ANALYSIS OF HOSPITAL ADMISSIONS DUE TO RESPIRATORY DISEASES BETWEEN 2010 TO 2017 IN HONG KONG

QINYI ZHANG

Mphil

The Hong Kong Polytechnic University

2020

Initial Submission for Examination Purpose

The Hong Kong Polytechnic University

Department of Applied Mathematics

Analysis of hospital admissions due to respiratory diseases between 2010 to 2017 in Hong Kong

Qinyi Zhang

A thesis submitted in partial fulfilment of the requirements

for the degree of Master of Philosophy

July 2020

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_________________(Signed)

_____Qinyi Zhang_____(Name of student)

Dedicate to my parents.

Abstract

Hospital admissions due to respiratory diseases (HARD) has been widely discussed in the past three decades, and has been linked to air pollutants, media information, and dynamic weather conditions, which can be observed daily and act as smooth trajectories. Classical research works mainly analyzed HARD through cross-sectional studies. It is also interesting and challenging to detect the effects of environmental conditions on HARD from the new perspective of functional data analysis. Motivated by the aforementioned problem, the thesis aims to two targets based on functional data analysis: one is to diagnose the risk of HARD through media information and weather conditions; the other is to explore how air pollutants and weather conditions impact on HARD through a new functional regression model.

Part I focuses on improving diagnosis of high- or low- hospital admissions by combining media information with weather conditions, the multiple functional markers. There is rich literature in combining scalar markers to improve diagnostic accuracy, but they are inapplicable for functional markers. We propose a scalar feature to represent the original functional curve, so that existing scalar combination methods can be applied.

Part II tries to explore a new functional additive regression model to characterize the complicated influence of weather conditions and air pollutants on HARD. I suggest some estimation procedures for the coefficients in the new functional model with hospital admissions as response. Such investigation from functional data anal-

ysis perspective can be also applied to other real worlds data problems that have intense daily records over many years.

Acknowledgements

Studying and researching is one of the most interesting work in this world, I am so lucky to turn on this road. Here, I would like to thank all those who have helped me.

First of all, I would like to extend my most sincere gratitude to my supervisor, Dr. Catherine Liu. It is she who has recognized my potential in research, and has provided me precious opportunities to move forward in scientific research. From the topic selection to the writing of the paper, Dr. Catherine Liu has supervised and encouraged to push my progress. I gradually understand the scientific spirit and how to do research infected by her research attitude and devotion.

Next, I want to show my appreciation to Dr. Haiqiang Ma for his very beneficial discussion and guidance during my research exploration. He never hesitated to provide me warm and effective instructions. His solid training in functional data analysis has guided me into the field of functional data analysis from scratch.

In addition, I'm indebted to my elder academic brotherhood, Dr. Michael Jin Yang and Dr. Ian Sheng Xu, for their suggestions and discussions. My thanks also go to my postgraduate fellows, Ying Sun, Rui Zhou, Jianfeng Luo, and all other fellows for the happy time together.

Furthermore, I'm grateful to the Hong Kong Hospital Authority for providing the data of daily hospital admissions due to respiratory system diseases.

Specially, I would like to show my deep love to my parents for their selfless affection and support. Last but not least, I will say thank you to my fiancée Lydia, for her long lasting trust and love that always gives me peace and power. I have to appreciate the lord for making her such an idiot that ignored all my weaknesses during more than eight years. Lydia makes me much happier than what I ever imagined I could be, and if she let me, I'll spend the rest of my life trying to make her feel in this way.

Contents

Certificate of Originality **v**

Abstract **ix**

Acknowledgements **xi**

List of Figures **xv**

List of Tables **xvii**

1 Introduction **1**

1.1 Introduction 1

1.2 Organization of the thesis 3

2 Combination of multiple functional markers to improve diagnostic accuracy **5**

2.1 Introduction 5

2.2 Methodology 8

2.2.1 Dimension reduction 9

2.2.2 Combination of features 11

2.2.3 Mathematical procedures 15

2.3 Simulations 18

2.4 Application in predicting high- or low- admissions due to respiratory 23

2.5 Conclusion 29

3 Functional semivarying-coefficient additive model 33

3.1 Motivation 33

3.2 Functional regression modeling with hospital admissions response . . 34

3.3 Estimation Procedures 38

3.4 Simulation study 41

4 Conclusion and Discussion 45

Bibliography 47

List of Figures

2.1 Hospital admissions bewteen 2010 to 2017 (left), in 2010 (middle) and in Jan, 2010 (right) 26

2.2 Temperature, dew temperature and searching records during 2010 to 2017 27

2.3 ROC curve of different diagnosis given by different markers 31

2.4 Histogram of hospital admissions on Sunday 31

3.1 Estimated (dashed) and true (solid line) function $g_1(x,u)/x$ (left) and $g_2(x,u)/x$ (right) in u, and their 95% CI (gray) with sample size $n = 300$ in design 1. 44

3.2 Estimated (dashed) and true (solid line) function $g_1(x,u)/x$ (left) and $g_2(x,u)/x$ (right) in u, and their 95% CI (gray) with sample size $n = 300$ in design 2. 44

List of Tables

2.1 Mean 5-fold CV AUC, YI, Time and their SEs (beneath in parentheses) for $p = 5$. . . . 24

2.2 Mean 5-fold CV AUC, YI, Time and their SEs (beneath in parentheses) for $p = 10$. . . . 25

2.3 Mean 5-fold CV AUC, YI, Time and their SEs (beneath in parentheses) for $p = 1$. . . . 26

2.4 Mean 5-fold CV AUC, YI, Time and their SEs (beneath in parentheses) for $p = 2$. . . . 27

2.5 Mean 5-fold CV AUC, YI, Time and their SEs (beneath in parentheses) for $p = 3$. . . . 28

2.6 The AUC derived by each curve and combinations by extreme quantile 0.15 and 0.85. . . . 30

2.7 The YI derived by each curve and combinations by extreme quantile 0.15 and 0.85. . . . 30

2.8 The computing time for each curve and combinations by extreme quantile 0.15 and 0.85. . . . 30

3.1 Biases for $\hat{\theta}_1$ to $\hat{\theta}_5$ and MSE for $\hat{Y}$ in design 1. The results are shown as averages over 100 replicates with standard errors listed in the parentheses. . . . 43

3.2 Biases for $\hat{\theta}_1$ to $\hat{\theta}_5$ and MSE for $\hat{Y}$ in design 2. The results are shown as averages over 100 replicates with standard errors listed in the parentheses. . . . 43

Chapter 1

Introduction

1.1 Introduction

Respiratory disease has been one of the most common reasons for mortality and hospitalization in Hong Kong. High hospital admissions due to respiratory diseases (HARD) will cause heavy burden to the demand for medical services and governmental financial budget (Hernandez et al. (2009)). The HARD has been broadly studied and linked to air pollutants such as sulphur dioxide (SO_2), nitrogen dioxide (NO_2), respirable suspended particulates (RSP) and ozone (O_3), dynamic weather conditions such as temperature, relative humidity and dew point temperature, and media information, by statistical community and environmental epidemiology community.

The thesis is motivated by the data arisen from the HARD associated with air pollutants, dynamic weather conditions and media information between 2010 and 2017 in Hong Kong. The data of hospital admissions was provided by the Hong Kong Hospital Authority, the data of air pollutants and weather conditions were downloaded on the Hong Kong Observatory website, and the data of media information was obtained through google trends. The data were collected daily in Hong Kong from January 1st, 2010, to December 31st, 2017. In statistical community, a data set collected during 1994 to 1997 analog to the HARD data during 2010 to 2017 in Hong Kong has been widely studied and demonstrated by time series data analysis and varying-coefficient modeling. For instance, the early study may be tracked to Fan and Zhang (1999) where they studied the impact of three air pollutants on daily hospital admissions response by a linear varying-coefficient model. As Xia et al. (2002) pointed out and Xia et al. (2004) modeled, they may incorporate up to 42 covariates to characterize the association between hospital admissions response and environmental and weather factors, if one tries to incorporate all possible correlation among variables based on daily observations over several years. Therefore, if the study is based on cross-sectional data, the statistical analysis looks tedious and complicated. Still, one could hardly assure all information were involved.

Motivated by the real climate conditions in Hong Kong, and suggestions from Xia et al. (2002) and Shao et al. (2009), we treat the daily measurements every four weeks as a functional curve observation for every environmental and weather covariate. Such functional data is dense since it includes twenty eight observations on every trajectory. Compared to the pairwise scalar observations, the functional curves are not partitioned and can keep the inherent correlation among observations on a random curve naturally. To the best of our knowledge, functional data analysis has not been applied to study within the HARD data setting. This has led to our research interest in two aspects.

The first target is driven by forecasting high- or low- daily hospital admissions based on environmental factors and other information in the previous four weeks, so that it may help governmental hospitalization management. It is known that weather conditions influence hospital admissions (Pun et al. (2014)), and contemporary media information from social network and internet may also play an important role in forecasting people's behavior (Cook et al. (2011)). Therefore we plan to combine media information with environmental and weather information to forecast the risk of hospital admissions. Statistically, this turns into the problem of improving diagnostic accuracy by combining multiple functional markers. There is rich literature for combining multiple scalar markers to improve diagnostic accuracy (Su and Liu (1993), Liu et al. (2011)), and Xu et al. (2015), among others). Also there is no statistical publication in combining functional markers, although there is a need in application (Zhou et al. (2015)). However, all these scalar-based methods can not be applied to combination of functional markers because of the intrinsic infinite dimensionality of functional data. We aim to apply the tool of functional principal components to do dimension reduction, so as to obtain a scalar feature to represent the original functional curve. Then the existing combination methods for scalar markers can be applied directly.

The second target is driven by exploring the relationship between hospital admissions and environmental and weather factors. In environmental epidemiology community, generalized linear models have been mainly employed to model the effects of weather conditions and air pollutants on HARD (Souza et al. (2014)). In statistical community, researchers presented various regression models by techniques of time series analysis based on pairwise cross-sectional data (Xia and Tong (2006)). These cross-sectional based models were complicated and suffered computational infeasibility and/or difficulty in inference. To the best of our knowledge, there is no publication of applying functional regression to model HARD data in environmental epidemiology. Therefore, from the insight of functional data analysis, we explore to propose a new functional additive regression model with versatile covariates types and accumulated hospital admissions over a time period as the response. We investigate the identifiability of the model, and suggest possible estimation procedures.

1.2 Organization of the thesis

The remaining of the thesis is organized as follows. Chapter 2 introduced the methodology of combining multiple functional markers to improve diagnostic accuracy. It was applied to predict the high- or low- hospital admissions by combination of weather and media information. Chapter 3 introduced our exploration of a new functional additive regression model. We focused on model identifiability and potential estimation procedures. Chapter 4 is the discussion and conclusion.

Chapter 2 is based on the manuscript Ma et al. (2020), and has been accepted.

Chapter 2

Combination of multiple functional markers to improve diagnostic accuracy

2.1 Introduction

Combination of multiple biomarkers is meaningful to improve the diagnostic accuracy, and is attractive for practitioners, clinical therapists and researchers. For a continuous scalar marker that helps diagnosis, its diagnostic performance is usually portrayed by the receiver operating characteristic (ROC) curve. Assuming that the individuals with higher value of diagnostic marker are more prone or with higher risk to be diseased, the subjects can be classified to diseased group with level of biomarker higher than a certain threshold. With a certain threshold, the sensitivity is the probability that a diseased subject is correctly diagnosed, and the specificity is the probability that a non-diseased subject is correctly diagnosed. Then the ROC curve is a plot of its sensitivity versus 1-specificity with thresholds varying on the whole real line. The performance of the diagnostic marker is usually measured by area or partial area under the ROC curve (AUC/pAUC, Bamber (1975)) and Youden index (YI, Youden (1950)), with larger values of the two indices being better. In more detail, AUC can be interpreted by average sensitivity of all values of specificity, and the vice versa. It can also be regarded as the probability that a diseased subject has a greater marker than a non-diseased subject. AUC summarizes the average performance of a diagnostic marker for all values of threshold. Besides, YI is defined as $J = \max_{c \in \mathbb{R}}\{Sensitivity(c) + Specificity(c) - 1\}$ where c is the threshold. YI can be explained as the maximum overall correct classification rate that a marker can attain. It can not only summarize the performance of the ROC curve, but can also be applied to select a certain threshold for classification.

Nowadays, with development of modern techniques, functional markers such as curves or images play an important role in diagnosis. For instance, diffusion tensor imaging can significantly affect the diagnosis for the central nervous system disease (Alexander et al. (2007)), arterial oxygen saturation of hemoglobin can help metabolic syndrome diagnosis (Inácio et al. (2016), Inácio et al. (2012)), and functional magnetic resonance imaging can be applied to diagnose for the Alzheimer's disease (Duc et al. (2020)) and for lympho-associated benign and malignant lesions of the parotid gland (Zhu et al. (2019)). To the best of our knowledge, there are only sporadic statistical works discussing functional markers to make the diagnosis, among which, Inácio et al. (2016) and Inácio et al. (2012) applied one functional marker to make the diagnosis by semiparametric and nonparametric functional regression analysis. Furthermore, there does not exist any work studying combination of functional markers to improve the diagnostic accuracy in the literature until now.

There exists rich works developing combinations for continuous scalar markers to improve the diagnostic accuracy. Su and Liu (1993) proposed the best linear combination when the diagnostic markers all come from Gaussian distribution. Because the normal assumption can be violated, other combination methods were also studied. For instance, Pepe and Thompson (2000) obtained the best empirical linear combination when there are two continuous diagnostic markers. Furthermore, when there are three or more scalar markers, Chen et al. (2015) proposed a empirical likelihood ratio (ELR) method to obtain the best linear combination, and Kang et al. (2016) proposed a stepwise method to successfully obtain the optimal empirical linear combination. Besides, min-max combination proposed by Liu et al. (2011) can provide both high diagnostic accuracy and efficient computation. The methods aforementioned above all focus on maximizing the AUC to derive the optimal combinations, while there are also works regarding YI as the objective function and the main measurement for the ROC curve. Among them, Yin and Tian (2014) followed the idea of stepwise and min-max methods and obtained the best empirical linear combination and min-max combination yielding optimal value of YI. Moreover, abandoning the linear structure, Xu et al. (2015) proposed a flexible non-linear combination method to improve the diagnostic accuracy assessed by YI. Unlike the combination methods based on AUC or YI, logistic regression for the binary disease status can also provide an appropriate combination. All the methods had been well developed and are worthy of reference.

It is impossible to apply scalar combination methods to the multiple functional markers directly because of infinite dimensionality of functional markers. To address this challenge, we want to find a bridge to connect functional markers and existing scalar combination methods. In this article, we propose a scalar feature motivated by square loss distance, as an alternative of the original functional curve in the sense that, it can retain information to the most extent. The square loss distance is defined as the function of projection scores generated from functional principal component Such a dimension-reduction procedure is conducted by commonly used functional principal component analysis (FPCA). Then the existing scalar combination methods can be applied to the scalar feature to improve the diagnostic accuracy. Finally, a mathematical procedure is performed to summarize our methodology.

