ABSTRACT

CHEN CHUN CHEN

Classification Methods for Circular-Linear Data Using Periodic Functions

(Under the direction of Dr. Varghese George)

In many fields such as medicine, agriculture and environmental studies, data are collected over time which can have some repeated pattern within a certain time period. Those data with the linear responses or measures such as blood pressure or solar energy with circular predictor, are called circular-linear data. The data having repeated measures over time are usually analyzed using longitudinal analysis methods. However, applying classical longitudinal data analysis to circular-linear data is generally inappropriate since the circular pattern of time would be treated as a simple continuous variable. Parametric approaches for circular-linear data have been developed using various modeling methods. We propose a Bayesian non-parametric MCMC circular smoothing splines approach, which is not only appropriate but also adds more flexibility for modeling and classification for circular-linear data. We first fit the circular-linear data on an estimated circle, to elicit functional pattern from the data, and then classify the patterns. In the development of the classification procedure, we use functional data analysis and some widely used dimension reduction classification methods such as the principal component analysis and support vector machine. We evaluate the performance of the proposed modelling and classification methods through extensive simulation, and demonstrate using the 2005-2006 NHANES physical activity monitor data on insomnia patients.

In simulation study, the non-parametric Bayesian smoothing splines method coupled with support vector machine approach yields best performance in classification in terms of concordance rate. Our proposed nonparametric approach performed slightly better than the established parametric methods. Also, the initial data fitting procedures using a periodic regression function to reduce the noise in the data are shown to improve the performance in the classification problem. The result in the analysis of the NHANES data is consistent with simulation.

KEYWORDS: Supervised Classification, Circular Linear Data, Bayesian Approach for Periodic Smoothing Splines

CLASSIFICATION METHODS FOR CIRCULAR-LINEAR DATA USING PERIODIC FUNCTIONS

By

CHEN CHUN CHEN

Submitted to the Faculty of the School of Graduate Studies

of Augusta University in partial fulfillment

of the Requirements of the Degree of

Doctor of Philosophy in Biostatistics

August

2016

CLASSIFICATION METHODS FOR CIRCULAR-LINEAR DATA USING PERIODIC FUNCTION

This thesis/dissertation is submitted by Chen Chun Chen and has been examined and approved by an appointed committee of the faculty of the Graduate School of Augusta University.

The signatures which appear below verify the fact that all required changes have been incorporated and that the thesis/dissertation has received final approval with reference to content, form and accuracy of presentation.

This thesis/dissertation is therefore in partial fulfillment of the requirements for the degree of (Master of Science/Master of Science in Nursing/Master of Health Education/Doctor of Philosophy).

_____ _____

Date Major Advisor

_____

Departmental Chairperson

(Nursing Only)_____

Associate Dean for Graduate Programs

_____

Dean, School of Graduate Studies

ACKNOWLEDGEMENTS

I would like to express my sincere thanks and appreciation to my major advisor Dr. Varghese George for his invaluable guidance and support for my doctoral training. I also would like to extend my thanks and appreciation to my co-advisors Drs. Duchwan Ryu and Ashis SenGupta for their valuable advise and encouragement that helped me to complete this dissertation.

I am also extremely thankful to Drs. Jie Chen and Hongyan Xu for their helpful comments and suggestions as members of my dissertation committee, and to Dr. Daniel Linder for serving as my reader and offering valuable suggestions. I would not have reached here without their great efforts.

I also like to thank my classmates and friends who always cheered me up and supported me throughout my graduate training.

Last but not least, I would like to thank my wife, Susan, my son, Austin, and my family in Taiwan who are my greatest mental supporters. They provided me the energy to keep working hard and moving on.

Table of Contents

Page

List of Figures

Page

List of Tables

Page

CHAPTER 1

INTRODUCTION

In many fields such as medicine, agriculture, biology and environmental studies, researches encounter data that would have repeated pattern within a certain time period. Those data can be represented on the perimeter of a unit circle and are called circular data. Most of the repeated measurements are likely continuous such as heart beats counts, blood pressure or some other laboratory readings along time, and are traditionally referred to as longitudinal data or functional data. However, when the repeated pattern is observed in the data along with time, the time should be considered as a circular predictor instead of a linear predictor. Such data with circular predictors with continuous and linear responses are called circular-linear data. For example, in a variety of investigations in biology and medicine, sometimes responses are often related to some biological rhythms such as blood pressure, heart beats, and body temperature [1]. Measuring those readings from individuals, similar patterns are likely to be observed over twenty-four hour time period. Some statistical approaches for analyzing circular-linear data are provided in Fisher [2], Jammalamadaka and SenGupta [3], and Mardia and Jupp [4].

When the response is circular, classical longitudinal or functional data analysis may not be appropriate and results in the misleading conclusion. In the case of topology, the difference between line and circle is the main cause of the misleading. For example, we consider the short arm on a clock as a simple illustration. Calculating the mean direction between 1 o'clock and 11 o'clock, the correct answer should be 12 o'clock. However, simply taking average of 1 o'clock and 11 o'clock, we will get 6 o'clock which the wrong answer. When the predictor is circular, the response should be analyzed after taking into account the periodicity.

