_j$, and specified basis functions ϕ_i, $i = 1, \cdots, k$, spanning L_k, we construct functions of the form $u_j = \sum_{i=1}^k \xi_{j,i}\phi_i$. This is equivalent to generate a sequence of functions $u_1, \ldots, u_J$ on $U^\infty \cap L_k$. Then we calculate the projection index of our data corresponding to u_j as

$$Q_j = Q(u_j).$$

The Q_j's may then be ranked, and we denote the largest M of them as $Q_{(1)} \geq \cdots \geq Q_{(M)}$. We denote the low-discrepancy points that produce $Q_{(m)}$ by $\boldsymbol{\xi}_{(m)}$, where $m = 1, \ldots, M$.

In a second step, we apply an optimization procedure to maximize $Q(v)$, for v continuous in local regions of the unit sphere centered at the initial points $\boldsymbol{\xi}_{(m)}$, $m = 1, \cdots, M$. The choice of optimization method is flexible, and common choices include the L-BFGS-B algorithm proposed by Byrd et al. (1995), particle swarm optimization (Clerc, 2010), and conjugate gradient descendant (Fletcher and Reeves, 1964). Let $\tilde{\boldsymbol{\xi}}_{(m)}$ denote the point at which $Q(\cdot)$ is optimized starting from the function on the unit sphere determined by the initial point $\boldsymbol{\xi}_{(m)}$. Then our final estimated vector of coefficients $\hat{\boldsymbol{\xi}}$ is determined as

$$\hat{\boldsymbol{\xi}} = \{\tilde{\boldsymbol{\xi}}_{(m)} : Q(\tilde{\boldsymbol{\xi}}_{(m)}) = \max_{m=1,\cdots,M} Q(\tilde{\boldsymbol{\xi}}_{(m)})\}.$$

For

$$\hat{u} = \sum_{i=1}^k \hat{\xi}_i \hat{v}_i, \tag{2.4.2}$$

the estimated maximum of the projection index Q is given by

$$\hat{Q}^{L,k} = Q(\hat{u}). \tag{2.4.3}$$

This procedure is similar to the coarse-to-fine optimization schemes popular in the machine learning community(see, for example, Pedersoli et al. (2015) and Charniak and Johnson (2005) for two applications in computer vision and natural language processing). Our algorithm is summarized in Algorithm 2.4.1. We should notice that we present a two-step optimization scheme here. However, when the problem has a more considerable complexity, one can repeat these two steps accordingly until satisfactory results are obtained.

Algorithm 2.4.1: Two-Step Approximation Algorithm for $\hat{Q}^{L,k}$

Input: $x_1, \cdots, x_n$, $\phi_1, \cdots, \phi_k$
Result: $\hat{Q}^{L,k}$
generate $\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_J$;
for $j = 1$ **to** J **do**
    generate $u_j = \sum_{l=1}^{k} \xi_{jl}\phi_l$;
    calculate $Q_j = Q(u_j)$;
end
rank $Q_1, \cdots, Q_J$ in decreasing order as $Q_{(1)}, \cdots, Q_{(J)}$;
for $m = 1$ **to** M **do**
    find $\boldsymbol{\xi}_{(m)}$ corresponding to $Q_{(m)}$;
    find the spherical coordinate $\{1, \theta_{(m),1}, \cdots, \theta_{(m),k-1}\}$ of $\boldsymbol{\xi}_{(m)}$;
    fix the $(k-1)$-dimension box in a small neighborhood of $(\theta_{(m),1}, \cdots, \theta_{(m),k-1})$;
    find optimized $\tilde{\boldsymbol{\xi}}_{(m)}$ in this box that maximizes $Q(\cdot)$;
end
let $\hat{\boldsymbol{\xi}} = \{\tilde{\boldsymbol{\xi}}_{(m)} : Q(\tilde{\boldsymbol{\xi}}_{(m)}) = \max_{m=1,\cdots,M} Q(\tilde{\boldsymbol{\xi}}_{(m)})\}$;
construct $\hat{u} = \sum_{l=1}^{k} \hat{\xi}_l \phi_l$;
calculate $\hat{Q}_n^{L,k} = Q(\hat{u})$.

In later chapters of this thesis, we adopt this algorithm to different applications, and show that projection pursuit method can lead to an efficient and robust dimension reduction for functional data.

Chapter 3

Projection pursuit based tests of normality with functional data

3.1 Introduction

The much related problem of testing for normality in multivariate data enjoys an enormous literature dating back at least to the 1960's. A myriad of techniques are now available, and, crudely, they can be categorized into four groups based on two characteristics. The first is how departures from normality in the data are measured, in which typically either moment based measures are used, such as the sample skewness and kurtosis, or goodness-of-fit tests involving the empirical distribution or characteristic function are employed. The second is how information is aggregated across the coordinates of the data, which usually amounts to either pooling/averaging the information across coordinates, or searching for linear combinations of the coordinates that maximize a given measure of non-Gaussianity. Approaches following the later paradigm are often termed "projection pursuit" methods, since finding such a linear combination can be framed as a classical projection pursuit problem as put forward in Kruskal (1972), and Friedman and Tukey (1974). Canonical test statistics based on moment methods of each type are Mardia's multivariate skewness (Mardia et al. (1979)), which aggregates the skewness across coordinates, and the skewness measure of Malkovich and Afifi (1973), which is the maximal sample skewness among all linear combinations of the coordinates. One test is expected to be preferable to the other depending on how "sparse" the non-Gaussianity is in the data: data for which all linear combinations of the coordinates are non-Gaussian should be more apparently non-Gaussian by considering aggregation based methods, while non-Gaussianity that can be explained

by only a few linear combination of the coordinates would typically be more easily detected using projection pursuit methods. Some examples of multivariate projection pursuit based normality tests can be found in Liang et al. (2000), Henze and Wagner (1997), Baringhaus and Henze (1991), Zhu et al. (1995a),Zhu et al. (1995b), and general reviews of tests for multivariate normality are given in Mecklin and Mundfrom (2004), Henze (2002), and Szekely and Rizzo (2005b).

In contrast, testing for normality of functional data objects has received considerably less attention. Methods based on random projections and subsequent Cramér-von Mises and Kolmogorov–Smirnov type goodness-of-fit tests are proposed and reviewed in Cuesta-Albertos et al. (2006), Cuesta-Albertos et al. (2007), Bugni et al. (2009), and Cuevas (2014). To date and to the best of our knowledge, the only test available for this purpose based on moment methods was put forward in Górecki et al. (2018), henceforth referred to as the GHHK test. Their approach involves projecting the functional data onto the span of the first several functional principal components estimated from the data, and then applying a test based on combining Mardia's skewness and kurtosis to the vectors of coefficients defining these projections, i.e. applying a multivariate Jarque-Bera test (Jarque and Bera (1980a)) to the projection scores. They also extend their method to serially correlated functional data. While this method proves to be effective in many cases, it evidently might be improved upon in several others. One is if the non-Gaussian components of the data are sparse among the leading principal components, analogously to the multivariate setting, but another is if the non-Gaussian components of the data are orthogonal to the leading principal components, in which case the GHHK test would not be expected to have more than trivial power. As we see below, this latter situation might occur more often than one might think, as it can arise from simply misspecifying the basis used to smooth/generate functional data objects from raw data and/or estimate the functional principal components. Although one may argue that this situation could be avoided by including more principal components, as later shown in a data example in Section 3.4.3, increasing the number of principal components incorporated into the GHHK test does not help solve the problem.

In this chapter, we propose and study an alternative normality test for functional data based on projection pursuit that overcomes some of these challenges. We consider as test statistics the maximal sample skewness and sample kurtosis among all scalar projections of the data onto a user selected compact subset of the unit ball, and hence the proposed test can be thought of as a functional generalization of the tests of Malkovich and Afifi (1973) and Baringhaus and Henze (1991). We show that the compact subset selected can be taken to be relatively high dimensional, and can also be generated by the functional principal components, which gives the test complimentary strengths to the GHHK test. A complete asymptotic theory is developed for the proposed statistics, and computational tools are

introduced to conduct the required high-dimensional projection pursuit. In addition to providing a test for Gaussianity, this projection pursuit method also furnishes a way to decompose functional data into a direct sum of approximately Gaussian and non-Gaussian components useful for data visualization or subsequent analyses, which we demonstrate in Section 3.4.3. This latter application builds upon some recent efforts to develop projection pursuit methods for functional data; see for example Bali et al. (2011). We study the proposed methods and compare them to the GHHK method in a simulation study, as well as in three applications to real data sets, which show the complimentary strengths of the two tests.

The rest of this chapter is organized as follows: In Section 3.2, we define our projection pursuit-based test statistics, and present their asymptotic properties. In Section 3.3, we detail several computational methods useful for calculating the proposed statistics and their critical values, and also describe and present the results of a simulation study. The results of the data analyses are presented in Section 3.4. The proofs of all technical results are contained in Section B.

3.2 Problem statement, definition of test statistics, and their asymptotic properties

Suppose that $x_1, ..., x_n$ is a simple random sample of size n of functional data sharing the same distribution as X. We assume throughout that each functional observation is then an independent stochastic process, whose sample path is in $L^2([0,1], \mathbb{R})$. More generally, we could consider a simple random sample of elements from a general, separable, Hilbert space, but because of the type of data applications we present in this chapter we consider the space $L^2([0,1], \mathbb{R})$ for clarity of presentation. Given this data, we are interested in testing the null hypothesis

$$\mathcal{H}_0 : \ X \text{ is a Gaussian process in } L^2([0,1], \mathbb{R}).$$

By definition, $\mathcal{H}_0$ can be equivalently stated as

$$\mathcal{H}_0 : \ \text{For each nonzero } v \in L^2([0,1], \mathbb{R}), \text{ the scalar random variable } \langle X, v \rangle \text{ is Gaussian.}$$

As discussed in the multivariate setting in Malkovich and Afifi (1973), the latter formulation motivates developing test statistics aiming to find the "least Gaussian" projection

of X. Indeed, if the distribution of such a projection does not significantly deviate from normality, then the same apparently holds for the entire process. In order to evaluate the normality of the projection of the data onto the direction v, a natural measure is the squared skewness and/or the absolute kurtosis:

$$S_n(v) = \frac{1}{n^2\hat{\sigma}^6(v)}\left[\sum_{i=1}^{n}(\langle x_i, v\rangle - \langle \bar{x}, v\rangle)^3\right]^2,$$

and

$$K_n(v) = \left|\frac{1}{n\hat{\sigma}^4(v)}\sum_{i=1}^{n}(\langle x_i, v\rangle - \langle \bar{x}, v\rangle)^4 - 3\right|.$$

Above we use $\bar{x}(t) = (1/n)\sum_{i=1}^{n} x_i(t), t \in [0,1]$, to denote the sample mean function and $\hat{\sigma}^2(v)$ to denote the sample variance of the scalar observations $\langle x_1, v\rangle, ..., \langle x_n, v\rangle$. Though here we consider "Jarque-Bera" moment based evaluations of normality, one could also consider projection pursuit methods based on other measures, for instance those surveyed in Thadewald and Büning (2007). Some benefits of using such moment based measures in this setting stem from their affine invariance and asymptotic properties, which, as we shall see below, are crucial in deriving feasible computational techniques to carry out a projection pursuit test in high dimensions.

Letting $U^\infty = \{u \in L^2([0,1], \mathbb{R}) \ : \ \|u\| = 1\}$ denote the unit sphere in $L^2([0,1], \mathbb{R})$, the least Gaussian projection may be calculated by evaluating the test statistics

$$S_n = \sup_{v \in U^\infty} S_n(v), \text{ and } K_n = \sup_{v \in U^\infty} K_n(v).$$

As discussed in Section 2.4, these test statistics are not necessarily well defined, and an obvious solution is to restrict the search for projections of the data to compact subsets of U^∞, which is finite dimensional. Let the chosen subspace $L_k = \text{span}(\phi_1, ..., \phi_k)$, for some orthonormal basis elements $\phi_1, ..., \phi_k$ chosen by the practitioner, we then instead consider the statistics

$$S_n^{L_k} = \sup_{v \in U^\infty \cap L_k} S_n(v), \text{ and } K_n^{L_k} = \sup_{v \in U^\infty \cap L_k} K_n(v). \tag{3.2.1}$$

Effectively, these statistics are measuring for multivariate normality in the subspace L_k based on the third and fourth order moments.

In the case when one would like a parsimonious finite dimensional representation of the observed functional data, functional principal component analysis is often employed. Let the first k estimated functional principal components introduced in Section 2.2 be $\hat{v}_{1,PCA}, \ldots, \hat{v}_{k,PCA}$. We note that the test statistic proposed in Górecki et al. (2018) is of the form

$$\text{GHHK}_k = \sum_{i=1}^{k} [S_n(\hat{v}_{i,PCA}) + K_n^2(\hat{v}_{i,PCA})],$$

which, under the condition that the first k eigenvalues in (2.2.1) are bounded away from zero and with suitable normalization, converges in distribution to a χ^2-random variable under $\mathcal{H}_0$. Letting $\hat{P}_k = \text{span}(\hat{v}_{1,PCA}, \ldots, \hat{v}_{k,PCA})$, one might alternatively test for normality in the principal component subspace by considering the statistics

$$S_n^{\hat{P},k} = \sup_{v \in U^\infty \cap \hat{P}_k} S_n(v), \text{ and } K_n^{\hat{P},k} = \sup_{v \in U^\infty \cap \hat{P}_k} K_n(v), \tag{3.2.2}$$

or

$$M_n^{\hat{P},k} = \max_{1 \le i \le k} \frac{n}{6} \left(S_n(\hat{v}_i) + \frac{1}{4} K_n^2(\hat{v}_i) \right), \tag{3.2.3}$$

i.e. the Jarque-Bera type test statistic, in which the maximal sum of the skewness and kurtosis is evaluated only over the first k principal component directions.

3.2.1 Large sample properties

The asymptotic properties of each of these statistics under $\mathcal{H}_0$ are detailed by the following two results.

Theorem 3.2.1. Suppose $x_1, ..., x_n$ are independent and identically distributed elements of $L^2([0,1], \mathbb{R})$ such that

1. $\mathcal{H}_0$ holds, and
2. $\inf_{v \in U^\infty \cap L_k} E\langle x_i, v \rangle^2 > 0$.

Then, with $S_n^{L_k}$ and $K_n^{L_k}$ defined in (3.2.1),

$$(nS_n^{L_k}, \sqrt{n}K_n^{L_k})^\top \xrightarrow{\mathcal{D}} (\sup_{v \in U^\infty \cap L_k} Z_1^2(v), \sup_{v \in U^\infty \cap L_k} |Z_2(v)|)^\top,$$

where Z_1 and Z_2 are independent mean zero Gaussian processes defined on $U^\infty \cap L_k$, whose covariance functions, defined as

$$\begin{aligned} \rho_1(v,r) &= 6(\boldsymbol{\xi}^\top(v)\boldsymbol{\xi}(r))^3, \text{ and} \\ \rho_2(v,r) &= 24(\boldsymbol{\xi}^\top(v)\boldsymbol{\xi}(r))^3 \end{aligned} \tag{3.2.4}$$

respectively, depend only on k.

This result may be proven in a similar fashion to the main theorem of Baringhaus and Henze (1991). We also note here that an asymptotic result of this type can easily be established under the more general condition that the projections of X onto L_k are elliptically symmetric, but we do not pursue that here. The asymptotic distribution presented in Theorem 3.2.1 can be used to estimate valid critical values for each test statistic under $\mathcal{H}_0$ using simulation. Furthermore, the form of this distribution shows that the tests based on $S_n^{L_k}$ and $K_n^{L_k}$ are asymptotically independent, which is useful in calculating a p value for $\mathcal{H}_0$ using both statistics jointly.

In order to derive similar results when the subspace used to define the test statistics is random and generated from the principal component basis, we make the following assumption.

Assumption 3.2.1. *The eigenvalues λ_i defined in (2.2.1) satisfy $\lambda_1 > \cdots > \lambda_k > \lambda_{k+1} \geq 0$.*

Assumption 3.2.1 implies that the principal component subspaces are asymptotically one dimensional, and in particular it implies that the estimated principal components are consistent up to a sign. This assumption could likely be relaxed to the one that only requires $\lambda_k > \lambda_{k+1}$ at the expense of some simplicity in the proof.

Theorem 3.2.2. Suppose Assumption 3.2.1 holds, and that $x_1, ..., x_n$ satisfy $\mathcal{H}_0$ and are independent and identically distributed. Then with $S_n^{\hat{P},k}$ and $K_n^{\hat{P},k}$ defined in (3.2.2),

$$(nS_n^{\hat{P},k}, \sqrt{n}K_n^{\hat{P},k})^\top \xrightarrow{\mathcal{D}} (\sup_{v \in U^\infty \cap P_k} Z_1^2(v), \sup_{v \in U^\infty \cap P_k} |Z_2(v)|)^\top,$$

where Z_1 and Z_2 are independent mean zero Gaussian processes defined on $U^\infty \cap P_k$. Their covariance functions are defined in (B.1.2). Furthermore,

$$M_n^{\hat{P},k} \xrightarrow{\mathcal{D}} \max_{1\le i\le k} \chi_i^2(2),$$

where $\chi_i^2(2),\ i = 1,\dots,k$, denote independent and identically distributed χ^2 random variables with two degrees of freedom.

Theorem 3.2.2 shows that at least when the principal component subspaces are fixed and one dimensional, the distribution of the maximal skewness and kurtosis is not asymptotically affected by the error in estimating the principal components. This comes basically as a result of the continuity of the functions $S_n(v)$ and $K_n(v)$. This result also shows that a test of asymptotic size α is obtained by rejecting $\mathcal{H}_0$ when $M_n^{\hat{P},k}$ exceeds $\chi^2([1-\alpha]^{1/k}, 2)$, where $\chi^2(\beta, 2)$ is the β^{th} quantile of the χ^2 distribution with two degrees of freedom.

3.3 Implementation and a Simulation Study

Practical evaluation of the estimates of the test statistics $\hat{S}_n^{L_k}$ and $\hat{K}_n^{L_k}$ defined in (3.2.1) requires maximizing the objective functions $S_n(v)$ and $K_n(v)$ over a potentially high-dimensional unit sphere, which presents a difficult optimization problem. This could be solved with the functional projection pursuit method introduced in Section 2.4, by letting the projection index $Q(\cdot)$ to be the skewness measure or kurtosis measure. In the rest of this sub-section we only use $\hat{S}_n^{L_k}$ as an example, since $\hat{K}_n^{L_k}$ can be evaluated similarly.

The projection pursuit algorithm for functional normality test consists the following two steps. First, we generate a low discrepancy sequence of length J on $U^\infty \cap L_k$. Denote the generated sequence by $u_1,\dots,u_J$. Then we calculate the skewness of the projection of our data onto u_j as

$$Sk_j = S_n(u_j), \quad j = 1,\dots,J.$$

We then rank Sk_j's in the descending order and denote the largest M of them as $Sk_{(1)} \ge \dots \ge Sk_{(M)}$. The Sk_j's may then be ranked, and we denote the largest M of them as $Sk_{(1)} \ge \dots \ge Sk_{(M)}$. We denote the corresponding points on the unit sphere that produce $Sk_{(m)}$ by $u_{(m)}$, where $m = 1,\dots,M$. In a second step, to maximize $S_n(v)$ we apply M times a local optimization procedure where as initial points we use $u_{(m)}$, $m = 1,\dots,M$.

In our implementation of the method we have used the L-BFGS-B algorithm proposed by Byrd et al. (1995), which allows the user to specify constraints on the domain over which the objective function is optimized. We also tried other optimization techniques, such as

particle swarm optimization (Clerc, 2010) and conjugate gradient descendant (Fletcher and Reeves, 1964), but the results were almost identical. Therefore, we only report results from L-BFGS-B method, since it performs slightly faster. Let $\hat{u}$ denote the point at which S_n is optimized, our estimated test statistic is then given by

$$\hat{S}_n^{L_k} = S_n(\hat{u}). \tag{3.3.1}$$

This algorithm is summarized in Algorithm 3.3.1.

Algorithm 3.3.1: Two-Step Approximation Algorithm for $\hat{S}_n^{L_k}$

Input: $x_1, \ldots, x_n$, $\phi_1(t), \ldots, \phi_k(t)$
Result: $\hat{S}_n^{L_k}$
generate $u_1, \ldots, u_J$;
for $j = 1$ **to** J **do**
    calculate $Sk_j = S_n(u_j)$;
end
rank $Sk_1, \ldots, Sk_J$ in decreasing order as $Sk_{(1)}, \ldots, Sk_{(J)}$;
for $m = 1$ **to** M **do**
    find $u_{(m)}$ corresponding to $Sk_{(m)}$;
    search for $\tilde{u}_{(m)}$ that maximize $S_n(\cdot)$ in a small neighborhood of $u_{(m)}$;
end
let $\hat{u} = \{\tilde{u} : S_n(\tilde{u}) = \max_{m=1,\ldots,M} S_n(u_{(m)})\}$;
calculate $\hat{S}_n^{L_k} = S_n(\hat{u})$.

This procedure necessitates the selection of two tuning parameters: the length of the low discrepancy sequence J and the number of initial points M. Our recommended procedure is to start from some initial values, like those we propose below, and stop as soon as we observe that the hypothesis testing decision and/or p-values are not sensitive to increasing values of these parameters. This is equivalent to checking that the statistic calculated and null quantiles estimated achieve stability as M and J increase. We have conducted a number of simulations to investigate what choices for these parameters are appropriate in practice. The results of some of these experiments are discussed and shown in the appendix of this chapter. In terms of stability in estimating the quantiles of the test statistics defined in (3.2.1), we have found that reasonable choices of these parameters in a dimension of 21 or less are $J = 3 \times 10^4$ and $M = 5$. We also illustrate here how one might choose the dimension of the subspace k in practice, and the discussion is also presented in the appendix. A natural idea is to perform the test for a range of values of k in order to further understand how any non-Gaussianity is manifested in the data, or if k should potentially

be increased. For Gaussian data, one expects that, as a function of k, the p-values of the test applied for different choices of k will fluctuate as, quite dependent, uniform random variables on $[0, 1]$, while for non-Gaussian data the p-values as a function of k should at some point become small.

In order to estimate the null distributions of $S_n^{L_k}$ and $K_n^{L_k}$, we utilize the fact that their limiting distributions are pivotal under $\mathcal{H}_0$ and estimate their critical values by simulation. In particular, letting q_α^S and q_α^K denote the α quantiles of $S_n^{L_k}$ and $K_n^{L_k}$ respectively, these are approximated by generating n k-dimensional multivariate normally distributed random vectors, $\boldsymbol{y}_i = (y_{i,1}, ..., y_{i,k})^\top$, $i = 1, ..., n$, with mean zero and identity covariance matrix. A functional sample $Y_i(t) = \sum_{j=1}^k y_{i,j}\phi_j(t)$ can be constructed from these vectors, to which we apply Algorithm 3.3.1 to calculate the statistics $S_{n,1}^{L_k}$ and $K_{n,1}^{L_k}$. By repeating this simulation B times we obtain a sample from statistics $S_{n,j}^{L_k}$ and $K_{n,j}^{L_k}$, $j = 1, ..., B$, and then we take q_α^S and q_α^K to be the α empirical quantiles of these respective samples. Below we take $B = 2000$ for estimating critical values. We found that this number is suitable for estimating the 1% and 5% critical values, although we recommend that it be increased if one wishes to consider values even further in the tail of the distribution.

We can estimate a p-value for the test based on these statistics as follows. With $\hat{S}_n^{L_k}$ and $\hat{K}_n^{L_k}$ denoting the test statistics estimated from the data, we take

$$\begin{aligned} p = P(S_n^{L_k} > \hat{S}_n^{L_k} \cup K_n^{L_k} > \hat{K}_n^{L_k}) &= 1 - P(S_n^{L_k} \leq \hat{S}_n^{L_k} \cap K_n^{L_k} \leq \hat{K}_n^{L_k}) \\ &\approx 1 - P(S_n^{L_k} \leq \hat{S}_n^{L_k})P(u_n^{L_k} \leq \hat{K}_n^{L_k}), \end{aligned}$$

where the last approximation is justified by the asymptotic independence of $S_n^{L_k}$ and $K_n^{L_k}$. The probabilities $P(S_n^{L_k} \leq \hat{S}_n^{L_k})$ and $P(K_n^{L_k} \leq \hat{K}_n^{L_k})$ can be estimated from the empirical CDF estimated from the simulation described above.

3.3.1 Simulation study

In order to evaluate the performance of the tests and the numerical methods proposed above, we conducted a simulation study, the results of which we now present. The synthetic data that we considered was generated from the basic model

$$x_i = \sum_{j=1}^{D} \epsilon_{i,j} f_j, \tag{3.3.2}$$

where $D = 101$, and $f_1, \ldots, f_{101}$ are the first 101 Fourier basis functions defined on the common domain $t \in [0,1]$ as $f_1 = 1$, $f_j(t) = \sqrt{2}\cos(\frac{j}{2}\pi t)$ for $j = 2, 4, \ldots, 100$, and $f_j(t) = \sqrt{2}\sin(\frac{j-1}{2}\pi t)$ for $j = 3, 5, \ldots, 101$. We also studied the case in which the basis elements f_j were non-smooth Haar basis elements. Our results, which for completeness are presented in the appendix, suggested that the performances of the tests were very similar to those in the smooth case.

We produced raw discrete data from the model (3.3.2) by evaluating $x_i(t)$ at 100 equally spaced points in the unit interval. To simulate data following $\mathcal{H}_0$, we generated the coefficient vectors $\epsilon_i = (\epsilon_{i,1}, ..., \epsilon_{i,D})^\top$ from a multivariate normal distribution with mean zero and covariance matrix $\Sigma = \Sigma_{D\times D}$. We considered three different types of the covariance structure. In two cases, Σ was diagonal,

$$\Sigma = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_D^2),$$

where we either took the diagonal elements to decay quickly, so that

$$\sigma_w^2 = \frac{1}{w^2}, \quad w = 1, \ldots, D,$$

and the resulting covariance matrix was labeled Σ_{fast}, or more slowly, in which case we took

$$\sigma_w^2 = \begin{cases} \frac{1}{\sqrt{w}} \text{ for } w = 1, 2, 3 \\ \\ \frac{2.4065}{w^2} \text{ for } w \geq 4, \end{cases}$$

and then the resulting covariance matrix was labeled Σ_{slow}. The normalizing constant 2.4065 was computed so that for both covariance matrices

$$\frac{\sum_{i=1}^{7} \sigma_i^2}{tr(\Sigma)} \approx 0.9,$$

which is a common threshold when using the total variance explained (TVE) in principal component analysis to select the number of components to retain. We should note that the TVE level could be arbitrary, for example in Górecki et al. (2018) the authors use 85%. In the rest of this chapter, we use 90% as the threshold for TVE.

For these diagonal covariance matrices the first d population level principal components of the observations $x_i, i = 1, \ldots, n$, are the functions f_1,..,f_d. Since initial Fourier basis elements do not fluctuate too much, the first d principal components can be estimated quite accurately. In order to investigate the situation in which the test might be sensitive

to the estimation of the principal components, we also considered generating data having a randomly constructed covariance matrix Σ_{ran} in the following way: we represent $\Sigma = P\Lambda P^{-1}$, where P is a $D \times D$ matrix whose columns are orthonormal to each other, and Λ is a diagonal matrix. We generate P by applying a QR decomposition to a $D \times D$ matrix filled by independent and identically distributed normal random variables with zero mean and unit variance, and we take $\Lambda = diag(101, 100, \ldots, 1)$. In this case the leading principal components of x_i in (3.3.2) are equally likely to be any of the functions $f_1, ..., f_D$, or linear combinations of them, and the eigenvalues of the covariance matrix decay quite slowly.

In order to generate data under $\mathcal{H}_A$, we consider three alternatives, which we label as $L1$, $L3$, and $M10$. For the alternative $L1$, we assume that the leading error term $\epsilon_{i,1}$ in (3.3.2) follows a scaled t-distribution with 5 degrees of freedom, mean zero, and variance equal to $\Sigma(1,1)$. In $L3$, the first three leading coefficients $\epsilon_{i,1}$, $\epsilon_{i,2}$, $\epsilon_{i,3}$ follow independently a scaled t-distribution with 5 degrees of freedom and variances $\Sigma(1,1)$, $\Sigma(2,2)$, and $\Sigma(3,3)$, respectively. In the last case $M10$, we assume $\epsilon_{i,10}$ follows the scaled t-distribution with 5 degrees of freedom and variance equal to $\Sigma(10,10)$. In both of the cases L_1 and L_3, the non-Gaussianity of the observations is contained in the leading principal components, and hence the methods based on PCA are expected to perform well. In contrast, for the alternative $M10$ the non-Gaussian component is orthogonal to the PCA subspaces of dimensions nine or less.