Our methodology are verified and illustrated by a simulation study and real data analysis. In our numerical study, diagnostic accuracy and computational efficiency of existing scalar combination methods on our proposed features are taken into comparison. Besides, to evaluate the effect of our proposed feature maintaining information, logistic regression on functional observations and on the features are also compared. In real data analysis, we diagnosed high- or low- admissions due to respiratory disease in Hong Kong between 2010 to 2017 by several functional markers, including weather conditions and media information. Moreover, we also provide an R function for convenient application (`https://github.com/Qinyi-Zhang/FunctionalMarkersCombination`). One can choose various methods to make the diagnosis through combinations of functional markers. The numerical results showed that our proposed dimension-reduced features do maintain the information of their corresponding functional markers to most extent.

The rest of this article is organized as follows. In Section 2.2, we propose a dimension-reduced scalar feature for functional markers, summarize some potential useful combination methods, and construct a computational procedure. The simulation studies are conducted to assess the performance (including diagnostic accuracy and computational efficiency) of the scalar feature in Section 2.3. We further analyzed the high- or low- admissions due to respiratory diseases between 2010 and 2017 in Hong Kong by combining weather conditions and media information, which are regarded as functional markers in section 2.4. Section 2.5 contains a summary and discussion.

2.2 Methodology

In this section, we introduce our methodology in detail. For each functional marker, we propose a scalar feature as an alternative, which can maintain information to the most extent, and apply existing scalar combination methods to the scalar features. To be more specific, we apply FPCA method to provide a group of basis functions, and then obtain the projection scores of the functional markers. By truncating the finite sum of Karhunen-Loève (K-L) expansion, the infinite-dimension functional markers can be mapped into finite-dimension projection vectors. Based on the projection vectors, we construct a square-based scalar distance, which is indeed a function of the projection scores and retains information of the functional markers to the most extent. After obtaining the scalar features, we have transformed the functional markers into scalar features, thus, various existing methods can be applied to improve the diagnostic accuracy. In this section, empirical likelihood ratio method (Chen et al. (2015)), stepwise method (Kang et al. (2016)), logistic regression, non-linear combination (Xu et al. (2015)) and min-max combination (Liu et al. (2011)) are of our interest.

2.2.1 Dimension reduction

Suppose there are p functional markers denoted by $M = (M_1, \cdots, M_p)$ on each individual, where $M_k \in L^2(\mathcal{I})$ and $\mathcal{I}$ is a compact time interval. Conditioning on the binary disease status G, the functional markers are denoted by $X = (X_1, \cdots, X_p)$ for a non-diseased subject ($G = 0$) and by $Y = (Y_1, \cdots, Y_p)$ for a diseased subject ($G = 1$). Due to the intrinsic infinite dimension of functional markers, it is impossible to make the diagnosis by them directly. Furthermore, combining the functional observations to make the diagnosis may lose much information, therefore, it seems necessary to reduce their dimension. Specifically, for a functional marker M_k, let $\{\phi_{kj}\}_{j=1}^{\infty}$ denotes an orthogonal basis, the projection score of M_k can be then obtained by

$$\xi_{kj} = \int_{\mathcal{I}} M_k(t)\phi_{kj}(t)dt, \;\; j = 1, 2, \cdots .$$

Likewise, for the functional markers on a diseased or non-diseased subject, respectively, their corresponding projection scores are

$$\zeta_{kj} = \int_{\mathcal{I}} X_k(t)\phi_{kj}(t)dt, \quad \eta_{kj} = \int_{\mathcal{I}} Y_k(t)\phi_{kj}(t)dt, \ \ k = 1, \cdots, p, \ \ j = 1, 2, \cdots . \tag{2.1}$$

with the means of projections $\mu_{kj} = E(\zeta_{kj})$, $\nu_{kj} = E(\eta_{kj})$. Thus, $X_k(t)$ and $Y_k(t)$ can be represented as

$$X_k(t) = \sum_{j=1}^{\infty} \zeta_{kj}\phi_{kj}(t), \quad Y_k(t) = \sum_{j=1}^{\infty} \eta_{kj}\phi_{kj}(t), \ \ k = 1, \cdots, p, \ \ j = 1, 2, \cdots , \tag{2.2}$$

In order to make accurate diagnosis through functional markers $M = (M_1, \cdots, M_p)$, it is of critical importance to remain the utility of every subject in implementing the dimension reduction methods. Therefore, we try to define a scalar feature measuring the difference of distances between the subject and the two populations. To be more specific, for functional marker M_k, based on the square loss, we construct the following 'distance'

$$D_k = \sum_{j=1}^{\infty} \{(\xi_{kj} - \nu_{kj})^2 - (\xi_{kj} - \mu_{kj})^2\}, \tag{2.3}$$

where ν_{kj} is the mean score for non-diseased group, and μ_{kj} is the mean score for diseased group. Since the projection score ξ_{kj} is indeed a random variable, the expectation of D_k can be represented by

$$E(D_k) = \sum_{j=1}^{\infty} \{E(\xi_{kj} - \nu_{kj})^2 - E(\xi_{kj} - \mu_{kj})^2\} = (2G - 1)\sum_{j=1}^{\infty} (\mu_{kj} - \nu_{kj})^2.$$

Thus, if a subject with functional markers $M = (M_1, \cdots, M_p)$ comes from diseased population, the expected value of scalar D_k is larger than 0. Otherwise, the expected value of D_k is less than 0. This indicates that the scalar feature, motivated by square loss distance, can maintain information of the functional marker to the most extent and can act as a scalar marker that help diagnosis. Then the p-dimensional scalar features $D = (D_1, \cdots, D_p)^\top$ can thus be utilized to improve the diagnostic accuracy.

2.2.2 Combination of features

One can always obtain dimension-reduced features D for a subject with functional markers M by the approach proposed in section 2.2.1. Note that D is a p-variate vector, it can be regarded as 'new' scalar markers, and existing scalar combination methods can be applied directly. For a non-diseased subject, its scalar features are denoted by $\mathbf{D}^{[0]} = \{D_1^{[0]}, \cdots, D_p^{[0]}\}$. Likewise, for a diseased subject, its features are denoted by $\mathbf{D}^{[1]} = \{D_1^{[1]}, \cdots, D_p^{[1]}\}$. The scalar features $\mathbf{D}^{[0]}$ and $\mathbf{D}^{[1]}$ are indeed comprehensive functions of projection scores, the distributions of which are unknown without additional assumptions, indicating that the distribution of them are inaccessible. Thus, without additional knowledge on distributions of $\mathbf{D}^{[0]}$ and $\mathbf{D}^{[1]}$, we mainly refer to the linear combinations, non-linear combination and min-max combination.

Linear combination of features is given by

$$l(M) = \lambda_1 D_1 + \cdots + \lambda_p D_p = \boldsymbol{\lambda}^\top \mathbf{D}, \tag{2.4}$$

where $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_p)^\top$, and $\mathbf{D} = (D_1, \cdots, D_p)^\top$. Additionally, min-max combination is

$$m(M) = D_{\max} + \lambda D_{\min} \tag{2.5}$$

where $D_{\max} = \max_{1 \leqslant k \leqslant p} D_k$, $D_{\min} = \min_{1 \leqslant k \leqslant p} D_k$. Linear combination yields the areas under the ROC curves

$$A = Pr(\lambda_1(D_1^{[1]} - D_1^{[0]}) + \cdots + \lambda_p(D_p^{[1]} - D_p^{[0]}) > 0), \tag{2.6}$$

and min-max combination yields the AUC

$$A = Pr((D_{\max}^{[1]} - D_{\max}^{[0]}) + \lambda(D_{\min}^{[1]} - D_{\min}^{[0]}) > 0), \tag{2.7}$$

respectively, where $D_{\max}^{[1]} = \max_{1\leqslant k\leqslant p} D_k^{[1]}$ and $D_{\min}^{[1]} = \min_{1\leqslant k\leqslant p} D_k^{[1]}$. Besides, the non-linear combination yields Youden index

$$J = \max_{c\in\mathbb{R}} Pr(h(\mathbf{D}^{[1]}) > c) + Pr(h(\mathbf{D}^{[0]}) \leqslant c) - 1, \tag{2.8}$$

Equations (2.6) and (2.7) do have their maximizers with some mild assumptions, and their maximizers are both unique under stronger conditions (Vexler et al. (2006)). Moreover, Su and Liu (1993) proposed the optimal value for $\lambda_1, \cdots, \lambda_p$ as follows:

$$(\lambda_1, \cdots, \lambda_p)^\top = (\Sigma^{[1]} + \Sigma^{[0]})^{-1}(\mu^{[1]} - \mu^{[0]}), \tag{2.9}$$

with assumption that $\mathbf{D}^{[0]} \sim N(\mu^{[0]}, \Sigma^{[0]})$ and $\mathbf{D}^{[1]} \sim N(\mu^{[1]}, \Sigma^{[1]})$ both come from normal distribution.

However, the distribution of the features $\mathbf{D}$ is hard to assume, and the optimal linear combination aforementioned is sensitive to assumptions. Thus, the optimal linear combination can be derived by maximizing the Mann-Whitney statistic (Pepe and Thompson (2000)) as follows:

$$\begin{aligned} W(\boldsymbol{\lambda}) &= \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} I\left\{\boldsymbol{\lambda}^\top(\mathbf{D}_i^{[1]} - \mathbf{D}_j^{[0]}) \geqslant 0\right\} \\ &= \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} I\left\{\sum_{k=1}^{p}\lambda_k(D_{i,k}^{[1]} - D_{j,k}^{[0]}) > 0)\right\}, \end{aligned} \tag{2.10}$$

where n_0 and n_1 denote the numbers of non-diseased and diseased subjects respectively, $D_{i,k}^{[1]}$ denotes the corresponding scalar feature of the k-th functional marker for the i-th diseased subject, $D_{j,k}^{[0]}$ denotes the corresponding scalar feature of the

k-th functional marker for the j-th non-diseased subject, $\mathbf{D}_i^{[G]} = (D_{i,1}^{[G]}, \cdots, D_{i,p}^{[G]})$ for $G = 0, 1$. Actually, Mann-Whitney statistic is not concave for $\boldsymbol{\lambda}$, which implies the optimization problem is difficult to solve. Thus, Kang et al. (2016) proposed a stepwise method to obtain the optimal coefficients $\boldsymbol{\lambda}$ step by step. Although the final solutions of the coefficients are not truly optimal, the method is at least computationally accessible.

Besides, the optimal coefficients for linear combination can also be obtained by maximizing the empirical likelihood ratio (Chen et al. (2015)):

$$L(A) = \sup\left\{ \prod_{i=1}^{n_0+n_1} \tilde{p}_i \middle| \sum_{i=1}^{n_0+n_1} \tilde{p}_i = 1, \sum_{i=1}^{n_0+n_1} \tilde{p}_i \tilde{v}_i(\boldsymbol{\lambda}) = A \right\} \tag{2.11}$$

where

$$\tilde{\mathbf{p}} = (\tilde{p}_1, \cdots, \tilde{p}_{n_0+n_1}) = (p_1^{[0]}, \cdots, p_{n_0}^{[0]}, p_1^{[1]}, \cdots, p_{n_1}^{[1]}),$$

$$\tilde{\mathbf{v}}(\boldsymbol{\lambda}) = (\tilde{v}_1(\boldsymbol{\lambda}), \cdots, \tilde{v}_{n_0+n_1}(\boldsymbol{\lambda})) = (v_1^{[0]}(\boldsymbol{\lambda}), \cdots, v_{n_0}^{[0]}(\boldsymbol{\lambda}), v_1^{[1]}(\boldsymbol{\lambda}), \cdots, v_{n_1}^{[1]}(\boldsymbol{\lambda})),$$

$$v_{i_G}^{[G]}(\boldsymbol{\lambda}) = \frac{1}{n_0 + n_1}\{\sum_{j=1}^{n_0+n_1} K_{\tilde{h}}(\boldsymbol{\lambda}^\top \mathbf{D}_{i_G}^{[G]} - \boldsymbol{\lambda}^\top \tilde{\mathbf{D}}_j)\},$$

and

$$\tilde{\mathbf{D}} = (\tilde{\mathbf{D}}_1, \cdots, \tilde{\mathbf{D}}_{n_0+n_1}) = (D_1^{[0]}, \cdots, D_{n_0}^{[0]}, D_1^{[1]}, \cdots, D_{n_1}^{[1]})$$

for $G = 0, 1$, $K_{\tilde{h}}(\cdot)$ is a symmetric kernel function with $K_{\tilde{h}}(x) = \int_{-\infty}^{x/\tilde{h}} k(u)du$ and $\tilde{h} > 0$ is the bandwidth.

When the number of functional markers p is extremely large, which implies the aforementioned methods are computational inefficient, min-max combination can be applied instead. Denote that $D_{i,\max}^{[G]} = \max_{1 \leqslant k \leqslant p} D_{i,k}^{[G]}$, $D_{j,\min}^{[G]} = \min_{1 \leqslant k \leqslant p} D_{j,k}^{[G]}$ for $G = 0, 1$, the optimal min-max combination can also be estimated by maximizing the corresponding Mann-Whitney statistics:

$$W(\lambda) = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I \left\{ (D_{i,\max}^{[1]} - D_{j,\max}^{[0]}) + \lambda (D_{i,\min}^{[1]} - D_{j,\min}^{[0]}) > 0 \right\}, \tag{2.12}$$

The optimal value of λ in (2.12) can be obtained by interpolation method.

When the obtained features are heteroscedastic in the diseased and non-diseased groups, non-linear combination is a powerful tool. Instead of maximizing (2.8), Xu et al. (2015) used the extended ϕ_δ−loss $L_\delta(u) = \min\{\frac{1}{\delta}(\delta - u)_+, 1\}$ (Hedayat et al. (2015)) to overcome the discrete property of indicator function, then it is equivalent to minimize the model-free estimation framework for $(h(\cdot), c)$:

$$\min_{h \in \mathcal{H}_K, c \in \mathbb{R}} \frac{1}{n_0} \sum_{i_0=1}^{n_0} L(c - h(\mathbf{D}_{i_0}^{[0]})) + \frac{1}{n_1} \sum_{i_1=1}^{n_1} L(h(\mathbf{D}_{i_1}^{[1]}) - c) + \lambda_0 \mathcal{J}(h),$$

where $\lambda_0 > 0$ is a tuning parameter, $\mathcal{H}_K$ is a reproducing kernel Hilbert space (RKHS, Wahba (1990)) associated with a pre-specified kernel function $K(\cdot, \cdot)$, and thus $\mathcal{J}(h) = \frac{1}{2}\|h\|_{\mathcal{H}_K}^2$ is the RKHS norm penalizing the complexity of $h(\cdot)$.

The representer theorem (Wahba (1990)) implies that the optimal non-linear combination must be of the form $\hat{h}(\mathbf{M}) = \sum_{i=1}^{n_0+n_1} a_i K(\tilde{\mathbf{M}}_i, \mathbf{M})$, and thus $\|h\|_{\mathcal{H}_K}^2 = \mathbf{a}^\top \mathbf{K} \mathbf{a}$ with $\mathbf{a} = (a_1, \cdots, a_{n_0+n_1})^\top$ and $\mathbf{K} = \{K(\tilde{\mathbf{M}}_i, \tilde{\mathbf{M}}_j)\}_{i,j=1}^{n_0+n_1}$. Then the optimal non-linear combination can be obtained by minimizing

$$\frac{1}{n_0} \sum_{j=1}^{n_0} L\{c - \sum_{i=1}^{n_0+n_1} a_j K(\tilde{\mathbf{D}}_i, \tilde{\mathbf{D}}_j)\} + \frac{1}{n_1} \sum_{j=1}^{n_1} L\{ \sum_{i=1}^{n_0+n_1} a_j K(\tilde{\mathbf{D}}_i, \tilde{\mathbf{D}}_j) - c\} + \frac{\lambda_0}{2} \mathbf{a}^\top \mathbf{K} \mathbf{a}, \tag{2.13}$$

where $\mathbf{a} \in \mathbb{R}^{n_0+n_1}, c \in \mathbb{R}$. This minimization problem can be solved by applying difference convex algorithm (Le Thi Hoai and Tao (1997)).