The circular-linear data analysis can be broadly applied to various fields including the researches in medicine which usually involve potential circular-linear data. The

circular-linear data approach can provide additional information that traditional approach may miss. The motivating example is the diagnosis of insomnia. Insomnia can include symptoms like difficulty or lack of sleep, fatigue, lack of daytime activity and inadequate sleeping patterns [5]. However, the diagnosis of insomnia can be subjective due to inaccuracy of individual report or physicians' experience. When there are accurate records of activity for a whole day in a long term, all symptoms related to activity can be observed systematically. Analytical inspection of daily activity pattern can lead to better inference of diagnosis of insomnia. People tend to have similar activity pattern every day due to work shift, sleep habit, dining time or other personal activity preference. Identifying circular features in activity patterns is natural and essential to have correct classification of insomnia patients. When observing both groups of insomnia and non-insomnia people's daily activity, we use mathematical model to retrieve the estimated pattern of their activity by considering their demographic information, disease status or other interested variables. By controlling these covariates, we can extract the reference activity patterns from each group and use the adjusted patterns for classification problems. Furthermore, some inference and interpretation can be made after the reference curves are derived in both groups.

Our major interest is classification of individuals into the group of normal individuals and the group of insomnia patients based on their daily activity intensities. To model the periodic daily activity intensity, we consider a circular-linear model for the linear response with circular covariates. In parametric approaches, we consider the circular-linear regression model. In nonparametric approaches, we also propose a Bayesian non-parametric regression model with periodic cubic smoothing splines. The periodic cubic smoothing splines replace the natural end conditions by periodic continuity conditions, and provide and periodic estimated curves after smoothing. The proposed Bayesian non-parametric regression approach, noted as Bayesian periodic cubic smoothing splines, have more flexibility in the estimation of curves by choosing different prior distributions for unknown parameters. An alternative approach to the circular-linear regression model are proposed by SenGupta and Ugwuowo [6] for periodic complications. For more flexibility in the circular-linear regression model, we can assume different distributions for the residuals in the model as explained in the extension

work by Bhattacharya and SenGupta [7]. For classification problem, we use the fitted values from the modeling procedure instead of original values as inputs. Several classification criteria have been considered such as support vector machines, distance-based classification, and functional data analysis based classification.

In chapter 2, we do a review to the literatures related to classification and inference in circular linear data. In chapter 3, proposed non-parametric and the extension to include reference time and linear predictor are described in detail. Also, the modifications for circular-linear regression model are described in the same chapter. In chapter 4, we conduct extensive simulation to evaluate the validity on our proposed method. In chapter 5, we use our proposed method to analysis the NHANES insomnia data.

CHAPTER 2
LITERATURE REVIEW

In section 2.1, we provide an extensive background and review of the circular linear regression models which are commonly used for dealing with circular-linear data under the parametric approach. In section 2.2, the additive model for longitudinal data under Bayesian framework. The idea will be extended and incorporated in developing the approach based on Bayesian nonparametric cubic smoothing splines. In section 2.3, the natural cubic smoothing splines is reviewed and described which is one of the basis methods in smoothing splines approach and is extensively used in several fields. Finally, some classification methods which are popular in handling the high-dimensional data are reviewed in the section 2.4.

2.1 Circular-Linear regression model

In parametric approaches, asymmetric circular-linear regression model has been developed by SenGupta and Ugwuowo [6]. The model is used to predict environmental characteristics based on some circular and linear predictors. One good example is the

solar energy studied in the paper. The response of interest is the solar energy absorbed, which is measured by the real numbers on a linear scale. The circular predictor considered is the twenty-four hours period in which the time points in a day are converted into angles. Other linear predictors they included are ambient temperature and control temperature which are also measured on linear scale. The primary purpose of the multivariate multiple regression model is to predict the linear response using a circular predictor and a set of linear predictors. A simple cosine function regression model is considered for the analysis. Let $Y_i, i = 1, \dots, m$, denote the response variable, x_{ip} denote the set of linear independent variables, $p = 1, \dots, P$, where P is the total number of linear independent variables and t_i is the circular independent variable, mostly time, within a time period T. Then, given the acrophase t_0 which indicates the crest or peak of the time period, the model is given by

$$Y_i = M + \sum_{p=1}^{P} \beta_p x_{ip} + A cos\omega(t_i - t_0) + \varepsilon_j \qquad (1)$$

where M is the mean level of responses, β_p is the regression coefficient, A and ω are the amplitude and the frequency of the angular function, respectively, and ε_i is the random error term. I.e., $\omega = \frac{2\pi}{T}\ or \frac{360^o}{T}$.

When T is known, ω is also known.

We can expand this to incorporate the generalized cosine model, which is a trigonometric polynomial model that contains the angular frequencies that are multiples of ω, given by $\omega t, 2\omega t, \dots, m\omega t$. The function contains smaller portion of the overall time period like $\frac{T}{2}, \frac{T}{3}$, ..., which fit into the overall period. The generalized model is

$$Y_i = M + \sum_{p=1}^{P} \beta_p x_{ip} + g(t_i) + \varepsilon_i \qquad (2)$$

where $g(t_i) = A_1 \cos(\omega t_i - \phi_1) + A_2 \cos(2\omega t_i - \phi_2) + \cdots + A_m \cos(m\omega t_i - \phi_m)$, and $\phi_1, \dots, \phi_m$ are the acrophases in the smaller portions of overall time period. The general period could possibly contain multiple partitioned periods in it.

When the pattern has the major or minor peaks and troughs, the non-linear periodic functions can also be introduced in the model. When the peaks and troughs do not connect to each other, the patterns are expected to be skewed, and the non-linear model for this situation is given by

$$Y_i = M + \sum_{p=1}^{P} \beta_p x_{ip} + Acos(\varphi_i + vcos\varphi_i) + \varepsilon_i \quad (3)$$

where $\varphi_i = \omega t_i - \omega t_0$ and v is the parameter for the skewness. In SenGupta and Ugwuowo [6], it has been shown that v usually lies between $-30^o \leq v \leq 30^o$ and the original simple cosine function will be obtained when $v = 0$.