To conduct the simulations, for each setting we generated 1000 samples of lengths $n = 150$, 450 and 900. For each sample of curves, to estimate the test statistics $S_n^{L,21}$ and $K_n^{L,21}$ we applied the approximation method described in Algorithm 3.3.1. For linear spaces L_{21}, we considered $F_{21} = \text{span}(f_1, \ldots, f_{21})$, where f_i are the Fourier basis elements described above, and $B_{21} = \text{span}(b_1, \ldots, b_{21})$, where b_i are ortho-normalized B-splines constructed from 75 equally spaced knots of order 4.

We considered the following tests:

1. **PP-F-21**: Projection pursuit test with the subspace spanned by F_{21}.
2. **PP-B-21**: Projection pursuit test with the subspace spanned by B_{21}.
3. **PP-PF-7**: Projection pursuit test with the subspace spanned by the first 7 functional principal components estimated by initially smoothing the raw data using the Fourier basis.
4. **PP-PB-7**: Projection pursuit test with the subspace spanned by the first 7 functional principal components estimated by initially smoothing the raw data using the B-spline basis.

5. **GHHK-F**: GHHK test where we smooth the data using the first 75 Fourier basis functions and then estimate the pricipal componets from the coefficients. We use the 90% TVE criterion to select the number of principal components included.

6. **GHHK-B**: GHHK test where we smooth the data using 75 B-spline basis functions and then estimate the pricipal componets from the coefficients. We use the 90% TVE criterion to select the number of principal components included.

7. **MAX-F**: MAX test defined in (3.2.3) with data smoothed by Fourier basis. We use the 90% TVE criterion to select the number of principal components included.

8. **MAX-B**: MAX test defined in (3.2.3) with data smoothed by B-spline basis. We use the 90% TVE criterion to select the number of principal components included.

The percentage of rejections from the 1000 simulations at levels 5% and 1% are presented in Tables 3.1–3.3 for each covariance structure. The numbers in the Null column show the test sizes for different methods, while the numbers in L1, L3, and M10 columns show the power of each test under these three scenarios. The results can be summarized as follows:

- Each test exhibited reasonable size. The GHHK test and the MAX type tests were a bit oversized for large n, while the projection pursuit based tests tended to be a bit undersized.

- For the covariance structures Σ_{fast} and Σ_{slow} and the alternatives $L1$ and $L3$, the GHHK and MAX type tests performed superiorly and worked well regardless of the basis used to smooth the data. The projection pursuit based tests exhibited good power and consistency in these cases. By comparing the results for **PP-PF-7** and **PP-F-21**, one can get a sense of the sacrifice in power that is made by increasing the dimension of the search space, which can be quite severe: when the dimension increased from 7 to 21, the power was roughly halved at the significance levels of 5% and 1%.

- As expected, in the case $M10$ the GHHK test, the MAX type test, and the projection pursuit tests based on functional principal components have no more than trivial power, while the power of the other projection pursuit tests is very similar to what was observed under the alternative $L1$.

- When the covariance matrix used to generate the data was Σ_{ran}, then the performance of the GHHK test was strongly affected by the choice of basis used to smooth the raw data. When the Fourier basis was used, the GHHK test still exhibited strong, although somewhat diminished, power. On the other hand, when orthogonal B-splines where used to smooth the data, then the power was strongly diminished. This can be explained by the fact that the non-Gaussian signal in these cases often ends up in the Fourier basis elements that cannot be well represented by the first seven principal components calculated after initially smoothing the raw data using the orthogonal B-splines. In this case, the projection pursuit type tests are essentially unaffected by the choice of the basis, since even when the non-Gaussian component of the data is not well represented in the early principal components, it remains present in some linear combinations of the coordinates of the full data and can be essentially recovered without loss by the projection pursuit optimization.

Table 3.1: Percentage of rejections under the fast decaying covariance matrix Σ_{fast}.

level		$\alpha = 5\%$				$\alpha = 1\%$			
	method	Null	L1	L3	M10	Null	L1	L3	M10
n = 150	PP-F-21	5.3	13.7	29.9	15.1	1.6	8.7	22.2	9.4
	PP-B-21	2.9	15.0	34.4	15.8	1.0	9.8	24.5	9.3
	PP-PF-7	3.1	34.5	65.1	3.8	0.8	23.6	48.9	1.5
	PP-PB-7	3.7	33.0	65.7	11.0	0.8	20.4	50.2	6.6
	GHHK-F	4.7	69.1	97.4	4. 9	1.5	59.4	95.8	1.4
	GHHK-B	4.0	66.0	96.9	5.9	1.6	57.7	94.3	3.2
	MAX-F	7.2	74.0	97.1	7.1	2.7	66.4	95.3	2.8
	MAX-B	7.7	73.7	96.9	9.1	3.3	65.8	95.0	4.9
n = 450	PP-F-21	4.9	34.1	70.8	35.5	1.1	25.4	60.9	26.0
	PP-B-21	3.9	38.3	75.6	38.6	0.7	28.7	64.5	29.0
	PP-PF-7	3.3	79.1	99.1	3.7	0.3	65.2	95.1	0.4
	PP-PB-7	4.4	80.4	99.5	10.5	0.5	66.1	95.8	6.2
	GHHK-F	5.5	98.3	100	5.5	2.0	96.3	100	2.0
	GHHK-B	6.0	97.6	100	6.9	2.4	95.7	100	2.5
	MAX-F	6.4	99.0	100	6.4	2.8	97.5	100	4.2
	MAX-B	8.1	98.8	100	8.9	3.9	96.9	100	3.2
n = 900	PP-F-21	4.4	65.0	94.7	66.4	0.7	50.2	85.7	50.7
	PP-B-21	4.7	72.5	97.6	71.8	0.6	56.1	89.3	55.1
	PP-PF-7	4.2	98.6	99.9	4.6	0.7	96.9	99.9	0.8
	PP-PB-7	4.4	98.5	100	7.9	0.9	97.2	99.9	4.0
	GHHK-F	6.8	100	100	6.8	1.5	100	100	1.5
	GHHK-B	6.3	100	100	7.5	1.7	99.9	100	2.2
	MAX-F	7.9	100	100	8.0	3.2	100	100	3.2
	MAX-B	8.1	100	100	8.8	3.5	100	100	4.4

Table 3.2: Percentage of rejections under the slow decaying covariance matrix Σ_{slow}.

level		$\alpha = 5\%$				$\alpha = 1\%$			
	method	Null	L1	L3	M10	Null	L1	L3	M10
n = 150	PP-F-21	6.4	15.0	31.6	17.2	1.4	9.7	22.7	9.2
	PP-B-21	2.8	15.0	34.4	16.5	1.0	9.8	24.8	9.4
	PP-PF-7	3.0	32.7	65.2	3.9	0.5	20.6	49.7	1.2
	PP-PB-7	3.3	32.8	66.3	4.2	0.7	21.5	51.3	1.5
	GHHK-F	4.4	66.4	92.6	4.8	2.0	57.9	88.9	2.1
	GHHK-B	4.8	64.1	91.6	10.6	2.0	55.6	87.3	6.8
	MAX-F	7.3	73.5	92.8	7.3	3.5	65.3	88.9	3.5
	MAX-B	8.1	72.5	92.4	14.8	4.1	64.4	88.6	10.2
n = 450	PP-F-21	5.5	37.6	71.6	35.0	1.5	29.0	61.0	28.1
	PP-B-21	4.1	38.3	76.3	38.5	0.8	28.7	64.6	29.0
	PP-PF-7	4.1	79.6	99.4	6.8	0.4	65.1	96.6	3.1
	PP-PB-7	4.5	80.3	99.5	7.6	0.4	65.6	97.0	3.4
	GHHK-F	6.4	98.1	100	6.9	2.3	95.5	100	2.4
	GHHK-B	5.9	97.0	100	12.9	2.6	94.6	100	8.1
	MAX-F	6.7	98.7	100	6.8	3.5	96.6	100	3.4
	MAX-B	8.8	98.6	100	15.6	4.3	96.6	100	10.8
n = 900	PP-F-21	4.7	67.7	95.1	67.4	1.1	52.3	85.6	51.2
	PP-B-21	4.6	72.5	98.0	70.6	0.5	56.1	89.6	54.9
	PP-PF-7	4.0	98.5	100	6.7	1.2	97.3	100	3.8
	PP-PB-7	4.6	98.5	100	7.2	1.5	97.4	100	4.0
	GHHK-F	6.6	100	100	6.2	1.5	99.9	100	1.3
	GHHK-B	7.1	100	100	10.7	1.7	99.9	100	5.0
	MAX-F	7.3	100	100	7.2	2.8	100	100	2.9
	MAX-B	8.4	100	100	12.9	3.3	100	100	7.3

Table 3.3: Percentage of rejections under the random covariance matrix Σ_{ran}.

level		$\alpha = 5\%$				$\alpha = 1\%$			
	method	Null	L1	L3	M10	Null	L1	L3	M10
n = 150	PP-F-21	3.8	13.4	30.7	14.4	0.7	7.8	21.4	8.3
	PP-B-21	4.9	15.9	36.0	17.1	1.0	7.6	20.9	8.2
	PP-PF-7	3.1	30.9	66.0	23.6	0.8	19.8	48.9	15.6
	PP-PB-7	3.6	10.8	26.5	10.8	0.3	5.6	16.4	5.5
	GHHK-F	5.5	41.0	74.1	15.7	2.5	34.2	67.1	11.1
	GHHK-B	5.9	9.8	17.8	10.3	2.6	5.7	13.1	5.9
	MAX-F	13.3	51.4	80.6	26.5	6.8	44.3	75.0	18.9
	MAX-B	28.6	34.9	42.4	34.7	18.1	23.7	32.1	23.0
n = 450	PP-F-21	3.8	38.8	73.3	38.1	1.2	29.3	60.4	27.8
	PP-B-21	6.5	42.6	78.7	39.9	0.9	31.3	64.4	29.0
	PP-PF-7	3.3	77.1	99.2	56.0	0.6	63.0	95.5	44.1
	PP-PB-7	5.5	30.4	59.6	28.3	0.8	21.1	47.4	18.1
	GHHK-F	5.2	77.5	98.5	54.0	2.3	71.3	97.1	45.5
	GHHK-B	5.2	19.3	38.6	16.6	2.0	13.1	29.3	10.5
	MAX-F	12.0	89.8	99.1	52.5	4.6	86.1	98.3	24.8
	MAX-B	21.7	38.4	59.7	37.3	12.8	27.4	47.9	70.8
n = 900	PP-F-21	3.7	68.8	96.5	67.5	0.6	50.5	87.1	49.4
	PP-B-21	7.9	76.7	98.3	72.9	0.9	55.9	89.6	51.5
	PP-PF-7	5.6	97.8	100	70.0	1.1	95.4	100	64.5
	PP-PB-7	6.0	52.4	84.9	50.1	1.7	43.5	79.3	40.3
	GHHK-F	5.8	98.6	100	71.5	1.7	97.9	100	62.6
	GHHK-B	5.6	28.4	59.9	26.8	1.5	20.7	51.6	19.9
	MAX-F	9.2	99.3	100.0	77.1	3.6	98.6	99.9	70.8
	MAX-B	15.2	44.5	73.9	41.4	6.6	33.4	63.9	31.8

3.4 Data Analysis

In this section, we apply our proposed normality test to several real datasets, with the main objective of comparing its performance with that of some of the existing methods. While for some data sets all tests give similar results, in one case the proposed test leads to different conclusions than those implied by the existing methods. In addition, we also explain how the proposed projection pursuit method can be used for identifying and visualizing the non-Gaussian components of functional data.

3.4.1 Fertility rate in Australia

We first consider Australian fertility rate data from 1921 to 2006 among women aged from 15 to 49. The dataset has been collected by the Australian Bureau of Statistics and is available in the R package `rainbow` (Shang and Hyndman, 2016). In the left panel of Figure 3.1 each curve represents the distribution of the number of births per 1000 females at each age. From the rainbow plot, and some further analysis, we have found that the second order differencing of the curves is sufficient to remove the prevalent trend in the sequence of curves. The detrended curves are depicted in the right panel of Figure 3.1. After applying the GHHK-F test described in Section 3.3.1 to the detrended data, we have obtained a p-value equal to 0.826, which suggests that these curves are reasonably Gaussian. Using the proposed PP-F-21 we have obtained values of the test statistics $\hat{S}_n = 114.825$ and $\hat{K}_n = 23.895$, while the 95% level critical values are 132.915 and 32.792, respectively. The corresponding empirical p-value is 0.325, which is in apparent agreement with the GHHK test.

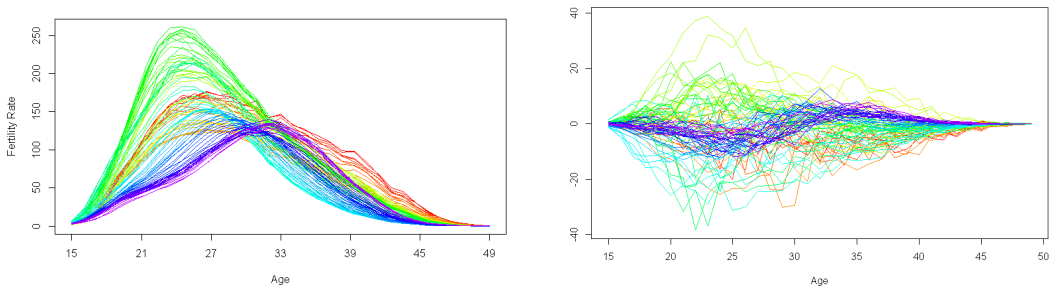

Figure 3.1: Fertility rate by age in Australia from 1921 to 2006.

3.4.2 Conditional intra-day stock prices

In modern finance, Brownian bridges arise naturally as conditioned Brownian motions in the context of the Black-Scholes model for option pricing. But there are numerous other applications of Brownian bridges, and more generally conditioned diffusion processes. For example, in applications that involve modeling of the flow of information in the market, like in Brody et al. (2008), a Brownian bridge represents the noise in the information about a future market event. In Cartea et al. (2016) the authors utilize a randomized Brownian bridge to model the mid-price of an asset with a random end-point that follows a distribution that is not necessary Gaussian. Such models can be justified, to some

extent, by the fact that Brownian motion is not the only diffusion process that produces a Brownian bridge when conditioned on its terminal value (Benjamini and Lee, 1997).

In this example we test whether a conditioned log-price of a traded security follows a Gaussian process, which is a less stringent requirement than the assumption that the price follows a geometric Brownian motion. To this end, we utilize the intra-day stock prices of IBM from 06/15/2006 to 04/02/2007, which are available in the R package `fChange` (Sonmez et al., 2018). The closing prices of one share of IBM stock were recorded from 9 a.m. to 4:30 p.m. at a one-minute resolution, and hence there are 390 observations each day. By analogy to the well-known construction for the Brownian bridge (e.g., Karlin and Taylor (1981)), we have transformed the observed prices to conditioned prices in the following way. Suppose the observed intra-day prices on a given day i are denoted $x_i(t_1), \ldots, x_i(t_n)$, and $y_{i,j} = \log x_i(t_j), j = 1, \ldots, 390$. We denote the straight line connecting $y_{i,1}$ and $y_{i,390}$ as $L_i(t)$. Then the bridged log prices are defined as $Y_i(t_1) = (y_{i,1} - L_i(t_1)), \ldots, Y_i(t_{390}) = (y_{i,390} - L_i(t_{390}))$. The widely used Black-Scholes model assumes that log-prices follow a Brownian motion, and hence these transformed price curves should follow a Brownian bridge, which is a Gaussian process. The daily curves of the transformed prices are shown in Figure 3.2. The p-value calculated from GHHK-F test is 3.78×10^{-8}, which suggests that these curves are non-Gaussian. Our PP-F-21 test generates test statistics $\hat{S}_n = 180.965$ and $\hat{K}_n = 42.401$, while the 95% level critical values are 101.635 and 29.381 respectively. The corresponding empirical p-value is essentially 0, and hence it is in agreement with the GHHK test.

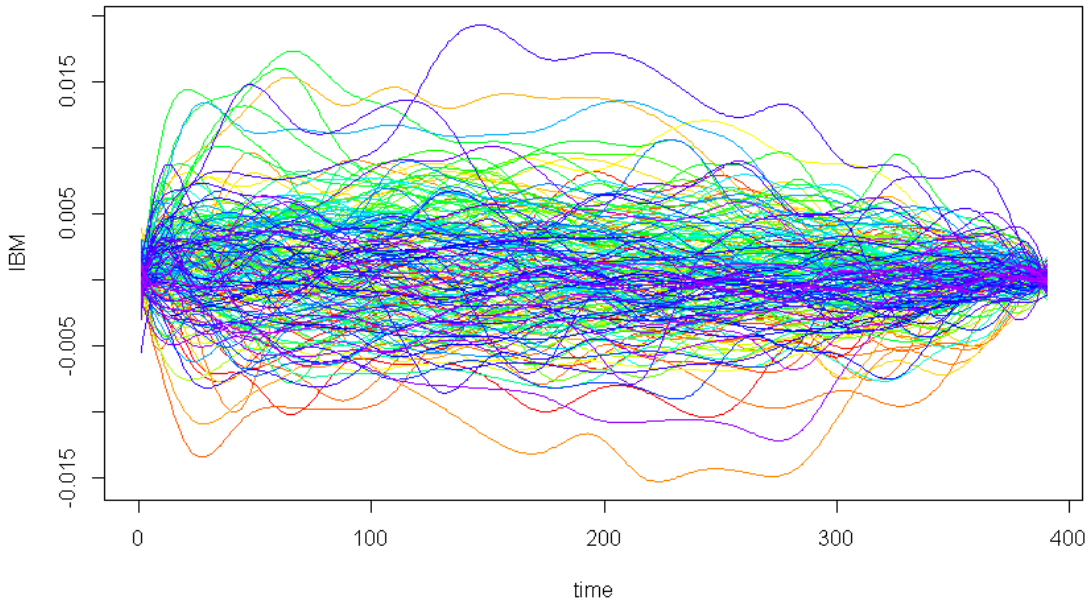

Figure 3.2: Daily curves of the transformed IBM prices from 06/15/2006 to 04/02/2007.

3.4.3 Yearly lower temperature profiles in Australia

In this final example we consider data comprised of the daily lowest temperature recorded in the Gayndah Post Office from 1893 to 2009, which is available both from the Australian Government Bureau of Meteorology and the R package `fChange` (Sonmez et al., 2018). Gayndah is a small town in Queensland, Australia, which is approximately 200km northwest of Brisbane. The settlement was established in 1849, and the Post Office was established at Gayndah in 1850. We analyze temperature records from 1894 to 2008, as the records prior to 1894 are not complete. In this case each functional observation x_i is defined to be the daily lowest temperature recorded in the Post Office for day $t = 1, 2, \ldots, 365$, in year $i = 1894, \ldots, 2008$. For leap years a 366^{th} data point is added. Since these yearly records have different lengths, we scale the data to the unit interval and smooth the curves using 21 Fourier basis. We than evaluate these curves on 365 equally spaced points in the unit interval. Figure 3.3 shows a rainbow plot of the data.

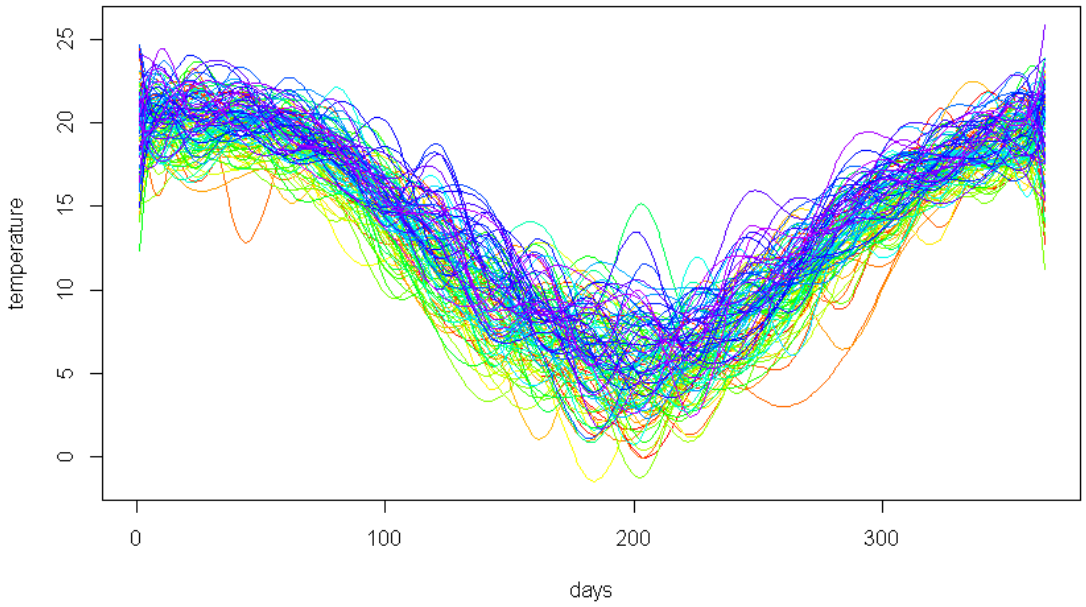

Figure 3.3: Daily lowest temperature at Gayndah Australia from 1894 to 2008.

The p-value of the GHHK test applied to this data is 0.928, which suggests that these temperature curves are plausibly realizations of a Gaussian process. However, in this case our projection pursuit based method suggests that these curves have components that are both skewed and heavy-tailed. The estimated test statistics for our PP-F-21 test are $\hat{S}_n^{L_k} = 185.49$ and $\hat{K}_n^{L_k} = 47.32$, which both exceed the corresponding estimated 95% critical values (130.59 and 33.97 respectively). The empirical p-value has been estimated as 0.002.

Letting p_1 denote the function that maximizes the skewness (or kurtosis) defined in (2.4.2), we can estimate the skewed (or leptokurtic) direction of each curve x_i as

$$g_{i1} = \langle x_i, p_1 \rangle p_1.$$

One can further remove this non-Gaussian component by point-wise subtraction to obtain the residual

$$x_i^{\text{new}} = x_i - g_{i1}.$$

Subsequent tests for Gaussianity may be applied to the sample $x_1^{\text{new}}, \ldots, x_n^{\text{new}}$ to find further directions $p_2,\ p_3, \ldots$ that will maximize the kurtosis or skewness. Suppose after m steps we are no longer able to reject the null hypothesis that the residuals are Gaussian processes. Then the curve x_i can be decomposed into two parts: an approximate non-Gaussian component $g_i = g_{i1} + \ldots + g_{im}$, and an approximate Gaussian component $r_i = x_i - g_i$.

These two components for the Gayndah temperature curves are presented in Figure 3.4, where we find 2 directions with excessive kurtosis and 1 direction with excessive skewness. We notice that most variants of the direction tend to vary more prominently at either end of the function, which corresponds to the summer in Australia.

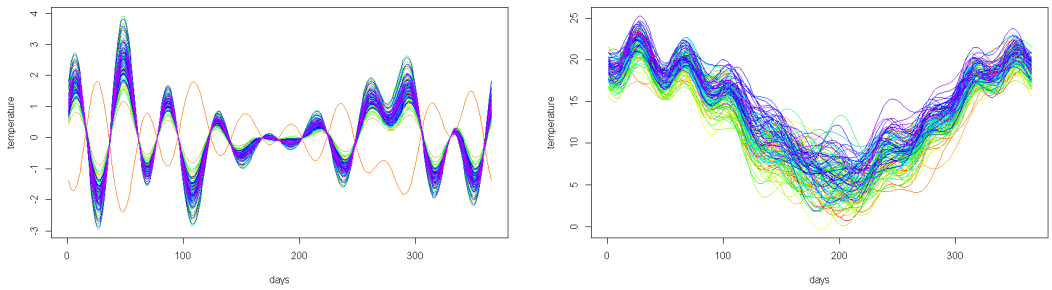

Figure 3.4: The left panel shows the non-Gaussian components we found when running the projection pursuit based test. The right panel are the residuals of the daily low temperature profile after removing these non-Gaussian components.

We ran the GHHK-F test again on both the estimated non-Gaussian components and and the residuals. The p-values were 0.000 and 0.678, respectively. The total variance explained (TVE) of the non-Gaussian components was around 3%, which is quite small relative to the usual TVE thresholds used to select the number of FPCs. While one could in general try to increase the TVE used to select the number of components in the GHHK test with the aim of discovering such a sparse non-Gaussian component, we point out that intuitively the GHHK method is a joint Jarque-Bera test applied to the projections onto FPCs. Therefore, increasing the number of FPCs will typically lead to a loss of overall testing power. A case in point is this example, in which after increasing the TVE threshold to 99% for the GHHK test, the test still fails to reject the Gaussianity of the curves at the 0.05% level with p-value equals to 0.723.

P-values of the test applied to the temperature data as a function of k are displayed in Figure 3.5 below, and show that strong non-Gaussianity is evident in the data after

projecting onto the first 3 Fourier basis elements, and then becomes essentially zero for k greater than 7.

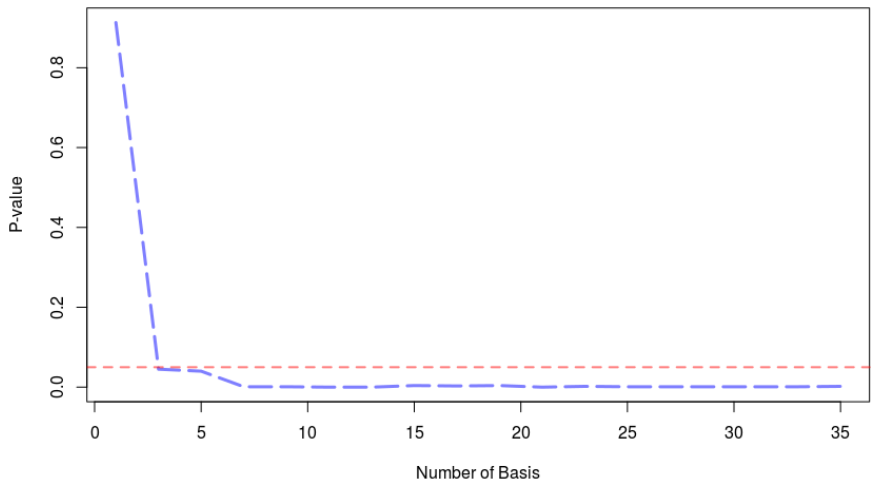

Figure 3.5: The p-values of proposed normality test under different number basis functions to construct the subspace with daily low temperature at Gayndah, Australia. The red horizontal dash-line is positioned at p=0.05.

Chapter 4

Functional Time Series Forecasting via Projection Pursuit

4.1 Introduction

The literature on forecasting functional time series is quite well developed and growing. Existing methods for predicting functional time series include functional auto regressive (FAR) models, the functional Yule-Walker equations proposed in Bosq (2012), non-parametric kernel based methods proposed in Besse et al. (2000), and linear wavelet based methods proposed in Antoniadis and Sapatinas (2003). In Kargin and Onatski (2008), the authors consider forecasting within an order one FAR (FAR(1)) framework by reducing the dimension of the autoregressive operator using the principal components of the lagged autocovariance operator.

Another class of methods to forecast functional time series are based on an initial dimension reduction of the observed data. The basic belief underlying this approach is that each curve of a functional time series can be decomposed as

$$x_i = \mu + \sum_{j=1}^{d} \beta_{ij} v_j + e_i, \tag{4.1.1}$$

where μ is the mean function, $v_j, j = 1, \ldots, d$, are d basis functions used for the purpose of dimension reduction, the β_{ij}'s are the corresponding projection scores, and the e_i are model errors. In Hyndman and Ullah (2007) and Shang (2013), the authors propose to use the leading functional principal components, which are the eigenfunctions of the sample

covariance operator at lag zero, as the basis functions for the dimension reduction. Motivated by the Gaussian case, in which the uncorrelated principal component scores are independent, the authors then propose to model each time series $\beta_{1j}, \ldots, \beta_{nj}, j = 1, \ldots, d$, independently with univariate time series models, for instance using Autoregressive Integrated Moving Average (ARIMA) models. Assuming that the h-step forward forecasted scores are $\hat{\beta}_{n+h,1}, \ldots, \hat{\beta}_{n+h,d}$, and the estimated mean function is $\hat{\mu}$, then the forecasted $(n+h)^{th}$ function may be expressed as

$$\hat{x}_{n+h} = \hat{\mu} + \sum_{j=1}^{d} \hat{\beta}_{n+h,j} v_j.$$

In Aue et al. (2015), the authors further extend this approach by modeling the p-variate principal component scores using a vector autoregressive (VAR) process. They demonstrate that with this adaptation one can handle potential cross-sectional lagged covariance structure within the principal component scores. As a result, this method offers some improvement in terms of forecasting accuracy when compared to some of the existing methods such as those proposed in Kargin and Onatski (2008) and Bosq (2012). Such dimension reduction based forecasting methods are straightforward to apply, and their basic application is summarized in Figure 4.1.