Selection of combination is impacted by the sample size n_0, n_1 and number of functional markers p. On the one hand, when p is relatively small, the linear combinations can be expected to perform well and achieves rather high diagnosis accuracy with efficient computation; when the dimension p of functional markers is very large, in order to reduce the computation time, min-max combination can be applied to save the computing time. On the other hand, when the sample size n_0 and n_1 are small, computational efficiency of non-linear combination is fair, and thus non-linear combination can be applied to make accurate diagnosis.

2.2.3 Mathematical procedures

Suppose there are n_0 non-diseased subjects and n_1 diseased subjects with functional markers $X_{i_0}(t) = \{X_{i_01}(t), \cdots, X_{i_0p}(t)\}$, for $i_0 = 1, \cdots, n_0$, and $Y_{i_1}(t) = \{Y_{i_11}(t), \cdots, Y_{i_1p}(t)\}$, for $i_1 = 1, \cdots, n_1$. Without loss of generality, the compact time interval can be assumed to be $[0, 1]$. In practice, since the functional observations are only measured and recorded at discrete time points, the entire functional markers cannot be completely observed. Thus, the observed data is of the form

$$W^0_{i_0kl} = X_{i_0k}(T_{i_0kl}) + \epsilon^0_{i_0kl}, \quad W^1_{i_1kl} = Y_{i_1k}(T_{i_1kl}) + \epsilon^1_{i_1kl},$$

for $k = 1, \cdots, p$, $l = 1, \cdots, N$, where N is the number of observed times, T_{i_Gkl} is the lth observation time for functional marker k on subject i_G from group $G = 0$ or 1. $\epsilon^0_{i_0kl}$ and $\epsilon^1_{i_1kl}$ are independent zero-mean measurement errors.

For subjects with different diseased status $G = 0$ and $G = 1$, we shall estimate their mean function $\mu^G_k(t)$, and covariance function $C^G_k(s,t)$ for the kth functional marker, $k = 1, \cdots, p$, by local linear smoothing approach as follows:

$$(\hat{b}^G_{1k}, \hat{b}^G_{2k}) = \mathop{\arg\min}_{(b_{1k}, b_{2k})} \frac{1}{n_G} \sum_{i=1}^{n_G} \frac{1}{N} \sum_{l=1}^{N} \{W^G_{ikl} - b_{1k} - b_{2k}(T_{i_Gkl} - t)\}^2 K\left(\frac{T_{i_Gkl} - t}{h^G_{1k}}\right), \quad (2.14)$$

$$\begin{aligned}(\hat{c}^G_{0k}, \hat{c}^G_{1k}, \hat{c}^G_{2k}) \quad &= \mathop{\arg\min}_{(c_{0k}, c_{1k}, c_{2k})} \tfrac{1}{n_G} \textstyle\sum_{i=1}^{n_G} \tfrac{1}{N(N-1)} \textstyle\sum_{1 \leqslant l_1 \neq l_2 \leqslant N} \{W^G_{ikl_1} W^G_{ikl_2} - c_{0k} \\ &\quad -c_{1k}(T_{i_Gkl_1} - s) - c_{2k}(T_{i_Gkl_2} - t)\}^2 K\left(\tfrac{T_{i_Gkl_1} - s}{h^G_{2k}}\right) K\left(\tfrac{T_{i_Gkl_2} - t}{h^G_{2k}}\right), (2.15)\end{aligned}$$

for $G = 0, 1$, where $K(\cdot)$ is the symmetric kernel function and h_{1k}^G, h_{2k}^G are the bandwidths which can be selected by cross-validation. Then the estimator of the mean function is $\hat{\mu}_k^G(t) = \hat{b}_{1k}^G(t)$, and of covariance function is $\hat{R}_k^G(s,t) = \hat{c}_{0k}^G - \hat{\mu}_k^G(s)\hat{\mu}_k^G(t)$, for $G = 0, 1$. The estimate of joint covariance for the two populations can be then obtained by

$$\hat{R}_k(s,t) = \frac{n_0}{n_0 + n_1}\hat{R}_k^0(s,t) + \frac{n_1}{n_0 + n_1}\hat{R}_k^1(s,t). \tag{2.16}$$

Based on the estimate of joint covariance function $\hat{R}_k(s,t)$, we can derive the corresponding eigenvalues $\{\hat{\lambda}_{kj}\}_{j=1}^{\infty}$ and eigenfunctions $\{\hat{\phi}_{kj}(t)\}_{j=1}^{\infty}$, respectively. Thus, for a new subject with functional markers $M(t) = (M_1(t), \cdots, M_p(t))^T$, its corresponding projection scores can be estimated by $\hat{m}_{kj} = \int_{\mathcal{I}} M_k(t)\hat{\phi}_{kj}(t)$, $k = 1, \cdots, p$ and $j = 1, \cdots, \infty$.

One often observes the functional markers intermittently with potential measurement errors, therefore, we first need to smooth the observations of X_{i_0k} and Y_{i_1k} by the local linear smoothing technique,

$$(\hat{a}_{0i_0k}, \hat{a}_{1i_0k}) = \underset{(a_{0i_0k}, a_{1i_0k})}{\arg\min} \sum_{l=1}^{N} \left\{W_{i_0kl}^0 - a_{0i_0k} - a_{1i_0k}(T_{i_0kl} - t)\right\}^2 K\left(\frac{T_{i_0kl}-t}{\tilde{h}_{i_0k}^0}\right),$$

$$(\hat{d}_{0i_1k}, \hat{d}_{1i_1k}) = \underset{(d_{0i_1k}, d_{1i_1k})}{\arg\min} \sum_{l=1}^{N} \left\{W_{i_1kl}^1 - d_{0i_1k} - d_{1i_1k}(T_{i_1kl} - t)\right\}^2 K\left(\frac{T_{i_1kl}-t}{\tilde{h}_{i_1k}^1}\right),$$

leading to the local linear estimators $\hat{X}_{i_0k}(t) = \hat{a}_{0i_0k}$, $\hat{Y}_{i_1k}(t) = \hat{d}_{0i_1k}$, where $K(\cdot)$ is a symmetric kernel function and $\tilde{h}_{i_0k}^0, \tilde{h}_{i_1k}^1$ are the bandwidths. The smoothed trajectories can be then regarded as a fully observed random processes. Then the projection scores ζ_{i_0kj}, η_{i_1kj} and their means μ_{kj}, ν_{kj} can be estimated by

$$\begin{aligned} \hat{\zeta}_{i_0kj} &= \int_0^1 \hat{X}_{i_0k}(t)\hat{\phi}_{kj}(t)dt, \quad \hat{\mu}_{kj} = \frac{1}{n_0}\sum_{i=1}^{n_0}\hat{\xi}_{ikj}, \\ \hat{\eta}_{i_1kj} &= \int_0^1 \hat{Y}_{i_1k}(t)\hat{\phi}_{kj}(t)dt, \quad \hat{\nu}_{kj} = \frac{1}{n_1}\sum_{i=1}^{n_1}\hat{\eta}_{ikj}, \end{aligned} \tag{2.17}$$

with $i_0 = 1, \cdots, n_0$, $i_1 = 1, \cdots, n_1$, $k = 1, \cdots, p$.

In order to address the difficulty caused by the infinite dimensionality of projection scores in (2.3), we will approximate the distance with number of projection scores truncated at S_k, $1 \leqslant k \leqslant p$. Thus, from (2.3) and (2.17), one can construct the estimate of the marker (2.3) as follows

$$\hat{D}_k = \sum_{j=1}^{S_k}\{(\hat{\xi}_{kj} - \hat{\nu}_{kj})^2 - (\hat{\xi}_{kj} - \hat{\mu}_{kj})^2\}, \quad 1 \leqslant k \leqslant p. \tag{2.18}$$

For the random sample $\{X_{i_0}(t)\} = (X_{i_01}(t), \cdots, X_{i_0p}(t))_{i_0=1}^{n_0}$ and $\{Y_{i_1}(t)\} = (Y_{i_11}(t), \cdots, Y_{i_1p}(t))_{i_1=1}^{n_1}$, their corresponding scalar markers are denoted by $\hat{\mathbf{D}}_{i_0}^{[0]}$ and $\hat{\mathbf{D}}_{i_1}^{[1]}$, which can be obtained by replacing the projection scores $\{\hat{\xi}_{kj}\}_{j=1}^{S_k}$ with the projection scores $\{\hat{\zeta}_{i_0kj_0}\}_{j_0=1}^{S_k^0}$ and $\{\hat{\eta}_{i_1kj_1}\}_{j_1=1}^{S_k^1}$ in (2.18), where S_k^0 and S_k^1 are the truncated numbers for functional markers $X_{i_0k}(t)$ and $Y_{i_1k}(t)$ respectively, where $i_0 = 1, \cdots, n_0$, $i_1 = 1, \cdots, n_1$, $k = 1, \cdots, p$.

Combinations of these scalar features can be applied to make the diagnosis. Because the obtained feature $\hat{D}_{i_Gk}^{[G]}$ is a comprehensive function for the projection scores, its distribution is inaccessible. Thus, we only apply combination methods without additional assumptions to make the diagnosis. The optimal combinations can be obtained by maximizing (2.10), (2.11), (2.12), or by minimizing (2.13) respectively. A very important problem in these optimization problems is the choice of the truncated numbers S_k^0 and S_k^1, $k = 1, \cdots, p$. Three criteria (AIC, BIC, and fraction of variance explained (FVE)) are commonly used in the functional data analysis (Li et al. (2010), Yao et al. (2005)), and the AIC and BIC are defined as in Yao et al. (2005). In this paper, we adopt the FVE method to select the truncated number of projection scores for each functional marker. As for the bandwidths h_{1k}^G, $\tilde{h}_{i_0k}^0$ and $\tilde{h}_{i_1k}^1$ for the mean or the smoothed curve estimators, the generalized cross-validation can be used, while the bandwidths h_{2k}^G for the covariance estimators are chosen by a 10-fold cross-validation to save computing time.

2.3 Simulations

Simulation studies are conducted to investigate the empirical performance of existing combination methods and the merit of the proposed scalar features. Overall, we compare the performance of four combination methods, namely, Kang *et al.*'s stepwise method, Chen *et al.*'s ELR method, Liu *et al.*'s min-max method, and Xu *et al*'s non-linear method. Besides, logistic regression methods for binary disease status on features (denoted by 'logistic') and on functional observations (denoted by 'logistic*') respectively are also compared to evaluate how well the scalar features retain the information.

Since we study diagnostic accuracy by combining multiple functional markers, conventional indices such as bias and mean square error (MSE) for estimators are not preferred to evaluate the proposed methodology. In practice, either AUC or YI is used to assess the diagnostic accuracy depending on the purpose of globally inspecting or focusing on a specific cut-off point for classification respectively. Computing time is utilized to compare the computational efficiency of all the methods aforementioned. Besides, since the dimension reduction procedure is necessary to obtain the scalar features, only computing time of all the scalar combinations is recorded. Besides, since the dimension reduction procedure is necessary to obtain the scalar features, only computing time of all the scalar combinations is recorded. Since logistic* is included to assess the effect of scalar features that maintaining information, its computing time is not of our interest. In this section, calculation of linear regression is based on R package 'glmnet'. Optimization problems maximizing Mann-Whitney statistic are all addressed by interpolation methods. Calculation of ELR linear combination is mainly on the basis of on R package 'emplik', and the non-linear combination is obtained through a combination of some comprehensive algorithms including difference convex algorithm.

Five different settings of the distributions of principal component scores were studied with sample size $n_0 = n_1 = 150$. Scenarios $p = 5$ and 10 are considered to illustrate the performances of the combination methods, and scenarios $p = 1, 2, 3$ are studied to serve for the real data analysis, in which only three functional markers are applied.

In each simulation case, 100 Monte Carlo samples are generated to obtain the mean and standard error (SE) of the $5-$fold cross-validation (CV) AUC and YI. Here we use 5-fold CV AUC (Kang et al. (2016)) and YI to assess the capability of the methodology of diagnosis for new individuals. The truncated number S_k of functional principal component analysis well be selected by FVE$= 0.99$. In addition, in most of time, we can only have observations of functional markers for only a short time interval, therefore, only part of the generated data are included to make the diagnosis.

One first generates

$$V_{ij}^1(t) = \sum_{k=1}^{50} \tilde{\zeta}_{ijk}\phi_{jk}(t), \quad V_{ij}^0(t) = \sum_{k=1}^{50} \tilde{\eta}_{ijk}\phi_{jk}(t),$$

for $j = 1, \cdots, p$, $p = 1, 2, 3, 5$ or 10, and $t \in [0, 1]$, $\phi_{jl}(t)$ are derived from Fourier basis, $\phi_{j,2l-1}(t) = \sqrt{2}\cos\{(2l-1)\pi t\}$ and $\phi_{j,2l}(t) = \sqrt{2}\sin\{(2l-1)\pi t\}$ for $l = 1, \cdots, 25$,

$\tilde{\zeta}_{ijl}$ and $\tilde{\eta}_{ijl}$ are both from zero-mean distributions. The functional markers are then observed

$$\begin{aligned} Y_{ij}(t) &= \mu_j^1(t) + V_{ij}^1(t) + 0.5(V_{i1}^1(t) + V_{i2}^1(t)) + \epsilon_{ijt}^1, \\ X_{ij}(t) &= \mu_j^0(t) + V_{ij}^0(t) + 0.5(V_{i1}^0(t) + V_{i2}^0(t)) + \epsilon_{ijt}^0, \end{aligned} \tag{2.19}$$

where $\mu_j^1(t) = 2t$, $\mu_j^0(t) = 0$, $\epsilon_{ijt}^G \sim N(0,1)$ are independent of other variables for $G = 0, 1$.

Five different cases of the distributions of the principal component scores, $\tilde{\zeta}_{ijk}$ and $\tilde{\eta}_{ijk}$, are generated as the following cases.

Case 1 (Gaussian distribution with the same variance): Let $\tilde{\zeta}_{ijk}$ and $\tilde{\eta}_{ijk}$ independently come from identical normal distribution. In other words, the difference between Y and X comes from the mean curves $\mu_{ij}^1(t)$ and $\mu_{ij}^0(t)$ only. For $i = 1, \cdots, m$, $j = 1, \cdots, p$ and $k = 1, \cdots, 50$, $\tilde{\zeta}_{ijk}^0$ come from normal distribution $N(0,1)$, and $\tilde{\zeta}_{ijk} = \tilde{\zeta}_{ijk}^0/k$. Similarly, for $i = 1, \cdots, n$, $j = 1, \cdots, p$ and $k = 1, \cdots, 50$, $\tilde{\eta}_{ijk}^0$ also come from normal distribution $N(0,1)$, and $\tilde{\eta}_{ijk} = \tilde{\eta}_{ijk}^0/k$.