In our approach, both linear and non-linear least square are considered in parameter estimation. Also, the goodness of fit of the model can be assessed by several diagnostic tools that are commonly used for assessing linear regression models.

2.2 Regression model for longitudinal data

In longitudinal data analysis, Linear Mixed Models (LMMs)[8] are well known and widely used for modeling the relationship between longitudinal observation and risk factor of interest. The covariate effects in the model are estimated parametrically and the model is capable of allowing for both continuous and discrete covariates. As one extension, Semiparametric Mixed Models (SPMMs)[9] consider a nonparametric function for time effects and parametric function for other covariates. Given an example, let $\boldsymbol{Y}_j$ be the reading vector for jth individual where $j = 1, \dots, n$

$$\boldsymbol{Y_j} = \boldsymbol{X}_j^T \boldsymbol{\beta} + \boldsymbol{g}_j(t) \quad (11)$$

where X_j is the design matrix and $\boldsymbol{\beta}$ is the vector of fixed effects, $\boldsymbol{g}_j(t)$ are differentiable smooth functions of time.

In SPMMs, the additional random effects are considered by another set of vector and covariates which is usually denoted as $\boldsymbol{Z}_j^T \boldsymbol{b}$ where $\boldsymbol{Z}_j$ are the covariates and $\boldsymbol{b}$ is a vector of random effects. For simplicity, only the fixed effects are considered in the

proposed non-parametric model and further extension to include random effect can be achieved using Bayesian approach in longitudinal data.

Several nonparametric approaches for estimating the function of time effect have shown their appropriateness and robustness in previous works with different modeling schemes [10-12]. Natural cubic smoothing splines and P-splines are some of the major considerations for the estimation of nonparametric functional time effect under Bayesian framework. Other choice for functional time effect $\boldsymbol{g}_j(t)$ can be a polynomial form like $\boldsymbol{g}_j(t) = \theta_1 t + \theta_2 t^2 + \cdots + \theta_p t^p$ in the parametric regression settings [12]. Regardless of how functional time effects are constructed, the additive models are the most common models considered in modeling longitudinal data. The time effect is an additive effect which is usually independent from other fixed or random effects. However, the time effects are based on the discrete time points on a plane without considering the period or rhythm of the responses. In order to capture the periodic feature of daily activity for individual, further consideration of the circular form of the time effect in additive model must take place.

2.3.1 Natural cubic smoothing splines

Natural cubic smoothing splines is a well-known smoothing method using spline basis functions with boundary conditions. The nonparametric nature makes it not limited to model-based dependence of response variables on independent variables [13]. The natural cubic smoothing splines regression finds the smoothing splines regression function $f(x)$, which minimizes the penalized residual sum of squares. For the ordered design points $a \leq x_1 \leq x_2 \leq \cdots \leq x_n \leq b$ and the corresponding response measurements $y = (y_1, y_2, \dots, y_n)^T$, The spline basis functions are defined by the internal knot points. The cubic smoothing splines require functions $f(x)$ that are expressed by the basis functions and corresponding regression coefficients and are twice differentiable at each knot point. In natural cubic smoothing splines, it is necessary to assign boundary conditions. It is shown that the natural cubic smoothing splines is the unique minimizer of the general form of the penalized residual sum of squares, given by

$$\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \alpha\int_a^b \{f''(x)\}^2 dx, \tag{4}$$

where α is a smoothing penalty for the regression function $f(x)$. The boundary points a and b are arbitrary as long as it contains all the data points. The smoothing splines beyond the boundary are linear which satisfy $f''(x_1) = f''(x_n) = 0$ as the natural end conditions. The first term in the formula (4) determines the closeness of the smoothing curve to the data, and the second term penalizes the curvature of the smoothing function. Hence, the choice of α determines the roughness of the smoothing function. By taking the design points as knot points, we can express the regression curve $\boldsymbol{f} = [f(x_1), \dots, f(x_n)] = \boldsymbol{X\beta}$ on $\boldsymbol{X} = [\mathrm{N_j}(x_i)]$, $i, j = 1, \dots, n$, with regression coefficients $\boldsymbol{\beta} = (\beta_1, \dots \beta_n)^T$, where $N_j(x_i)$ is the jth spline basis function evaluated at x_i, and we can calculate the roughness of the curve such that

$$\int \{f(x)''\}^2 dx = \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{\Omega}_{ij}\boldsymbol{X\beta} \tag{5}$$

$$\textbf{where } \boldsymbol{\Omega}_{ij} = \int N_i''(x)N_j''(x)dx, \tag{6}$$

and $\boldsymbol{\Omega}$ is a $n \times n$ matrices of rank $n - 2$.