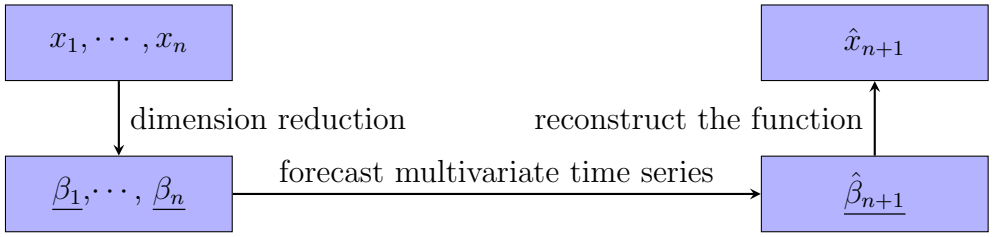

Figure 4.1: A basic schematic for forecasting functional time series through dimension reduction.

Notably, many of the dimension reduction based methods proposed to date for the purpose of forecasting rely on functional principal component analysis (fPCA). For such approaches, $\hat{v}_{1,PCA}, \ldots, \hat{v}_{d,PCA}$ are taken to be the eigenfunctions of the sample covariance operator of the data, namely they are the eigenfunctions of the kernel integral operator with kernel

$$\hat{C}(t,s) = \frac{1}{n}\sum_{i=1}^{n}[x_i(t) - \bar{x}(t)][x_i(s) - \bar{x}(s)], \text{ with } \bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t) \text{ for } t \in [0,1]. \quad (4.1.2)$$

For a survey on the topic, we refer to Shang (2014). Although principal component analysis decomposes data along directions where the variance of the projection scores is maximized, the projection scores with high variance might not always coincide with the predictable components of the series. To illustrate this problem, consider the following toy example. Suppose the functional time series to be forecast is generated as

$$x_i(t) = \alpha_i\sqrt{2}\sin(2\pi t) + \beta_i\sqrt{2}\cos(2\pi t),\ t \in [0,1],\ i = 1,\ldots,$$

where $\alpha_i,\ i = 1,2,\ldots,$ follow a scalar AR(1) process with marginal variance σ_α, and $\beta_i,\ i = 1,\ldots,$ is an independent white noise with variance σ_β. If $\sigma_\beta > \sigma_\alpha$, and one applies dimension reduction based methods using the leading functional principal component as the projection direction, then the resulting projection scores will be asymptotically (as the number of observations increases) the white noise sequence $\beta_i,\ i = 1,\ldots,$ which is not predictable. This kind of lack of efficacy of PCA for the purpose of forecasting can, as we also demonstrate in some examples below, happen in a more nuanced and impactful way with real datasets. Moreover, the calculation of the principal components relies on estimating the covariance kernel in (4.1.2), which can be affected by the presence of non-stationarity or outliers in the series.

In order to obtain a better subspace for forecasting, the dimension reduction step should focus on minimizing prediction errors rather than maximizing the variance of projection scores. Any such dimension reduction method will evidently depend on the forecasting method used to predict the projected series.

In this chapter, we consider the problem of finding a subspace that is tailored to a given finite dimensional forecasting method and a loss metric used to evaluate the prediction performance. Specifically, we approximate the forecasting loss for a given dimension reduction subspace using time series cross-validation. We then implement a projection pursuit technique to search for the subspace that minimizes the estimated loss. This is achieved by the development of novel computational tools to overcome the burden of this potentially high dimensional optimization problem. By using the same univariate ARIMA or vector AR forecasting methods as in Hyndman and Ullah (2007) and Aue et al. (2015), we show that this technique can significantly improve forecasting accuracy, especially when the functional time series is non-stationary or has the presence of outliers.

The rest of this chapter is organized as follows: In Section 4.2 we introduce our forecasting method. Section 4.3 contains the results of a Monte Carlo simulation study, and in Section 4.4 we present the application of two real-data sets. We present some additional results pertaining to the selection of hyperparameters involved in the proposed method in the appendix.

4.2 Methodology

4.2.1 Functional projection pursuit for forecasting

Suppose the observed functional time series is $x_1, \ldots, x_n$, where each x_i is an element of $L^2([0,1], \mathbb{R})$, and our goal is to predict the future curve x_{n+h}, where h is the desired forecasting horizon. Under the dimension reduction based forecasting framework, assume that we have found d orthogonal directions $v_1, \ldots, v_d$ that span a d-dimensional forecasting subspace as in (4.1.1). We denote the projection scores of x_i onto this subspace by $\beta_{i,j} = \langle x_i, v_j \rangle$, $i = 1, \ldots, n$, and $j = 1, \ldots, d$. Suppose, with some prediction function F_q, $1 \leq q \leq h$, one can forecast the d-variate projection scores q steps ahead. Let us denote the forecasted scores as

$$(\hat{\beta}_{n+q,1}, \ldots, \hat{\beta}_{n+q,d}) = F_q(\beta_{i,j} : i = 1, \ldots, n,\ j = 1, \ldots, d),$$

where at this point we are not imposing any restrictions on the prediction function F_q. Below, we use ARIMA models when comparing our approach with Hyndman and Ullah (2007) and Shang (2013), and vector autoregression when comparing with Aue et al. (2015), although one might also consider, for example, exponential smoothing, among many other options. With the forecasted scores, one may then construct the forecasted function as

$$\hat{x}_{n+q} = \sum_{j=1}^{d} \hat{\beta}_{n+q,j} v_j.$$

As discussed in the introduction, a natural method of selecting the directions $v_1, \ldots, v_d$ is fPCA, although this choice is not directly linked to the performance of any forecasting method. Here we propose instead a functional projection pursuit technique to search for such directions. To explain the main idea behind this method, we first consider the case when $d = 1$. Let v be an element of the unit sphere U^∞ and $\beta_i^{(v)} = \langle x_i, v \rangle$ the projection score of x_i onto v.

Denoting the h-step forward future curve as x_{n+h}, then the h-step forward forecasting error based on projection onto the direction v can be measured by

$$S_h(v) \coloneqq \frac{1}{h}\sum_{q=1}^{h} E\big[\|x_{n+q} - \hat{\beta}_{n+q}^{(v)} v\|^2\big]. \tag{4.2.1}$$

Instead of using the loss $L(x,y) = E\|x-y\|^2$ in (4.2.1), one could consider here an arbitrary loss function $L(\cdot,\cdot)$; for example $L(x,y) = \sup_{0\le t\le 1}|x(t)-y(t)|$. However, to simplify exposition and enable comparison with Hyndman and Ullah (2007) and Aue et al. (2015), in this study we use the expected integrated squared error. Ideally, one would like to minimize (4.2.1) with respect to v. However, without making strong assumptions about the data generating mechanism like, for example, assuming a functional autoregressive structure as in Kargin and Onatski (2008), the loss defined in (4.2.1) cannot be directly minimized as a function of v. We instead consider approximating $S_h(v)$ using time series cross-validation. Specifically, for some index i such that $i+h \le n$, let

$$\tilde{S}_h(i,v) \coloneqq \frac{1}{h}\sum_{q=1}^{h} \|x_{i+q} - \hat{\beta}_{i+q}^{(v)} v\|^2 \tag{4.2.2}$$

denote the accumulated loss in predicting x_{i+1} through x_{i+h} based on the data $x_1,\dots,x_i$ and projection onto the direction v. When it is reasonable to assume that the losses $\|x_{i+q} - \hat{\beta}_{i+q}^{(v)} v\|^2$ form a stationary series, one can approximate (4.2.1) by averaging $\tilde{S}_h(i,v)$ over a validation set. Let $r \in (0,1)$ denote the proportion of the validation set to the whole dataset, and $w = \lfloor r\cdot n\rfloor$ be the size of the validation set. Letting $\mathcal{V} = \{n-w-h+1,\dots,n-h\}$, so that $|\mathcal{V}| = w$, one can approximate the prediction loss defined in (4.2.1) by

$$\hat{S}_h(v) \equiv \hat{S}_{h,r}(v) = \frac{1}{w}\sum_{i\in\mathcal{V}} \tilde{S}_h(i,v). \tag{4.2.3}$$

A proper selection of the proportion r is an important aspect of the method, and we discuss it in Section 4.2.2.

Having an approximation for $S_h(v)$, we may then approximate the optimal forecasting direction by

$$\hat{v}_1 = \underset{v\in U^\infty\cap L_k}{\operatorname{argmin}}\ \hat{S}_h(v), \tag{4.2.4}$$

where $L_k = \text{span}\{\phi_1, \ldots, \phi_k\}$ is a finite dimensional linear subspace of $L^2([0,1], \mathbb{R})$ spanned by some orthonormal functions $\phi_1, \ldots, \phi_k$. The purpose of restricting the search for optimal forecasting directions to finite dimensional subsets of the unit ball is twofold, as discussed in Section 2.4. Firstly, for such subsets, $\hat{v}_1$ is well defined, as then $U^\infty \cap L_k$ is compact for all k, and $\hat{S}_h(v)$ is a continuous function of v. Secondly, conducting optimization to estimate $\hat{v}_1$ becomes feasible, as it can be parameterized in a finite dimensional space.

In order that $\hat{v}_1$ approximately minimizes (4.2.1) though, a large value of k is desired. Consequently, computing $\hat{v}_1$ as defined in (4.2.4) leads to a high-dimensional optimization problem, which we carry out using the projection pursuit algorithm introduced in Section 2.4 with the projection index set to be the approximated prediction loss $\hat{S}_h(v)$.

In the first step, we evaluate $\hat{S}_h(v)$ over a low-discrepancy sequence of length J on the $k-$dimensional subset of the unit sphere $U^\infty \cap L_k$. In the second step, a subset of size M of the points for which $\hat{S}_h(v)$ is smallest are selected, and a fine search for the optimal direction is conducted in a neighborhood of each point from this set using the L-BFGS-B algorithm. This optimization algorithm is summarized in Algorithm 4.2.1.

Algorithm 4.2.1: Two-Step optimization algorithm for finding most predictable direction.

Input: $x_1, \cdots, x_n$
Result: $\hat{v}$
generate low-discrepancy sequence $u_1, \ldots, u_J \in U^\infty \cap L_k$;
for $j = 1$ **to** J **do**
    calculate $S_j = \hat{S}_h(u_j)$ where $\hat{S}_h(\cdot)$ is defined in (4.2.3);
end
rank $S_1, \ldots, S_J$ in increasing order as $S_{(1)}, \cdots, S_{(J)}$;
for $m = 1$ **to** M **do**
    find the $u_{(m)}$ corresponding to $S_{(m)}$;
    search for $\tilde{u}_{(m)}$ that minimize $\hat{S}_h(\cdot)$ using L-BFGS-B algorithm in a neighborhood of $u_{(m)}$;
    compute $\hat{S}_h(\tilde{u}_{(m)})$;
end
set $\hat{v}_1 = \{\tilde{u} : \hat{S}_h(\tilde{u}) = \min_{m=1,\cdots,M} \hat{S}_h(\tilde{u}_{(m)})\}$;
return $\hat{v}_1$.

Finding multiple projection directions

While the discussion above focuses on the case when $d = 1$, in practice a higher dimensional forecasting subspace with $d > 1$ is often desired. In this case, our goal is to find a sequence of directions $v_j,\ j = 1, 2, \ldots, d$, on the unit sphere that collectively minimize the forecasting error. We suggest to approximate these iteratively as follows:

$$\hat{v}_1 = \underset{v \in U^\infty \cap L_k}{\operatorname{argmin}} \hat{S}_{h,1}(v), \text{ and}$$
$$\hat{v}_j = \underset{\substack{v \in U^\infty \cap L_k, \\ \langle v_j, v_l \rangle = 0 \text{ for } 1 \le l < j}}{\operatorname{argmin}} \hat{S}_{h,j}(v) \text{ for } j = 2, 3, \ldots, d,$$

where $\hat{S}_{h,1}(v)$ is defined in the same way as in (4.2.3), and $\hat{S}_{h,j}(v),\ j = 2, 3, \ldots, d$, are given by

$$\hat{S}_{h,j}(v) = \frac{1}{w} \sum_{i \in \mathcal{V}} \frac{1}{h} \sum_{q=1}^{h} \left\| x_{i+q} - \sum_{m=1}^{j-1} \hat{\beta}_{i+q}^{(\hat{v}_m)} \hat{v}_m - \hat{\beta}_{i+q}^{(v)} v \right\|^2.$$

Then Algorithm 4.2.1 can be repeatedly applied with $\hat{S}_h$ replaced by $\hat{S}_{h,j}$ to find each projection direction, $\hat{v}_1, ..., \hat{v}_d$.

4.2.2 Details of implementation

Tuning forecasting hyperparameters

In a typical forecasting problem with functional data, the dimension d of the forecasting subspace is an unknown parameter, and must be selected from the data. While there are undoubtedly a multitude of methods one might conceive to select d, here we introduce two: the sequential Goodness-of-Fit (SGF) stopping rule, and the elbow stopping rule. In the SGF stopping rule, we keep searching for new directions until we find a direction v_m such that if one applies a test for white noise to the projected time series $\beta_1^{(v_m)}, \ldots, \beta_n^{(v_m)}$, such as the Ljung-Box test (see e.g. Chapter 3 of Shumway and Stoffer (2017)), one cannot reject the hypothesis that the univariate projection scores are white noises. This approach is tailored to ARIMA modeling of the projected series, since if the projections onto the most predictive directions cannot be rejected as a white noise, then sensible ARIMA models fitted to these projections that are white noise models, and do not affect the forecast. In this chapter we use the SGF stopping rule unless otherwise stated, and stop our search as

soon as the p-value of the Ljung-Box test with maximal lag 3 applied to the projection scores on the last estimated direction is greater than 0.05.

In the elbow stopping rule, we consider the sequence $\hat{S}_{h,d}(\hat{v}_d)$ as a function of d. According to their respective definitions, $\hat{S}_{h,d}(\hat{v}_d)$ is, up to variations in the high-dimensional optimization scheme, a decreasing function of d, and it estimates the cross-validated loss from selecting the dimension d for forecasting. Typically, this function decreases quickly for small d and then flattens out for large d, as eventually further projections contribute little to improve the forecasting accuracy. As a result, the plot of $\hat{S}_{h,d}(\hat{v}_d)$ against d would typically exhibit an "elbow" around the point where this transition occurs, and then this value d may be as the elbow point. An example of such a pattern is shown in Figure C.1 in the context of forecasting pollution curves. See Arlot (2019) for a detailed survey of elbow methods in model selection.

Our proposed method also requires the choice of the number k of initial orthonormal basis functions to span L_k. Although these in principle could be any k orthonormal basis functions, e.g. spline functions or Fourier basis functions, what we recommend is to use the k leading principal components estimated from the data. The justification for this choice is the fact that the functional principal components are efficient in summarizing the original curves in terms of $L^2([0,1],\mathbb{R})$ normed loss. Given that leading principal components may not be ideal for forecasting, the resulting estimated forecasting directions are then linear combinations of a possibly large number of the k functional principal components that are approximately optimal for forecasting. We suggest taking k to be a much larger number than one would typically consider in fPCA. Letting k_v be the number of functional principal components needed to explain $\nu\%$ of the total variance of the observed data, we select $k = 3k_\nu$, and as a default set $\nu = 0.9$.

Tuning optimization hyperparameters

In the optimization step, one also needs to set r, which determines the size of the validation set by $w = \lfloor rn \rfloor$. In a number of simulated experiments, we investigated the impact of different values of $r \in (0.01, 0.6)$ on the forecasting loss. The results are presented in Section C.2. Based on these, we found that $r \in [0.05, 0.1]$ performed well across many data examples. In the simulations and real data examples below we set $r = 0.05$.

In the proposed method, two other optimization hyperparameters are J, the size of the low-discrepancy sequence generated in the first step of the optimization, and M, the number of potential candidates for optimal directions in the second step. In our implementations, we took $J = 10^3$ and $M = 3$, and did not observe any significant sensitivity of

our results to other choices of these parameters. We refer to Section C.2 for an additional discussion about the selection of these parameters.

Forecasting the multivariate scores

We forecast the projection scores of the functional time series using univariate ARIMA models as in Hyndman and Ullah (2007), or vector AR models as in Aue et al. (2015). To compare our approach with Hyndman and Ullah (2007), we fit d independent univariate ARIMA processes using the `auto.arima` function in the `forecast` package by Hyndman et al. (2019a), with the default setting for maximum orders left unchanged. The forecasted scores are obtained from the `forecast` function in the `forecast` package with the desired horizon. Suppose that these h-step forward forecasted coefficients are $\hat{\beta}_{n+h,1}, \ldots, \hat{\beta}_{n+h,d}$. Then the h-step forward forecasted curve is

$$\hat{x}_{n+h} = \sum_{j=1}^{d} \hat{\beta}_{n+h,j} \hat{v}_j. \tag{4.2.5}$$

To fit the d-dimensional vector AR model, the order p is determined using the functional FPE criterion discussed in Aue et al. (2015), with the maximum order limited to 5 and the dimension d calculated as described above. The VAR(p) model is then fitted using the `VAR` function in the `vars` package developed by Pfaff (2008), and the forecast is made using the `predict` function from the same package. The forecasted curve is constructed in the same way as in (4.2.5).

4.3 Simulation Study

In this section we present the results of a simulation study we conducted to evaluate the proposed method. We first introduce the data generating processes (DGPs) that we considered, and then compare the prediction accuracy of the proposed methods with the above referenced dimension reduction based competitors.

Recall from Bosq (2012) that a functional ARMA (FARMA) model with orders p and q is defined as

$$x_k = \Phi_1(x_{k-1}) + \ldots + \Phi_p(x_{k-p}) + w_k + \Theta_1(w_{k-1}) + \ldots + \Theta_q(w_{k-q}), k \in \mathbb{Z}, \tag{4.3.1}$$

where the $w_k's$ are a strong white noise innovation sequence in $L^2([0,1],\mathbb{R})$, and the Φ_i's and Θ_j's are Hilbert-Schmidt kernel integral (linear) operators mapping $L^2([0,1],\mathbb{R})$ to $L^2([0,1],\mathbb{R})$, so that $\Phi_i(f)(t) = \int_0^1 \phi_i(t,s)f(s)ds$, and $\Theta_j(f)(t) = \int_0^1 \theta_j(t,s)f(s)ds$, with $\|\phi_i\|, \|\theta_j\| < \infty$. In our simulation study, we assume the curves are generated in a D-dimensional function space spanned by the standard orthonormal Fourier basis functions $\psi_j,\ j = 1,\ldots,D$, *i.e.* each sample curve x_i and innovation w_i can be represented as

$$x_i = \sum_{j=1}^{D} \beta_{i,j}\psi_j, \quad w_i = \sum_{j=1}^{D} \omega_{i,j}\psi_j, \tag{4.3.2}$$

where $\beta_{i,j} = \langle x_i, \psi_j\rangle$, $\omega_{i,j} = \langle w_i, \psi_j\rangle$, $\boldsymbol{\beta}_i = [\beta_{i,j},\ j = 1,...,D]$, and $\boldsymbol{\omega}_i = [\omega_{i,j},\ j = 1,...,D]$. Since in this case the kernels ϕ_i and θ_j can be expanded as

$$\phi_i(t,s) = \sum_{j,k=1}^{\infty} \tilde{\Phi}[j,k]\phi_j(t)\phi_k(s),$$

with a similar expression holding for θ_j, then the operators Φ_k's and Θ_j applied to x_i and w_i of the form (4.3.2) can be represented as matrices $\tilde{\Phi} = \{\tilde{\Phi}[j,k],\ \ 1 \le j,k \le D\} \in \mathbb{R}^{D\times D}$. Hence, when working with these operators in our examples below, we simply associate them with a $D \times D$ matrix of coefficients defining their expansions based on the first D Fourier basis elements. Each FARMA process is generated after discarding a burn-in period of length equal to half of the desired sample size. Each curve in the synthetic sample is generated accordingly as in (4.3.2), and evaluated discretely at 75 equally spaced points in the unit interval to produce the raw data that is then forecasted.

In this study, we consider the following data generating processes:

- **FAR(2) process (FAR(2))**: In this DGP we simulate a standard FAR(2) process with D = 31. The matrix operators are $\tilde{\Phi}_1 = 0.5\Psi$, and $\tilde{\Phi}_2 = 0.2\Psi$, where Ψ is a $D \times D$ matrix generated in the following way. We first generate $\tilde{\Psi}$ such that $\tilde{\Psi}_{ij} = N(0,(i\times j)^{-1})$, and then take $\Psi = \tilde{\Psi}/\|\tilde{\Psi}\|$. This is similar to the DGP used in Aue et al. (2015). The innovation sequence is simulated according to $\boldsymbol{\omega}_i \sim N_D(\mathbf{0}, diag(\boldsymbol{\sigma}))$, where $\boldsymbol{\sigma}' = (j^{-1}, j = 1,\ldots,31)$.

- **FAR(1) process with a predictable component orthogonal to the leading principal components (FAR-PredCompOrth)**: In this DGP we let D = 11, and

x_i follows an FAR(1) model with

$$\tilde{\Phi}_1 = \begin{bmatrix} O_{10\times 10} & O_{10\times 1} \\ O_{1\times 10} & 0.8 \end{bmatrix},$$

where $O_{k\times l}$ is a $k \times l$ matrix filled with 0's, and $\boldsymbol{\omega}_i \sim N_D(\mathbf{0}, diag(\boldsymbol{\sigma}))$ with $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_{11})$, $\sigma_1 = \ldots = \sigma_{10} = 1$, and $\sigma_{11} = 0.2$. This DGP is constructed in such a way that when we select the number of principal components that can explain at least 90% of the total variance, then the only predictable component that corresponds to the 11^{th} Fourier basis element will be orthogonal to the leading principal components.

- **FAR(1) process with cross-sectional covariance structure (FAR-CrossSecCov)**: This scenario is similar to the one discussed in Section 6.2 of Aue et al. (2015). In this DGP,

$$\Psi = \begin{bmatrix} \Omega_{2\times 2} & O_{2\times 29} \\ O_{29\times 2} & O_{29\times 29} \end{bmatrix}, \text{ where } \Omega = \begin{bmatrix} 0.1 & 0.8 \\ 0.8 & 0.1 \end{bmatrix},$$

and the covariance matrix of the white noise is $\Sigma = diag(\boldsymbol{\sigma})$, where $\sigma_1 = \sigma_2 = 1$, and $\sigma_3 = \cdots = \sigma_{31} = 0.5$. This DGP generates scenarios where the cross-sectional covariance, modeled by Ω, dominates the autocovariance operator.

- **FAR(1) process with temporary change in white noise variance (FAR-VarShock)**: In this experiment we simulate scenarios where there is a temporary shock in the variance of the white noise. This DGP is similar to the **FAR-CrossSecCov** process with $D = 13$. That is,

$$\Psi = \begin{bmatrix} \Omega_{2\times 2} & O_{2\times 11} \\ O_{11\times 2} & O_{11\times 11} \end{bmatrix}, \text{ where } \Omega = \begin{bmatrix} 0.1 & 0.8 \\ 0.8 & 0.1 \end{bmatrix},$$

and the white noise vectors $\boldsymbol{\omega}_i \sim N_D(\mathbf{0}, \Sigma_i)$, where $\Sigma_i = diag(\boldsymbol{\sigma}_i)$, with $\boldsymbol{\sigma}_i = [\sigma_{1,i}, \cdots, \sigma_{13,i}]$ depends on i, the time series index of the curves. Specifically, when $1 \leq i \leq 100$ or $111 \leq i \leq 250$, $\sigma_{1,i} = \cdots = \sigma_{13,i} = 1$. However, when $101 \leq i \leq 110$, $\sigma_{1,i} = \sigma_{2,i} = \sigma_{8,i} = \ldots = \sigma_{13,i} = 1$ but $\sigma_{3,i} = \cdots \sigma_{7,i} = 10$. Hence, while the predictable components in the time series are stationary, their variances are inflated from the 101^{st} to the 110^{th} observation. A similar situation appears in the mortality data set we analyze in Section 4.4.2.

- **FAR(2) process with a polynomial trend (FAR-PolyTrend)**: In this scenario, $y_i, i = 1, 2, \ldots$ follow the DGP of the **FAR(2)** case, and $x_i(t) = (0.05i)^2 +$

$y_i(t)$, for $t \in [0,1]$. This DGP simulates an FAR(2) process with a polynomial trend.

- **FARMA(1,1) process (FARMA)**: In this DGP we generate an FARMA(1,1) process for which the AR operator of the coefficients is $\tilde{\Phi}_1 = 0.5\Psi$, and the MA operator is $\tilde{\Theta}_1 = 0.3\Psi$, with Ψ generated in the same way as for the **FAR(2)** case.

- **FARIMA(1,1,1) process (FARIMA)**: In this DGP, $(1-B)x_i$ follows the **FARMA** process, where B is the pointwise backshift operator, i.e. $Bx_i(t) = x_{i-1}(t)$, $t \in [0,1]$.

- **FARIMA(1,1,1) process with predictable components coinciding with the first two leading principal components (FARIMA-PCA)**: In this DGP we generate the coefficients $\beta_{i,1}$ and $\beta_{i,2}$ in the expression (4.3.2) following two univariate independent ARIMA(1,1,1) processes:

$$(\beta_{i,1} - \beta_{i-1,1}) = 0.5(\beta_{i-1,1} - \beta_{i-2,1}) + 0.3\omega_{i-1}^{(1)} + \omega_i^{(1)}$$
$$(\beta_{i,2} - \beta_{i-1,2}) = 0.4(\beta_{i-1,2} - \beta_{i-2,2}) - 0.2\omega_{i-1}^{(2)} + \omega_i^{(2)},$$

  where $\omega_t^{(1)}$ and $\omega_t^{(2)}$ are two Gaussian white noises with variance 1, and $\beta_{l,t}, l = 3, \ldots, 31$, follow a normal distribution with mean 0 and standard deviation 0.5. This DGP generates functional time series whose predictable part is in the leading functional principal components and is orthogonal to the white noise model errors. This DGP is specifically designed to be favorable for prediction using the fPCA based method of Hyndman and Ullah (2007) and Shang (2013).

To conduct our simulation study, for each setting we generate 100 samples of length $n = 250$. For each sample of functional time series, we forecast the last ten curves $x_{241}, \ldots, x_{250}$. For 1-step forward forecasts we predict the curve x_{i+1} using $x_1, \ldots, x_i$ for $i = 240, \ldots, 249$. For 10-step forward forecasts we make a 10-step forward forecasting and predict $x_{241}, \ldots, x_{250}$ at once using $x_1, \ldots, x_{240}$.

We denote our projection pursuit based method using independent univariate ARIMA models to forecast the projection scores as **PP-I**, and the one using vector AR model to forecast the curves jointly as **PP-J**. We compare our proposed methods with the method proposed in Hyndman and Ullah (2007), which we denote by **fPCA-I**, and with the method based on jointly forecasting the principal component scores that proposed in Aue et al. (2015), which we denote by **fPCA-J**. To the best of our knowledge, the former method is the only existing method suitable for forecasting non-stationary functional time series. While there are other competitive forecasting methods for stationary functional time series, such as those proposed by Bosq (2012) and Kargin and Onatski (2008), they have

previously been compared with **fPCA-J**. In Didericksen et al. (2012) and Aue et al. (2015), the authors suggest that **fPCA-J** is generally competitive or improves upon the above-mentioned methods for stationary functional time series forecasting. Therefore we did not compare them with the proposed methods.

For the **fPCA-I** and **fPCA-J** methods, we select the number of principal components such that their total variance explained (TVE) is at least equal to 90%. The functional principal components are approximated based on equally spaced discrete observations of the functional objects at a resolution of 1/75, as discussed in Hyndman and Ullah (2007). For the **fPCA-J** method, we apply the automatic selection method proposed in Aue et al. (2015) to choose the autoregressive order.