Case 2 (Gaussian distribution with different variance): In this case, the difference between Y and X not only comes from the mean curve, but also comes from the distribution of $\tilde{\zeta}_{ijk}$ and $\tilde{\eta}_{ijk}$. Let $\tilde{\zeta}_{ijk}$ also come from the identical distribution as in Case 1, but the variance of $\tilde{\eta}_{ijk}$ is different. For $i = 1, \cdots, m$, $j = 1, \cdots, p$ and $k = 1, \cdots, 50$, $\tilde{\eta}_{ijk}^0$ comes from normal distribution $N(0, 1 + u_j)$ with u_j randomly sampled from $\{0.5, 1, 1.5, 2, 2.5, 3\}$, and $\tilde{\eta}_{ijk} = \tilde{\eta}_{ijk}^0/k$.

Case 3 (Gamma-Normal distribution): Let the principal components coming from the distribution of normal and gamma with equal allocation. The principal components of V_{ij}^0 and V_{ij}^1 are obtained by adding scaled centered gamma distribution to the multivariate normal distributions in Case 2. To be more specific, let $\tilde{\zeta}_{ijk}^1$ were generated by adding a centered gamma variate with a shape 0.1, then $\tilde{\zeta}_{ijk} = 0.5(\tilde{\zeta}_{ijk}^0 + \tilde{\zeta}_{ijk}^1)/k$; let $\tilde{\eta}_{ijk}^1$ were generated by a centered gamma variate with shape $0.2v_j$, where v_j is randomly sampled from $\{3, 4, 5, 6\}$, and then $\tilde{\eta}_{ijk} = 0.5(\tilde{\eta}^0_{ijk} + \tilde{\eta}^1_{ijk})/k$.

Case 4 (Student-t distribution with the same variance): Let the distribution of principal components are symmetric and heavy-tailed t-distributions with the same variance. Let the degree of freedom of the t-distributions is 3. For $i = 1, \cdots, m$, $j = 1, \cdots, p$ and $k = 1, \cdots, 50$, let $\tilde{\zeta}^0_{ijk} \sim t(3)$, and $\tilde{\zeta}_{ijk} = \tilde{\zeta}^0_{ijk}/(\sqrt{3}k)$. At the same time, for $i = 1, \cdots, n$, $j = 1, \cdots, p$ and $k = 1, \cdots, 50$, let $\tilde{\eta}^0_{ijk} \sim t(3)$, and $\tilde{\eta}_{ijk} = \tilde{\eta}^0_{ijk}/(\sqrt{3}k)$.

Case 5 (Student-t distribution with the different variance): Let the distribution of principal components are symmetric and heavy-tailed t-distributions with different variance. $\tilde{\zeta}_{ijk}$ in this case is obtained identically as in case 4. Let $\tilde{\eta}^{t_0}_{ijk} \sim t(3)$, and $\tilde{\eta}_{ijk} = \sqrt{1 + u_j}\tilde{\eta}^{t_0}_{ijk}/(\sqrt{3}k)$.

The empirical means and SEs of 5-fold CV AUC, YI and computing time for combination of each settings for $p = 5, 10, 1, 2$ and 3 are shown in Tables 2.1-2.5 respectively. To explain empirical performances of different approaches, the results for scenario $p = 5$ in Table 2.1 are chosen for detailed illustration. Similar conclusions can be obtained from Table 2.2. To illustrate the merit of our proposed scalar features, logistic method and logistic* are compared. The high AUC and YI in all the scenarios indicates that the obtained dimension-reduced features can maintain the information of the functional markers to the most extent. For cases 1 and 4, where the distributions of the principal components are symmetric and have the same variance, linear combinations have better performances. For the other cases, where the distributions of the principal component are asymmetric and/or have different variance, the min-max and non-linear combination slightly performs better. This may due to min-max and non-linear combination have more power to maintain sensitivity and specificity when the variance of estimated projection scores from two groups are different. In other cases, non-linear method is dominated by some other methods, this may be due to the non-linear method mainly focuses on penalized YI rather than AUC or YI themselves.

Table 2.3 contains comprehensive numerical results in our simulation study when there is only one functional marker that helps diagnosis. In this table, results of ELR, min-max and logistic method are set to be missing, while the computing time for all the methods (including logistic* method) except non-linear combination is also omitted. The reasons why we set these cells to be missing are as follows. We add logistic* method into comparison to reflect the merit of our proposed feature, thus, its computing time is not of our interest and is set to be missing in all the tables. There is only one functional marker, indicating that the stepwise, ELR and logistic methods are indeed all the same and need not to be computed because they are all linear combinations, thus, the cells for ELR and logistic methods are omitted, and the computing time of them are also set to be missing. Min-max method requires at least two scalar markers to make the combination, thus, it cannot be applied for only one scalar feature, and the cells for min-max combination are set to be missing.

According to Tables 2.1-2.2, the scalar feature can retain more information than combinations of functional observations, indicating that functional analysis has its merit for diagnoses. By comparing the results in Tables 2.3-2.5, it is shown that combination of multiple functional markers does improve the diagnostic accuracy. At the same time, as p increases, which implies the number of coefficients in linear combinations grows up, the computing time (including stepwise method, ELR method, and penalized logistic regression) also increases. Besides, since there is only one coefficient to optimize, min-max combination always provide high computational efficiency. Computational efficiency of non-linear combination is dominated by all other methods. In most of the cases, computing efficiency of non-linear combination may not be affected by the number of functional markers, while it would be significantly impacted in some other cases. This may due to the number of coefficients for non-linear combination mainly depends on the sample size but not the number of functional markers. Furthermore, ELR method can provide higher diagnostic accuracy than stepwise method when p is large, while the advantage of YI is higher than AUC. This may due to the stepwise method maximizes the Mann-Whitney statistic directly, thus, its AUC may be larger than ELR method when p is small. However, when p is large, the coefficients obtained by stepwise method may be a little far from the optimal coefficients, while the ELR method can always provide true optimizers.

In summary, when the number of functional markers is small, we suggest to use linear or min-max combinations. When the number of functional markers is large, the performance for ELR method may exceed the stepwise method, therefore, ELR method or min-max combination method are both recommended for their high accuracy and efficient computation. When the number of functional markers is extremely large, one can choose logistic regression, min-max combination or non-linear combination depending on their empirical performance. When the principal components are symmetric distributed with equal variance, the linear combinations are recommended, otherwise, the min-max combination is recommended.

2.4 Application in predicting high- or low- admissions due to respiratory

In this section, we shall utilize the stepwise, ELR, min-max and non-linear, logistic, and logistic* methods to predict high- or low- hospital admissions due to respiratory diseases in Hong Kong between 2010 and 2017. Weather conditions, including temperature, dew point temperature do have significant contribution for respiratory diseases (Davis et al. (2016)). Since patients with respiratory diseases and their family are prone to search keyword 'influenza', searching records of this keyword may

Methods	Stepwise	ELR	Min-max	Logistic*	Logistic	Non-linear
			Case 1: Gaussian distribution with the same variance			
AUC	0.8362 (0.0027)	0.8317 (0.0023)	0.8111 (0.0027)	0.8281 (0.0027)	0.8350 (0.0025)	0.8176 (0.0028)
YI	0.5473 (0.0045)	0.5379 (0.0044)	0.5025 (0.0048)	0.5311 (0.0050)	0.5442 (0.0050)	0.5222 (0.0050)
Time	17.778 (1.2137)	5.1037 (0.1114)	4.3243 (0.2948)	-	4.2469 (0.3248)	26.512 (1.3641)
			Case 2: Gaussian distribution with the different variance			
AUC	0.7610 (0.0034)	0.7592 (0.0029)	0.8453 (0.0033)	0.7555 (0.0037)	0.7628(0.0033)	0.8535 (0.0031)
YI	0.4563 (0.0052)	0.4500 (0.0045)	0.5741 (0.0060)	0.4466 (0.0056)	0.4583 (0.0053)	0.5835 (0.0062)
Time	17.030 (0.4936)	4.9647 (0.1224)	4.2311 (0.1610)	-	3.7488 (0.1357)	31.831 (0.5139)
			Case 3: Gamma-Normal distribution			
AUC	0.9473 (0.0013)	0.9458 (0.0013)	0.9300 (0.0015)	0.9399 (0.0015)	0.9468 (0.0013)	0.9393 (0.0016)
YI	0.7749 (0.0039)	0.7736 (0.0035)	0.7317 (0.0040)	0.7560 (0.0039)	0.7753 (0.0038)	0.7623 (0.0044)
Time	17.325 (0.5026)	4.0360 (0.0811)	4.1914 (0.1416)	-	4.1225 (0.1567)	31.533 (0.5464)
			Case 4: Student-t distribution with the same variance			
AUC	0.8563 (0.0024)	0.8616 (0.0028)	0.8161 (0.0029)	0.8450 (0.0024)	0.8545 (0.0023)	0.8587 (0.0022)
YI	0.6039 (0.0044)	0.6121 (0.0047)	0.5453 (0.0054)	0.5833 (0.0043)	0.6025 (0.0040)	0.5895 (0.0046)
Time	16.596 (0.6275)	5.7831 (0.1396)	4.1209 (0.1646)	-	3.7711 (0.1651)	31.665 (0.8114)
			Case 5: Student-t distribution with different variance			
AUC	0.7912 (0.0030)	0.7917 (0.0035)	0.8236 (0.0027)	0.7789 (0.0033)	0.7892 (0.0029)	0.8414 (0.0024)
YI	0.5121 (0.0048)	0.5133 (0.0059)	0.5501 (0.0048)	0.4877 (0.0054)	0.5093 (0.0047)	0.5655 (0.0050)
Time	16.844 (0.8242)	6.5262 (0.1498)	4.4276 (0.2494)	-	4.0764 (0.2621)	30.030 (0.9559)

Table 2.1: Mean 5-fold CV AUC, YI, Time and their SEs (beneath in parentheses) for $p = 5$.

also reflect the varying of hospital admissions for respiratory diseases. In practice, temperature, dew point temperature and searching records of keyword 'influenza' through google in the past 4 weeks are regarded as functional diagnostic markers.

Throughout the data, patients with respiratory diseases in Hong Kong are divided by districts, genders, and ages. Residents from nineteen districts are contained in the hospital admissions dataset, and are classified by age less than 65 or not. Hospital admissions due to respiratory diseases for any genders and ages are collected by each day and district, in which the numbers less than 5 have been truncated. The whole number of hospital admissions due to respiratory diseases are plotted in figure 2.1. According to figure 2.1, obviously there exists an annual periodicity for the behavior of hospital admissions. Hospital admissions are higher in Spring and Autumn, while

Methods	Stepwise	ELR	Min-max	Logistic*	Logistic	Non-linear
	Case 1: Gaussian distribution with the same variance					
AUC	0.8881 (0.0022)	0.9081 (0.0018)	0.8402 (0.0023)	0.8843 (0.0019)	0.9057 (0.0017)	0.8851 (0.0024)
YI	0.6424 (0.0047)	0.6779 (0.0045)	0.5519 (0.0046)	0.6348 (0.0041)	0.6794 (0.0041)	0.6403 (0.0051)
Time	10.839 (0.1364)	16.729 (0.5903)	1.2063 (0.0175)	-	1.4269 (0.0211)	119.68 (5.7710)
	Case 2: Gaussian distribution with the different variance					
AUC	0.8080 (0.0031)	0.8350 (0.0028)	0.9087 (0.0024)	0.8116 (0.0032)	0.8330 (0.0027)	0.9259 (0.0020)
YI	0.5285 (0.0053)	0.5692 (0.0044)	0.6871 (0.0055)	0.5318 (0.0053)	0.5667 (0.0049)	0.7267 (0.0051)
Time	19.469 (0.9475)	20.525 (0.8209)	2.1733 (0.1142)	-	2.4415 (0.1070)	12.807 (0.7327)
	Case 3: Gamma-Normal distribution					
AUC	0.9805 (0.0007)	0.9846 (0.0007)	0.9570 (0.0010)	0.9718 (0.0010)	0.9844 (0.0005)	0.9809 (0.0007)
YI	0.8743 (0.0026)	0.8939 (0.0026)	0.8023 (0.0030)	0.8459 (0.0032)	0.8913 (0.0022)	0.8747 (0.0028)
Time	39.253 (0.8466)	6.7141 (0.1884)	4.5696 (0.1423)	-	6.0583 (0.1972)	28.005 (0.5403)
	Case 4: Student-t distribution with the same variance					
AUC	0.9043 (0.0021)	0.9145 (0.0020)	0.8210 (0.0031)	0.8967 (0.0023)	0.9130 (0.0017)	0.9131 (0.0021)
YI	0.6971 (0.0047)	0.7254 (0.0040)	0.5553 (0.0056)	0.6738 (0.0048)	0.7183 (0.0042)	0.6862 (0.0050)
Time	39.372 (0.9088)	15.636 (0.4788)	4.2307 (0.1427)	-	5.2823 (0.1680)	28.209 (0.6618)
	Case 5: Student-t distribution with different variance					
AUC	0.8390 (0.0029)	0.8537 (0.0030)	0.8635 (0.0024)	0.8382 (0.0032)	0.8558 (0.0028)	0.9022 (0.0022)
YI	0.5929 (0.0056)	0.6167 (0.0051)	0.6227 (0.0046)	0.5859 (0.0058)	0.6216 (0.0053)	0.6774 (0.0049)
Time	29.011 (0.8438)	16.055 (0.5843)	3.2195 (0.1126)	-	3.6093 (0.1161)	22.261 (3.2906)

Table 2.2: Mean 5-fold CV AUC, YI, Time and their SEs (beneath in parentheses) for $p = 10$.

lower in Summer and Winter. Besides the annual periodicity, there also exists weekly periodicity on the hospital admissions.

Temperature and dew point temperature each day are displayed on the website `https://www.hko.gov.hk/contentc.htm`. The searching records has been scaled by each month with the highest day of searching frequency measured as 100. Temperature, dew point temperature and searching records in Hong Kong are utilized to predict high- or low- admissions due to respiratory diseases. The plots of temperature, dew point temperature and searching records are plotted in figure 2.2.

Because of the shorter work hours on Sunday in Hong Kong, we mainly focus on each Sunday to overcome the weekly periodicity for the number of hospital admissions. According to the histogram of hospital admissions on Sundays in figure

Methods	Stepwise	ELR	Min-max	Logistic*	Logistic	Non-linear
	Case 1: Gaussian distribution with the same variance					
AUC	0.7144 (0.0030)	-	-	0.7047 (0.0035)	-	0.6738 (0.0037)
YI	0.3553 (0.0052)	-	-	0.3448 (0.0055)	-	0.3315 (0.0063)
Time	-	-	-	-	-	30.444 (0.5659)
	Case 2: Gaussian distribution with the different variance					
AUC	0.6653 (0.0037)	-	-	0.6523 (0.0044)	-	0.6917 (0.0036)
YI	0.3329 (0.0047)	-	-	0.3126 (0.0059)	-	0.3404 (0.0058)
Time	-	-	-	-	-	31.153 (0.3670)
	Case 3: Gamma-Normal distribution					
AUC	0.8286 (0.0026)	-	-	0.8195 (0.0027)	-	0.7883 (0.0030)
YI	0.5280 (0.0046)	-	-	0.5147 (0.0048)	-	0.5108 (0.0051)
Time	-	-	-	-	-	31.054 (1.0188)
	Case 4: Student-t distribution with the same variance					
AUC	0.7609 (0.0029)	-	-	0.7477 (0.0033)	-	0.7422 (0.0032)
YI	0.4428 (0.0050)	-	-	0.4219 (0.0053)	-	0.4234 (0.0057)
Time	-	-	-	-	-	33.818 (0.5359)
	Case 5: Student-t distribution with different variance					
AUC	0.7045 (0.0035)	-	-	0.6844 (0.0045)	-	0.7171 (0.0032)
YI	0.3885 (0.0052)	-	-	0.3533 (0.0064)	-	0.3842 (0.0059)
Time	-	-	-	-	-	31.623 (0.3566)

Table 2.3: Mean 5-fold CV AUC, YI, Time and their SEs (beneath in parentheses) for $p = 1$.