Then the criterion can be rewritten in the matrix form as,

$$(\boldsymbol{y} - \boldsymbol{f})^T(\boldsymbol{y} - \boldsymbol{f}) + \alpha\boldsymbol{\beta}^T\boldsymbol{\Omega\beta}. \tag{7}$$

Setting the derivative with respect to $\boldsymbol{\beta}$ equal to zero gives,

$$(\boldsymbol{X}^T\boldsymbol{X} + \alpha\boldsymbol{\Omega})\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{y}. \tag{8}$$

Therefore, the natural cubic smoothing splines can be written as

$$\boldsymbol{f} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \alpha\boldsymbol{\Omega})^{-1}\boldsymbol{X}^T\boldsymbol{y} = (\boldsymbol{I} + \alpha\boldsymbol{K})^{-1}\boldsymbol{y}, \tag{9}$$

where $\boldsymbol{K} = \boldsymbol{X}^{-T}\boldsymbol{\Omega}\boldsymbol{X}^{-1}$. In matrix notation, over all candidate fitted vector $\boldsymbol{f}$ and $\boldsymbol{K}$, the cubic smoothing splines minimizes

$$(\boldsymbol{y} - \boldsymbol{f})^T(\boldsymbol{y} - \boldsymbol{f}) + \alpha\boldsymbol{f}^T\boldsymbol{K}\boldsymbol{f}. \tag{10}$$

It is appropriate to denote the later term $\boldsymbol{f^T K f}$ as the roughness since it can be represented in a quadratic form in the second derivative.

2.3.2 Bayesian natural cubic smoothing splines

Hastie and Tibshirani[13], provides the details of the Bayesian approach that is analogous to the frequentist natural cubic smoothing splines. Following the idea in natural cubic smoothing splines approach, Bayesian natural cubic smoothing splines uses all x_i as the knot points and assign prior distributions for all unknown parameters including regression coefficients, nuisance parameters as well as the smoothing penalty.

Under the Bayesian model the following assumptions are made:

(1) The data have a Gaussian distribution with mean $\boldsymbol{X\beta}$ and variance $\sigma^2\boldsymbol{I}$.

(2) $\boldsymbol{\beta}$ has a multivariate Gaussian prior distribution with mean 0 and variance $\tau^{-1}\boldsymbol{\Omega}^-$ where $\tau^{-1} = \frac{\sigma^2}{\alpha}$

The resulting posterior distribution of $\boldsymbol{\beta}$ is a multivariate Gaussian distribution with mean $E(\boldsymbol{\beta}|y) = \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \alpha\boldsymbol{\Omega})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ and covariance matrix $cov(\boldsymbol{\beta}|\boldsymbol{y}) = (\tau\boldsymbol{\Omega} + \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

From the posterior distribution of $\boldsymbol{\beta}$, the posterior inference of all characteristics of the cubic splines $\sum_{i=1}^{n} N_i(x)\beta_i$, parameterized by β, can be obtained. In addition, the posterior mean and variance of $\boldsymbol{f} = \boldsymbol{X\beta}$ can be computed as well. The posterior mean is $\hat{\boldsymbol{f}} = \boldsymbol{Sy}$, where $\boldsymbol{S} = (\boldsymbol{I} + \alpha\boldsymbol{K})^{-1}$ is the hat matrix, and the posterior covariance is $\boldsymbol{S}\sigma^2$. However, estimation based on this model is known to be biased.

A more direct approach is to put the prior on f itself. The prior for f corresponding to that for β above has mean zero and covariance $\tau^{-1}\boldsymbol{X}^T\boldsymbol{\Omega}^-\boldsymbol{X} = \tau^{-1}\boldsymbol{K}^-$, where $\boldsymbol{K} = \boldsymbol{N}^{-T}\boldsymbol{\Omega}\boldsymbol{N}^{-1}$ and $\boldsymbol{K}^-$ is the generalized inverse of $\boldsymbol{K}$. The priors for functional values are evaluated at all n design points, $\boldsymbol{f}$, the smoothing parameter, τ and the nuisance parameter σ^2, as given by below:

$$\boldsymbol{f}|\alpha, \sigma^2 \sim Singular\ Normal(0, \tau^{-1}\boldsymbol{K}^-)$$

$$\tau \sim Gamma(A_\tau, B_\tau)$$

$$\sigma^2 \sim Inverse\ Gamma(A_\sigma, B_\sigma)$$

The full conditional distributions derived from the prior distributions and likelihood of Y are given by

$$\boldsymbol{f}|\alpha, \sigma^2, \boldsymbol{Y} \sim Normal((\boldsymbol{I} + \alpha\boldsymbol{K})^{-1}\boldsymbol{Y}, (\boldsymbol{I} + \alpha\boldsymbol{K})^{-1}\sigma^2)$$

$$\tau|\sigma^2, \boldsymbol{Y}, \boldsymbol{f} \sim Gamma(n - 2 + \frac{A_\tau}{2}, \boldsymbol{f^T K f} + \frac{B_\tau}{2})$$

$$\sigma^2|\tau, \boldsymbol{Y}, \boldsymbol{f} \sim Inverse\ Gamma(n + \frac{A_s}{2}, (\boldsymbol{Y} - \boldsymbol{f})^T(\boldsymbol{Y} - \boldsymbol{f}) + \frac{B_s}{2})$$

For Bayesian inference, the Gibb's sampler is used to generate samples of $\boldsymbol{f}, \tau, \sigma^2$. For practical calculation aspect, the construction of $\boldsymbol{K} = \boldsymbol{Q}\boldsymbol{R}^{-1}\boldsymbol{Q}^{\boldsymbol{T}}$ is as suggested by de Boor [14] with a piecewise polynomial representation. Here, $\boldsymbol{Q}^{\boldsymbol{T}}$ is a $(n-2) \times n$ matrix and $\boldsymbol{R}$ is a $(n-2) \times (n-2)$ matrix which are given by

$$\boldsymbol{Q}^{\boldsymbol{T}} = \begin{pmatrix} \frac{1}{x_2 - x_1} & -\frac{1}{x_3 - x_2} - \frac{1}{x_2 - x_1} & \frac{1}{x_3 - x_2} & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & & \frac{1}{x_{n-1} - x_{n-2}} & -\frac{1}{x_n - x_{n-1}} - \frac{1}{x_{n-1} - x_{n-2}} & \frac{1}{x_n - x_{n-1}} \end{pmatrix}$$

$$6\boldsymbol{R} = \begin{pmatrix} x_2 - x_1 & x_3 - x_2 & & \cdots & 0 \\ x_3 - x_2 & 2(x_4 - x_2) & x_4 - x_3 & & \\ & & & \ddots & \\ 0 & & & \cdots & x_{n-2} - x_{n-3} & (x_{n-1} - x_{n-3}) & x_{n-1} - x_{n-2} \\ & & & & x_{n-1} - x_{n-2} & x_n - x_{n-1} \end{pmatrix}$$