In this simulation study, we compare the integrated squared errors of our proposed methods with existing methods for 1-step and 10-step forward forecasting. Let the last 10 simulated curves in the b^{th} sample be $x^b_{241}, \dots, x^b_{250}$, and the corresponding forecasted curves be $\hat{x}^{b,1}_{241}, \dots, \hat{x}^{b,1}_{250}$ for 1-step forward forecasting, or $\hat{x}^{b,10}_{241}, \dots, \hat{x}^{b,10}_{250}$ for 10-step forward forecasting. Then the average integrated squared error is calculated as

$$IMSE_1 = \frac{1}{100}\sum_{b=1}^{100}\frac{1}{10}\sum_{i=241}^{250}\int_0^1 \left[x_i^b(t) - \hat{x}_i^{b,1}(t)\right]^2 dt,$$

and

$$IMSE_{10} = \frac{1}{100}\sum_{b=1}^{100}\frac{1}{10}\sum_{i=241}^{250}\int_0^1 \left[x_i^b(t) - \hat{x}_i^{b,10}(t)\right]^2 dt,$$

in the case of 10-step ahead forecasting. The average integrated squared error for each forecasting method and DGP are presented in Table 4.1.

Table 4.1: Average integrated squared error over 100 simulated datasets for each forecasting method and DGP considered. Cells are left blank (-) for methods **PP-J** and **fPCA-J**, which assume stationarity of the projected series, when applied to non-stationary DGPs.

	horizon	**PP-I**	**PP-J**	**fPCA-I**	**fPCA-J**
FAR(2)	$IMSE_1$	1.41	1.44	1.40	1.38
	$IMSE_{10}$	1.68	1.69	1.76	1.68
FAR-PredCompOrth	$IMSE_1$	10.63	10.59	10.84	10.80
	$IMSE_{10}$	10.75	10.67	10.79	10.79
FAR-CrossSecCov	$IMSE_1$	3.86	3.80	7.56	3.61
	$IMSE_{10}$	6.57	6.54	9.63	6.33
FAR-VarShock	$IMSE_1$	14.27	15.53	17.34	16.55
	$IMSE_{10}$	17.21	17.62	19.44	18.82
FAR-PolyTrend	$IMSE_1$	1.57	-	3.25	-
	$IMSE_{10}$	2.54	-	4.97	-
FARMA	$IMSE_1$	3.03	3.08	2.96	2.96
	$IMSE_{10}$	3.33	3.31	3.32	3.31
FARIMA	$IMSE_1$	4.03	-	20.66	-
	$IMSE_{10}$	21.39	-	38.20	-
FARIMA-PCA	$IMSE_1$	14.21	-	11.91	-
	$IMSE_{10}$	48.27	-	47.50	-

For stationary functional time series with forecastable components not orthogonal to the leading principal components, such as the **FAR** and **FARMA** DGPs, we see that both **PP-I** and **PP-J** performed quite similarly to the **fPCA-I** and **fPCA-J** methods. On the other hand, when the forecastable components were orthogonal to the leading principal components, like in the **FAR-PredCompOrth** case, **PP-I** or **PP-J** both outperformed the fPCA based methods. For the **FAR-CrossSecCov** process the performances of **PP-I** and **PP-J** were close to that of **fPCA-J**, while they all outperformed **fPCA-I**, which apparently was negatively impacted by the off-diagonal covariance structure. Interestingly in this case, the projection pursuit within the **PP-I** method apparently found projection directions that are able to capture the cross-sectional dependence structure with independent component forecasts to a similar degree as the joint methods.

When there exists a change in the innovation variance, as in the **FAR-VarShock** process, the proposed methods significantly outperformed the fPCA based methods. This can be attributed to the fact that the principal component estimates themselves are perturbed by the shock in such a way that they lose alignment with the predictable components of the process. The proposed methods based on functional projection pursuit appear to be more robust against such changes.

For the two non-stationary functional time series, **FAR-PolyTrend** and **FARIMA**, **PP-I** outperformed **fPCA-I**. In the **FARIMA-PCA** case, which is tailored to the **fPCA-I** method, we observed that **fPCA-I** indeed produced better 1-step and 10-step forward forecasting results, although the forecasts based on **PP-I** were similar. We did not implement the **fPCA-J** or **PP-J** to forecast non-stationary functional time series, since these methods have been designed for stationary series only.

To facilitate a comparison of the variability of the forecasts for each method, we plot the histogram of loss ratios of integrated squared errors between different methods based on all 100 samples. The ratios are defined as

$$r_1 = \frac{IMSE_1(\textbf{PP-I})}{IMSE_1(\textbf{fPCA-I})} \text{ and } r_2 = \frac{IMSE_1(\textbf{PP-J})}{IMSE_1(\textbf{fPCA-J})}.$$

A ratio smaller than 1 indicates that the forecasting error from projection pursuit based method works better than the analogous fPCA based method, with the opposite conclusion for ratios larger than 1. Figure 4.2 illustrates the distributions of loss ratios for selected DGPs, and Table 4.2 presents the percentage of time our proposed projection pursuit based methods outperform the corresponding fPCA based methods (i.e. when the ratio is smaller than 1). From these plots and Table 4.2 we can infer that when the functional time series is stationary and the DGP meets the assumptions of fPCA based methods, like in the **FAR** or **FARMA** case, our projection pursuit methods perform similarly to the existing methods. However, for non-stationary functional time series, or when the assumptions of fPCA based methods no longer satisfied, the projection pursuit based methods have smaller forecasting errors.

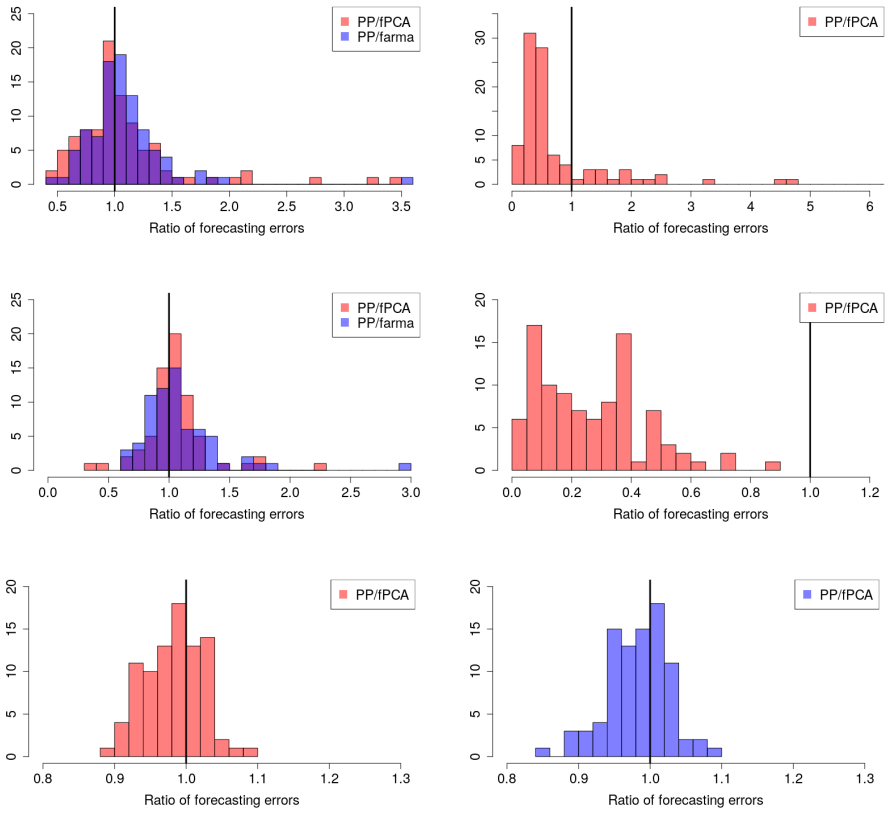

Figure 4.2: Histograms of ratios of 1-step forward forecasting errors of 100 simulated samples. The red bars present the ratios between **PP-I** and **fPCA-I**, and blue bars present the ratios between **PP-J** and **fPCA-J**. The two top panels correspond to **FAR** and **FAR-PolyTrend**, while the two middle panels correspond to **FARMA** and **FARIMA**. In the bottom panel, the left plot shows the ratios between **PP-I** and **fPCA-I** corresponding to **FAR-PredCompOrth**, while the right plot shows the ratios between **PP-J** and **FPCA-J**. Ratios less than one indicate an improved performance for the projection pursuit based techniques.

Table 4.2: Percentage of 1-step forward forecasting error ratios less than 1 in 100 simulated samples.

DGP	r1	r2
FAR	54%	42%
FAR-PolyTrend	80%	-
FARMA	40%	44%
FARIMA	100%	-
FAR-PredCompOrth	65%	61%

4.4 Data Analysis

In this section we apply the proposed projection pursuit based forecasting methods to two datasets: Daily PM10 concentration curves and age specific mortality curves.

4.4.1 PM10 concentration data

The first dataset we study is comprised of 30 minute resolution measurements of the concentration in air of particulate matter pollution with a diameter of less than 10 μm recorded in Graz-Mitte, Austria, from October 1, 2010, to March 31, 2011. Such particulate matter pollution is abbreviated as "PM10", and is known to have a negative effect on human health. Therefore, it is desirable for both policy makers and researchers to understand and be able to forecast the dynamics of PM10 concentration throughout the day. More details about PM10 can be found in Stadlober et al. (2008). The specific data that we have used is available in the `ftsa` package by Hyndman and Shang (2019). As the PM10 concentration is measured every 30 minutes and exhibits clear daily cycles, we can think of the data as discrete evaluations of daily pollution curves that we wish to forecast. The raw data is converted to full curves using linear interpolation. In total we have 182 such curves obtained over 182 consecutive days. A square-root transformation is applied to stabilize the intraday variance. These 182 curves are depicted in the top panel of Figure 4.3. From the plot we can tell that there is no obvious trend in the PM10 concentration curves, and the air pollution level is generally highest near noon and into the late afternoon, and the lowest in the late evening/early morning. We also plot the daily PM10 concentration at 9.a.m. for all 182 days in the bottom panel of Figure 4.3. The graphs suggest that the data are reasonably stationary.

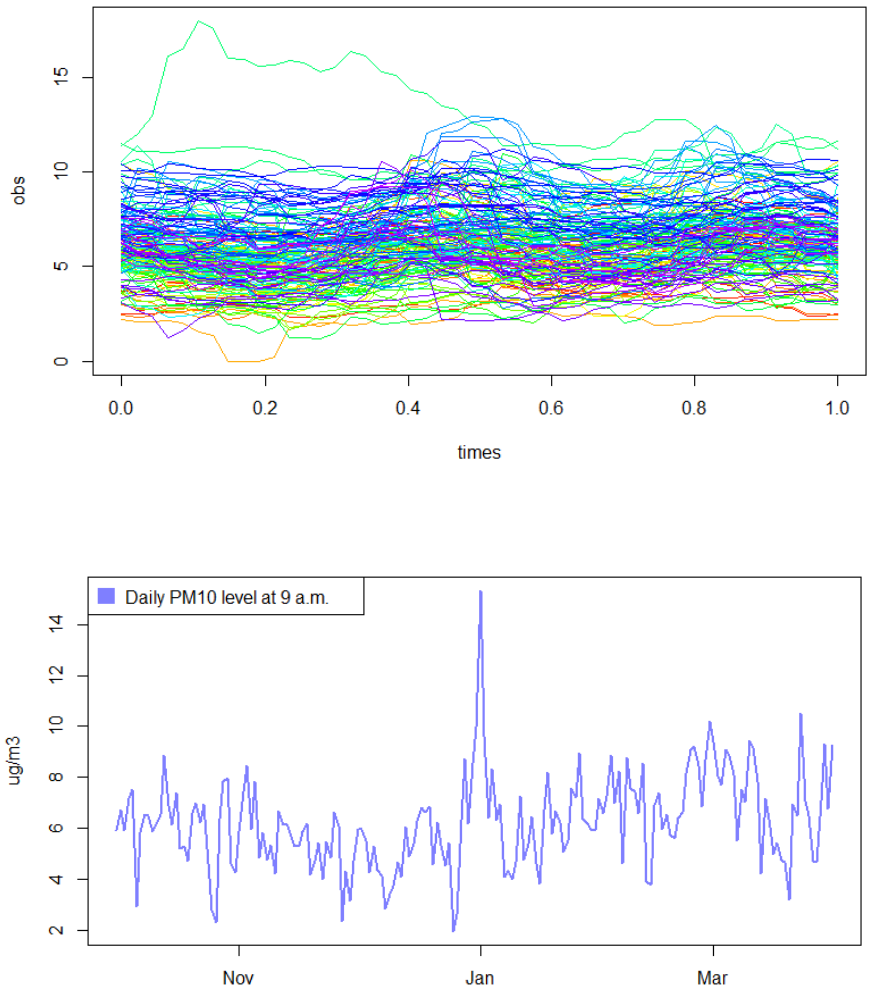

Figure 4.3: PM10 concentration in Graz-Mitte, Austria, from October 1, 2010, to March 31, 2011 (top panel) and daily PM10 concentration at 9 a.m.(bottom panel).

We split the data into two parts: we use the first 172 curves as the observed data, and compare the forecasting performance of **PP-J**, **fPCA-I**, and **fPCA-J** on the last 10 curves. For 1-step forward forecasting, we use an expanding window to predict the next day's PM10 concentration level curve from March 22 to March 31, 2011. For the 10-step

forward forecasting, we forecast PM10 concentration level curves from March 22 to March 31, 2011, all at once. The 1-step forward and 10-step forward prediction errors are both evaluated with integrated squared error. The prediction errors are presented in Table 4.3.

Table 4.3: Comparison of integrated squared prediction errors of three forecasting methods under different forecasting horizons for PM10 data.

	PP-J	**fPCA-I**	**fPCA-J**
$h = 1$	1.28	1.25	1.14
$h = 10$	1.25	2.39	1.42

We find that the 1-step prediction errors from the three different methods are quite close, while the 10-step prediction error from **fPCA-I** is much greater than the other two methods. The forecasted curves for March 22, 2011, from the 1-step forward predictions are presented in the top panel of Figure 4.4. We applied the Diebold-Mariano test, see Diebold and Mariano (2002), to the series of sequentially computed, 1-step ahead, integrated squared errors to asses whether differences in the performances of the forecasting methods are significant. The p-values of the Diebold-Mariano test between **PP-J** and **fPCA-I** and between **PP-J** and **fPCA-J** are 0.92 and 0.82 respectively, which would indicate that the performances of these 3 methods are not significantly different in terms of the 1-step forward forecasting.

The forecasted curves on March 31, 2011, from 10-step forward prediction are presented in the bottom panel of Figure 4.4. In this setting we expect some degree of cross-sectional autocovariance at lags greater than zero. For example, when the PM10 level is high in the late afternoon of the previous day, the particles in the air will not dissipate overnight, which may result in generally higher PM10 concentrations in the next morning. In such situations, the **fPCA-I** method tends to perform poorly, but projection pursuit seems to be flexible enough to capture this effect to a similar degree as the **fPCA-J** method. While the difference is not obvious for the 1-step forward prediction, we observe a much greater difference in the 10-step forward predictions.

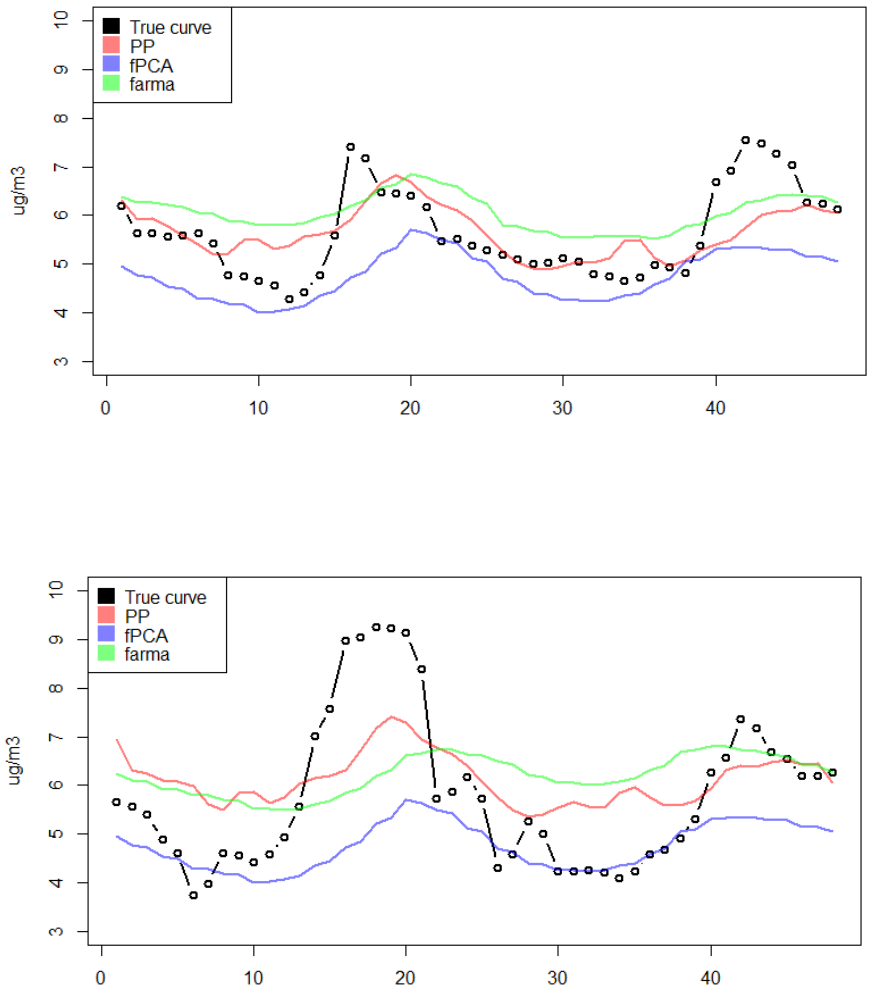

Figure 4.4: Forecasted PM10 concentration level curve on March 22, 2011, from 1-step forward forecasting (top panel) and forecasted PM10 concentration level curve on March 31, 2011, from the 10-step forward forecasting (bottom panel).

4.4.2 French male mortality rate data

The French male mortality rate data contains 191 years of age-specific mortality rates from 1816 to 2006. For simplicity we only consider the records for males aged between 0 and 100 years old. The dataset is available in the `demography` package by Hyndman et al. (2019b). To better emphasize the changes in the rate, we transform the data to the log scale. The transformed data is presented in Figure 4.5. As in Section 4.4.1, we split the data into two parts: we treat the first 181 curves as observed data, and based on these we make 1-step and 10-step forecasts for the last 10 curves. There is an obvious trend in the plot, and therefore the assumption of stationarity required for **fPCA-J** and **PP-J** is not satisfied. Hence in this data example we only compare results from the **PP-I** and **fPCA-I** methods.

The 1-step forward and 10-step forward prediction accuracy of these two methods presented in Table 4.4 are evaluated similarly as in Section 4.4.1. In this data example, we found that **PP-I** was more accurate compared to the **fPCA-I** method, with a p-value of 0.066 from the Diebold-Mariano test applied to the series of successive 1-step ahead forecast errors. The forecasted curves for year 1997 from the 1-step froward forecasting and year 2006 from the 10-step forward forecasting are depicted in Figure 4.6 and Figure 4.7.

Observe that in the 10-step ahead forecast of 2006, a noticeable difference in the predicted mortality rate curve is for males between 18 and 45 years old. We zoom in on this segment in Figure 4.8. Revisiting Figure 4.5 we can see some outlier curves for males around these ages in the time periods corresponding to World War I, the Spanish Flu, and World War II. Thus, a possible reason for this difference could be that while the overall trend for mortality rate is downward, those anomalous years have a significant impact on the estimated covariance operator, thereby affecting the estimated principal components. The **fPCA-I** method based on these estimated principal components then leads to an overestimation of mortality rates for these ages. In contrast, our method is less sensitive to these abnormal large variations, since they apparently contribute little to forecastability.

We further investigate the form of the 95% pointwise prediction interval curves resulting from these two methods, which we obtained by simulating from the model residuals of the component ARIMA models. While **PP-I** and **fPCA-I** generate similar 1-step forecasts, the prediction intervals are quite different. This difference is more obvious for the 10-step forward forecasting results. As shown in Figure 4.8, for the forecasted log mortality rates for males age between 18 and 45, **fPCA-I** has a much wider prediction interval. This could also be explained by the outliers caused by the two world wars. However, for males age above 60, we notice that the true log mortality rate is lower than the lower bound from **fPCA-I** method, while our **PP-I** method gives a reasonable lower bound, as illustrated in Figure 4.9. This might be due to the fact that there exists a strong decreasing trend

in log mortality rate for males at these ages that is not captured by the fPCA method. A similar pattern can also be observed for log mortality rates for males age between 0 and 15. Overall, among the 101 true log mortality rates, 41 (40.6%) of them are outside the 95% prediction interval generated from the **fPCA-I** method, while only 6 (5.9%) are outside the prediction interval from **PP-I**. Therefore, in this example, our method appears to be more robust against outliers that may cause large variability and strong non-stationarity in the functional time series.

Table 4.4: Comparison of average integrated squared prediction errors (on the order of 10^{-2}) of **PP-I** and **fPCA-I** under different forecasting horizons for French male log mortality data.

	PP-I	**fPCA-I**
$h = 1$	0.68	5.28
$h = 10$	2.68	5.75

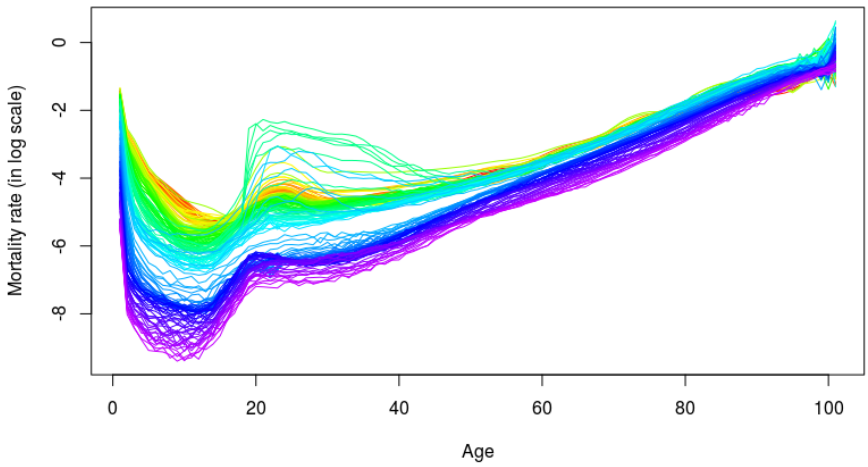

Figure 4.5: Log mortality rate for French Male between 0 and 100 years old from 1816 to 2006.

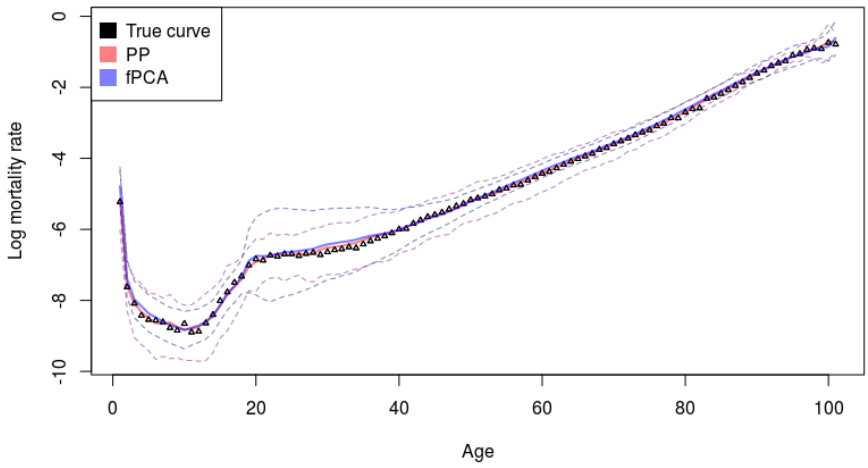

Figure 4.6: Forecasted French male log mortality rate curve in 1997 from 1-step forward forecasting. The dotted lines indicate 95% prediction interval curves for each method.

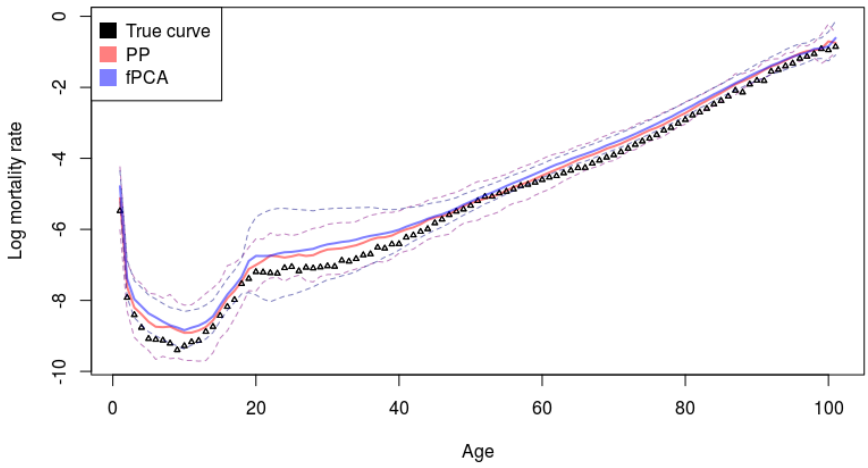

Figure 4.7: Forecasted French male log mortality rate curve in 2006 from the 10-step forward forecasting (bottom panel). The dotted lines indicate 95% prediction interval.

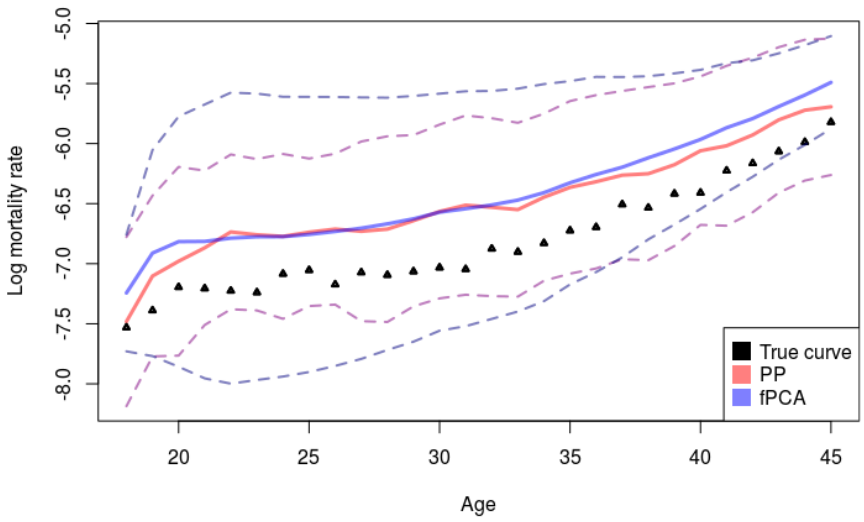

Figure 4.8: Forecasted log mortality rate curve for French male between 18 and 45 years old in 2006 with 95% pointwise prediction intervals resulting from the respective models.

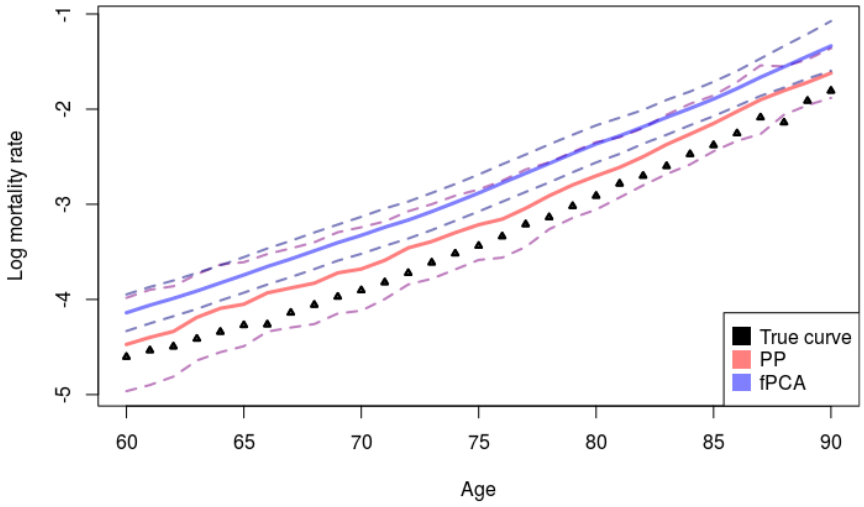

Figure 4.9: Forecasted log mortality rate curve for French male between 60 and 90 years old in 2006 with 95% pointwise prediction intervals resulting from the respective models.