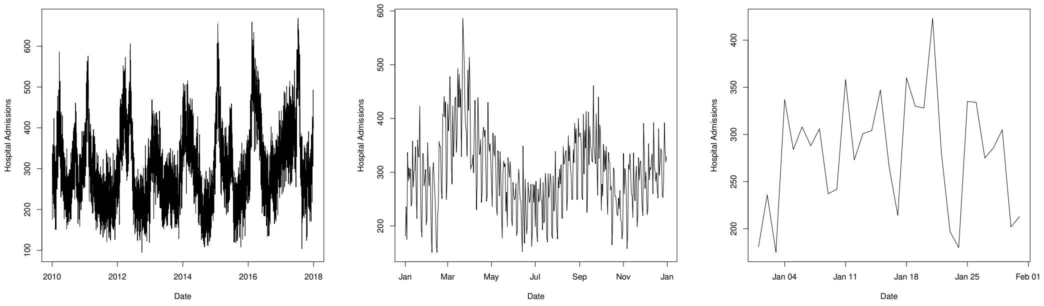

Figure 2.1: Hospital admissions bewteen 2010 to 2017 (left), in 2010 (middle) and in Jan, 2010 (right)

Methods	Stepwise	ELR	Min-max	Logistic*	Logistic	Non-linear
Case 1: Gaussian distribution with the same variance						
AUC	0.7351 (0.0031)	0.7274 (0.0028)	0.7328 (0.0032)	0.7266 (0.0033)	0.7305 (0.0032)	0.7045 (0.0037)
YI	0.3819 (0.0049)	0.3736 (0.0050)	0.3789 (0.0053)	0.3734 (0.0054)	0.3775 (0.0052)	0.3589 (0.0059)
Time	2.3923 (0.1200)	1.7082 (0.0825)	2.4591 (0.1406)	-	1.7655 (0.0985)	17.556 (0.9116)
Case 2: Gaussian distribution with the different variance						
AUC	0.6864 (0.0034)	0.6678 (0.0036)	0.7291 (0.0033)	0.6728 (0.0044)	0.6810 (0.0036)	0.7479 (0.0035)
YI	0.3534 (0.0050)	0.3265 (0.0046)	0.4034 (0.0052)	0.3344 (0.0063)	0.3479 (0.0053)	0.4133 (0.0059)
Time	4.3864 (0.1288)	1.8558 (0.1581)	4.6836 (0.1484)	-	3.4487 (0.1185)	32.345 (0.4624)
Case 3: Gamma-Normal distribution						
AUC	0.8543 (0.0025)	0.8478 (0.0028)	0.8535 (0.0026)	0.8462 (0.0025)	0.8517 (0.0025)	0.8267 (0.0029)
YI	0.5758 (0.0045)	0.6626 (0.0040)	0.5773 (0.0049)	0.5613 (0.0048)	0.5723 (0.0046)	0.5614 (0.0049)
Time	4.5574 (0.2833)	1.5315 (0.0836)	4.8230 (0.3040)	-	4.3062 (0.2998)	30.470 (0.9769)
Case 4: Student-t distribution with the same variance						
AUC	0.7746 (0.0029)	0.7687 (0.0029)	0.7710 (0.0031)	0.7642 (0.0032)	0.7712 (0.0030)	0.7679 (0.0033)
YI	0.4699 (0.0051)	0.5284 (0.0049)	0.4647 (0.0052)	0.4487 (0.0052)	0.4613 (0.0055)	0.4507 (0.0056)
Time	3.9752 (0.1380)	1.7968 (0.0586)	4.1541 (0.1382)	-	3.2716 (0.1293)	33.542 (0.6207)
Case 5: Student-t distribution with different variance						
AUC	0.7156 (0.0034)	0.7088 (0.0038)	0.7507 (0.0029)	0.7004 (0.0041)	0.7094 (0.0038)	0.7545 (0.0031)
YI	0.3960 (0.0049)	0.4509 (0.0056)	0.4395 (0.0050)	0.3712 (0.0060)	0.3893 (0.0054)	0.4282 (0.0052)
Time	4.1538 (0.1236)	2.3717 (0.1072)	4.5683 (0.1454)	-	3.4572 (0.1119)	33.169 (0.5927)

Table 2.4: Mean 5-fold CV AUC, YI, Time and their SEs (beneath in parentheses) for $p = 2$.

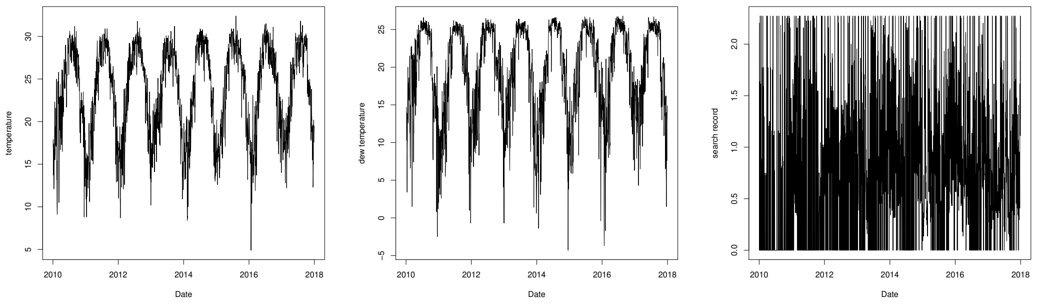

Figure 2.2: Temperature, dew temperature and searching records during 2010 to 2017

Methods	Stepwise	ELR	Min-max	Logistic*	Logistic	Non-linear
Case 1: Gaussian distribution with the same variance						
AUC	0.7725 (0.0027)	0.7750 (0.0029)	0.7661 (0.0025)	0.7654 (0.0029)	0.7715 (0.0027)	0.7489 (0.0037)
YI	0.4385 (0.0047)	0.4503 (0.0050)	0.4281 (0.0043)	0.4270 (0.0053)	0.4375 (0.0048)	0.4205 (0.0052)
Time	8.2934 (0.2626)	2.7989 (0.0893)	4.2788 (0.1665)	-	3.5171 (0.1525)	33.641 (0.5778)
Case 2: Gaussian distribution with the different variance						
AUC	0.7105 (0.0038)	0.7092 (0.0033)	0.7841 (0.0034)	0.7048 (0.0038)	0.7092(0.0037)	0.7903 (0.0040)
YI	0.3820 (0.0053)	0.3855 (0.0048)	0.4787 (0.0057)	0.3719 (0.0054)	0.3785 (0.0052)	0.4757 (0.0069)
Time	10.826 (0.5577)	3.2868 (0.1182)	5.7857 (0.3180)	-	4.6966 (0.2856)	32.563 (1.1229)
Case 3: Gamma-Normal distribution						
AUC	0.8956 (0.0017)	0.8971 (0.0020)	0.8886 (0.0018)	0.8879 (0.0020)	0.8949 (0.0018)	0.8816 (0.0021)
YI	0.6569 (0.0043)	0.6626 (0.0040)	0.6438 (0.0043)	0.6404 (0.0046)	0.6571 (0.0043)	0.6467 (0.0047)
Time	6.2395 (0.2454)	2.4987 (0.0677)	3.0796 (0.1033)	-	2.8551 (0.1381)	25.729 (0.6669)
Case 4: Student-t distribution with the same variance						
AUC	0.8127 (0.0023)	0.8117 (0.0026)	0.7990 (0.0026)	0.8054 (0.0025)	0.8114 (0.0022)	0.8108 (0.0026)
YI	0.5295 (0.0044)	0.5284 (0.0049)	0.5103 (0.0047)	0.5151 (0.0043)	0.5257 (0.0044)	0.5170 (0.0049)
Time	8.6609 (0.2902)	3.3166 (0.1371)	4.4444 (0.1860)	-	3.6472 (0.1627)	34.019 (0.9702)
Case 5: Student-t distribution with different variance						
AUC	0.7480 (0.0031)	0.7497 (0.0037)	0.7878 (0.0027)	0.7377 (0.0035)	0.7455 (0.0032)	0.7965 (0.0030)
YI	0.4493 (0.0051)	0.4509 (0.0056)	0.4971 (0.0048)	0.4275 (0.0055)	0.4451 (0.0052)	0.4881 (0.0059)
Time	8.6076 (0.2970)	3.4655 (0.1506)	4.3373 (0.1363)	-	3.5444 (0.1385)	33.498 (0.7119)

Table 2.5: Mean 5-fold CV AUC, YI, Time and their SEs (beneath in parentheses) for $p = 3$.

2.4, there exists a heavy tail for high hospital admissions. Thus, we mainly consider high- and low- admission days (higher than 85%- and lower than 15% empirical quantiles, respectively). The truncation numbers in the FPCA procedure are selected by FVE=0.99.

Temperature, dew point temperature and searching records observed are regarded as functional markers to predict high- or low- hospital admissions. For each diagnostic method, we assess the performances of each single process respectively, then the combinations of each two processes, and finally the combinations of all the three processes. Table 2.6 and 2.7 respectively reflects the 5-fold cross-validation AUC and YI based on 100 replications with standard errors listed in the parentheses. Computing time of all the combinations are listed in Table 2.8. ROC curves obtained by temperature; temperature & dew point temperature and the three functional markers respectively are displayed in Figure 2.3, with the mean AUC shown in the figure for instance, and the corresponding YI can be found in Table 2.7.

We first assess the performances of our proposed features. Linear combination of features outperforms linear combination of observed data, indicating that the dimension-reduced feature maintained most of the correlation information within one stochastic process. The performances of the three functional markers are then compared respectively. Among the three functional markers, temperature and searching records provide the highest and lowest diagnostic accuracy respectively. Moreover, as an additional functional marker is added to a combination, the diagnostic accuracy will be improved with higher AUC.

The empirical performance of each combination method is considered as follows. The stepwise linear combination has the best performance among all the methods, no matter which marker(s) is (are) applied. The more functional markers are considered in diagnosis, the less is the difference between stepwise and ELR linear combinations. The non-linear combination is not recommended in our real data analysis because of its low diagnostic accuracy and computational efficiency. Since the number of functional markers is not very large, the computing efficiency of any combination methods is all efficient except non-linear combination. Thus, despite the efficient computing of min-max combination, the stepwise linear combination is most recommended for this problem.

2.5 Conclusion

In this paper, we combine the functional markers to improve the diagnostic accuracy, and applied the combinations to diagnose for high- or low- hospital admissions between 2010 and 2017 in Hong Kong. To be more specific, we construct a scalar

Methods	Stepwise	ELR	Min-max	Logistic*	Logistic	Non-linear
temp	0.8823 (0.0007)	-	-	0.8658 (0.0011)	-	0.8258 (0.0016)
dew	0.8342 (0.0006)	-	-	0.8265 (0.0009)	-	0.8263 (0.0017)
influenza	0.7145 (0.0012)	-	-	0.5840 (0.0046)	-	0.7016 (0.0025)
temp & influenza	0.8899 (0.0010)	0.8786 (0.0013)	0.9060 (0.0008)	0.8699 (0.0012)	0.8899 (0.0006)	0.8268 (0.0015)
dew & influenza	0.8591 (0.0013)	0.8454 (0.0018)	0.9061 (0.0009)	0.8250 (0.0014)	0.8599 (0.0008)	0.8199 (0.0016)
temp & dew	0.9148 (0.0016)	0.9006 (0.0021)	0.8884 (0.0011)	0.8670 (0.0011)	0.9042 (0.0009)	0.8379 (0.0015)
temp, dew & influenza	0.9351 (0.0010)	0.9327 (0.0016)	0.9031 (0.0010)	0.8688 (0.0014)	0.9267 (0.0008)	0.8411 (0.0016)

Table 2.6: The AUC derived by each curve and combinations by extreme quantile 0.15 and 0.85.

Methods	Stepwise	ELR	Min-max	Logistic*	Logistic	Non-linear
temp	0.6943 (0.0003)	-	-	0.6780 (0.0019)	-	0.6931 (0.0005)
dew	0.6516 (0.0009)	-	-	0.6476 (0.0021)	-	0.6423 (0.0013)
influenza	0.3814 (0.0024)	-	-	0.2019 (0.0064)	-	0.3634 (0.0033)
temp & influenza	0.6548 (0.0015)	0.6486 (0.0021)	0.6990 (0.0010)	0.6582 (0.0027)	0.6502 (0.0015)	0.6923 (0.0007)
dew & influenza	0.5744 (0.0017)	0.5729 (0.0021)	0.6940 (0.0026)	0.5638 (0.0025)	0.5724 (0.0015)	0.6437 (0.0012)
temp & dew	0.7137 (0.0011)	0.7071 (0.0016)	0.6250 (0.0014)	0.6816 (0.0021)	0.7084 (0.0013)	0.6904 (0.0007)
temp, dew & influenza	0.7739 (0.0032)	0.7610 (0.0072)	0.6958 (0.0006)	0.6534 (0.0028)	0.7377 (0.0026)	0.6909 (0.0011)

Table 2.7: The YI derived by each curve and combinations by extreme quantile 0.15 and 0.85.

Methods	Stepwise	ELR	Min-max	Logistic*	Logistic	Non-linear
temp	0.5042 (0.0042)	-	-	-	1.1842 (0.0093)	51.513 (0.2882)
dew	0.5067 (0.0045)	-	-	-	1.1676 (0.0089)	54.443 (0.3047)
influenza	0.5135 (0.0033)	-	-	-	1.1439 (0.0099)	55.319 (0.2646)
temp & influenza	0.5017 (0.0061)	1.4120 (0.2246)	0.5314 (0.0068)	-	1.1881 (0.0118)	64.904 (0.5599)
dew & influenza	0.4954 (0.0056)	1.7640 (0.2182)	0.5216 (0.0063)	-	1.1619 (0.0135)	64.684 (0.5667)
temp & dew	0.4772 (0.0083)	16.740 (0.7582)	0.5086 (0.0080)	-	1.2863 (0.0185)	66.090 (0.9314)
temp, dew & influenza	1.8780 (0.0419)	4.8752 (0.4383)	0.9622 (0.0221)	-	0.8718 (0.0212)	19.481 (0.4387)

Table 2.8: The computing time for each curve and combinations by extreme quantile 0.15 and 0.85.