The $\boldsymbol{R}$ matrix can be derived by dividing all elements in matrix $6\boldsymbol{R}$ by 6 for construction of $\boldsymbol{K}$.

2.3.3 Periodic cubic smoothing splines

Although natural smoothing splines are used extensively and also show the appropriateness for smoothing various type of data, they can not provide adequate result when used to smooth design points belong to the periodic data. The dissatisfaction of

using natural smoothing splines to estimate the periodic data can be shown especially in graphics application. When estimating the closed curves, natural smoothing splines give the unacceptable result majorly because the natural end conditions force the estimation on the boundary becoming linear. This brings the estimated curves closed up with the obvious discontinuity or sometimes even not close at all. In Graham [15], the method provides the periodic end condition to replace the natural end condition and achieve the predetermined closeness of fitting the cubic smoothing splines for periodic data.

Let $(x_i, y_i), i = 1, \dots, n$ be the n design points with corresponding responses. The cubic splines on $[x_1, x_n]$ with knots at $x_1, \dots, x_n$ are functions f that construct with third order polynomial f_i on each interval $[x_i, x_{i+1}], i = 1, \dots, n-1$. The functions f are continuous and also continuous on first and second derivatives over the entire interval $[x_1, x_n]$. Therefore, for each $i = 1, \dots, n-1$, one set of the spline coefficients of f is defined and derived for every x in the interval $[x_1, x_n]$.

$$f(x) = f_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3$$

The continuity of f and their first and second derivatives on $[x_1, x_n]$ lead to the continuity condition at each interior knots $x_k, k = 2, \dots, n-1$ which is given by

$$f_{n-1}(x_k) = f_k(x_k)$$

$$f'_{n-1}(x_k) = f'_k(x_k)$$

$$f''_{n-1}(x_k) = f''_k(x_k)$$

When predetermining the periodic end condition, the knots points at the boundary need to satisfy the condition as follow,

$$f(x_n) = f(x_1)$$

$$f'(x_n) = f'(x_1)$$

$$f''(x_n) = f''(x_1)$$

There is one thing to notice that the major difference between natural cubic smoothing splines and periodic cubic smoothing splines where natural cubic smoothing splines has natural end condition which refers to $f''(x_n) = f''(x_1) = 0$.

The smoothing splines f are derived from the determination of $4(n-1)$ coefficients of periodic smoothing splines with an arbitrary constant M. The deriving of the coefficients is given by f which gives the minimal of total curvature $G(f) = \int_{x_1}^{x_n} f''(x)^2 \, dx$ and f satisfies the weighted, distance-squared constraint with respect to the given points $H(f) = \sum_{i=1}^{n} \left[\frac{f(x_i)-y_i}{w_i}\right]^2 \leq M$. The weighting factor w_i gives the inverse importance to the points. Also, the M is the degree of smoothness which is analog to the smoothing penalty in natural cubic smoothing splines method. Further detail of the deriving the solution of the splines coefficients and the efficient calculation methods are also provided in this paper. This method shows the appropriate construction of the smoothing curves respect to periodic design points and yields visually acceptable closed curves as expected. This method can adept as few as only three design points to form a closed curve and up to 250 design points before the round-off error shows.

2.4.1 Classification using Principal Component Analysis

Principal Component Analysis (PCA) is one popular methods to overcome the problems of high-dimensionality and multicollinearity by reducing the number of predictor variables without losing predictive information [16]. It uses sophisticated underlying mathematical principles to transform a number of possibly correlated variables into a smaller set of variables, which could be functions of the original variables, called principal components [16]. The origin of PCA is from multivariate data analysis which is designed for dealing with the high-dimensional response variables. However, it now has a board range of other applications as well as adoptions by combining with other methods. PCA has been well known as one of the most important application of applied linear algebra [17], and possibly the most one of the most common uses of it is its role as the first step in analyzing large data sets.

The general idea of principal component analysis is to use a vector space transform to reduce the dimensionality of high-dimensional data sets. By applying mathematical projection, the original data set, which usually includes many variables, can often be summarized in just a few functions of variables, which are called the principal

components that account for most of the variability in the response variable. Under this dimension reduction approach, the analysis becomes far more easy and interpretable than would have been possible without performing the PCA.

In the following description, the dataset is represented by a matrix $\boldsymbol{X}$, which is a $N \times p$ matrix with p predictors and N p-variate observations. Principal component analysis transforms the set of p-variate inputs $\boldsymbol{X_1}, \ldots, \boldsymbol{X_N}$ into another set of vectors $\boldsymbol{T_1}, \ldots, \boldsymbol{T_N}$. $\boldsymbol{Ts}$ maintain most of the information in the original data through the principal component scores. Each principal component is a linear combination of the original variables and they are orthogonal to each other, which reduces the problem of multicollinearity. Such linear transformation of the matrix $\boldsymbol{X}$ is specified by a $p \times p$ matrix $\boldsymbol{P}$, so that the transformed variables $\boldsymbol{T}$ are given by:

$$\boldsymbol{T} = \boldsymbol{XP}$$

or alternatively $\boldsymbol{X}$ is given by

$$\boldsymbol{X} = \boldsymbol{TP^T}$$

where $\boldsymbol{P}$ is called the *loadings matrix*. The columns of the loadings matrix $\boldsymbol{P}$ can be calculated as the eigenvectors of the matrix $\boldsymbol{X^T X}$ [18].