Chapter 5

Change-points Detection in Functional Data

5.1 Introduction

As discussed in Section 1.2.3, the existing functional change point detection methods are limited in their applicability, as most of them focus on detecting changes of a specific type. In Berkes et al. (2009), the authors propose an approach based on functional principal component analysis (fPCA). The basic premise underlying this approach is that each functional observation x_i can be decomposed as

$$x_i(t) = \mu_i(t) + e_i(t), i = 1, \ldots, n, \text{ for } t \in [0, 1],$$

where the mean function μ_i changes after the K^{th} observation in the following way: $\mu_1 = \ldots = \mu_K \neq \mu_{K+1} = \ldots = \mu_n$. The functions $e_i, i = 1, \ldots, n$, are model errors representing random fluctuations such that $\int_0^1 e_i^2(t)dt < \infty$ and $E(e_i) = 0$. While it is not feasible to work on infinite dimensional data directly, the authors suggest that one can instead calculate the cumulative sum (CUSUM) statistic of the estimated leading principal components scores of the original data. In Aue et al. (2009), the asymptotic properties of this detection method are derived, and subsequent works by Aston and Kirch (2012) and Aston et al. (2012) extend this idea to dependent functional data. In Aue et al. (2014) and Gromenko et al. (2017a), the authors further extend this approach to functional linear models and spatially distributed functional data, respectively.

However, there is an unavoidable dimension-reduction loss when we summarize func-

tional data with leading principal components. In Aue et al. (2018) the authors propose a fully functional approach that is based on a functional version of CUSUM statistic and does not require any dimension reduction.

Very recently, researchers have started focusing on higher order structural breaks, especially on a change in the covariance structure of the functional data. In such cases, μ_i's are equal for $i = 1, \ldots, n$. However, the covariance structure of $e_i(t), t \in [0, 1]$, and therefore the covariance structure of $x_i(t), t \in [0, 1]$, changes after the K^{th} observations. In Aue et al. (2020), the authors suggest to apply CUSUM statistic to the eigenvalues and trace of empirical covariance operator truncated to a finite number of terms to detect changes in the covariance structure. In Stoehr et al. (2019) and Sharipov and Wendler (2019) the authors propose similar ideas to detect change points in the covariance structures of functional time series, and in Dette and Kutta (2019) the authors extend the method to detect the change points in both eigenvalues and eigenfunctions.

Notably, there are several problems with the existing methods. For instance, these methods can only be used to detect a specific type of change point, and would fail if the assumption about the type of change point is incorrect. To illustrate this problem, consider the following example. Suppose the functional data are generated as

$$x_i(t) = \alpha_i \sin(2\pi t), \ t \in [0, 1], i = 1, \ldots, 100;$$
$$x_i(t) = \beta_i \sin(2\pi t), \ t \in [0, 1], i = 101, \ldots, 200,$$

where $\alpha_i \sim N(0, 1)$ and $\beta_i \sim EXP(1) - 1$. While evidently there is a change point between $x_1, \ldots, x_{100}$ and $x_{101}, \ldots, x_{200}$, as the distributions of the coefficients are different in the two segments, methods for detecting changes in mean level or covariance operators will fail to identify the true change point, since the mean levels and covariance operators are constant. This type of distributional change could be more subtle in the context of real datasets.

In this chapter, we propose a more general change point detection framework for functional data based on projection pursuit. Suppose we have selected a metric to measure a difference in distributions between two groups of scalar values. Then for each potential location of the change point, we search for a projection direction such that this metric when applied to the projection scores before and after this location is maximized. We then check through all possible locations of the change point and identify a change point location that leads to the maximization of our metric applied to projection scores.

Since the above method is based on a flexible choice of metric, it can be adapted to different types of change points. In addition to locating the change point, this method

can also simultaneously identify the changing component of the functional objects. We further provide novel computational tools for an efficient implementation of our proposed algorithm. Using both simulation studies and real data examples, we demonstrate that our new method is versatile in detecting different types of change points in functional data, and can provide some insight into the potential cause of the change points.

We should notice that the proposed method can be easily modified to change point detection methods in other settings, although we do not pursue this direction. For example, one can replace a functional basis with the standard basis in the Euclidean space to detect change points in multivariate data.

The rest of this chapter is organized as follows: In Section 5.2 we introduce our functional change point detection method. In Section 5.3 we present results of our simulation study, and in Section 5.4 we apply our proposed method to two real-data sets.

5.2 Methodology

5.2.1 Functional projection pursuit for change point detection

We first define a functional change point as follows:

Definition 5.2.1. (i) Two random functions X and Y from $L^2([0,1],\mathbb{R})$ have the same distributions if and only if for any $u \in U^\infty$, the scalar random variables $\langle X, u\rangle$ and $\langle Y, u\rangle$ have the same distributions.

(ii) We say that K is a change point in functional realizations $z_1, \ldots, z_n$ if there exists some $u \in U^\infty$ such that the scalars $\langle z_1, u\rangle, \ldots, \langle z_K, u\rangle$ and $\langle z_{K+1}, u\rangle, \ldots, \langle z_n, u\rangle$ have different distributions.

Let u be an element on the unit sphere U^∞, and let $z_i^{(u)} = \langle z_i, u\rangle$ be the projection score of z_i onto u. For an integer $r \in \{2, \ldots, n-1\}$, we denote by $d(r,u)$ a measure of a difference between the common distribution of $z_1^{(u)}, \ldots, z_r^{(u)}$ and the common distribution of $z_{r+1}^{(u)}, \ldots, z_n^{(u)}$. Suppose that our goal is to find the change point K, as described in Definition 5.2.1, under the assumption that such a point exists and is unique. One strategy to locate the point is to examine all possible r, where for each r we search for the direction $u \in U^\infty$ such that $d(r,u)$ is maximized. Then the estimated change point is the location η that corresponds to the largest $d(\eta, v_\eta)$. That is,

$$\eta, v_\eta = \underset{r\in\{2,\cdots,n-1\}}{\operatorname{argmax}} \ \underset{u\in U^\infty}{\operatorname{argsup}}\, d(r,u). \tag{5.2.1}$$

The intuition behind (5.2.1) is as follows. By Definition 5.2.1, there exists some $u \in U^\infty$ such that the corresponding K is also a change point for the scalar projection scores. We quantify a distributional change in projection scores using a pre-selected metric, and the location of the change point should be the one that maximizes the metric function. Our proposed method aims to find a location η and the corresponding projection direction v_η such that the distributional change is maximal in view of the criterion in (5.2.1).

Due to the fact that U^∞ is not a compact set, the values η and v that solve the optimization problem in (5.2.1) may not be well defined. A possible solution is to restrict our search to some compact subset of U^∞, as discussed in Section 2.4. Such a finite dimensional subset can be constructed as the intersection between U^∞ and a linear subspace L_k spanned by k basis functions $\{\psi_j, j = 1, \dots k\}$, chosen by the practitioner, or

$$L_k = \text{span}\{\psi_j, j = 1, \dots k\}. \tag{5.2.2}$$

Then η^k and v_η^k that solve the problem

$$\eta^k, v_\eta^k = \underset{r \in \{2, \cdots, n-1\}}{\text{argmax}} \ \underset{u \in U^\infty \cap L_k}{\text{argsup}} \ d(r, u) \tag{5.2.3}$$

are well defined.

It is important to note that the choice of the linear subspace L_k is flexible, as long as the following assumption is satisfied.

Assumption 5.2.1. *Suppose K is the true change point, then there exists some $v \in L_k$ such that $\langle z_1, v\rangle, \dots, \langle z_K, v\rangle$ and $\langle z_{K+1}, v\rangle, \dots, \langle z_n, v\rangle$ have different distributions.*

This assumption implies that if we choose the linear subspace L_k such that it is not orthogonal to the subspace where the true change happens, then the functional change point detection problem is equivalent to a univariate change point detection problem after finding a proper projection direction v. Therefore, to locate the change point in functional observations one can first project the functional data onto L_k to obtain the k-variate scores, and then apply the projection pursuit method discussed in Section 2.4 to find a univariate projection scores such that an existing change-point detection method for univariate case would work.

One might argue that we can apply an arbitrary change point detection method for multivariate data to the k-variate projection scores. However, for the following reasons we instead implement a functional projection pursuit based method: First, to detect a change, one would expect large values of k so that the corresponding linear subspace is

arbitrarily close to the original function space. However, large values of k will likely lead to the so-called "curse of dimensionality", or the observed data are too sparse for multivariate methods to effectively detect the change point (see, for example Zhu et al. (2019)). Second, one might be interested not only in detecting a change point but also in understanding (or identifying) the aspect of the distribution that changes at this point. As demonstrated later in Section 5.4, our proposed change point detection method based on functional projection pursuit technique can help with the interpretation of the causes of the change in functional observations.

We would also like to note that one can project functional observations onto some random direction instead of searching for an optimal direction at each location r, so long as this random direction is not orthogonal to the components where the change happens. However, such an approach will not help us gain more information about the mechanism behind the change point, and the change point would be more difficult to detect when compared with our proposed projection pursuit based method.

5.2.2 Approximate optimal direction

As mentioned in Section 5.2.1, in order for our proposed method to be sufficiently powerful to detect the change point, large k that defines the dimension of the space L_k in 5.2.2 is often desired. Therefore, (5.2.3) poses a significant practical challenge as a high-dimensional optimization problem. In this chapter, we apply the projection pursuit technique introduced in Section 2.4. Suppose that r is fixed, and in the first step we evaluate $d(r, u)$ over a set of u comprising a low discrepancy sequence $u_1, \ldots, u_J$ generated on the k dimensional unit sphere in $U^\infty \cap L_k$.

In the second step of our approach, we select M points from $u_1, \ldots, u_J$ such that the corresponding $d(r, u)$ are the largest. Then a fine search is conducted in a small neighborhood of each of these selected points to obtain the optimal direction.

We repeat the above two steps for each possible r, and the final result would be a pair of r and u such that $d(r, u)$ is maximized. Two optimization hyperparameters are J, the length of the low-discrepancy sequence generated on the unit sphere, and M, the size of the subset we selected for finer search. In our implementation of the method, we used $J = 100$ and $M = 3$, and did not observe significant sensitivity in the results beyond this choice, in both simulation study and real-data analysis. We summarize the algorithm in Algorithm 5.2.1.

Algorithm 5.2.1: Multi-layer optimization algorithm for functional change-point detection.

Input: $z_1, \ldots, z_n$
Result: $\hat{\eta}^k$, $\hat{v}_\eta^k$
generate $u_1, \ldots, u_J \in U^\infty \cap L_k$;
for $r = 2$ **to** $n-1$ **do**
  for $j = 1$ **to** J **do**
    calculate $d_j^{(r)} = d(r, u_j)$;
  end
  rank $d_1^{(r)}, \ldots, d_J^{(r)}$ in decreasing order as $d_{(1)}^{(r)}, \ldots, d_{(J)}^{(r)}$;
  for $m = 1$ **to** M **do**
    find $u_{(m)}$ corresponding to $d_{(m)}^{(r)}$;
    search for $\tilde{u}_{(m,r)}$ that maximize $d(r, \cdot)$ in a small neighborhood of $u_{(m)}$;
  end
  let $\hat{v}^{(r)} = \{\tilde{u} : d(r, \tilde{u}) = \max_{m=1,\cdots,M} d(r, \tilde{u}_{(m,r)})\}$;
end
let $\hat{\eta}^k = \text{argmax}_{r \in \{2,\ldots,n-1\}} d(r, \hat{v}^{(r)})$ and $\hat{v}_\eta^k = \hat{v}^{(\hat{\eta})}$;
return $\hat{\eta}^k$, $\hat{v}_\eta^k$.

5.2.3 Choice of $d(r, u)$

In the context of the proposed method, the choice of $d(r, u)$ is flexible. Existing methods for multivariate data often assume that the type of change that occurs is known. For example, if we want to detect the change in the mean level, the t-test statistic would be a natural choice. That is,

$$d(r, u) = \frac{\left| {}^1\!/_r \sum_{i=1}^r z_i^{(u)} - {}^1\!/_{n-r} \sum_{j=r+1}^n z_j^{(u)} \right|}{\sqrt{s^2({}^1\!/_r + {}^1\!/_{n-r})}}$$

where

$$s^2 = \frac{\sum_{i=1}^r \left(z_i^{(u)} - \bar{z}_i{}^{(u)}\right)^2 + \sum_{j=r+1}^n \left(z_j^{(u)} - \bar{z}_j{}^{(u)}\right)^2}{n-2}$$

is the pooled sample variance of the projection scores of the functional observations onto u. The metric can be constructed in a similar way for detecting change points in higher moments.

However, a more realistic scenario is the one where we do not have prior knowledge about the type of change point. In this case, one could consider a metric that measures a distributional difference between two samples in a more general way. In this chapter, we quantify this difference using the weighted squared difference between the corresponding empirical characteristic functions. Characteristic functions, which characterize uniquely probability distributions by including information of all moments, can detect in principle any distributional change in observations, including but not limited to changes in mean level or covariance operators.

We denote by $\phi_u^{(r)}$ and $\varphi_u^{(r)}$ the characteristic functions of the underlying random variables of $z_1^{(u)}, \ldots, z_r^{(u)}$ and that of $z_{r+1}^{(u)}, \ldots, z_n^{(u)}$, respectively. In this chapter, we consider the following measure of difference between $\phi_u^{(r)}$ and $\varphi_u^{(r)}$:

$$d(r,u) = \frac{r(n-r)}{n} \int |\phi_u^{(r)}(t) - \varphi_u^{(r)}(t)|^2 \omega(t) dt,$$

where $\omega(\cdot)$ is some weight function which decays to zero, as $t \to \infty$, quickly enough for the above integral to be finite. The normalizer $r(n-r)/n$ is smaller when r is close to the beginning or end of the sequence, and larger when it is in the center, which helps to compensate for the effect of unbalanced sample sizes, similarly to the CUSUM statistic.

For a random variable X, its characteristic function $\phi_X(t) = E[e^{itX}]$ can be estimated with

$$\hat{\phi}_X(t) = 1/n \sum_{h=1}^{n} e^{itx_h},$$

where i is the imaginary number, and $x_1, \ldots, x_n$ are independent observations of X. In practice, the true characteristic functions $\phi_u^{(r)}$ and $\varphi_u^{(r)}$ are unknown, and one can replace them with empirical characteristic functions $\hat{\phi}_u^{(r)}$ and $\hat{\varphi}_u^{(r)}$. Then the distance $d(r,u)$ between two characteristic functions is approximated by

$$\hat{d}(r,u) = \frac{r(n-r)}{n} \int |\hat{\phi}_u^{(r)}(t) - \hat{\varphi}_u^{(r)}(t)|^2 \omega(t) dt. \tag{5.2.4}$$

We then estimate η^k and v_η^k as

$$\hat{\eta}^k, \hat{v}_\eta^k = \underset{r\in\{2,\cdots,n-1\}}{\operatorname{argmax}} \underset{u\in U^\infty\cap L_k}{\operatorname{argmax}} \hat{d}(r,u). \tag{5.2.5}$$

We further notice that empirical characteristic functions converge to their population counterpart more quickly when $|t|$ is close to 0, and it is a common choice to select $\omega(t)$ such that it decays to 0 fast when $|t| \to 0$. In this chapter we adopt the approximation proposed in Szekely and Rizzo (2005a). Specifically, we choose

$$\omega(t) = \Big(\frac{2\pi^{1/2}\Gamma(1/2)}{2\Gamma(1)}|t|^2\Big)^{-1},$$

and for which Szekely and Rizzo (2005a) and Székely et al. (2007) show that the integral in (5.2.4) can be approximated by

$$\begin{aligned}\int |\phi_u^{(r)}(t) - \varphi_u^{(r)}(t)|^2\omega(t)dt \approx{}& \frac{2}{r(n-r)}\sum_{i=1}^{r}\sum_{j=r+1}^{n}|z_i^{(u)} - z_j^{(u)}| \\ &- \binom{r}{2}^{-1}\sum_{1\le i<l\le r}|z_i^{(u)} - z_l^{(u)}| \\ &- \binom{n-r}{2}^{-1}\sum_{n-r+1\le j<l\le n}|z_j^{(u)} - z_l^{(u)}|. \end{aligned}\tag{5.2.6}$$

A similar approximation is also implemented in Matteson and James (2014). For simplicity, in the rest of the chapter we use this empirical characteristic function based metric. For other choices of the weight function ω, one can always evaluate the integral in (5.2.4) using more general quadrature methods.

5.2.4 Statistical significance of estimated change point

The statistical significance of the estimated change point can be tested through bootstrapping. Suppose that for the observed functional data $z_1, \ldots, z_n$, our projection pursuit method finds the change point $\hat{\eta}^k$, the projection pursuit direction $\hat{v}_\eta^k$, and the corresponding distance between empirical characteristic functions is $\hat{d}(\hat{\eta}^k, \hat{v}_\eta^k)$. In the bootstrapping test, one can resample the same number of curves with replacements, and assume the resampled observations are $z_1^{(b)}, \ldots, z_n^{(b)}$ in the b^{th} sample. Then with the location of the

change point $\hat{\eta}^k$ fixed, one can search for the optimal projection pursuit direction $\hat{v}^{(b)}$ and calculate the corresponding distance measure $\hat{d}^{(b)} = \hat{d}(\hat{\eta}^k, \hat{v}^{(b)})$. If there is indeed a change point at $\hat{\eta}^k$, $\hat{d}^{(b)}$ is more likely to be smaller than $\hat{d}(\hat{\eta}^k, \hat{v}^k_\eta)$, since the distributional difference between $z_1, \ldots, z_{\hat{\eta}^k}$ and $z_{\hat{\eta}^k+1}, \ldots z_n$ is supposed to be greater than the distributional difference between $z^{(b)}_1, \ldots, z^{(b)}_{\hat{\eta}^k}$ and $z^{(b)}_{\hat{\eta}^k+1}, \ldots z^{(b)}_n$. On the other hand, if there is no change point in the original observations, we should expect $\hat{d}^{(b)}$ to be close to $\hat{d}(\hat{\eta}^k, \hat{v}^k_\eta)$.

Suppose we repeat this procedure B times, and denote by $\hat{q}^B_\alpha$ as the $1-\alpha^{th}$ empirical quantile of $\hat{d}^{(b)}, b = 1, \ldots, B$. Then our estimated change point $\hat{\eta}^k$ is statistically significant if $\hat{d}(\hat{\eta}^k, \hat{v}^k_\eta) \geq \hat{q}^B_\alpha$.

5.3 Simulation Study

To construct simulated data, we follow a similar data generating process (DGP) as the one implemented in Aue et al. (2018). In each sample, $n = 250$ functional data objects are generated using $D = 21$ Fourier basis functions $\psi_1, \cdots, \psi_D$ on the interval $[0, 1]$. Independent curves are generated as

$$x_i = \sum_{w=1}^{D} \alpha_{i,w} \psi_w, \tag{5.3.1}$$

where $\boldsymbol{\alpha}_i = (\alpha_{i,w} : w = 1, \cdots, D) \sim MVN(\boldsymbol{\mu}, \Sigma), i = 1, \ldots, n$, are independent vectors of coefficients of the basis functions. We take 101 equally spaced realizations for each curve to simulate discrete observations of functional data. We should notice that in practice the bases that we typically use for smoothing discrete observations are never the same as that from the true DGP, and we mimic this mis-specification by spanning L_k with 15 B-spline basis functions.

The functional change point is introduced by a change in the distribution of the coefficients. We first define $\boldsymbol{\mu}_1 = \{\mu_{1,w} = 0, w = 1, \ldots, D\}$ as the baseline mean values of the coefficients. To introduce a change in the mean level, we define $\boldsymbol{\mu}_2$ such that $\mu_{2,1} = \mu_{2,2} = \mu_{2,3} = 0.5$ and $\mu_{2,4} = \ldots = \mu_{2,D} = 0$, which corresponds to the case where changes happen in the leading principal components, as explained below. We further construct $\boldsymbol{\mu}_3$ such that $\mu_{3,1} = \ldots = \mu_{3,10} = \mu_{3,12} = \mu_{3,D} = 0$ and $\mu_{3,11} = 0.1$. This mean vector mimics the scenario where the relevant for detection of change signal is weak.

The covariance matrix is generated as

$$\Sigma = P\Lambda P^{-1}$$

where P is a $D \times D$ squared matrix whose columns are orthonormal to each other, and Λ is a diagonal matrix. We generate P by applying QR decomposition to a $D \times D$ squared matrix filled with iid normally distributed random numbers with zero mean and unit variance. There are two possible choices for eigenvalues of Σ:

- $\Lambda_1 = \text{diag}(3^{-w} : w = 1, \cdots, D)$, which mimics fast decaying eigenvalues;
- $\Lambda_2 = \text{diag}(w^{-1} : w = 1, \cdots, D)$, which mimics slowly decaying eigenvalues.

In our study, we consider the following 8 types of change-point in functional observations:

- **Mean function-Strong-Fast**(**MSF**): Change in the mean function with fast decaying eigenvalues:

$$\boldsymbol{\alpha}_i \sim MVN(\boldsymbol{\mu}_1, P\Lambda_1 P^{-1}), i = 1, \cdots, 150;$$
$$\boldsymbol{\alpha}_i \sim MVN(\boldsymbol{\mu}_2, P\Lambda_1 P^{-1}), i = 151, \cdots, 250.$$

- **Mean function-Strong-Slow**(**MSS**): Same as the **MSF** case except the covariance matrix has slowly decaying eigenvalues:

$$\boldsymbol{\alpha}_i \sim MVN(\boldsymbol{\mu}_1, P\Lambda_2 P^{-1}), i = 1, \cdots, 150;$$
$$\boldsymbol{\alpha}_i \sim MVN(\boldsymbol{\mu}_2, P\Lambda_2 P^{-1}), i = 151, \cdots, 250.$$

- **Mean function-Weak**(**MW**): In this scenario we consider the more extreme case when the change in the mean function is very weak and away from the leading principal components:

$$\boldsymbol{\alpha}_i \sim MVN(\boldsymbol{\mu}_1, P\Lambda_2 P^{-1}), i = 1, \cdots, 150;$$
$$\boldsymbol{\alpha}_i \sim MVN(\boldsymbol{\mu}_3, P\Lambda_2 P^{-1}), i = 151, \cdots, 250.$$

- **Eigenvalue-Strong**(**EVS**): Change in all eigenvalues of covariance operator:

$$\boldsymbol{\alpha}_i \sim MVN(\boldsymbol{\mu}_1, P\Lambda_1 P^{-1}), i = 1, \cdots, 150;$$
$$\boldsymbol{\alpha}_i \sim MVN(\boldsymbol{\mu}_1, P\Lambda_2 P^{-1}), i = 151, \cdots, 250.$$

- **Eigenvalue-Weak**(**EVW**): Change in the non-leading eigenvalues of covariance operator. In this scenario, Λ_3 is a $D \times D$ diagonal matrix such that

$$diag(\Lambda_3) = \begin{cases} i^{-1}, i = 1, \ldots, 10, \\ 3^{-i}, i = 11, \ldots, 21. \end{cases}$$

  Then the coefficients are generated as

$$\boldsymbol{\alpha}_i \sim MVN(\boldsymbol{\mu}_1, P\Lambda_1 P^{-1}), i = 1, \cdots, 150;$$
$$\boldsymbol{\alpha}_i \sim MVN(\boldsymbol{\mu}_1, P\Lambda_3 P^{-1}), i = 151, \cdots, 250.$$

- **Eigenfunction-Strong**(**EFS**): Change in all eigenfunctions of covariance operator. In this scenario, we generate two random matrices P_1 and P_2, and then calculate two different covariance matrices with the slowly decaying eigenvalues:

$$\boldsymbol{\alpha}_i \sim MVN(\boldsymbol{\mu}_1, P_1\Lambda_1 P_1^{-1}), i = 1, \cdots, 150;$$
$$\boldsymbol{\alpha}_i \sim MVN(\boldsymbol{\mu}_1, P_2\Lambda_1 P_2^{-1}), i = 151, \cdots, 250.$$

- **Eigenfunction-Weak**(**EFW**): Similarly to the **EFS** case, we first generate two random matrices P_1 and P_2. We then construct another matrix P_3 such that the first 3 columns of P_3 are the same as the first 3 columns of P_1, and the remaining columns are the same as the corresponding columns of P_2. Then,

$$\boldsymbol{\alpha}_i \sim MVN(\boldsymbol{\mu}_1, P_1\Lambda_1 P_1^{-1}), i = 1, \cdots, 150;$$
$$\boldsymbol{\alpha}_i \sim MVN(\boldsymbol{\mu}_1, P_3\Lambda_1 P_3^{-1}), i = 151, \cdots, 250.$$

- **Distribution-Fast**(**DF**): Change in distribution of coefficients. We generate sample curves using Fourier basis. In this case we simulate 500 curves instead of 250. For $i = 1, \cdots, 300$, the coefficients are generated from iid $N(0, 1)$ distributions, while for $i = 301, \cdots, 500$, the coefficients are generated independently from $Gamma(1, 1)$ distributions and subtracted by 1 so that $E(\alpha_i) = 0$ and $Var(\alpha_i) = 1$ for all coefficients.

The change point detection method we mainly focus on in this chapter is the one based on the distance between empirical characteristic functions as the measure of difference in distributions (denoted as **change_PP**). For detecting change in mean functions, we also include results from the method based on t-test statistic (denoted as **change_PP_t**). We

compare our proposed method with several existing methods that are discussed in the introduction. For detecting a change in the mean function, we compare with **change_FF** proposed by Aue et al. (2018) and **change_fPCA** proposed by Aue et al. (2009). For detecting a change in the eigenvalues and eigenfunctions of the covariance operator, we compare our method with Aue et al. (2020) (denoted as **change_ev**) and Stoehr et al. (2019) (denoted as **change_ef**) respectively.

For each method, we present in Table 5.1 the average estimated location of the change point and the corresponding standard deviation based on 100 repetitions. To evaluate the performance of each change point detection method, we use the Rand index proposed in Fowlkes and Mallows (1983) and Hubert and Arabie (1985). In this study, we calculate the Rand index as the proportion of correctly estimated change point location, adjusted for correction-by-chance. Therefore, Rand index is a value between 0 and 1, and a larger score indicates a more accurate change point detection method. For each test, in Table 5.2 we present the Rand index, as well as the corresponding standard deviation.

Table 5.1: Average locations of change point based on different detection methods. Standard deviations are included in the parentheses.

	change_PP	**change_PP_t**	**change_FF**	**change_fPCA**
MSF	150.0(0.000)	150.0(0.000)	150.1(0.722)	149.3(6.263)
MSS	150.4(6.617)	150.0(0.681)	146.6(17.470)	150.2(2.187)
MW	150.0(0.200)	150.0(0.100)	138.4(34.165)	128.8(49.434)
	change_PP	**change_ev**	**change_ef**	
EVS	151.6(2.739)	150.8(1.855)	-	
EVW	147.9(6.929)	139.8(38.705)	-	
EFS	150.1(1.292)	-	150.0(4.007)	
EFW	144.0(20.525)	-	122.7(49.670)	
DF	291.3(30.183)	-	-	

Table 5.2: Average adjusted Rand indices for estimated locations of change point based on different detection methods. Standard deviations are included in the parentheses.

	change_PP	**change_PP_t**	**change_FF**	**change_fPCA**
MSF	1.000(0.000)	1.000(0.000)	0.998(0.011)	0.983(0.077)
MSS	0.944(0.081)	0.995(0.010)	0.908(0.187)	0.981(0.029)
MW	1.000(0.003)	1.000(0.002)	0.696(0.297)	0.466(0.301)
	change_PP	**change_ev**	**change_ef**	
EVS	0.974(0.042)	0.987(0.029)	-	
EVW	0.955(0.113)	0.632(0.296)	-	
EFS	0.992(0.019)	-	0.975(0.054)	
EFW	0.854(0.227)	-	0.444(0.301)	
DF	0.869(0.164)	-	-	

Table 5.1 and Table 5.2 suggest that the proposed projection pursuit based method performs better than the existing methods in all scenarios. For detecting the change in mean functions, our proposed method exhibits both better accuracy and better consistency compared with the two existing methods (**change_FF** and **change_fPCA**). Especially when the changing component has weaker signals, i.e. in the **MW** case, our method still shows outstanding performance, while two other methods have much smaller power in detecting such kind of change in the mean level. For the change in covariance operators, **change_PP** works as well as the existing methods when the signal of change is strong. However, when the changing component is not in the leading principal components (in **EVW** and **EFW** cases), our method still shows robust performance while the respective competitors fail. Furthermore, **change_PP** shows exceptional performance in detecting the change in the distribution of coefficients, while there is currently no competing method. In summary, the presented results suggest that **change_PP** is a powerful and versatile functional change point detection method that can be used without assumptions about the type of change point or data generating process. In all the studied scenarios, projection pursuit based methods are proven to be a very competitive and robust method.