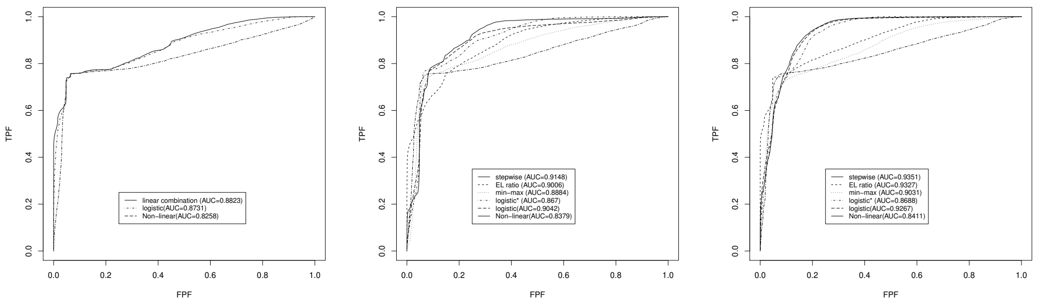

Figure 2.3: ROC curve of different diagnosis given by different markers

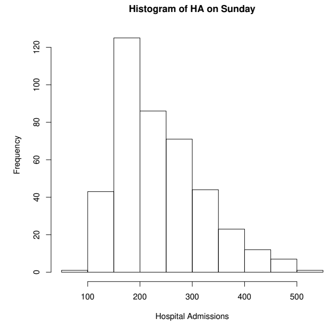

Figure 2.4: Histogram of hospital admissions on Sunday

feature motivated by square loss distance, which is an alternative of the original functional marker and can maintain the information. Then existing scalar combination methods are applied to make the diagnosis. In detail, common FPCA is utilized in the dimension reduction procedure. Linear combinations (including stepwise and ELR methods), non-linear combination and min-max combination are applied on the scalar features to improve the diagnostic accuracy. In addition, we provide an implement procedure and an R function for convenient application.

Simulations and real data analysis are studied to illustrate existing scalar combination methods and effectiveness of our proposed feature maintaining information of the original functional marker. On the basis of numerical analysis, the logistic regression on the proposed features outperforms logistic regression on functional observations directly, indicating the dimension reduction procedure retains the information to the most extent. Linear combinations have almost the best performance, with high diagnostic accuracy and fair computational efficiency. Min-max combination can efficiently save computing time and provide high diagnostic accuracy. Nonlinear combination costs a lot more time than linear and min-max combinations, while its performance is better than linear combinations in a few cases. Thus, all the combination methods can be applied to make the diagnosis, and their empirical performances can help the selection of them.

There exist several directions for future research. Firstly, one can use some other method to reduce the dimension of functional data rather than functional principal components analysis. For instance, instead of obtaining projections on basis functions derived by FPCA, one can obtain the projections on some other basis functions. Secondly, some other distance can be constructed, while other combinations of functional markers can also be studied. Thirdly, when functional and scalar markers are observed simultaneously, how to combine the functional markers and scalar markers is also an interesting problem for further research, although it has been studied comprehensively already. Finally, functional markers may not be necessarily transformed to scalars, some other extension are also worth a study for combination.

Chapter 3

Functional semivarying-coefficient additive model

3.1 Motivation

High hospital admissions will cause burden for healthcare system. Thus, analysis and prediction for hospital admissions may help hospitalization management.

Respiratory hospital admissions associated with weather conditions and air pollutants have been broadly studied. In environmental epidemiology community, hospital admissions have been mainly modeled by generalized linear model, which do not contain dynamic structure, and the effect of time on hospital admissions would not be discussed. Thus, in statistical community, cross-sectional time series models were mainly used to study hospital admissions data, such as varying coefficient models (Fan and Zhang (1999)), semi-varying coefficient models (Xia et al. (2004)), semi-parametric partially linear single-index models (Xia and Härdle (2006)), and generalized additive models (Zhang et al. (2015)). Although a functional additive cumulative time series model were proposed by Kong et al. (2010), the model was reduced to cross-sectional generalized additive model with cumulative effects (Xia and Tong (2006)) in their application for studying hospital admissions data. To the best of our knowledge, there is no publications focusing on hospital admissions from the perspective of functional regression models.

In this topic, we aim to explore a new functional regression model to study hospital admissions associated with weather conditions and air pollutants in the previous four weeks on respiratory hospital admissions. In our data set, hospital admissions, weather conditions including temperature, dew point temperature and relative humidity, and air pollutants including SO_2, NO_2, RSP and O_3 are observed daily between 2010 to 2017 in Hong Kong. Motivated by the real climate conditions in Hong Kong, and suggestions from Xia et al. (2002) and Shao et al. (2009), we treat the daily measurements for air pollutants every four weeks as functional curves. The functional trajectories are dense since they include twenty eight observations on each curve.

There is rich literature studying functional regression models of versatile types (Wang et al. (2016)). However, existing models may not be suitable enough for our real world problem. Although a functional additive cumulative time series model were proposed by Kong et al. (2010), the model was reduced to cross-sectional generalized additive model with cumulative effects (Xia and Tong (2006)) in their application for studying hospital admissions data. In our study, we aim to explore the format of functional regression model based on potential influential factors other than air pollutants and weather conditions, and how these influential factors impact on hospital admissions.

3.2 Functional regression modeling with hospital admissions response

Since the accumulative hospital admissions over a month, the response of our interest, has a annual period, the month of the year can influence the monthly hospital admissions. Besides, weather conditions may have a varying impact on hospital admissions based on the value of them. Moreover, air pollutants, the functional covariates, may not only affect hospital admissions with nonliearity, but also have interaction with weather conditions. Thus, the following model is explored:

$$\begin{aligned} Y &= \mathbf{Z}^\top \boldsymbol{\theta} + \sum_{k=1}^{q} g_k\left(\int_{\mathcal{I}} X_k(t)\gamma_k(t)dt, U\right) + \epsilon, \\ E(\epsilon|\mathbf{Z}, \mathbf{X}, U) &= 0, \quad Var(\epsilon|\mathbf{Z}, \mathbf{X}, U) = \sigma^2, \end{aligned} \tag{3.1}$$

where $Y \in \mathbb{R}$ is a scalar response, $\mathbf{Z} \in \mathbb{R}^p$ is the scalar covariates, U is a univariate variable to avoid the "curse of dimensionality", $X_k \in L^2(\mathcal{I})$ is the functional covariate for $k = 1, \cdots, q$, $g_k(\cdot, \cdot)$ are unknown functions, $\mathcal{I}$ is the impact time interval, and ϵ is the error term.

In our application, Y is the accumulative hospital admissions over a month, $\mathbf{Z}$ are the dummy variables indicating the month of the year, $U \in \mathbb{R}$ is the average temperature, relative humidity or dew point temperature over the previous month, and $X_k(t)$ are the four air pollutants curves for $k = 1, \cdots, 4$.

We call the model (3.1) a functional semi-varying coefficient additive model (FSV-CAM). On the one hand, if the unknown nonlinear function $g_k(x, u) = x \cdot \alpha_k(u)$, $\gamma_k(t) = 1$ and $X_k(t) = x_k$ for any $t \in \mathcal{I}$, then model (3.1) is reduced to the semi-varying coefficient model (Xia et al. (2004)):

$$\begin{aligned} y &= \mathbf{Z}^\top \boldsymbol{\theta} + \sum_{k=1}^{q} \alpha_k(U) x_k + \epsilon, \\ E(\epsilon|U, \mathbf{x}, \mathbf{Z}) &= 0, \quad Var(\epsilon|U, \mathbf{x}, \mathbf{Z}) = \sigma^2(U), \end{aligned} \tag{3.2}$$

where $\mathbf{x} = (x_1, \cdots, x_q)^\top$. On the other hand, if $g_k(x, u) = g_k(x)$, model (3.1) is reduced to the functional additive cumulative effects model (Kong et al. (2010)):

$$Y = \mathbf{Z}^\top \boldsymbol{\theta} + \sum_{k=1}^{q} g_k\left(\int_{\mathcal{I}} X_k(t)\gamma_k(t)dt\right) + \epsilon. \tag{3.3}$$

Since the model (3.1) has not been studied before, we need to explore some conditions for model identifiability. To be more specific, the nonidentifiability may come from several aspects. First, the nonlinear unknown functions $g_k(\cdot,\cdot)$ may be added to some non-zero constants, and the model identifiability is broken. Second, the coefficients γ_k and index functions $g_k(\cdot,\cdot)$ would also make the model not identifiable by multiply/divided a non-zero constant.

To address this problem, motivated by the conditions suggested by Shen et al. (2014), Fan et al. (2015), and Kong et al. (2010), we explore the following conditions for identifiability:

C0. The functional covariates $\mathbf{X}$ are conditional independent to $\mathbf{Z}$ given U.

C1. The response and scalar covariates are both centered:

$$E(Y) = 0 \quad \text{and} \quad E(\mathbf{Z}) = \mathbf{0}.$$

C2. The nonlinear function $g(\cdot,\cdot)$ is also centered given U:

$$E\left[g_k\left\{\int_{\mathcal{I}} X_k(t)\gamma_k(t)dt, U\right\}\middle| U\right] = 0.$$

C3. The functional coefficients were scaled: $\int_{\mathcal{I}} \gamma_k^2(t)dt = 1$ and $\int_{\mathcal{I}} \gamma_k(t)dt > 0$ for $k = 1, \cdots, q$.

Remark. *Condition C0 is a strong condition requiring the conditional independence between scalar covariates $\mathbf{Z}$ and functional covariates $\mathbf{X}$ given U. If $g_k(x,u) = h(x)$ do not depend on u, independence between functional and scalar covariates are not common used in functional regression analysis. This condition may be possible to be loosen in future work.*

Condition C1 is a general conditions for semi-varying coefficient model (Xia et al. (2004), Shen et al. (2014)) to center all the random variables in the model.

Condition C2 is a general assumption explored by the common used condition for functional additive model (Fan et al. (2015)), which was also applied in Kong et al. (2010). In Fan et al. (2015) and Kong et al. (2010), they assumed that the index functions are centered. In this thesis, condition C2 means that the index functions are centered condition on U. The condition C2 is equivalent to $E\big[g_k\big\{\int_{\mathcal{I}} X_k(t)\gamma_k(t)dt, c\big\}\big] = 0$ for any constant $c \in \mathbb{R}$. If condition C2 changes to

$$E\{g_k(\int_{\mathcal{I}} X_k(t)\gamma_k(t)dt, U)\} = 0,$$

suppose $g_1(x,u)$, $\cdots$,$g_4(x,u)$ are functions satisfying our explored model (3.1), then another group of functions $g_1^*(x,u) = g_1(x,u) + \alpha(u)$, $g_2^*(x,u) = g_2(x,u) - \alpha(u)$, $g_3^*(x,u) = g_3(x,u)$ and $g_4^*(x,u) = g_4(x,u)$ satisfying both model (3.1) and Condition C2, where $\alpha(\cdot)$ is a non-zero function satisfying $E\{\alpha(U)\} = 0$. The model identifiability will be broken. Thus, the condition C2 seems necessary for model identifiability.

Condition C3 is to scale the functional parameter $\gamma_k(t)$ without loss of generality. Similar condition $\int_{\mathcal{I}} \gamma_k(t)dt = 1$ was commonly used in functional regression models (Hall et al. (2007), Kong et al. (2010)).

Proposition 3.1. *With conditions C0 - C3, the model (3.1) is identified.*

Proof. Assume there are both sets of parameters and functions $\{\boldsymbol{\theta}, \{g_k\}_{k=1}^q, \{\gamma_k\}_{k=1}^q\}$ and $\{\boldsymbol{\theta}^*, \{g_k^*\}_{k=1}^q, \{\gamma_k^*\}_{k=1}^q\}$ satisfy conditions C0 to C3 and the model (3.1), then we shall prove that $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, $g_k(x,u) = g_k^*(x,u)$ for $k = 1, \cdots, q$, and $\gamma_k(t) = \gamma_k^*(t)$ almost surely for $k = 1, \cdots, q$.

We shall prove $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ first. By conditions C0 and C2, the conditional expectation of Y given $\mathbf{Z}$ and U would be

$$\begin{aligned} E(Y|\mathbf{Z}, U) &= E(\mathbf{Z}^\top \boldsymbol{\theta}|\mathbf{Z}, U) + E\bigg[g_k\bigg(\int_{\mathcal{I}} X_k(t)\gamma_k(t)dt, U\bigg)\bigg|\mathbf{Z}, U\bigg] + E(\epsilon|\mathbf{Z}, U) \\ &= \mathbf{Z}^\top \boldsymbol{\theta} + E(\epsilon|\mathbf{Z}, U), \end{aligned}$$

and similarly,

$$E(Y|\mathbf{Z},U) = \mathbf{Z}^\top\boldsymbol{\theta}^* + E(\epsilon|\mathbf{Z},U),$$

which indicates that $\mathbf{Z}^\top\boldsymbol{\theta} = \mathbf{Z}^\top\boldsymbol{\theta}^*$ for any $\mathbf{Z} \in \mathbb{R}^p$, and thus $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

Then we shall show that $g_k(x,u) = g_k^*(x,u)$ for any x,u, and $\gamma_k(t) = \gamma_k^*(t)$ almost surely. Since $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, for any fixed U, let $h_k(x) = g_k(x,U)$ and $h_k^*(x) = g_k^*(x,U)$, then we have

$$\sum_{k=1}^q h_k\left\{\int_\mathcal{I} X_k(t)\gamma_k(t)dt\right\} = \sum_{k=1}^q h_k^*\left\{\int_\mathcal{I} X_k(t)\gamma_k(t)dt\right\}$$

with $E\big[h_k\{\int_\mathcal{I} X_k(t)\gamma_k(t)dt\}\big] = E\big[h_k^*\{\int_\mathcal{I} X_k(t)\gamma_k(t)dt\}\big] = 0$. According to Fan et al. (2015), $h_k(x) = h_k^*(x)$ for any x, and $\gamma_k(t) = \gamma_k^*(t)$ almost surely by conditions C2 and C3. Thus, $g_k(x,u) = g_k^*(x,u)$ for any $x,u \in \mathbb{R}$. The model (3.1) is then confirmed to be identified. □

3.3 Estimation Procedures

The estimation procedure is as follows. Firstly, orthonormal functions $\{\phi_{km}(t)\}_{m=1}^\infty$ for $k = 1,\cdots,q$ and $m = 1,2,\cdots$ can be found and act as basis functions in $L^2(\mathcal{I})$ space. Basis functions derived by FPCA are applied in our method. Based on the obtained orthonormal basis functions $\{\hat\phi_{km}(t)\}_{m=1}^{S_k}$ where S_k is the truncation point which can be selected by specified fraction of variance explained (FVE) threshold (for instance), $\{\mathbf{X}_{ik},\gamma_k,g_k\}_{k=1}^q$ can be approximated by

$$X_{ik}(t) - E\{X_{ik}(t)\} \approx \sum_{m=1}^{S_k}\hat\xi_{ikm}\hat\phi_{km}(t),\ \gamma_k(t) \approx \sum_{m=1}^{S_k}\gamma_{km}\hat\phi_{km}(t),\ g_k(x,U) \approx \boldsymbol{\eta}_k(U)^\top\boldsymbol{\psi}(x), \tag{3.4}$$

where $\hat\xi_{ikm} = \int_\mathcal{I}[X_{ik}(t) - E\{X_{ik}(t)\}]\hat\phi_{km}(t)dt$ is the principal component score for $\mathbf{X}_i$, $\gamma_{km} = \int_\mathcal{I}\gamma_k(t)\hat\phi_{km}(t)dt$, $\boldsymbol{\eta}_k \in \mathbb{R}^d$ and $\boldsymbol{\psi}_k$ is a d-dimensional B-spline basis function, where d is the number of basis functions selected. When number of functional predictors q is low, the unknown function $g_k(x, U)$ can also be approximated by kernel method. However, to avoid curse of dimensionality as q increases, we prefer the orthogonal basis functions to approximate the unknown functions. Besides, B-spline is only a try and an exploration of in the estimation procedure. Other basis functions (such as Fourier basis and P-spline basis) can also be utilized be potentially useful. Based on the results of basis expansions (3.4), the model can be rewritten as

$$\begin{gathered} Y_i \approx \mathbf{Z}_i^\top \boldsymbol{\theta} + \sum_{k=1}^{q} \boldsymbol{\eta}_k(U_i)^\top \boldsymbol{\psi}_k\bigg(\sum_{m=1}^{S_k} \hat{\xi}_{ikm}\gamma_{km}\bigg) + \epsilon_i, \\ E(\epsilon_i|\mathbf{Z}, U, \mathbf{X}(t) \text{ for } t \in \mathcal{I}) = 0,\ Var(\epsilon_i|\mathbf{Z}, U, \mathbf{X}(t) \text{ for } t \in \mathcal{I}) = \sigma^2. \end{gathered} \tag{3.5}$$

In addition, the proposed conditions for model identifiability indicates that $E\{\boldsymbol{\psi}_k(\sum_{m=1}^{S_k} \hat{\xi}_{ikm}\gamma_{km})\} \approx 0$, $\sum_{m=1}^{S_k} \gamma_{km}^2 \approx 1$ and $\sum_{m=1}^{S_k} \gamma_{km} > 0$ for $k = 1. \cdots, q$.