Assume that the columns of $\boldsymbol{X}$, which are the original predictor variables, have been centered to have mean 0. The singular value decomposition (SVD) of $\boldsymbol{X}$ can be represented as

$$\boldsymbol{X} = \boldsymbol{UDV^T}$$

where $\boldsymbol{U}$, $\boldsymbol{D}$, and $\boldsymbol{V^T}$ are $N \times m$, $m \times m$, and $m \times p$, $m = min(N - 1, p)$ is the rank of $\boldsymbol{X}$, $\boldsymbol{D}$ is a diagonal matrix that contains the singular values d_j, the columns of $\boldsymbol{U}$ are the principal components $u_1, u_2, \ldots, u_m$ and they are assumed to be ordered, therefore $d_1 \geq d_2 \geq \cdots \geq d_m \geq 0$ [19]. Letting $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m)$ where $\boldsymbol{u}_1$ is called the first supervised principal component of **X**, and so on. Therefore, we can fit a univariate linear regression model with response **y** and predictor u_1

$$\hat{\boldsymbol{y}} = \gamma_0 + \boldsymbol{u}_1 \hat{\gamma} \tag{12}$$

For simplicity, the demonstration here uses only the first principal component but it can be generalized into more than one component. Then, $\boldsymbol{u_1}$ is a left singular vector with mean zero and unit variance. Therefore, $\hat{\gamma} = \boldsymbol{u}_1^T \boldsymbol{y}$ and the intercept is γ_0 is the mean of $\boldsymbol{y}$. Note that the number of principal components chosen can vary depending on the different criteria used to derive them and the proportion of variation explained by on the data. Consider the re-arranged SVD,

$$\boldsymbol{U} = \boldsymbol{XVD^{-1}} = \boldsymbol{XW} \tag{13}$$

Therefore, $\boldsymbol{u}_1$ is a linear combination of $\boldsymbol{X}$ and $\boldsymbol{u}_1 = \boldsymbol{X}\boldsymbol{w}_1$. The linear regression model now can be written as a linear model using all of the predictors $\boldsymbol{X}$ $\boldsymbol{as}$,

$$\begin{aligned} \hat{\boldsymbol{y}} &= \gamma_0 + \boldsymbol{X}\boldsymbol{w_1}\hat{\gamma} \\ &= \gamma_0 + \boldsymbol{X}\hat{\boldsymbol{\beta}} \end{aligned} \tag{14}$$

where $\hat{\boldsymbol{\beta}} = \boldsymbol{w_1}\hat{\gamma}$.

After building the model from learning data as above, the model could be used later for testing the effect of $\boldsymbol{X}$. The prediction of the response variable $\hat{\boldsymbol{y}}_j^*$, given the centered predictor $\boldsymbol{X}^*$ using the regression model have the same steps as in learning step. The prediction model can be written as

$$\hat{\boldsymbol{y}}_j^* = \gamma_0 + \boldsymbol{X}_j^* \boldsymbol{w}_1 \hat{\gamma} = \gamma_0 + \boldsymbol{X}_j^* \hat{\boldsymbol{\beta}} \tag{15}$$

The model is also applicable, under the generalized regression setting, for classification problem. This process also has two steps where principal component analysis are carries out in the first step [20], and, in the second step, the linear regression using the chosen principal components is used for classification.

In the binary case, the principal components analysis can be used using logistic regression. Considering the prediction model (14) above in the logistic regression setting,

the regression process is to find the coefficient vector $\boldsymbol{\beta}$ best that would be the best fit for training dataset. Let

$$y = 1, \ \ if\ \gamma_0 + \boldsymbol{X\beta} > 0$$
$$y = 0, otherwise$$

Then, the logistic regression model can be written as

$$\ln\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right) = \gamma_0 + \boldsymbol{X\beta} \tag{16}$$

$where\ p(\boldsymbol{X}) = p(y = 1)$. In generalized linear model (GLM), the maximum likelihood estimates of $\boldsymbol{\beta}$ are found by an iterative searching process using the observed $\boldsymbol{X}$ and **y**. However, instead of searching for best $\boldsymbol{\beta}$, the model can actually search for best γ since $\boldsymbol{\beta} = \gamma \boldsymbol{w}_c$ and $\boldsymbol{w}_c$ is the first c principal components transformation which accounts for most of the variation. The model can be re-written as

$$\ln\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right) = \gamma_0 + \boldsymbol{X}\boldsymbol{w}_c\gamma \tag{17}$$

Under supervised classification, γ, γ_0 and $\boldsymbol{w_c}$ are determined from the training dataset. The predicted classifications use the input from testing dataset $\boldsymbol{X}^*$ and the prediction model can be written as

$$\ln\left(\frac{p(\boldsymbol{X}^*)}{1-p(\boldsymbol{X}^*)}\right) = \gamma_0 + \boldsymbol{X}^*\boldsymbol{w}_c\gamma \tag{18}$$