We further investigate the size and power of our proposed change point detection method using the simulation approach discussed in Section 5.2.4 to estimate the p-value. In this experiment we compare our method with **change_FF** under the null case where there is no change point, and **MSF** and **MW** case discussed above. The percentage of rejections from the 100 simulations at levels 5% and 10% are presented in Table 5.3. The results suggest that our proposed method has a reasonable size when under the null case. In addition, the proposed projection pursuit based method has strong power against alter-

natives regardless of the strength of the true signal, while the power of **change_FF** drops significantly when the change in mean function is weak.

Table 5.3: Percentage of rejections under the null, **MSF**, and **MW**.

	level	**change_PP**	**change_FF**	**change_fPCA**
null	$\alpha = 5\%$	4	6	3
	$\alpha = 10\%$	9	7	7
MSF	$\alpha = 5\%$	100	78	86
	$\alpha = 10\%$	100	91	90
MW	$\alpha = 5\%$	100	35	8
	$\alpha = 10\%$	100	50	16

5.4 Data Examples

In this section we present two data examples to show how our projection pursuit based change point detection method can be applied on real world datasets. For both cases, the L_k is spanned by 15 B-spline basis functions. Comparing with the **change_FF** and **change_fPCA** mentioned above, our new method can detect different types of change and can provide more insights about the nature of the change.

5.4.1 Australian fertility data

We first analyze the Australian Fertility Data recorded by the Australian Bureau of Statistics from 1921 to 2015. It includes the fertility rate of Australian females aged from 15 to 49. The dataset is available in the R package `rainbow` (Shang and Hyndman, 2016). From Figure 5.1 we can see a clear trend in the curves, which suggests that these functional observations would not be homogeneous. The methods **change_PP**, **change_FF**, and **change_fPCA** all report a significant change point in the mean function near the 54^{th} curve, and the estimated change functions are also very similar (see the left panel of Figure 5.2). The right panel of Figure 5.2 shows the projection scores of Australian Fertility Data for the estimated change function from **change_PP**. We can tell that instead of a structural break in the mean function, there might exist a nonlinear trend in the data. We further take the 1^{st} and 2^{nd} order point-wise differences respectively, and apply the same as before methods of change point detection.

The 1^{st} order difference represents the rate of change for the curves. The three different methods, **change_PP**, **change_FF**, and **change_fPCA**, suggest that there is a change point at the 50^{th}, 40^{th}, and 55^{th} curve, respectively. The direction found by **change_PP** on which the change is most significant, as well as the corresponding projection scores, are presented in the first row of Figure 5.3. From the projection scores in the right panel we can see that there might exist both a change in mean level and a change in the covariance, and the rate of change shifts from positive to negative after the 50^{th} curve. A possible explanation of these results is the fact that the fertility rate is generally increasing before the 70s, and decreasing thereafter. The corresponding direction presented in the left panel suggest that women younger than 27 contribute most to the decrease of rate of change after 1970, while women elder than 30 have opposite effect on the change. In other words, the fertility rate is first increasing and then decreasing for younger women, and first decreasing and then increasing for elder women.

For the 2^{nd} order differences, which represents the acceleration of the change, **change_FF** and **change_fPCA** suggest that there is no change point. The change point found by **change_PP** is located after the 54^{th} curve with p-value 0.03. From the projection scores, which are shown in the right panel of the bottom row in Figure 5.3, we can tell that the change should be dominated by a change in the covariance operator instead of the change in mean function. Furthermore, the variance of projection scores is much smaller after the change point, which indicates that the variation of fertility rate at each age is smaller after the year 1975. From the corresponding direction found by **change_PP**, which is depicted in the left panel, we can tell that women aged around 25 and 40 contribute most to such a change. These findings can be confirmed by inspecting the raw curves in Figure 5.1. In conclusion, our results strongly indicate that in addition to being versatile in detecting different types of change point, the new projection pursuit based method is not only capable of providing the location of the change point as accurately as the existing methods, but it also gives more information about how the change happens.

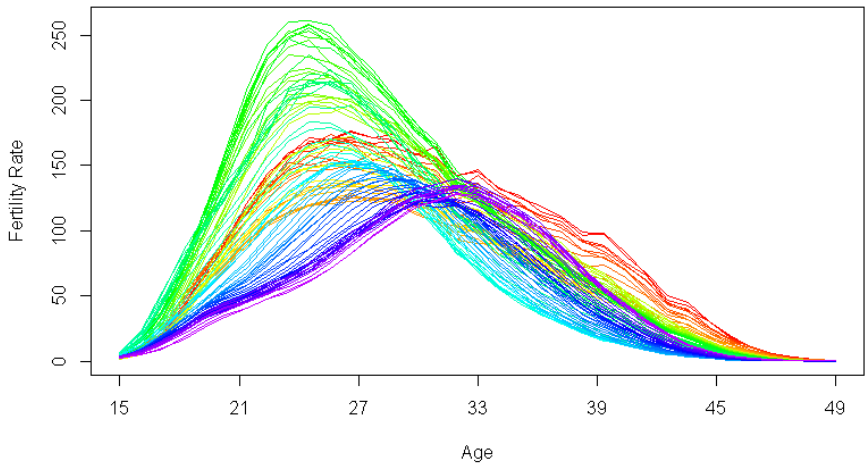

Figure 5.1: Australian fertility rate from 1921 to 2006.

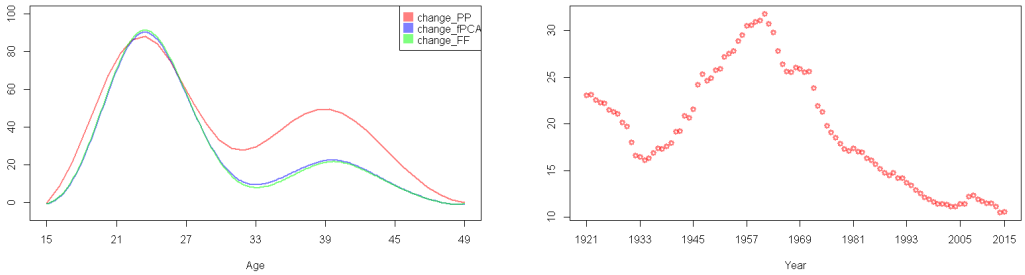

Figure 5.2: Estimated change functions from **change_FF** and **change_fPCA** and the scaled estimated direction from **change_PP** (left), and the projection scores on estimated change direction (right) for Australian Fertility Data.

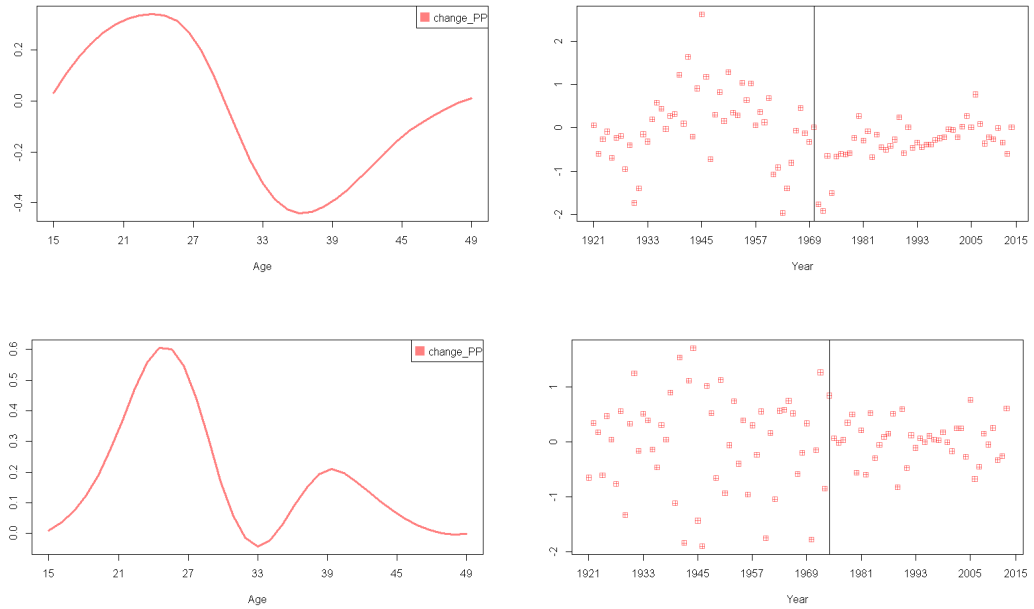

Figure 5.3: The direction on which the change is most significant (left) and the corresponding projection scores on estimated change direction (right) for Australian Fertility Data after taking 1^{st} order difference (top row) and 2^{nd} order difference (bottom row).

5.4.2 Daily low temperature profile in Gayndah

Another example we present here is the temperate data recorded in Gayndah, a small town in Northern Australia. The daily low temperature has been recorded from 1893 to 2009, however, the incomplete curves in 1893 and 2009 have been removed from this analysis. The full dataset is available in the R package `fChange` (Sonmez et al., 2018). The three methods **change_PP**, **change_FF**, and **change_fPCA** all suggest a change point for the daily low temperature curves, which we present in Figure 5.4. The methods **change_PP** and **change_FF** detect a change point around the 59^{th} curve, while **change_fPCA** suggests that the change point exists around the 69^{th} curve. Before we draw a conclusion about the location of the change point, we check the projection scores on the estimated change direction as shown in the right panel of Figure 5.5. A strong linear trend can be observed, which suggests that instead of a step-like change in the mean function, the detected change point could possibly be introduced by the trend.

After removing the linear trend by taking first order difference, we apply the three change point detection methods again. This time both **change_FF** and **change_fPCA** suggest there is no change point. However, our **change_PP** method locates a change point at the 65^{th} curve with p-value close to 0. By checking the corresponding projection scores, which we depict in the right panel of Figure 5.6, we can see a clear difference in the variance on the two sides of the detected change point. Therefore, we can conjecture that along side with a growing trend in the daily low temperature in Gayndah, there is also a change point after the year of 1959. This result not only provides additional evidence for the widely accepted fact of global warming, but it also identifies the impact of human activities on climate in modern age.

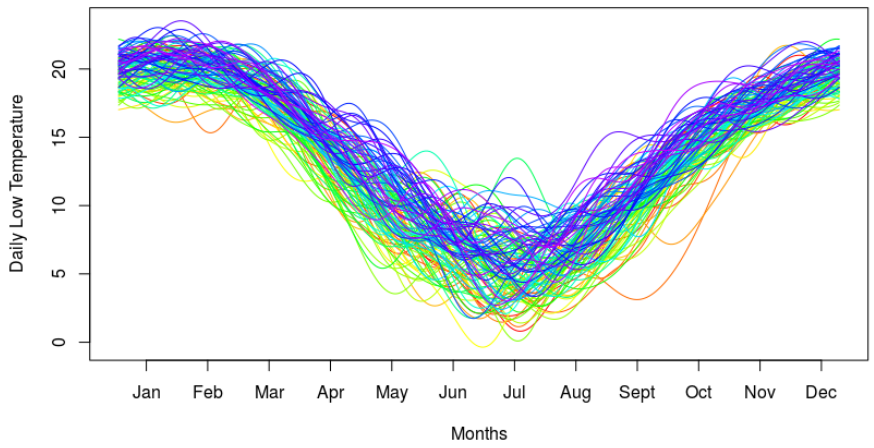

Figure 5.4: Daily low temperature profile in Gayndah, Australia from 1894 to 2008.

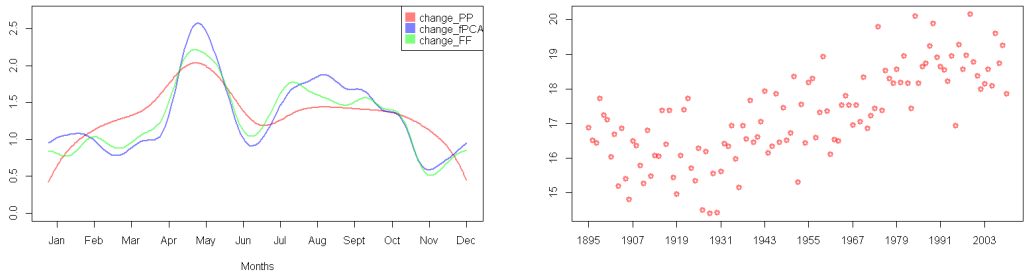

Figure 5.5: Estimated change functions from **change_FF** and **change_fPCA**, together with the scaled estimated change direction from **change_PP** (left), and the projection scores on estimated change direction (right) for Daily Low Temperature Profile in Gayndah.

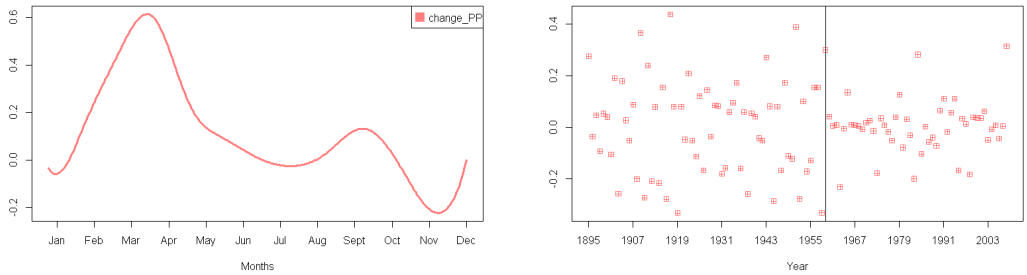

Figure 5.6: The direction on which the change is most significant (left) and the corresponding projection scores on estimated change direction (right) for Daily Low Temperature Profile in Gayndah after removing the linear trend.

Chapter 6

Concluding Remarks and Future Works

In this thesis, we present a versatile dimension reduction framework for functional and high-dimensional data rooted in the idea of projection pursuit. In particular, we propose a computational framework to search for the optimal direction in projection pursuit, and present three applications. In this chapter, we provide conclusion and future research questions for each component of this thesis.

6.1 Functional Projection Pursuit Algorithm

In Chapter 2 we discuss a new dimension reduction framework for functional data that utilizes projection pursuit techniques, and present a new computational tool to solve related high dimensional optimization problems.

While the focus of this thesis is on the functional data analysis, we should notice that the algorithm introduced in this chapter can be applied to the projection pursuit problem in an arbitrary space and for arbitrary problems related to dimension reduction. For example, we plan to implement the developed in this thesis projection pursuit technique to feature selection in regression problems, with the objective of finding a set of best linear combinations of variables that minimizes certain loss functions. We also plan to develop and publish software packages that can automate the projection pursuit steps discussed in this chapter. We plan to invite future researchers to extend the boundary of projection pursuit to more research problems.

6.2 Functional Normality Test

In Chapter 3, we show how projection pursuit can be applied to a normality test for functional data. The proposed normality test shows great potential in cases when the non-Gaussian components are orthogonal to the leading principal components.

Although in this chapter we specifically focus on the normality test, a more general goodness-of-fit test could be formulated with a properly defined projection index replacing the skewness and kurtosis measures. We believe that this idea will lead to interesting research problems.

In Section 3.4.3, we also show how to decompose functional data into Gaussian and non-Gaussian components. Some related research questions in our plan include: (i) Showing that the proposed decomposition also works for more general goodness-of-fit test; (ii) Developing schemes for analyzing the two separated components by applying to them more efficient existing methods (for example, Gaussian and non-Gaussian as discussed in this chapter).

6.3 Forecasting Functional Time Series

In Chapter 4, we discuss how to construct a subspace of the most predictable components of a functional time series. We compare the forecasting results with existing methods for functional ARIMA model and show that the proposed projection pursuit based methods can outperform exiting methods when forecasting non-stationary functional time series.

We should notice that in this chapter we focus on functional ARIMA model because this is a well-studied type of functional time series. However, the projection pursuit based method introduced in this chapter is not restricted to any specific type of functional time series. Therefore, an interesting related research problem would be to develop forecasting methods for other types of functional time series.

6.4 Change-points Detection in Functional Data

In Chapter 5, we introduce and study a general change point detection method for functional data based on empirical characteristic functions of projection scores. Using simulated data, as well as real data sets, we demonstrate that the new method can be applied to

detect various types of change points and can outperform existing methods in different scenarios.

However, in our study we only consider the case when there is a single change point in a sequence of independent data. We plan to continue developing methods that work for multiple change points in the dependent case. We also plan to explore some asymptotic properties of this change point detection method, as outlined in Pötscher and Prucha (1994) and Billingsley (1968).

References

Aliprantis, C. and Border, K. (2006). *Infinite Dimensional Analysis, A Hitchhiker's Guide*. Springer.

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American statistical association*, 49(268):765–769.

Antoniadis, A. and Sapatinas, T. (2003). Wavelet methods for continuous-time prediction using hilbert-valued autoregressive processes. *Journal of Multivariate Analysis*, 87(1):133–158.

Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. *Journal de la Société Française de Statistique*, 160(3):1–106.

Aston, J. A. and Kirch, C. (2012). Detecting and estimating changes in dependent functional data. *Journal of Multivariate Analysis*, 109:204–220.

Aston, J. A., Kirch, C., et al. (2012). Evaluating stationarity via change-point alternatives with applications to fmri data. *The Annals of Applied Statistics*, 6(4):1906–1948.

Aue, A., Gabrys, R., Horváth, L., and Kokoszka, P. (2009). Estimation of a change-point in the mean function of functional data. *Journal of Multivariate Analysis*, 100(10):2254–2269.

Aue, A., Hörmann, S., Horváth, L., and Hušková, M. (2014). Dependent functional linear models with applications to monitoring structural change. *Statistica Sinica*, pages 1043–1073.

Aue, A. and Horváth, L. (2013). Structural breaks in time series. *Journal of Time Series Analysis*, 34(1):1–16.

Aue, A., Norinho, D. D., and Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110(509):378–392.

Aue, A., Rice, G., and Sönmez, O. (2018). Detecting and dating structural breaks in functional data without dimension reduction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):509–529.

Aue, A., Rice, G., and Sönmez, O. (2020). Structural break analysis for spectrum and trace of covariance operators. *Environmetrics*, 31(1):e2617.

Bali, J. L., Boente, G., Tyler, D. E., Wang, J.-L., et al. (2011). Robust functional principal components: A projection-pursuit approach. *The Annals of Statistics*, 39(6):2852–2882.

Baringhaus, L. and Henze, N. (1991). Limit distributions for measures of multivariate skewness and kurtosis based on projections. *Journal of Multivariate Analysis*, 38(1):51–69.

Benjamini, I. and Lee, S. (1997). Conditioned diffusions which are brownian bridges. *Journal of Theoretical Probability*, 10(3):733–736.

Berkes, I., Gabrys, R., Horváth, L., and Kokoszka, P. (2009). Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):927–946.

Berkes, I., Horváth, L., and Kokoszka, P. (2004). Testing for parameter constancy in garch (p, q) models. *Statistics & probability letters*, 70(4):263–273.

Besse, P. and Cardot, H. (1996). Spline approximation of the forecast of a first-order autoregressive functional process. *CANADIAN JOURNAL OF STATISTICS-REVUE CANADIENNE DE STATISTIQUE*, 24(4):467–487.

Besse, P. C., Cardot, H., and Stephenson, D. B. (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, 27(4):673–687.

Billingsley, P. (1968). *Convergence of probability measures*. Wiley.

Bolton, R. J. and Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical science*, pages 235–249.

Bosq, D. (2000). Linear processes in function spaces: theory and applications.

Bosq, D. (2012). *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media.

Brauchart, J. S., Dick, J., and Fang, L. (2015). Spatial low-discrepancy sequences, spherical cone discrepancy, and applications in financial modeling. *Journal of Computational and Applied Mathematics*, 286:28–53.

Brockwell, P. J. and Davis, R. A. (2013). *Time Series: Theory and Methods*. Springer Science & Business Media.

Brody, D. C., Hughston, L. P., and Macrina, A. (2008). Information-based asset pricing. *International Journal of Theoretical and Applied Finance*, 11(01):107–142.

Bueno-Larraz, B. and Klepsch, J. (2017). Variable selection for the prediction of c [0, 1]-valued ar processes using rkhs. *arXiv preprint arXiv:1710.06660*.

Bugni, F. A., Hall, P., Horowitz, J. L., and Neumann, G. R. (2009). Goodness-of-fit tests for functional data. *The Econometrics Journal*, 12(s1):1–18.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, 45(1):11–22.

Cartea, Á., Jaimungal, S., and Kinzebulatov, D. (2016). Algorithmic trading with learning. *International Journal of Theoretical and Applied Finance*, 19(04):1650028.

Charniak, E. and Johnson, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 173–180. Association for Computational Linguistics.

Chen, J. and Gupta, A. K. (2011). *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer Science & Business Media.

Clerc, M. (2010). *Particle swarm optimization*, volume 93. John Wiley & Sons.

Connor, J. T., Martin, R. D., and Atlas, L. E. (1994). Recurrent neural networks and robust time series prediction. *IEEE transactions on neural networks*, 5(2):240–254.

Constantinou, P., Kokoszka, P., and Reimherr, M. (2017). Testing separability of space-time functional processes. *Biometrika*, 104(2):425–437.

Crainiceanu, C. M., Staicu, A.-M., and Di, C.-Z. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104(488):1550–1561.

Croux, C., Filzmoser, P., and Oliveira, M. R. (2007). Algorithms for projection–pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87(2):218–225.

Croux, C. and Ruiz-Gazen, A. (1996). A fast algorithm for robust principal components based on projection pursuit. In *Compstat*, pages 211–216. Springer.

Croux, C. and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1):206–226.

Cuesta-Albertos, J., del Barrio, E., Fraiman, R., and Matrán, C. (2007). The random projection method in goodness of fit for functional data. *Computational Statistics & Data Analysis*, 51(10):4814 – 4831.

Cuesta-Albertos, J. A., Fraiman, R., and Ransford, T. (2006). Random projections and goodness-of-fit tests in infinite-dimensional spaces. *Bulletin of the Brazilian Mathematical Society*, 37(4):477–501.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1 – 23.

Cuevas, A., Febrero, M., and Fraiman, R. (2004). An ANOVA test for functional data. *Computational Statistics and Data Analysis*, 47:111–122.

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

Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223–239.

Dette, H. and Kutta, T. (2019). Detecting structural breaks in eigensystems of functional time series. *arXiv preprint arXiv:1911.07580*.

Didericksen, D., Kokoszka, P., and Zhang, X. (2012). Empirical properties of forecasts with the functional autoregressive model. *Computational statistics*, 27(2):285–298.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144.

Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Ewers, M., Walsh, C., Trojanowski, J. Q., Shaw, L. M., Petersen, R. C., Jack, C. R., Feldman, H. H., Bokde, A. L., Alexander, G. E., Scheltens, P., Vellas, B., Dubois, B., Weiner, M., and Hampel, H. (2012). Prediction of conversion from mild cognitive impairment to alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. *Neurobiology of Aging*, 33(7):1203 – 1214.e2.

Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):1–28.

Fletcher, R. and Reeves, C. M. (1964). Function minimization by conjugate gradients. *The computer journal*, 7(2):149–154.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers*, 100(9):881–890.

Gabrys, R. and Kokoszka, P. (2007). Portmanteau test of independence for functional observations. *Journal of the American Statistical Association*, 102(480):1338–1348.

Gasser, T., Kneip, A., Binding, A., Prader, A., and Molinari, L. (1991). The dynamics of linear growth in distance, velocity and acceleration. *Annals of Human Biology*, 18(3):187–205.

Gasser, T., Ziegler, P., Kneip, A., Prader, A., Molinari, L., and Largo, R. (1993). The dynamics of growth of weight, circumferences and skinfolds in distance, velocity and acceleration. *Annals of Human Biology*, 20(3):239–259.

Geman, S. and Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, pages 401–414.

Georgieva, A. and Jordanov, I. (2009). Global optimization based on novel heuristics, low-discrepancy sequences and genetic algorithms. *European Journal of Operational Research*, 196(2):413–422.

Gomar, J. J., Conejero-Goldberg, C., Davies, P., and Goldberg, T. E. (2014). Extension and refinement of the predictive value of different classes of markers in ADNI: Four-year follow-up data. *Alzheimer's & Dementia*, 10(6):704 – 712.

González, J. P., San Roque, A. M., and Perez, E. A. (2018). Forecasting functional time series with a new hilbertian armax model: Application to electricity price forecasting. *IEEE Transactions on Power Systems*, 33(1):545–556.

Górecki, T., Hörmann, S., Horváth, L., and Kokoszka, P. (2018). Testing normality of functional time series. *Journal of time series analysis*, 39(4):471–487.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, 40(1):33–51.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press.

Gromenko, O., Kokoszka, P., and Reimherr, M. (2017a). Detection of change in the spatiotemporal mean function. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):29–50.

Gromenko, O., Kokoszka, P., Sojka, J., et al. (2017b). Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves. *The Annals of Applied Statistics*, 11(2):898–918.

Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659.

He, G., Müller, H., and Wang, J. (2000). Extending correlation and regression from multivariate to functional data. *Asymptotics in statistics and probability*, pages 301–315.

Henze, N. (2002). Invariant tests for multivariate normality: a critical review. *Statistical Papers*, 43(4):467–506.

Henze, N. and Wagner, T. (1997). A new approach to the bhep tests for multivariate normality. *Journal of Multivariate Analysis*, 62(1):1–23.

Horváth, L. and Kokoszka, P. (2012a). *Inference for functional data with applications*, volume 200. Springer Science & Business Media.

Horváth, L. and Kokoszka, P. (2012b). *Inference for Functional Data with Applications*. Springer.

Horváth, L. and Rice, G. (2014). Extensions of some classical methods in change point analysis. *Test*, 23(2):219–255.

Huber, P. J. (1985). Projection pursuit. *The annals of Statistics*, pages 435–475.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., and Yasmeen, F. (2019a). *forecast: Forecasting functions for time series and linear models*. R package version 8.7.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Hyndman, R. J., Booth, H., Tickle, L., and Maindonald., J. (2019b). *demography: Forecasting Mortality, Fertility, Migration and Population Data*. R package version 1.22.

Hyndman, R. J. and Shang, H. L. (2019). *ftsa: Functional Time Series Analysis*. R package version 5.5.

Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956.

Jarque, C. M. and Bera, A. K. (1980a). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3):255 – 259.

Jarque, C. M. and Bera, A. K. (1980b). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters*, 6(3):255–259.

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, pages 163–172.

Jarušková, D. (1997). Some problems with application of change-point detection methods to environmental data. *Environmetrics: The official journal of the International Environmetrics Society*, 8(5):469–483.

Jiang, S. and Xie, Y. (2020). Variable selection based on a two-stage projection pursuit algorithm. In *11th International Conference on Bioinformatics Models, Methods and Algorithms, BIOINFORMATICS 2020-Part of 13th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2020*, pages 188–193. SciTePress.

Jiang, S., Xie, Y., and Colditz, G. A. (2020). Functional ensemble survival tree: Dynamic prediction of alzheimer's disease progression accommodating multiple time-varying covariates. *bioRxiv*.

Johnson, R. A., Wichern, D. W., et al. (2002). *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ.

Jolliffe, I. (2011). *Principal component analysis*. Springer.

Kargin, V. and Onatski, A. (2008). Curve forecasting by functional autoregression. *Journal of Multivariate Analysis*, 99(10):2508–2526.

Karlin, S. and Taylor, H. E. (1981). *A Second Course in Stochastic Processes*. Academic Press.

Kawahara, Y. and Sugiyama, M. (2012). Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(2):114–127.

Kimura, S. and Matsumura, K. (2007). Improvement of the performances of genetic algorithms by using low-discrepancy sequences. *Transactions of the Society of Instrument and Control Engineers*, E-6(1):16–25.

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2017). A bayesian multivariate functional dynamic linear model. *Journal of the American Statistical Association*, 112(518):733–744.

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2019). Functional autoregression for sparsely sampled data. *Journal of Business & Economic Statistics*, 37(1):97–109.

Kruskal, J. B. (1969). Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new "index of condensation". In *Statistical Computation*, pages 427–440. Elsevier.

Kruskal, J. B. (1972). Linear transformation of multivariate data to reveal clustering. *Multidimensional scaling*, 1:101–115.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, 54(1-3):159–178.

LaFerla, F., Green, K., and Oddo, S. (2007). Intracellular amyloid-β in alzheimer's disease. *Nat Rev Neurosci*, 8:499–509.

Lai, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(4):613–644.

Largo, R., Gasser, T., Prader, A., Stuetzle, W., and Huber, P. (1978). Analysis of the adolescent growth spurt using smoothing spline functions. *Annals of Human Biology*, 5(5):421–434.

Lavielle, M. and Teyssiere, G. (2006). Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, 46(3):287–306.

Leobacher, G. and Pillichshammer, F. (2014). *Introduction to quasi-Monte Carlo integration and applications*. Springer.