Let $w_{ij} = \frac{K_h(U_i - U_j)}{\sum_{j_0=1}^{n}\sum_{i_0=1}^{n} K_h(U_{i_0} - U_{j_0})}$, $i, j \in \{1, 2, \cdots, n\}$, using the local linear approximation, the estimation will be constructed by minimizing:

$$\begin{aligned} \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} [Y_i - \mathbf{Z}_i^\top \boldsymbol{\theta} - b_{0j} - \mathbf{c}_{0j}^\top (U_i - U_j) \\ - \{\mathbf{b}_j + \sum_{k=1}^{q} \mathbf{c}_{kj}(U_{ik} - U_{jk})\}\boldsymbol{\psi}(\hat{\boldsymbol{\xi}}_i, \boldsymbol{\gamma})]^2 \cdot w_{ij}, \end{aligned} \tag{3.6}$$

with respect to $\boldsymbol{\theta}$, b_{0j}, c_{0j}, $\mathbf{b}_j$, $\mathbf{c}_{kj}$ and $\boldsymbol{\gamma}$, where $\boldsymbol{\xi}_i = \{\hat{\xi}_{ikm}\}_{\substack{k=1,\cdots,q \\ m=1,\cdots,S_k}}$, $\boldsymbol{\gamma} = \{\gamma_{km}\}_{\substack{k=1,\cdots,q \\ m=1,\cdots,S_k}}$, $\boldsymbol{\psi}(\hat{\boldsymbol{\xi}}_i, \boldsymbol{\gamma}) = \{\boldsymbol{\psi}_1(\sum_{m=1}^{S_1} \hat{\xi}_{i1m}\gamma_{1m})^\top, \cdots, \boldsymbol{\psi}_r(\sum_{m=1}^{S_q} \hat{\xi}_{irm}\gamma_{rm})^\top\}^\top = \{\boldsymbol{\psi}_1(\hat{\boldsymbol{\xi}}_{i1}^\top \boldsymbol{\gamma}_1)^\top, \cdots, \boldsymbol{\psi}_q(\hat{\boldsymbol{\xi}}_{ir}^\top \boldsymbol{\gamma}_q)^\top\}^\top$.

Based on the objective function (3.6), we propose the estimation procedure as follows:

Step 1. Apply the FPCA procedure to derive the FPC scores $\boldsymbol{\xi}$ and truncation numbers S_k for $k = 1, \cdots, q$ by FVE.

Step 2. Initialize $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ with model without nonlinear additive elements or varying coefficients. That is, the initial values $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}^*$ of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}^*$ respectively are derived by minimizing:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \mathbf{Z}_i^\top\boldsymbol{\theta} - \sum_{k=1}^{q}\sum_{m=1}^{S_k}\hat{\xi}_{ikm}\gamma_{km}\right\}^2 w_i, \tag{3.7}$$

with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. Obtained the minimizer $\hat{\boldsymbol{\gamma}}^*$, let $\hat{\boldsymbol{\gamma}}_k = \mathrm{sgn}(\sum_{m=1}^{S_k}\gamma_{km})\hat{\boldsymbol{\gamma}}_k^*/(\sum_{m=1}^{S_k}\hat{\gamma}_{km}^{*2})$ for $k = 1, \cdots, q$, where $\hat{\boldsymbol{\gamma}}_k^* = (\hat{\gamma}_{k1}^*, \cdots, \hat{\gamma}_{kS_k}^*)^\top$.

Step 3. In this step, we take the non-linear function into consideration, and then derive the estimation of $\boldsymbol{\theta}$, $\alpha(U_j)$ and $\boldsymbol{\eta}(U_j)$ by minimizing:

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}[Y_i - \mathbf{Z}_i^\top\hat{\boldsymbol{\theta}} - \{\mathbf{b}_j + \sum_{k=1}^{q}\mathbf{c}_{kj}(U_{ik} - U_{jk})\}^\top\boldsymbol{\psi}(\hat{\boldsymbol{\xi}}_i, \hat{\boldsymbol{\gamma}})]^2 \cdot w_{ij}, \tag{3.8}$$

with respect to $\mathbf{b}_j$ and $\mathbf{c}_{kj}$, while the estimator for $\boldsymbol{\eta}(U_j)$ are $\hat{\boldsymbol{\eta}}(U_j) = \hat{\mathbf{b}}_j$. The efficient estimation proposed by Xia et al. (2004) is applied in this step.

Step 4. We aim to estimate $\boldsymbol{\gamma}$ in this step. Direct estimation for $\boldsymbol{\gamma}$ is difficult due to the nonlinearity of the basis functions $\boldsymbol{\psi}_k(x)$. To overcome this difficulty, with the estimator of $\hat{\boldsymbol{\eta}}(U_j)$ from step 3 and current value $\boldsymbol{\gamma}_{k,old}$ of $\boldsymbol{\gamma}_k$, the nonlinear functions $\boldsymbol{\psi}_k(x)$ can be approximated by

$$\boldsymbol{\psi}_k(\hat{\boldsymbol{\xi}}_{ik}^\top\boldsymbol{\gamma}_k) \approx \boldsymbol{\psi}_k(\hat{\boldsymbol{\xi}}_{ik}^\top\boldsymbol{\gamma}_{k,old}) + \boldsymbol{\psi}_k'(\hat{\boldsymbol{\xi}}_{ik}^\top\boldsymbol{\gamma}_{k,old}) \cdot \hat{\boldsymbol{\xi}}_{ik}^\top(\boldsymbol{\gamma}_k - \boldsymbol{\gamma}_{k,old}),$$

and thus $\boldsymbol{\gamma}^*$ can be observed by minimizing

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\Bigg[Y_i - \mathbf{Z}_i^\top\hat{\boldsymbol{\theta}}$$

$$- \{\hat{\mathbf{b}}_j + \sum_{k=1}^{q}\hat{\mathbf{c}}_{kj}(U_{ik} - U_{jk})\}^\top \left\{\begin{matrix} \boldsymbol{\psi}_1(\hat{\boldsymbol{\xi}}_{i1}^\top\boldsymbol{\gamma}_{1,old}) + \boldsymbol{\psi}_1'(\hat{\boldsymbol{\xi}}_{i1}^\top\boldsymbol{\gamma}_{1,old}) \cdot \hat{\boldsymbol{\xi}}_{i1}^\top(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_{1,old}) \\ \vdots \\ \boldsymbol{\psi}_q(\hat{\boldsymbol{\xi}}_{ir}^\top\boldsymbol{\gamma}_{r,old}) + \boldsymbol{\psi}_q'(\hat{\boldsymbol{\xi}}_{ir}^\top\boldsymbol{\gamma}_{r,old}) \cdot \hat{\boldsymbol{\xi}}_{ir}^\top(\boldsymbol{\gamma}_r - \boldsymbol{\gamma}_{r,old}) \end{matrix}\right\}\Bigg]^2 \cdot w_{ij}, \tag{3.9}$$

with respect to $\boldsymbol{\gamma}$. Then the estimator $\hat{\boldsymbol{\gamma}}$ can be obtained by

$\hat{\boldsymbol{\gamma}}_k = \text{sgn}(\sum_{m=1}^{S_k} \gamma_{km})\hat{\boldsymbol{\gamma}}_k^* / (\sum_{m=1}^{S_k} \hat{\gamma}_{km}^{*2})$ for $k = 1, \cdots, q$. Similar algorithm for nonlinear functions was proposed in Fan et al. (2015).

Step 5. Repeat Step 3 and Step 4 until $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$ convergence.

3.4 Simulation study

In this section, we show the performance of our explore estimation procedure and the model identifiability by a simulation example. In particular, the unknown function g_k are set to be $g_k(x, u) = x\alpha_k(u)$ and we let $\boldsymbol{\psi}(x) = x$ in the estimation procedures.

The simulated data $\{Y_i\}_{i=1}^n$ are generated from the model

$$Y_i = \mathbf{Z}_i^\top \boldsymbol{\theta} + \sum_{k=1}^{q} g_k(\int_0^1 X_{ik}(t)\gamma_k(t)dt, U_i) + \epsilon_i = \mathbf{Z}_i^\top \boldsymbol{\theta} + \sum_{k=1}^{q} g_k\left(\sum_{l} \gamma_{kl}\xi_{ikl}, U\right) + \epsilon_i,$$

with $q = 2$ functional predictors, $p = 5$ dimensional covariates $\mathbf{Z}$, and univariate U; $\epsilon_1, \cdots, \epsilon_n$ are independent and identically distributed from $N(0, 0.25)$; $\boldsymbol{\theta}$ is the scalar vector for scalar covariates; $g_k(x, u) = x\alpha_k(u)$ are the nonlinear functions. The underlying regression function is $\gamma_k(t) = \sum_{m=1}^4 \gamma_{km}\phi_k(t)$, a linear combination of the eigenbasis. The scalar covariates $\mathbf{Z}_i = (Z_1, \cdots, Z_p)$ are generated by Bernoulli and Gaussian distributions, in our settings, $p = 5$. To be more specific, let $\tilde{\mathbf{Z}}_i = (\tilde{Z}_1, \cdots, \tilde{Z}_p)$ are jointly normal with zero mean, unit variance and AR(0.5) correlation structure. Then Z_{ij} are generated by Bernoulli distribution $Bernoulli(\Phi_0^{-1}(\tilde{Z}_{ij}))$ for $j = 1, 2$ where Φ_0 is the cumulative distribution function of standard normal distribution; $Z_{ij} = \tilde{Z}_{ij}$ for $j = 3, \cdots, p$. The univariate variable U_i are generated from uniform distribution $U[0, 1]$ for $i = 1, \cdots, n$. Next, we describe how to generate the predictors $X_{ik}(t)$. The functional predictors have mean zero and covariance function derived from the Fourier basis $\boldsymbol{\Phi}(t) =$

$(\phi_1(t), \phi_2(t), \phi_3(t), \phi_4(t))^\top = (\sqrt{2}\sin(\pi t), \sqrt{2}\cos(\pi t), \sqrt{2}\sin(3\pi t), \sqrt{2}\cos(3\pi t))^\top$. For $l = 1, \cdots, q$, define $V_{ik}(t) = \sum_{m=1}^4 \tilde{\xi}_{ikm}\phi_m(t)$, where $\{\tilde{\xi}_{ikm}\}_{i=1}^n$ are independent and identically distributed as $N(0, k^{-2})$ for different i and k. The two functional predictors are then derived through the linear combinations

$$X_{i1} = 1.5V_{i1} + 0.5V_{i2}, \quad X_{i2} = 0.5V_{i1} + 1.5V_{i2}.$$

Here, the two functional predictors are correlated with each other. The actual observations of $X_{ik}(t)$ are at 101 equally spaced times $\{t_{ikl} \in [0,1]\}_{l=1}^{100}$ with independent and identically distributed noise $\tilde{\epsilon}_{ikl} \sim N(0,1)$.

We use 100 Monte Carlo runs for model assessment. Bandwidth h is selected to be $\frac{1}{2}n^{-\frac{1}{5}}$, which is suggested to be $O(n^{-\frac{1}{5}})$ in Xia et al. (2004). Since the parameter $\boldsymbol{\theta}$, unknown function $g_k(x,u)$ are of our interest, we report the Monte Carlo averages of bias and its standard error for $\boldsymbol{\theta}$, and plot the estimated $\alpha_k(u)$ which is equivalent to estimation for $g_k(x,u)$ multiplied by x. The fitted values for Y_i is given by

$$\hat{Y}_i = \mathbf{Z}_i^\top \hat{\boldsymbol{\theta}} + \sum_{k=1}^{q} \hat{\alpha}_k(U_i) \sum_{m=1}^{S_k} \hat{\xi}_{ikm}\hat{\gamma}_{km}.$$

The mean square error (MSE) will also be reported with its standard error.

We propose two designs for our simulation study. In design 1, the sample size is set to be $n = 100, 200$ and 300. $\boldsymbol{\theta}_0 = (1, -1, 1, -1, 1)^\top$, $\alpha_1(u) = u^2$, $\alpha_2(u) = \sin(2\pi u)$. In design 2, the sample size is also set to be $n = 100, 200$ and 300. $\boldsymbol{\theta}_0 = (2, 1, 0.5, -1, -2)$, $\alpha_1(u) = \cos(\pi u)$, $\alpha_2(u) = \exp(u)$. Besides, in both designs, let $\boldsymbol{\gamma}_1^* = (1, 0.8, 0.5, 0.2)^\top$, $\boldsymbol{\gamma}_2^* = (0.7, 0.5, 0.3, 0.1)^\top$ and then $\boldsymbol{\gamma}_k = \text{sgn}(\sum_{l=1}^{101} \boldsymbol{\gamma}_k^\top \phi_k(t_{ikl}))\boldsymbol{\gamma}_k^*/\|\boldsymbol{\gamma}_k^*\|_2$ for $k = 1, 2$.