For each testing sample, the predicted probability of membership $p(\boldsymbol{X_j^*})$ can be calculated and testing samples can be assigned as follows:

$$y_j^* = 1, \quad if\ \ p\left(\boldsymbol{X_j^*}\right) > 0.5$$
$$y_j^* = 0, \quad otherwise$$

2.4.2 Classification Using Support Vector Machines

Support vector machines (SVM), the foundations for which were developed by Vapnik [21], are very popular because of their many attractive features and promising empirical performance. Their formulation is constructed using the structural risk minimization (SRM) principle, which has been shown as superior [22] to the traditional empirical risk minimization (ERM) principle applied by conventional neural networks. Contrary to ERM, which minimizes the error on the learning dataset, SRM minimizes the upper bound on the VC dimension. This modification gives SVM a greater ability for generalization, which is always a desirable goal in statistical learning. SVM were developed to solve classification problems, but they have also recently been used to extend to the domain of regression problems [23]. Classification problems in SVM can be reduced to the consideration of a two-class problem without loss of generality. To solve this problem, the goal is to separate the two classes by a function that is derived from available samples. The ultimate goal is to determine a classifier that will perform well on testing examples and will also generalize well. There are many other linear classifiers that can be utilized to separate the data, but only one aims for maximizing the margin, which maximize the distance between the margin and the nearest data point for each class. Such a linear classifier is usually called the optimal separating hyperplane. One may consider the following setting for the data set.

The given learning data with n samples can be expressed by $(\overrightarrow{x_1}, y_1), \dots, (\overrightarrow{x_n}, y_n)$, where y_i is either 1 or -1 and indicates the group which the $\overrightarrow{x_i}$ belongs to. The $\overrightarrow{x_i}$'s are p-dimensional vectors, and $\vec{x}$ is the set of the points. One hyperplane can be written as

$$\overrightarrow{w} \cdot \vec{x} - b = 0$$

where $\vec{w}$ is the normal vector of the hyperplane; and the parameter $\frac{b}{\|\vec{w}\|}$ carries out the offset of the hyperplane from the origin with the $\vec{w}$. If the training dataset is linearly separable, which is referred to as the hard-margin approach, we would select two parallel hyperplanes to separate the two groups of data. In this way, we can make the distance between the two groups as large as possible. The ideal maximum-margin hyperplane can be described as

$$\vec{w} \cdot \vec{x} - b = 1$$

and

$$\vec{w} \cdot \vec{x} - b = -1$$

Therefore, the distance between the two hyperplanes is $\frac{2}{\|\vec{w}\|}$, and in order to maximize this distance, the SVM seeks the minimization of $\|\vec{w}\|$. Also, to prevent the data points from falling into the margin, the following constraints must be met:

$$\vec{w} \cdot \vec{x_i} - b \geq 1, \text{if } y_i = 1$$
$$\vec{w} \cdot \vec{x_i} - b \leq -1, \text{if } y_i = -1$$

and these constraints can be written as

$$y_i(\vec{w} \cdot \vec{x_i} - b) \geq 1, for\ all\ 1 \leq i \leq n$$

Consequently, the optimization problem becomes

$$minimize\ \|\vec{w}\|\ suject\ to\ y_i(\vec{w} \cdot \vec{x_i} - b) \geq 1, for\ i = 1, \dots, n$$

The $\vec{w}$ and b that solve this minimization problem will determine the classifier, $\vec{x} \rightarrow sign(\vec{w} \cdot \vec{x} - b)$.

For a nonlinear separable case, an extension of SVM that is usually called the soft-margin approach is considered, using the hinge loss function

$$\max(0,1 - y_i(\vec{w} \cdot \vec{x_i} - b))$$

This loss function is zero when the $\vec{x_i}$ is on the correct side of the margin or group. Otherwise, the value of the loss function is proportional to the distance from the margin. From here, minimization of

$$\frac{1}{n}\sum_{i=1}^{n} \max(0,1 - y_i(\vec{w} \cdot \vec{x_i} - b)) + \lambda\|\vec{w}\|$$

is desirable. λ is a predetermined value that can be derived once the classifier problem is decided. Generally, the parameter λ controls for the tradeoff between ensuring that the $\vec{x_i}$ falls on the correct side of margin and increasing the size of the margin.

For the primal problem, Cortes [24] introduced the non-negative variables δ_i and a penalty function of

$$F(\delta) = \sum_{i=1}^{n} \delta_i$$

where the δ are a measure of the misclassification error, in order to allow the optimal separating hyperplane method to be generalized. Therefore:

$$\delta_i = \max\big(0,1 - y_i(\vec{w} \cdot \vec{x_i} - b)\big)$$

if and only if

$$y_i(\vec{w} \cdot \vec{x_i} - b) \geq 1 - \delta_i$$

and δ_i is the smallest non-negative number. Thus, the minimization problem now is equivalent to minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\delta_i + \lambda\|\vec{w}\|$$

$$suject\ to\ y_i(\vec{w}\cdot\vec{x_i} - b) \geq 1 - \delta_i\ and\ \delta_i \geq 0, for\ i = 1, \ldots, n$$

The classical Lagrangian duality permits the primal problem to be transformed into a dual problem. By solving the Lagrangian dual for the problem above, the problem becomes

$$\max(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i\alpha_i(\vec{x_i}\cdot\vec{x_j})y_j\alpha_j$$

which satisfies the constraints $\sum_{i=1}^{n} y_i\alpha_i\vec{x_i} = 0$ and $0 \leq \alpha_i \leq \frac{1}{2n\lambda}$ for all i. Therefore, the variables α_i are defined as $\vec{w} = \sum_{i=1}^{n}\alpha_i y_i$. As the dual problem is a quadratic function optimization of the α_i with linear constraints, it is solvable by quadratic programming algorithms.