Li, G. and Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and monte carlo. *Journal of the American Statistical Association*, 80(391):759–766.

Liang, J., Li, R., Fang, H., and Fang, K.-T. (2000). Testing multinormality based on low-dimensional projection. *Journal of Statistical Planning and Inference*, 86(1):129 – 141.

Lin, K., Sharpnack, J. L., Rinaldo, A., and Tibshirani, R. J. (2017). A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Advances in Neural Information Processing Systems*, pages 6884–6893.

Lung-Yut-Fong, A., Lévy-Leduc, C., and Cappé, O. (2011). Homogeneity and change-point detection tests for multivariate data using rank statistics. *arXiv preprint arXiv:1107.1971*.

Lütkepohl, H. (2013). *Introduction to multiple time series analysis*. Springer Science & Business Media.

Machado, S. (1983). Two statistics for testing for multivariate normality. *Biometrika*, 70(3):713–718.

Malfait, N. and Ramsay, J. O. (2003). The historical functional linear model. *Canadian Journal of Statistics*, 31(2):115–128.

Malkovich, J. F. and Afifi, A. (1973). On tests for multivariate normality. *Journal of the american statistical association*, 68(341):176–179.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. Academic Press.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345.

Mattson, M. (2004). Pathways towards and away from alzheimer's disease. *Nature*, 430:631–639.

Mecklin, C. J. and Mundfrom, D. J. (2004). An appraisal and bibliography of tests for multivariate normality. *International Statistical Review*, 72(1):123–138.

Mercer, J. (1909). Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446.

Monica, T., Rajasekhar, A., Pant, M., and Abraham, A. (2011). Enhancing the local exploration capabilities of artificial bee colony using low discrepancy sobol sequence. In *International Conference on Contemporary Computing*, pages 158–168. Springer.

Muggeo, V. M. and Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, 27(2):161–166.

Müller, H.-G. et al. (2008). Functional modeling of longitudinal data. In *Longitudinal data analysis*, pages 225–253. Chapman and Hall/CRC.

Müller, H.-G., Stadtmüller, U., et al. (2005). Generalized functional linear models. *the Annals of Statistics*, 33(2):774–805.

Müller, H.-G. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association*, 103(484):1534–1544.

Müller, U. K. and Watson, M. W. (2018). Long-run covariability. *Econometrica*, 86(3):775–804.

Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. Number 63 in CBMS-NSF Series in Applied Mathematics. SIAM, Philadelphia.

Ormoneit, D., Black, M. J., Hastie, T., and Kjellström, H. (2005). Representing cyclic human motion using functional analysis. *Image and Vision Computing*, 23(14):1264–1276.

Panaretos, V. M., Kraus, D., and Maddocks, J. H. (2010). Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *Journal of the American Statistical Association*, 105:670–682.

Pant, M., Thangaraj, R., Grosan, C., and Abraham, A. (2008). Improved particle swarm optimization with low-discrepancy sequences. In *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, pages 3011–3018. IEEE.

Pedersoli, M., Vedaldi, A., Gonzalez, J., and Roca, X. (2015). A coarse-to-fine approach for fast deformable object detection. *Pattern Recognition*, 48(5):1844–1853.

Pfaff, B. (2008). Var, svar and svec models: Implementation within R package vars. *Journal of Statistical Software*, 27(4).

Phillips, P. C. (1998). New tools for understanding spurious regressions. *Econometrica*, pages 1299–1325.

Poli, R., Kennedy, J., and Blackwell, T. (2007). Particle swarm optimization. *Swarm intelligence*, 1(1):33–57.

Pötscher, B. M. and Prucha, I. R. (1994). Generic uniform convergence and equicontinuity concepts for random functions: An exploration of the basic structure. *Journal of Econometrics*, 60(1-2):23–63.

Rabin, J. S., Klein, H., Kirn, D. R., Schultz, A. P., Yang, H.-S., Hampton, O., Jiang, S., Buckley, R. F., Viswanathan, A., Hedden, T., et al. (2019). Associations of physical activity and β-amyloid with longitudinal cognition and neurodegeneration in clinically normal older adults. *JAMA neurology*, 76(10):1203–1210.

Ramsay, J. (2005). Functional data analysis. *Encyclopedia of Statistics in Behavioral Science*.

Ramsay, J. and Bock, R. (2002). Functional data analyses for human growth. *McGill University: Unpublished manuscript*.

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

Ramsay, J. O. (2000). Functional components of variation in handwriting. *Journal of the American Statistical Association*, 95(449):9–15.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional data analysis*. Springer.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis*. Springer, New York.

Ramsay, J. O. and Silverman, B. W. (2007). *Applied functional data analysis: methods and case studies*. Springer.

Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of applied meteorology and climatology*, 46(6):900–915.

Riesz, F. and Sz.-Nagy, B. (1990). *Functional Analysis*. Dover.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, pages 220–238.

Scheipl, F., Staicu, A.-M., and Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2):477–501.

Shang, H. L. (2013). Functional time series approach for forecasting very short-term electricity demand. *Journal of Applied Statistics*, 40(1):152–168.

Shang, H. L. (2014). A survey of functional principal component analysis. *AStA Advances in Statistical Analysis*, 98(2):121–142.

Shang, H. L. and Hyndman, R. J. (2016). *rainbow: Rainbow Plots, Bagplots and Boxplots for Functional Data*. R package version 3.4.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Sharipov, O. S. and Wendler, M. (2019). Bootstrapping covariance operators of functional time series. *arXiv preprint arXiv:1904.06721*.

Shepstone, L., Rogers, J., Kirwan, J., and Silverman, B. (1999). The shape of the distal femur: a palaeopathological comparison of eburnated and non-eburnated femora. *Annals of the rheumatic diseases*, 58(2):72–78.

Shumway, R. H. and Stoffer, D. S. (2017). *Time series analysis and its applications: with R examples*. Springer.

Sidhu, G. S., Asgarian, N., Greiner, R., and Brown, M. R. (2012). Kernel principal component analysis for dimensionality reduction in fmri-based diagnosis of adhd. *Frontiers in systems neuroscience*, 6:74.

Smith, A. D., Heron, J., Mishra, G., Gilthorpe, M. S., Ben-Shlomo, Y., and Tilling, K. (2015). Model selection of the effect of binary exposures over the life course. *Epidemiology (Cambridge, Mass.)*, 26(5):719.

Sonmez, O., Aue, A., and Rice, G. (2018). *fChange: Change Point Analysis in Functional Data*. R package version 0.2.0.

Spokoiny, V. et al. (2009). Multiscale local change point detection with applications to value-at-risk. *The Annals of Statistics*, 37(3):1405–1436.

Stadlober, E., Hörmann, S., and Pfeiler, B. (2008). Quality and performance of a pm10 daily forecasting model. *Atmospheric Environment*, 42(6):1098–1109.

Stoehr, C., Aston, J. A., and Kirch, C. (2019). Detecting changes in the covariance structure of functional time series with application to fmri data. *arXiv preprint arXiv:1903.00288*.

Szekely, G. J. and Rizzo, M. L. (2005a). Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of classification*, 22(2):151–183.

Szekely, G. J. and Rizzo, M. L. (2005b). A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794.

Székely, G. J. and Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58 – 80.

Tartakovsky, A. G., Rozovskii, B. L., Blazek, R. B., and Kim, H. (2006). A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Transactions on Signal Processing*, 54(9):3372–3382.

Thadewald, T. and Büning, H. (2007). Jarque–bera test and its competitors for testing normality–a power comparison. *Journal of applied statistics*, 34(1):87–105.

Thorndike, R. L. (1953). Who belongs in the family. In *Psychometrika*. Citeseer.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Tuddenham, R. D. (1954). Physical growth of california boys and girls from birth to eighteen years. *University of California publications in child development*, 1:183–364.

Ullah, S. and Finch, C. F. (2013). Applications of functional data analysis: A systematic review. *BMC medical research methodology*, 13(1):43.

Vidal, R., Ma, Y., and Sastry, S. (2005). Generalized principal component analysis (gpca). *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1945–1959.

Viviani, R., Grön, G., and Spitzer, M. (2005). Functional principal component analysis of fmri data. *Human brain mapping*, 24(2):109–129.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.

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

Zhu, C., Yao, S., Zhang, X., and Shao, X. (2019). Distance-based and rkhs-based dependence metrics in high dimension. *arXiv preprint arXiv:1902.03291*.

Zhu, L.-X., Fang, K.-T., and Bhatti, M. I. (1997). On estimated projection pursuit-type cramer–von mises statistics. *journal of multivariate analysis*, 63(1):1–14.

Zhu, L.-X., Fang, K.-T., and Zhang, J.-T. (1995a). A projection nt-type test for spherical symmetry of a multivariate distribution. *New trends in probability and statistics*, 3:109–122.

Zhu, L.-X., Wong, H. L., and Fang, K.-T. (1995b). A test for multivariate normality based on sample entropy and projection pursuit. *Journal of statistical planning and inference*, 45(3):373–385.

APPENDICES

Appendix A

Appendix for Chapter 1

In this appendix we provide additional details about an application of functional data analysis that is described in Example 2. The presentation is based on Jiang et al. (2020)

A.1 Motivation

In the past few decades, Alzheimer's Disease (AD) has drawn numerous attention and resources from both academia and pharmaceutical industry, as it is one of the most prevalent diseases related to aging, with serious health consequences like memory loss and dementia (Mattson, 2004; LaFerla et al., 2007; Rabin et al., 2019). Past studies have shown that certain biomarkers, especially the expression level of certain genes, could be used for diagnosis of and development of AD (see, for example, Ewers et al. (2012) and Gomar et al. (2014)). The measuring techniques for biomarkers are progressing, and in more recent studies researchers have encountered longitudinal measures for multiple biomarkers. For example, in the Alzheimer's Disease Neuroimagining Initiative (ADNI) dataset, one could find the time-varying measures of a large collection of biomarkers beside baseline covariates. While these longitudinal observations may provide extra information to understand the potential factors that are related to AD, they also pose challenges for statistical analysis.

In Jiang et al. (2020), our goal is to propose an individualized dynamic prediction method for AD patient's survival probability using random forest. Random forest algorithms have many advantages, for instance they can incorporate nonlinear relationships, can conduct prediction and feature selection at the same time, and are robust against noises and outliers. However, traditional random forest algorithms cannot take these time-varying

biomarker measures as inputs. For example, as mentioned above, the measuring times are different across both individuals and biomarkers, and hence cannot be used directly with existing methods, which assume the inputs have the same dimension.

To address this issue, we propose to apply the functional principle component analysis to both reduce the dimension of these longitudinal covariates and characterize them via the principal component scores before feeding to the random survival forest. Specifically, we propose a multivariate functional principal component analysis described below to achieve the goal.

A.2 PACE Algorithm

For a single biomarker, a commonly implemented method to find the functional principal components is the Principal Analysis by Conditional Estimation (PACE) algorithm proposed by Yao et al. (2005a), which we briefly review in this section. We let $\boldsymbol{x}_i = (x_i(t_{i,1}), \ldots, x_i(t_{i,r_i}))'$ be the observed time-varying biomarkers for individual i, $i = 1, \ldots, n$, at time $t_{i,1}, \ldots, t_{i,r_i} \in [0, 1]$. We further assume that the observed trajectory of the i^{th} individual, $x_i(t), t \in [0, 1]$, is recorded with error, that is,

$$x_i(t) = z_i(t) + \epsilon_i(t), \tag{A.2.1}$$

where $z_i(t)$ denotes the de-noised mean value of $x_i(t)$ for $t \in [0, 1]$. The error term is assumed to have $E(\epsilon_i(t)) = 0$ and $Var(\epsilon_i(t)) = \sigma^2$ for $t \in [0, 1]$. Over the observed grid, $t_{i,k}, k = 1, \ldots, r_i$, the mean functions $z_i(t)$ and $\epsilon_i(t)$ are assumed to be mutually independent. Suppose that $z_i, i = 1, \ldots, n$, are realizations of a stochastic process $\{Z(t), t \in [0, 1]\}$ with mean function μ and covariance operator $C(t, s)$. One can estimate the discretized mean function and covariance operator in the following way. Assume we pool all observed time points $t_{i,1}, \ldots, t_{i,r_i}, i = 1, \ldots, n$, together, and obtain the pooled grid $t_1, \ldots, t_R$. Then, the estimated discretized mean vector is

$$\hat{\boldsymbol{\mu}} = (\hat{\mu}(t_1), \ldots, \hat{\mu}(t_R))',$$

where $\hat{\mu}(t)$ is the average value of all observations at time point t. In order to estimate the eigenfunctions and eigenvalues of the covariance operator $C(s, t)$, in Yao et al. (2005a) the authors suggest to first estimate the $R \times R$ empirical covariance matrix $\hat{\Sigma}$ from pooled discrete observations. Then one can remove the effect of the variance introduced by the error terms by first using a local linear smoother along the diagonal of $\hat{\Sigma}$, and then by applying the quadratic smoother to estimate the surface of $C(s, t)$ along the off-diagonal

direction. Then we can obtain the estimated leading eigenfunctions $\hat{\phi}_j$ and corresponding eigenvalues $\hat{\lambda}_j$, $j = 1, \ldots, p$, from $\hat{C}(s,t)$, where $\hat{C}(s,t)$ denotes the smoothed covariance operator.

For the i^{th} individual, let $\hat{\boldsymbol{\mu}}_i = (\hat{\mu}(t_{i,1}), \ldots, \hat{\mu}(t_{i,r_i}))$, $\hat{\phi}_{i,j} = (\hat{\phi}_j(t_{i,1}), \ldots, \hat{\phi}_j(t_{i,r_i}))$, and $\hat{\Sigma}_i$ be an $r_i \times r_i$ matrix composed of the corresponding entries of $\hat{\Sigma}$ where the i^{th} individual has observed values at the corresponding time. Then the functional principal component score for the i^{th} individual on the j^{th} principal component can be estimated as

$$\hat{\xi}_{i,j} = \hat{\lambda}_j \hat{\phi}_{i,j}^T \hat{\Sigma}_i^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_i) \,, \tag{A.2.2}$$

$j = 1, \ldots, p$, where the p is determined by Akaike information criterion (AIC) or the total variance explained (TVE).

A.3 Multivariate Functional Principal Component Analysis

While the PACE algorithm described above can effectively characterize a single biomarker, in order to incorporate the correlations between multiple biomarkers one have to use use the multivariate functional principal component analysis (MFPCA) proposed by Happ and Greven (2018).

Assuming that in total Q biomarkers are studied, then the i^{th} individual have functional observations x_i^q for $q = 1, \ldots, Q$ and $i = 1, \ldots, n$. The first step is to apply the PACE algorithm to the q^{th} biomarker across all individuals for $q = 1, \ldots, Q$, and then select the first M_q principal component terms. The corresponding estimated functional principal component scores for the i^{th} individual are $\hat{\xi}_{i,1}^q, \ldots, \hat{\xi}_{i,M_q}^q$. Let $M = \sum_{q=1}^{Q} M_q$ and $\hat{\Lambda} \in \mathbb{R}^{n \times M}$ be an $n \times M$ matrix for which the ith row is $\{\hat{\xi}_{i,1}^1, \ldots, \hat{\xi}_{i,M_1}^1, \ldots, \hat{\xi}_{i,1}^Q, \ldots, \hat{\xi}_{i,M_Q}^Q\}$.

In the multivariate setting we aim to perform a matrix eigenanalysis such that we can estimate the corresponding eigenvectors $\hat{v}_m$ from the empirical block matrix $\hat{G} = \frac{1}{n-1} \hat{\Lambda}^T \hat{\Lambda} \in \mathbb{R}^{M \times M}$, $m = 1, \ldots, M$. Note that MFPCA indirectly addresses the correlations among multiple biomarkers via correlation among the estimated functional principal component scores by pooling all estimated eigenvalues from the univariate biomarkers in the block matrix $\hat{G}$. Hence the eigenvectors $\hat{v}_m$ contain the information of correlations across different

time-varying biomarkers. As a result, the multivariate eigenfunctions are estimated as

$$\hat{\psi}_m^q(t) = \sum_{l=1}^{M_q} [\hat{v}_m]_l^q \hat{\phi}_k^q(t),\ t \in [0,1], \tag{A.3.1}$$

where $[\hat{v}_m]_l^q$ denotes the l^{th} entry in the qth block of $\hat{v}_m$, for $q = 1, \ldots, Q$, $m = 1, \ldots, M$. The corresponding individual-specific multivariate functional principal component scores can thus be estimated as

$$\hat{\rho}_{i,m} = \sum_{q=1}^{Q} \sum_{l=1}^{M_q} [\hat{v}_m]_l^q \hat{\xi}_{i,l}^q, m = 1, \ldots, M. \tag{A.3.2}$$

Similarly to the univariate setting, the optimal number of multivariate functional principal components d can also be chosen based on TVE or AIC.

Recall that in the original dataset, each individual has observations of Q time-varying biomarkers that are measured on irregular time grids. After applying the MFPCA, information of these biomarkers now is characterized by d MFPC scores that have the same length across different individuals and are ready to be fed to existing statistical analysis or machine learning methods.

Appendix B

Appendix for Chapter 3

B.1 Proof of Theorem 3.2.1 and Theorem 3.2.2

Below we let $\|\cdot\|_{E,k}$ denote the Euclidean norm in $\mathbb{R}^k$, and we let c_i denote unimportant absolute numeric constants.

Proof of Theorem 3.2.1. Let

$$Q_{S,n}(v) = \frac{1}{\sqrt{n}} \frac{1}{\hat{\sigma}^3(v)} \sum_{i=1}^{n} \langle X_i - \bar{X}, v\rangle^3,$$

and

$$Q_{K,n}(v) = \frac{1}{\sqrt{n}} \frac{1}{\hat{\sigma}^4(v)} \sum_{i=1}^{n} [\langle X_i - \bar{X}, v\rangle^4 - 3].$$

With these definitions

$$nS_n^{L_k} = \sup_{v \in U^\infty \cap L_k} Q_{S,n}^2(v), \text{ and } \sqrt{n}K_n^{L_k} = \sup_{v \in U^\infty \cap L_k} |Q_{K,n}(v)|.$$

For $v \in U^\infty \cap L_k$, v can be written as

$$v(t) = \sum_{i=1}^{k} \xi_i(v)\varphi_i(t), \quad \xi_i(v) = \langle \varphi_i, v\rangle, \text{ with } \sum_{i=1}^{k} \xi_i(v)^2 = 1.$$

Let $\boldsymbol{\xi}(v) = (\xi_1(v), ..., \xi_k(v))^\top \in \mathbb{R}^k$, and further let $\mathbf{Y}_i = (\langle X_i, \varphi_i\rangle, ..., \langle X_i, \varphi_k\rangle)^\top \in \mathbb{R}^k$. Under $\mathcal{H}_0$, $\mathbf{Y}_i \sim \mathcal{N}_k(\mu_k, \Sigma_k)$ for some mean vector μ_k and covariance matrix Σ_k. As a result of the second

assumption of the theorem, Σ_k is nonsingular. With this notation in place, we have that

$$Q_{S,n}(v) = Q_{S,n}(v; Y_1, ..., Y_n) = \frac{1}{\sqrt{n}} \frac{1}{[\boldsymbol{\xi}^\top(v) S_{n,Y} \boldsymbol{\xi}(v)]^{3/2}} \sum_{i=1}^{n} [\boldsymbol{\xi}^\top(v)(\mathbf{Y}_i - \bar{\mathbf{Y}})]^3,$$

where $\bar{\mathbf{Y}} = (1/n)\sum_{i=1}^n \mathbf{Y}_i$, and $S_{n,Y} = (1/n)\sum_{i=1}^n (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^\top$. Similarly

$$Q_{K,n}(v) = \frac{1}{n} \frac{1}{[\boldsymbol{\xi}^\top(v) S_{n,Y} \boldsymbol{\xi}(v)]^2} \sum_{i=1}^{n} \{[\boldsymbol{\xi}^\top(v)(\mathbf{Y}_i - \bar{\mathbf{Y}})]^4 - 3\}.$$

If $A \in \mathbb{R}^{k \times k}$ is a nonsingular matrix, and $b \in \mathbb{R}^k$, then

$$\begin{aligned} Q_{S,n}(v; AY_1 + b, ..., AY_n + b) &= \frac{1}{\sqrt{n}} \frac{1}{(\boldsymbol{\xi}(v)^\top A S_{n,Y} A^\top \boldsymbol{\xi}(v))^{3/2}} \sum_{i=1}^{n} [\boldsymbol{\xi}^\top(v) A(\mathbf{Y}_i - \bar{\mathbf{Y}})]^3, \\ &= \frac{1}{\sqrt{n}} \left(\frac{\boldsymbol{\xi}^\top(v) A}{\|A\boldsymbol{\xi}(v)\|_{E,k}} S_{n,Y} \frac{A^\top \boldsymbol{\xi}(v)}{\|A\boldsymbol{\xi}(v)\|_{E,k}} \right)^{-3/2} \sum_{i=1}^{n} \left[\frac{\boldsymbol{\xi}^\top(v) A}{\|A\boldsymbol{\xi}(v)\|_{E,k}} \left(\mathbf{Y}_i - \bar{\mathbf{Y}}\right) \right]^3. \end{aligned}$$

From this it is clear that

$$\sup_{v \in U^\infty \cap L_k} Q^2_{S,n}(v; \mathbf{Y}_1, ..., \mathbf{Y}_n) = \sup_{v \in U^\infty \cap L_k} Q^2_{S,n}(v; A\mathbf{Y}_1 + b, ..., A\mathbf{Y}_n + b),$$

and hence the distribution of $nS_n^{L_k}$ is invariant with respect to nonsingular affine transformations of $\mathbf{Y}_1, ..., \mathbf{Y}_n$. The same holds for $\sqrt{n}K_n^{L_k}$, and so we can assume without loss of generality that $Y_i \sim \mathcal{N}_k(0, I_{k\times k})$, where $I_{k\times k}$ is the identity matrix. The proof from here proceeds along similar lines as Baringhaus and Henze (1991). Let

$$Q^*_{S,n}(v) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\boldsymbol{\xi}^\top(v)\mathbf{Y}_i]^3 - 3\boldsymbol{\xi}^\top(v)\mathbf{Y}_i,$$

and

$$Q^*_{K,n}(v) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\boldsymbol{\xi}^\top(v)\mathbf{Y}_i]^4 - 3[2(\boldsymbol{\xi}^\top(v)\mathbf{Y}_i)^2 - 1].$$

We now aim to show

$$(Q^*_{S,n}(v), Q^*_{K,n}(r))^\top \overset{\mathcal{D}(U^\infty \cap L_k \times U^\infty \cap L_k)}{\longrightarrow} (Z_1(u), Z_2(r))\top, \tag{B.1.1}$$

where Z_1 and Z_2 are independent, mean zero Gaussian processes defines on $U^\infty \cap L_k$ with re-

spective covariance functions

$$\rho_1(v,r) = 6(\boldsymbol{\xi}^\top(v)\boldsymbol{\xi}(r))^3, \tag{B.1.2}$$

$$\rho_2(v,r) = 24(\boldsymbol{\xi}^\top(v)\boldsymbol{\xi}(r))^3,$$

and $\overset{\mathcal{D}(U^\infty\cap L_k\times U^\infty\cap L_k)}{\longrightarrow}$ denotes weak convergence in the product metric space of continuous functions on $U^\infty \cap L_k$ equipped with the supremum norm. In order to show this, first we observe that for each $v \in U^\infty \cap L_k$, $[\boldsymbol{\xi}^\top(v)\mathbf{Y}_i]^3 - 3\boldsymbol{\xi}^\top(v)\mathbf{Y}_i$ and $[\boldsymbol{\xi}^\top(v)\mathbf{Y}_i]^4 - 3[2(\boldsymbol{\xi}^\top(v)\mathbf{Y}_i)^2 - 1]$ have mean zero. It follows from straightforward calculation that for all $v, r \in U^\infty \cap L_k$,

$$\begin{aligned}
E\{[\boldsymbol{\xi}^\top(v)\mathbf{Y}_i]^3 - 3\boldsymbol{\xi}^\top(v)\mathbf{Y}_i\}\{[\boldsymbol{\xi}^\top(r)\mathbf{Y}_i]^3 - 3\boldsymbol{\xi}^\top(r)\mathbf{Y}_i\} &= \rho_1(v,r)\\
E\{[\boldsymbol{\xi}^\top(v)\mathbf{Y}_i]^3 - 3\boldsymbol{\xi}^\top(v)\mathbf{Y}_i\}\{[\boldsymbol{\xi}^\top(r)\mathbf{Y}_i]^4 - 3[2(\boldsymbol{\xi}^\top(r)\mathbf{Y}_i)^2 - 1]\} &= 0\\
E\{[\boldsymbol{\xi}^\top(v)\mathbf{Y}_i]^4 - 3[2(\boldsymbol{\xi}^\top(v)\mathbf{Y}_i)^2 - 1]\}\{[\boldsymbol{\xi}^\top(r)\mathbf{Y}_i]^4 - 3[2(\boldsymbol{\xi}^\top(r)\mathbf{Y}_i)^2 - 1]\} &= \rho_2(v,r).
\end{aligned}$$

We get from this and the central limit theorem that for $v_1, ..., v_m, r_1, ..., r_\ell \in U^\infty \cap L_k$, $a_1, ..., a_m, b_1, ..., b_\ell \in \mathbb{R}$,

$$\begin{aligned}
&\sum_{\nu=1}^m a_\nu Q^*_{S,n}(v_\nu) + \sum_{j=1}^\ell b_j Q^*_{K,n}(r_j)\\
&\quad= \frac{1}{\sqrt{n}}\sum_{i=1}^n \left[\sum_{\nu=1}^m \{a_\nu[\boldsymbol{\xi}^\top(v_\nu)\mathbf{Y}_i]^3 - 3\boldsymbol{\xi}^\top(v_n u)\mathbf{Y}_i\} + \sum_{j=1}^\ell b_j\{[\boldsymbol{\xi}^\top(r_j)\mathbf{Y}_i]^4 - 3[2(\boldsymbol{\xi}^\top(r_j)\mathbf{Y}_i)^2 - 1]\}\right]\\
&\quad\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sum_{\nu_1,\nu_2=1}^m a_{\nu_1}a_{\nu_2}\rho_1(v_{\nu_1}, v_{\nu_2}) + \sum_{j_1,j_2=1}^\ell b_{j_1}b_{j_2}\rho_2(r_{j_1}, r_{j_2})\right), \quad n\to\infty.
\end{aligned}$$

Hence by the Cramér-Wold theorem, the finite dimensional distributions of $(Q^*_{S,n}(v), Q^*_{K,n}(r))$ converge to those of $(Z_1(u), Z_2(r))$. The metric space $(L_k \cap U^\infty, \|\cdot\|)$ is isomorphic to $(S^{k-1}, \|\cdot\|_{E,k})$, where S^{k-1} is the boundary of the unit sphere in $\mathbb{R}^k$, and hence $(L_k \cap U^\infty, \|\cdot\|)$ satisfies the metric entropy condition

$$\int_0^1 \log^{1/2}(N(\epsilon))d\epsilon < \infty,$$

where $N(\epsilon)$, $\epsilon > 0$ is the ϵ-covering number of $U^\infty \cap L_k$. Furthermore, it follows as in the proof of Theorem 1 of Baringhaus and Henze (1991) that

$$(E|Q^*_{S,n}(v) - Q^*_{S,n}(r)|^2)^{1/2} \leq c_1\|v - r\|, \text{ and } (E|Q^*_{K,n}(v) - Q^*_{K,n}(r)|^2)^{1/2} \leq c_2\|v - r\|.$$

Hence (B.1.1) follows from Theorem 7.2.4 of Dudley (1999). Theorem 3.2.1 now follows upon showing that.

$$\sup_{v\in U^\infty\cap L_k} \left|Q_{S,n}(v) - Q^*_{S,n}(v)\right| = o_P(1), \text{ and } \sup_{v\in U^\infty\cap L_k} \left|Q_{K,n}(v) - Q^*_{K,n}(v)\right| = o_P(1), \tag{B.1.3}$$

We provide the details to show $\sup_{v\in U^\infty\cap L_k} \left|Q_{S,n}(v) - Q^*_{S,n}(v)\right| = o_P(1)$, and the approximation for $Q_{K,n}$ follows along similar lines. To begin, we note that

$$Q_{S,n}(v) - Q^*_{S,n}(v) = G_{1,n}(v) + G_{2,n}(v),$$

where

$$G_{1,n}(v) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{[\boldsymbol{\xi}^\top(v)\mathbf{Y}_i]^3 - 3\boldsymbol{\xi}^\top(v)\mathbf{Y}_i - [\boldsymbol{\xi}^\top(v)(\mathbf{Y}_i - \bar{\mathbf{Y}})]^3\right\},$$

and

$$G_{2,n}(v) = \frac{1 - [\boldsymbol{\xi}^\top(v)S_{n,Y}\boldsymbol{\xi}(v)]^{-3/2}}{\sqrt{n}}\sum_{i=1}^n [\boldsymbol{\xi}^\top(v)(\mathbf{Y}_i - \bar{\mathbf{Y}})]^3.$$