To show the merit of our explored model, similar results for estimation of our explored model (3.1) and (3.10) are also reported in Table 3.1 - 3.2. To illustrate the estimation results, we compare our proposed estimation with the general partial

Model	Bias					MSE
FSVCAM	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	
n=100	0.0062(0.0005)	0.0056(0.0005)	0.0061(0.0005)	0.0212(0.0008)	0.0144(0.0006)	0.171(0.051)
n=200	0.0042(0.0002)	0.0057(0.0002)	0.0082(0.0002)	0.0047(0.0003)	0.0026(0.0003)	0.220(0.026)
n=300	0.0025(0.0001)	0.0014(0.0001)	0.0004(0.0002)	0.0005(0.0002)	0.0007(0.0001)	0.227(0.021)
PFLM	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	
n=100	0.0118(0.0012)	0.0092(0.0012)	0.0067(0.0014)	0.0207(0.0015)	0.0036(0.0016)	0.723(0.124)
n=200	0.0061(0.0005)	0.0159(0.0005)	0.0101(0.0006)	0.0058(0.0008)	0.0014(0.0006)	0.845(0.117)
n=300	0.0010(0.0003)	0.0068(0.0004)	0.0076(0.0006)	0.0037(0.0004)	0.0098(0.0004)	0.878(0.087)

Table 3.1: Biases for $\hat{\theta}_1$ to $\hat{\theta}_5$ and MSE for $\hat{Y}$ in design 1. The results are shown as averages over 100 replicates with standard errors listed in the parentheses.

Model	Bias					MSE
FSVCAM	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	
n=100	0.0034(0.0003)	0.0094(0.0005)	0.0006(0.0005)	0.0198(0.0009)	0.0056(0.0007)	0.199(0.058)
n=200	0.0035(0.0002)	0.0036(0.0002)	0.0061(0.0003)	0.0027(0.0003)	0.0051(0.0003)	0.244(0.026)
n=300	0.0031(0.0001)	0.0015(0.0001)	0.0013(0.0002)	0.0007(0.0002)	0.0003(0.0001)	0.251(0.024)
PFLM	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_5$	
n=100	0.0034(0.0010)	0.0058(0.0011)	0.0037(0.0012)	0.0068(0.0020)	0.0151(0.0013)	0.626(0.117)
n=200	0.0028(0.0004)	0.0011(0.0004)	0.0145(0.0006)	0.0128(0.0007)	0.0030(0.0006)	0.751(0.080)
n=300	0.0022(0.0003)	0.0019(0.0003)	0.0037(0.0002)	0.0027(0.0004)	0.0028(0.0003)	0.742(0.071)

Table 3.2: Biases for $\hat{\theta}_1$ to $\hat{\theta}_5$ and MSE for $\hat{Y}$ in design 2. The results are shown as averages over 100 replicates with standard errors listed in the parentheses.

functional linear regression model (PFLM):

$$Y = \mathbf{Z}^{\top}\boldsymbol{\theta} + \sum_{k=1}^{q}\int_{0}^{1} X_k(t)\gamma_k(t)dt + \epsilon. \tag{3.10}$$

The bias of $\hat{\boldsymbol{\theta}}$, and MSE for fitted value by the model (3.10) are also reported in Table 3.1 - 3.2. According to the two tables, most of estimators for θ_1 to θ_5 by FSVCAM are more accurate with less biases, and more precise estimation with less standard errors than FPLM, when sample size n is large. Besides, the prediction for response Y by FSVCAM is much more accurate than prediction by FPLM in all cases. According to Table 3.1-3.2, the MSEs, which can be treated as estimators of σ^2, are increasing nearer to 0.25, the true value of σ^2, as sample size increases. the standard errors for all the estimators and prediction will decrease, and most of biases will also be less. The estimators of function $\hat{\alpha}_k$ and the 95% confidence interval (CI) in both designs with sample size $n = 300$ are plotted in Figure 3.1 - 3.2. According to the plotted figures, the estimated functions fit well for true functions.

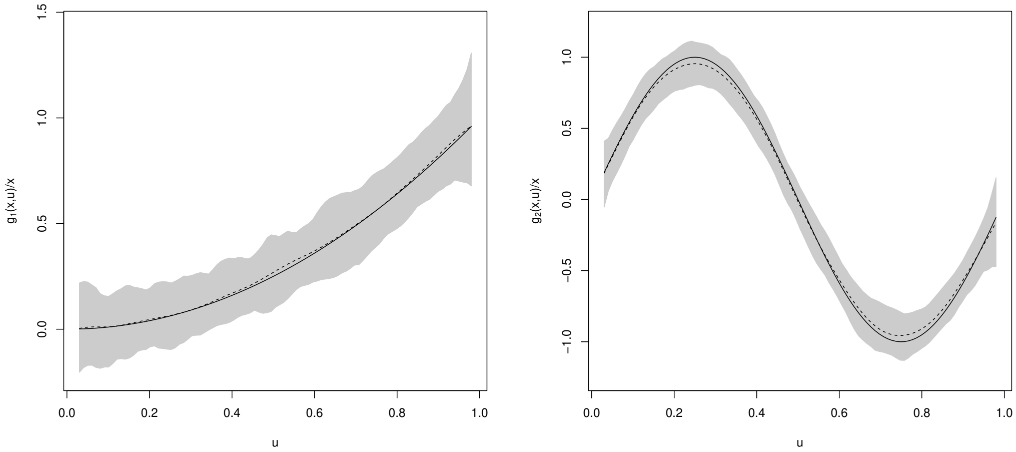

Figure 3.1: Estimated (dashed) and true (solid line) function $g_1(x,u)/x$ (left) and $g_2(x,u)/x$ (right) in u, and their 95% CI (gray) with sample size $n = 300$ in design 1.

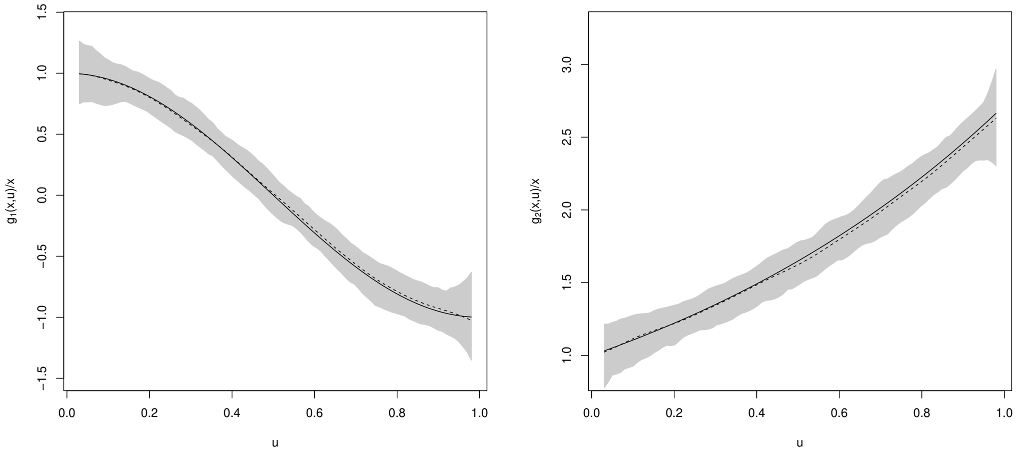

Figure 3.2: Estimated (dashed) and true (solid line) function $g_1(x,u)/x$ (left) and $g_2(x,u)/x$ (right) in u, and their 95% CI (gray) with sample size $n = 300$ in design 2.

Chapter 4

Conclusion and Discussion

The thesis focuses on improving diagnostics accuracy by combining multiple functional markers, together with the exploring of new functional regression modeling motivated from the HARD data setting with information of hospital admissions and environmental factors.

In Chapter 2, some uncertainties in our work need to be discussed. There is no unified criteria to distinguish high- or low- hospital admissions. Motivated by the effect modifier discussed in Katsouyanni et al. (2001), we treated daily higher/lower than 85/15 percent quantile as high- or low- hospital admissions. This classification may be sort of artificial because it is not a real binary classification. In a real dichotomous discriminant case, our methodology may behave better.

In Chapter 3, we tried to propose a functional additive regression model with versatile covariates types, including varying-coefficients, scalar covariates, and functional covariates. There are existing statistical models including both functional and scalar covariates (Lu et al. (2014), Kong et al. (2016)), however, none of them have discussed functional additive regression models with similar format to model (3.5). Thus, we need to make sure for model identifiability. Computational feasibility for functional regression models is another important problem. It is known nontrivial and deserves more future work in separate.

Bibliography

Alexander, A. L., Lee, J. E., Lazar, M. and Field, A. S. (2007) Diffusion tensor imaging of the brain. *Neurotherapeutics*, **4**, 316–329.

Bamber, D. (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, **12**, 387–415.

Chen, X., Vexler, A. and Markatou, M. (2015) Empirical likelihood ratio confidence interval estimation of best linear combinations of biomarkers. *Computational Statistics & Data Analysis*, **82**, 186–198.

Cook, S., Conrad, C., Fowlkes, A. and Mohebbi, M. H. (2011) Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic. *PLOS ONE*, **6**.

Davis, R. E., McGregor, G. R. and Enfield, K. B. (2016) Humidity: A review and primer on atmospheric moisture and human health. *Environmental research*, **144**, 106–116.

Duc, N. T., Ryu, S., Qureshi, M. N. I., Choi, M., Lee, K. H. and Lee, B. (2020) 3d-deep learning based automatic diagnosis of alzheimer¡s disease with joint mmse prediction using resting-state fmri. *Neuroinformatics*, **18**, 71–86.

Fan, J. and Zhang, W. (1999) Statistical estimation in varying coefficient models. *The Annals of Statistics*, **27**, 1491–1518.

Fan, Y., James, G. M., Radchenko, P. et al. (2015) Functional additive regression. *The Annals of Statistics*, **43**, 2296–2325.

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

Hedayat, A., Wang, J. and Xu, T. (2015) Minimum clinically important difference in medical studies. *Biometrics*, **71**, 33–41.

Hernandez, C., Jansa, M., Vidal, M., Nunez, M., Bertran, M. J., Garciaaymerich, J. and Roca, J. (2009) The burden of chronic disorders on hospital admissions prompts the need for new modalities of care: A cross-sectional analysis in a tertiary hospital. *QJM: An International Journal of Medicine*, **102**, 193–202.

Inácio, V., de Carvalho, M., Alonzo, T. A., González-Manteiga, W. et al. (2016) Functional covariate-adjusted partial area under the specificity-roc curve with an application to metabolic syndrome diagnosis. *The Annals of Applied Statistics*, **10**, 1472–1495.

Inácio, V., González-Manteiga, W., Febrero-Bande, M., Gude, F., Alonzo, T. A. and Cadarso-Suárez, C. (2012) Extending induced roc methodology to the functional context. *Biostatistics*, **13**, 594–608.

Kang, L., Liu, A. and Tian, L. (2016) Linear combination methods to improve diagnostic/prognostic accuracy on future observations. *Statistical methods in medical research*, **25**, 1359–1380.

Katsouyanni, K., Touloumi, G., Samoli, E., Gryparis, A., Le Tertre, A., Monopolis, Y., Rossi, G., Zmirou, D., Ballester, F., Boumghar, A., Anderson, H., Wojtyniak, B., Paldy, A., Braunstein, R., Pekkanen, J., Schindler, C. and Schwartz, J. (2001) Confounding and effect modification in the short-term effects of ambient particles on total mortality: Results from 29 european cities within the aphea2 project. *Epidemiology (Cambridge, Mass.)*, **12**, 521–31.

Kong, D., Xue, K., Yao, F. and Zhang, H. H. (2016) Partially functional linear regression in high dimensions. *Biometrika*, **103**, 147–159.

Kong, E., Tong, H. and Xia, Y. (2010) Statistical modelling of nonlinear long-term cumulative effects. *Statistica Sinica*, 1097–1123.

Le Thi Hoai, A. and Tao, P. D. (1997) Solving a class of linearly constrained indefinite quadratic problems by dc algorithms. *Journal of global optimization*, **11**, 253–285.

Li, Y., Hsing, T. et al. (2010) Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, **38**, 3321–3351.

Liu, C., Liu, A. and Halabi, S. (2011) A min–max combination of biomarkers to improve diagnostic accuracy. *Statistics in medicine*, **30**, 2005–2014.

Lu, Y., Du, J. and Sun, Z. (2014) Functional partially linear quantile regression model. *Metrika*, **77**, 317–332.

Ma, H., Yang, J., Xu, S., Liu, C. and Zhang, Q. (2020) Combination of multiple functional markers to improve diagnostic accuracy. *Journal of Applied Statistics*, DOI:10.1080/02664763.2020.1796945.

Pepe, M. S. and Thompson, M. L. (2000) Combining diagnostic test results to increase accuracy. *Biostatistics*, **1**, 123–140.

Pun, V. C., Yu, I. T., Qiu, H., Ho, K., Sun, Z., Louie, P. K. K., Wong, T. W. and Tian, L. (2014) Short-term associations of cause-specific emergency hospitalizations and particulate matter chemical components in hong kong. *American Journal of Epidemiology*, **179**, 1086–1095.

Shao, Q., Wong, H., Ip, W. and Li, M. (2009) Effect of ambient air pollution on respiratory illness in hong kong: a regional study. *Environmetrics*, **21**, 173–188.

Shen, S., Cui, J., Mei, C. and Wang, C. (2014) Estimation and inference of semi-varying coefficient models with heteroscedastic errors. *Journal of Multivariate Analysis*, **124**, 70–93.

Souza, J. B. d., Reisen, V. A. A., Santos, J. M. A. and Franco, G. C. A. A. (2014) Principal components and generalized linear modeling in the correlation between hospital admissions and air pollution. *Revista de SaÃPÃ*, **48**, 451 – 458.

Su, J. Q. and Liu, J. S. (1993) Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, **88**, 1350–1355.

Vexler, A., Liu, A., Schisterman, E. F. and Wu, C. (2006) Note on distribution-free estimation of maximum linear separation of two multivariate distributions. *Nonparametric Statistics*, **18**, 145–158.

Wahba, G. (1990) *Spline models for observational data*, vol. 59. Siam.

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

Xia, Y. and Härdle, W. (2006) Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*, **97**, 1162–1184.

Xia, Y. and Tong, H. (2006) Cumulative effects of air pollution on public health. *Statistics in medicine*, **25**, 3548–3559.

Xia, Y., Tong, H., Li, W. and Zhu, L. (2002) An adaptive estimation of optimal regression subspace. *J. Roy. Statist. Soc. Ser. B*, **64**, 363–410.

Xia, Y., Zhang, W. and Tong, H. (2004) Efficient estimation for semivarying-coefficient models. *Biometrika*, **91**, 661–681.

Xu, T., Fang, Y., Rong, A. and Wang, J. (2015) Flexible combination of multiple diagnostic biomarkers to improve diagnostic accuracy. *BMC medical research methodology*, **15**, 94.

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

Yin, J. and Tian, L. (2014) Optimal linear combinations of multiple diagnostic biomarkers based on youden index. *Statistics in medicine*, **33**, 1426–1440.

Youden, W. J. (1950) Index for rating diagnostic tests. *Cancer*, **3**, 32–35.

Zhang, W., Li, D. and Xia, Y. (2015) Estimation in generalised varying-coefficient models with unspecified link functions. *Journal of Econometrics*, **187**, 238–255.

Zhou, X., Wang, S., Xu, W., Ji, G., Phillips, P., Sun, P. and Zhang, Y. (2015) Detection of pathological brain in mri scanning based on wavelet-entropy and naive bayes classifier. *International Conference on Bioinformatics and Biomedical Engineering*, 201–209.

Zhu, L., Wang, J., Shi, H. and Tao, X. (2019) Multimodality fmri with perfusion, diffusion-weighted mri and 1h-mrs in the diagnosis of lympho-associated benign and malignant lesions of the parotid gland. *Journal of Magnetic Resonance Imaging*, **49**, 423–432.