By expanding the term $[\boldsymbol{\xi}^\top(v)(\mathbf{Y}_i - \bar{\mathbf{Y}})]^3$, we have that

$$\begin{aligned} G_{1,n}(v) = \frac{3}{n}\left[\sum_{i=1}^n (\boldsymbol{\xi}^\top(v)\mathbf{Y}_i)^2 - 1\right](\sqrt{n}\boldsymbol{\xi}^\top(v)\bar{\mathbf{Y}}) \\ + \sqrt{n}(\boldsymbol{\xi}^\top(v)\bar{\mathbf{Y}})^2\frac{3}{\sqrt{n}}\sum_{i=1}^n(\boldsymbol{\xi}^\top(v)\mathbf{Y}_i) + \sqrt{n}(\boldsymbol{\xi}^\top(v)\bar{\mathbf{Y}})^3 =: G^{(1)}_{1,n}(v) + G^{(2)}_{1,n}(v) + G^{(3)}_{1,n}(v). \end{aligned}$$

By the Cauchy-Schwarz inequality and the multivariate central limit theorem, $\sup_{v\in U^\infty\cap L_k} |\boldsymbol{\xi}^\top\bar{\mathbf{Y}}| \leq \|\bar{\mathbf{Y}}\|_{E,k} = O_P(1/\sqrt{n})$. This readily implies that $\sup_{v\in U^\infty\cap L_k} |G^{(i)}_{1,n}(v)| = o_P(1)$, $i = 2, 3$. Let

$\hat{\kappa}_1, ..., \hat{\kappa}_k$ denote the eigenvalues $S_{n,Y}$ in decreasing order. It is well known that $\max_{1\leq i\leq k} |\hat{\kappa}_i - 1| = o_P(1)$. One has that

$$\hat{\kappa}_k \leq \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\xi}^\top(v)\mathbf{Y}_i)^2 \leq \hat{\kappa}_1.$$

As a result,

$$\sup_{v\in U^\infty\cap L_k}\left|\frac{1}{n}\left[\sum_{i=1}^{n}(\boldsymbol{\xi}^\top(v)\mathbf{Y}_i)^2 - 1\right]\right| \leq \max_{1\leq i\leq k}|\hat{\kappa}_i - 1| = o_P(1),$$

which implies $\sup_{v\in U^\infty\cap L_k}|G^{(1)}_{1,n}(v)| = o_P(1)$. Hence $\sup_{v\in U^\infty\cap L_k}|G_{1,n}(v)| = o_P(1)$. With regard to $G_{2,n}$, it follows as above that

$$\sup_{v\in U^\infty\cap L_k}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\boldsymbol{\xi}^\top(v)(\mathbf{Y}_i - \bar{\mathbf{Y}})]^3\right| = O_P(1),$$

and

$$\sup_{v\in U^\infty\cap L_k}|1 - [\boldsymbol{\xi}^\top(v)S_{n,Y}\boldsymbol{\xi}(v)]^{-3/2}| \leq \max_{1\leq i\leq k}|\hat{\kappa}_i^{-3/2} - 1| = o_P(1),$$

implying that $\sup_{v\in U^\infty\cap L_k}|G_{2,n}(v)| = o_P(1)$, which establishes the first half of (B.1.3). The result for the kurtosis measure can be established using similar arguments. □

Proof of Theorem 3.2.2. Theorem 3.2.2 follows immediately from Theorem 3.2.1 upon showing that

$$\left|\sup_{v\in\hat{P}_k\cap U^\infty} Q^2_{S,n}(v) - \sup_{v\in P_k\cap U^\infty} Q^2_{S,n}(v)\right| = o_P(1), \tag{B.1.4}$$

and

$$\left|\sup_{v\in\hat{P}_k\cap U^\infty}|Q_{K,n}(v)| - \sup_{v\in P_k\cap U^\infty}|Q_{K,n}(v)|\right| = o_P(1). \tag{B.1.5}$$

We show (B.1.4), as (B.1.5) can be shown similarly. Let $\varepsilon > 0$, and define the event

$$Q_n^{\varepsilon} = \left\{ \left| \sup_{v \in \hat{P}_k \cap U^{\infty}} Q_{S,n}^2(v) - \sup_{v \in P_k \cap U^{\infty}} Q_{S,n}^2(v) \right| > \varepsilon \right\}.$$

Under Assumption 3.2.1, we have from Lemma 2.2 of Horváth and Kokoszka (2012b) that

$$\max_{1 \le i \le k} |\lambda_i - \hat{\lambda}_i| \overset{a.s.}{\to} 0 \text{ and } \max_{1 \le i \le k} \|\varphi_i - \hat{s}_i \hat{\varphi}_i\| \overset{a.s.}{\to} 0, \tag{B.1.6}$$

where $\hat{s}_i = sign(\langle \varphi_i, \hat{\varphi}_i \rangle$. Let $0 < \delta < \lambda_k^{3/2}$, and define the event

$$A_{\delta,n} = \left\{ \inf_{v \in (P_k \cup \hat{P}_k) \cap U^{\infty}} \hat{\sigma}^3(v) > \delta \right\}.$$

By Mercer's theorem

$$\hat{C}(t,s) = \sum_{j=1}^{\infty} \hat{\lambda}_j \hat{\varphi}_j(t) \hat{\varphi}_j(s),$$

from which it follows that

$$\hat{\sigma}^2(v) = \iint \hat{C}(t,s) v(t) v(s) dt ds \ge \hat{\lambda}_k \sum_{i=1}^{k} \langle \hat{\varphi}_i, v \rangle^2.$$

Evidently then for $v \in \hat{P}_k \cap U^{\infty}$, $\hat{\sigma}^2(v) \ge \hat{\lambda}_k$. For $v \in P_k \cap U^{\infty}$, v can be written as $v(t) = \sum_{j=1}^{k} v_j \varphi_j(t)$, where $\sum_{i=1}^{k} v_j^2 = 1$. It follows that for $v \in P_k \cap U^{\infty}$,

$$\hat{\sigma}^2(v) \ge \hat{\lambda}_k \sum_{i=1}^{k} \left(\sum_{j=1}^{k} v_j \hat{\theta}_{i,j} \right)^2, \tag{B.1.7}$$

where $\hat{\theta}_{i,j} = \langle \hat{\varphi}_i, \varphi_j \rangle$. Using (B.1.6), it follows that $\hat{\theta}_{i,i}^2 \overset{a.s.}{\to} 1$ for $i = 1, ..., k$, and $\hat{\theta}_{i,j} \overset{a.s.}{\to} 0$ for $1 \le i \ne j \le k$, giving that the right hand side of (B.1.7) converges almost surely to λ_k for all $v \in P_k \cap U^{\infty}$. Thus $P(A_{\delta,n}) \to 1$ as $n \to \infty$. We then write

$$P(Q_n^{(\varepsilon)}) = P(A_{n,\delta} \cap Q_n^{(\varepsilon)}) + P(A_{n,\delta}^c \cap Q_n^{(\varepsilon)}) \le P(A_{n,\delta} \cap Q_n^{(\varepsilon)}) + P(A_{n,\delta}^c).$$

On the set $A_{n,\delta}$, the random function $Q_{S,n}(v)$ is evidently continuous for all $v \in (P_k \cup \hat{P}_k) \cap U^{\infty}$,

being a bounded rational function of continuous functions. Since the sets $\hat{P}_k \cap U^\infty$ and $P_k \cap U^\infty$ are compact, there exist points $v'_n \in \hat{P}_k \cap U^\infty$ and $\hat{v}'_n \in P_k \cap U^\infty$ so that

$$\sup_{v \in \hat{P}_k \cap U^\infty} Q^2_{S,n}(v) = Q^2_{S,n}(\hat{v}'_n) \text{ and } \sup_{v \in P_k \cap U^\infty} Q^2_{S,n}(v) = Q^2_{S,n}(v'_n).$$

The existence of well defined (measurable) random elements $\hat{v}'_n$ and v'_n satisfying the above relation is guaranteed by Theorem 18.19 in Aliprantis and Border (2006), which is a form of the Kuratowski-Ryll-Nardzewski measurable selection theorem, since 1) the set correspondence from the underlying sample space $\omega \to \hat{P}_k \cap U^\infty$ is weakly measurable, and (non-empty) compact valued for all ω, and 2) the function $Q^2_{S,n}(v)$ is almost surely continuous. It follows that

$$\begin{aligned} P(A^c_{n,\delta} \cap Q^{(\varepsilon)}_n) \leq P\left(A^c_{n,\delta} \cap \left\{Q^2_{S,n}(v'_n) > Q^2_{S,n}(\hat{v}'_n) + \varepsilon\right\}\right) + P\left(A^c_{n,\delta} \cap \left\{Q^2_{S,n}(\hat{v}'_n) > Q^2_{S,n}(v'_n) + \varepsilon\right\}\right) \\ =: p_{1,n} + p_{2,n}. \end{aligned}$$

We show that $p_{1,n} \to 0$ as $n \to \infty$, and it can be shown similarly that $p_{2,n} \to 0$ as $n \to \infty$, and so we omit the details in this latter case. According to the definitions of v'_n and $\hat{v}'_n$,

$$v'_n(t) = \sum_{i=1}^k r_{i,n}\varphi_i(t), \text{ and } \hat{v}'_n(t) = \sum_{i=1}^k \hat{r}_{i,n}\hat{s}_i\hat{\varphi}_i(t). \tag{B.1.8}$$

Let

$$v''_n(t) = \sum_{i=1}^k \hat{r}_{i,n}\varphi_i(t), \text{ and } \hat{v}''_n(t) = \sum_{i=1}^k r_{i,n}\hat{s}_i\hat{\varphi}_i(t). \tag{B.1.9}$$

It follows from (B.1.6) that $\|\hat{v}'_n - v''_n\| \leq \max_{1\leq i\leq k} \|\varphi_i - \hat{s}_i\hat{\varphi}_i\| = o_P(1)$, and $\|\hat{v}''_n - v'_n\| = o_P(1)$.

$$\begin{aligned} P\big(Q^2_{S,n}(v'_n) > Q^2_{S,n}(\hat{v}'_n) + \varepsilon\big) &= P\left(Q^2_{S,n}(v'_n) - Q^2_{S,n}(\hat{v}''_n) + Q^2_{S,n}(\hat{v}''_n) > Q^2_{S,n}(\hat{v}'_n) + \varepsilon\right) \\ &\leq P\left(|Q^2_{S,n}(v'_n) - Q^2_{S,n}(\hat{v}''_n)| + Q^2_{S,n}(\hat{v}''_n) > Q^2_{S,n}(\hat{v}'_n) + \varepsilon\right) \\ &\leq P\left(|Q^2_{S,n}(v'_n) - Q^2_{S,n}(\hat{v}''_n)| > \varepsilon/2\right) + P\left(Q^2_{S,n}(\hat{v}''_n) > Q^2_{S,n}(\hat{v}'_n) + \varepsilon/2\right). \end{aligned}$$

According to the definitions of v'_n ,$\hat{v}'_n$ and $\hat{v}''_n$, $P\left(Q^2_{S,n}(\hat{v}''_n) > Q^2_{S,n}(\hat{v}'_n) + \varepsilon/2\right) = 0$. Therefore we now aim to show that

$$P\left(\left\{|Q^2_{S,n}(v'_n) - Q^2_{S,n}(\hat{v}''_n)| > \varepsilon\right\} \cap A_{\delta,n}\right) \to 0 \text{ as } n \to \infty.$$

This readily follows if

$$P\left(\left\{|Q_{S,n}(v'_n) - Q_{S,n}(\hat{v}''_n)| > \varepsilon\right\} \cap A_{\delta,n}\right) \to 0 \text{ as } n \to \infty. \tag{B.1.10}$$

One the set $A_{\delta,n}$ we have by the triangle inequality and mean value theorem that

$$\begin{aligned}
|Q_{S,n}(v'_n) - Q_{S,n}(\hat{v}''_n)| &\leq \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \langle X_i - \bar{X}, v'_n\rangle^3\right| \left|\frac{1}{\hat{\sigma}^3(v'_n)} - \frac{1}{\hat{\sigma}^3(\hat{v}''_n)}\right| \\
&\quad + \left|\frac{1}{\hat{\sigma}^3(\hat{v}''_n)}\right| \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \langle X_i - \bar{X}, v'_n\rangle^3 - \frac{1}{\sqrt{n}}\sum_{i=1}^n \langle X_i - \bar{X}, \hat{v}''_n\rangle^3\right| \\
&\leq \frac{3}{2\delta^{5/2}} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \langle X_i - \bar{X}, v'_n\rangle^3\right| \left|\hat{\sigma}^3(v'_n) - \hat{\sigma}^3(\hat{v}''_n)\right| \\
&\quad + \frac{1}{\delta}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \langle X_i - \bar{X}, v'_n\rangle^3 - \frac{1}{\sqrt{n}}\sum_{i=1}^n \langle X_i - \bar{X}, \hat{v}''_n\rangle^3\right| \\
&=: R_{1,n} + R_{2,n}.
\end{aligned}$$

By expanding third power, one has that

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^n \langle X_i - \bar{X}, v'_n\rangle^3 - \frac{1}{\sqrt{n}}\sum_{i=1}^n \langle X_i - \bar{X}, \hat{v}''_n\rangle^3 &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \langle X_i, v'_n\rangle^3 - \frac{1}{\sqrt{n}}\sum_{i=1}^n \langle X_i, \hat{v}''_n\rangle^3 \\
&\quad - \left[\frac{3}{\sqrt{n}}\sum_{i=1}^n \langle X_i, v'_n\rangle^2 \langle \bar{X}, v'_n\rangle - \frac{3}{\sqrt{n}}\sum_{i=1}^n \langle X_i, \hat{v}''_n\rangle^2 \langle \bar{X}, \hat{v}''_n\rangle\right] \\
&\quad + \left[\frac{3}{\sqrt{n}}\sum_{i=1}^n \langle X_i, v'_n\rangle \langle \bar{X}, v'_n\rangle^2 - \frac{3}{\sqrt{n}}\sum_{i=1}^n \langle X_i, \hat{v}''_n\rangle \langle \bar{X}, \hat{v}''_n\rangle^2\right] \\
&\quad - \sqrt{n}\left[\langle \bar{X}, v'_n\rangle^3 - \langle \bar{X}, \hat{v}''_n\rangle^3\right] \\
&=: T_{1,n}(v'_n, \hat{v}''_n) + T_{2,n}(v'_n, \hat{v}''_n) + T_{3,n}(v'_n, \hat{v}''_n) + T_{4,n}(v'_n, \hat{v}''_n).
\end{aligned}$$

We note that under $\mathcal{H}_0$, the random element $X_i \otimes X_i \otimes X_i \in L^2[0,1]^3$ has mean zero and satisfies

that

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \otimes X_i \otimes X_i \right\| = O_P(1).$$

With regards to $T_{1,n}(v_n', \hat{v}_n'')$, we have using (B.1.6), (B.1.8), (B.1.8), and the Cauchy-Schwarz inequality that

$$\begin{aligned}
|T_{1,n}(v_n', \hat{v}_n'')| &= \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j,p,\ell=1}^{k} r_{j,n} r_{p,n} r_{\ell,n} \langle X_i \otimes X_i \otimes X_i, \varphi_j \otimes \varphi_p \otimes \varphi_\ell - \hat{s}_j \hat{\varphi}_j \otimes \hat{s}_p \hat{\varphi}_p \otimes \hat{s}_\ell \hat{\varphi}_\ell \rangle \right| \\
&= \left| \sum_{j,p,\ell=1}^{k} r_{j,n} r_{p,n} r_{\ell,n} \langle \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \otimes X_i \otimes X_i, \varphi_j \otimes \varphi_p \otimes \varphi_\ell - \hat{s}_j \hat{\varphi}_j \otimes \hat{s}_p \hat{\varphi}_p \otimes \hat{s}_\ell \hat{\varphi}_\ell \rangle \right| \\
&\leq c_1 \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \otimes X_i \otimes X_i \right\| \max_{1 \leq i \leq k} \|\varphi_i - \hat{s}_i \hat{\varphi}_i\| = o_P(1).
\end{aligned}$$

One can establish similarly that $|T_{i,n}(v_n', \hat{v}_n'')| = o_P(1)$, $i = 2, 3$, and 4, giving that $R_{2,n} = o_P(1)$. Replacing $\hat{v}_n''$ with zero in the definition of $R_{2,n}$, we see from the above that

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \langle X_i - \bar{X}, v_n' \rangle^3 \right| = O_P(1).$$

Finally, for $v, r \in (\hat{P}_k \cup P_k) \cap U^\infty$, we have

$$|\hat{\sigma}^2(v) - \hat{\sigma}^2(r)| \leq \hat{\lambda}_1 \|v - r\|,$$

and hence $R_{1,n} = o_P(1)$ using (B.1.6). This implies (B.1.10), and thus completes the proof of the first part of the Theorem. The limit result for $M_n^{\hat{P},k}$ follows from Theorem A.2 in the supporting information of Górecki et al. (2018) and the continuous mapping theorem. □

B.2 Selection of Parameters J and M

Here we discuss how the parameters $J = 3 \times 10^4$ and $M = 5$ in Section 3.3 are determined. We experimented with different combinations of J and M, and selected simulation results related to calculating the 95% quantiles for $\hat{S}_n$ and $\hat{K}_n$ for various values of these parameters are shown in Figure B.1 with $n = 450$ and $k = 21$. The two top panels show that when

$M = 5$, the estimated quantiles of $\hat{S}_n$ and $\hat{K}_n$ increase when J increases and flattened when J is greater than 3×10^4. Similarly, when $J = 10^4$ or 3×10^4 the estimated quantiles increases as M increases but plateaued when $M \geq 5$. Therefore we believe $J = 3 \times 10^4$ and $M = 5$ is a reasonable combination for estimating our test statistics.

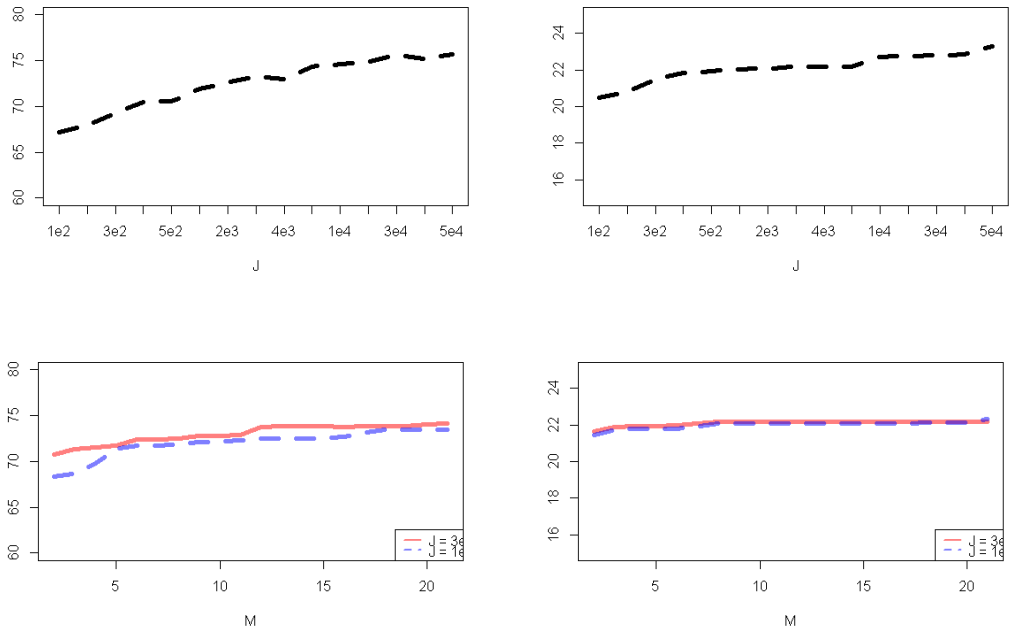

Figure B.1: Estimated 95% quantiles of $\hat{S}_n$ (left panels) and $\hat{K}_n$ (right panels) with $n = 450$ and $k = 21$. The top-left panel shows the estimated quantiles of $\hat{S}_n$ under different J's from 10^2 to 5×10^4 with $M = 5$, the top-right panel presents the estimated quantiles of $\hat{K}_n$ under different J's from 10^2 to 5×10^4 with $M = 5$. The bottom-left panel compares the estimated quantiles of $\hat{S}_n$ under different M from 2 to 21, with the number of iterations equal to $J = 10^4$ (in blue) or 3×10^4 (in red). The bottom-left panel compares the estimated quantiles of $\hat{K}_n$ under different M from 2 to 21, with the number of iterations equal to $J = 10^4$ (in blue) or 3×10^4 (in red).

B.3 Selection of Number of Basis Functions

Here we investigate the influence of the number of basis functions k on the results of the proposed normality test. First, we illustrate the impact of k on the test power through a simulation study. The data is generated with 101 Fourier basis, and the coefficients are generated with the slowly decaying covariance matrix. Then the 10^{th} coefficient is replace with a t-distributed random variable with same standard variance and $df = 5$. The projection pursuit based normality test is conducted with B-spline basis and the number of basis varies from 5 to 35. The size of low discrepancy sequence is fixed to be 3×10^4 From Figure B.2 we can see a clear pattern that the test power increases rapidly first, and then decreases slowly. This seems to be due to the fact that ones needs a certain number of basis functions in order to form the subspace on which the non-Gaussian signal can be captured. However, when the number of basis functions is too large, the high dimensionality begins to pose problem in the optimization step that tends to decrease the power.

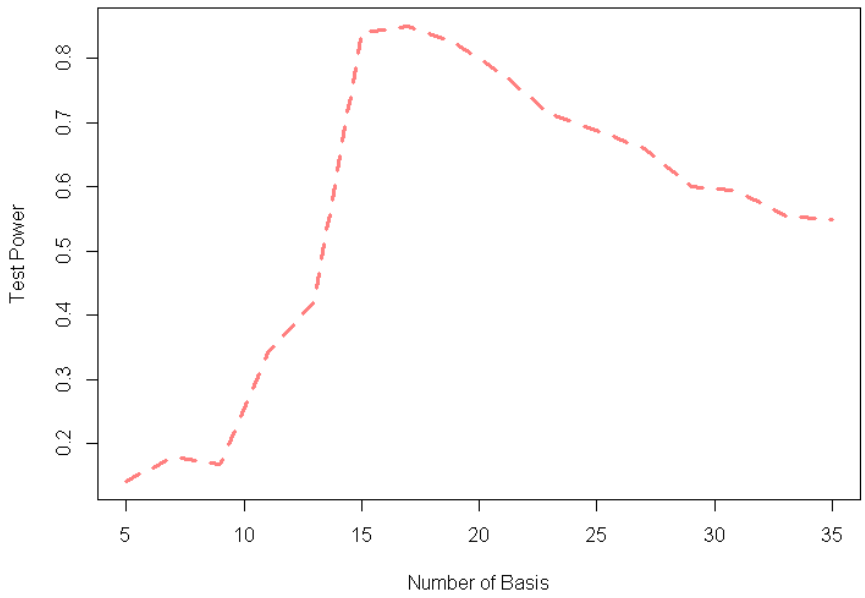

Figure B.2: The test power of proposed normality test under different number B-spline basis functions to construct the subspace with simulated data. The sample size is 900, and the covariance matrix is Σ_{slow} defined in Section 3.3.1.

B.4 Non-smooth Curves: Simulation Study

Here we show that our proposed method also works for data generated using non-smooth basis functions. The data generating process and testing setup are the same as in Section 3.3, except that we generate the functional objects using $K = 31$ Haar basis. Specifically, define

$$h(t) = \begin{cases} 1 & 0 < t \leq 1/2 \\ -1 & 1/2 < t \leq 1 \\ 0 & t \leq 0 \text{ or } 1 < t \end{cases}$$

and the basis functions are constructed as

$$f_k^j(t) = 2^{j/2} h(2^j t - k)$$

for $j = 0, \ldots, 4$ and $k = 0, \ldots, 2^j - 1$ for each j.

Table B.1: Percentage of rejections under the slow decaying covariance matrix Σ_{slow}.

level		$\alpha = 5\%$				$\alpha = 1\%$			
	method	Null	L1	L3	M10	Null	L1	L3	M10
n = 150	PP-F-21	3.1	16.9	33.5	12.5	1.3	10.8	21.9	6.7
	GHHK-F	6.4	58.4	85.5	7.6	2.5	49.7	79.9	3.4
n = 450	PP-F-21	3.8	36.5	72.9	36.9	0.2	29.0	60.2	27.4
	GHHK-F	5.8	95.0	99.9	8.9	2.3	90.7	99.8	4.1
n = 900	PP-F-21	3.8	67.4	96.6	64.3	0.5	53.6	88.8	49.1
	GHHK-F	6.0	99.9	100	10.3	1.7	99.8	100	4.4

Appendix C

Appendix for Chapter 4

C.1 Selection of the Stopping Criterion

We use the PM10 data to compare two stopping criteria: the elbow stopping rule and the SGF stopping rule. For the elbow stopping rule, we stop searching when an inclusion of another direction does not significantly reduce the forecasting error. In Figure C.1, we plot forecasting errors indexed by the number of possible directions found by our functional projection pursuit method discussed in Section 4.2.1. Following the elbow stopping rule, in this case one may want to stop after including the first and second directions.

The second stopping rule, namely the SGF stopping rule, will select the first 6 directions, which would suggest that by following this rule we will explore potential directions more thoroughly.

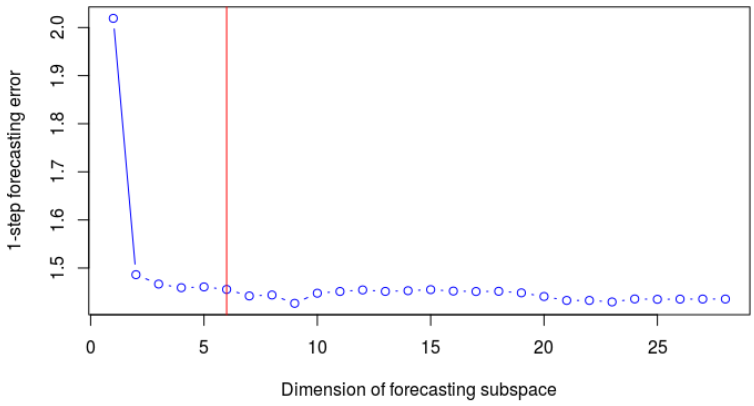

Figure C.1: The 1-step cross-validated forecasting errors for the PM10 data as a function of the number of dimensions included in the proposed projection pursuit method.

C.2 Selection of the Tuning Parameters

Here we present a small experiment with the objective of providing an insight into the problem of selection of the proportion of data r used as the test set in our projection index (4.2.3). In this experiment we generate data following **FAR-CrossSecCov** process, and try different possible values r. The simulation results for forecasting errors are presented in Figure C.2. The plot shows the 1-step and 10-step forecasting errors for different proportions of testing data. We observe a decreasing trend in the forecasting errors, and the curves flatten when $r > 0.05$. These findings would suggest that $r = 0.05$ is a reasonable choice. However, we should also mention that in cases when data include a change point in recent history, the errors will likely increase when r goes beyond certain level, suggesting that in such cases the above choice will no longer be valid.

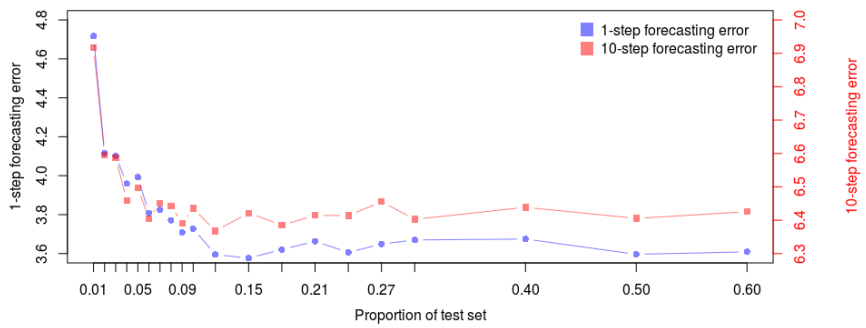

Figure C.2: The 1-step and 10-step forecasting errors for different values of r range from 0.01 to 0.1.