When under a situation for which a linear boundary is inappropriate, the SVM can be modified by mapping the input vector **x** into a high-dimensional feature space. Acceptable mappings include polynomials, radial basis functions, and certain sigmoid functions. By applying this mapping, which is also called the kernel trick, the optimization problem becomes

$$\max(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i\alpha_i k(\vec{x_i}\cdot\vec{x_j})y_j\alpha_j$$

where $k(\vec{x_i}\cdot\vec{x_j})$ is the kernel function performing the nonlinear mapping into the feature space; and the constraints are unchanged from the previous setting. Some common kernels include

- homogeneous polynomial: $k(\vec{x_i}\cdot\vec{x_j}) = (\vec{x_i}\cdot\vec{x_j})^d$
- inhomogeneous polynomial: $k(\vec{x_i}\cdot\vec{x_j}) = (\vec{x_i}\cdot\vec{x_j} + 1)^d$

- Gaussian radial basis function: $k(\overrightarrow{x_i} \cdot \overrightarrow{x_j}) = \exp\left(-\gamma \left|\left|\overrightarrow{x_i} - \overrightarrow{x_j}\right|\right|^2\right), for\ \gamma > 0.$

When we would like to train the nonlinear classification rule to correspond to a linear classification rule, we transform $\overrightarrow{x_i}$ into $\varphi(\overrightarrow{x_i})$ which satisfies $\overrightarrow{w} = \sum_{i=1}^{n} \alpha_i\, y_i \varphi(\overrightarrow{x_i})$. The kernels are related to the transformation which satisfies $k(\overrightarrow{x_i} \cdot \overrightarrow{x_j}) = \varphi(\overrightarrow{x_i}) \cdot \varphi(\overrightarrow{x_j})$.

The solution to this optimization problem is the same as the separable case using quadratic programming except for a modification of the bounds with the Lagrange multipliers. Once the α_i is solved, we then solve the offset b as

$$b = \overrightarrow{w} \cdot \varphi(\overrightarrow{x_i}) + y_i = \sum_{k=1}^{n} y_k \alpha_k k(\overrightarrow{x_k} \cdot \overrightarrow{x_i}) + y_i$$

When the new point comes in, all the information from the learning model will be stored and used, including $y_i \alpha_i, b, x_i$, and the choice of the kernel with the chosen parameter. The new point is

$$\vec{z} \rightarrow sign(\sum_{i=1}^{n} y_i \alpha_i k(\vec{z} \cdot \overrightarrow{x_i}) - b)$$

The remaining part of Cortes's approach is to determine the coefficient λ. The parameter provides additional capacity control within the classifier. In some situations, λ can be directly related to the regularization parameter [25, 26]. Blanz [27] uses a value of λ = 5, but ultimately λ must be chosen to reflect the knowledge of the noise on the data.

2.4.3 Classification Using Functional Generalized Linear Models

In multivariate covariates analysis, the famous generalized linear models (GLM) (McCullagh and Nelder [28]) are established to generalize the traditional linear regression by considering the linear model to be associated with a response variable Y and assuming the variable to be generated from a specific distribution in the exponential family, such as normal, binomial, and Poisson distributions in the form

$$p(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \tag{19}$$

We model the association between the predictor X and the response Y using

$$g(\mu) = \beta_0 + \beta_1^T X \tag{20}$$

where $\mu = E(Y;\theta,\phi) = b'(\theta)$; and g() refers to the link function in the generalized linear model.

The responses are connected through linear combination of the covariates using a pre-determined link function. For instance, the use of a logistic link for binary responses such as having a certain disease and the use of an identity link for continuous responses such as the number of patients in a treatment group are classic examples from medical research.

However, the GLM assumes that the predictors have a finite dimension. The functional generalized linear model (FGLM) extends the GLM to handle functional predictors that can be measured at different times as well as with different numbers of measurements for each individual. One of the features with these kinds of data sets is that measurements from the same individual will generally be correlated when predictors are functional. Many studies have invested a lot of effort on data with correlated outcomes, and various models have been established for the response variable. For example, Moyeed and Diggle [29] and Zeger and Diggle [30] model the relationship between a response $Y(t)$ and predictor $X(t)$ that are both measured over time, using the following model:

$$Y(t) = \alpha_0(t) + \beta_0^T X(t) + \epsilon(t)$$

where $\alpha_0(t)$ is a smooth function of t; β_0 is a fixed but unknown vector of the regression coefficients; and $\epsilon(t)$ is a zero mean stationary Gaussian process. This kind of model has been proposed to focus on situations where the predictors and responses are observed as pairs at the same time within a certain time period. However, under many situations, one

may wish to model the relationship between a single response and a functional predictor. For example, we might wish to construct a predictive model to determine whether an individual suffers from a mental disorder based on predictors measured over time. Also, we may like to compute the probability of a successful transplant operation based on measurements of a patient over time. In such situations, a single response is observed from the cases, but the predictors are functional because they are observed over time. The methods described above are not feasible or not easily applied to this kind of situation, because the models assume separate responses for each time that the predictors are observed. The functional generalized linear models are designed to directly model the relationship between a single response from any member of the exponential family of distributions and a functional predictor. The predicted coefficients for each individual are used in the linear portion of the link function to relate the predictors to the response. The general model of the functional generalized linear model between the predictor X and response Y, then, is

$$g(\mu) = \beta_0 + \int \omega_1(t)X(t)dt$$

where $\omega_1(t)$ is the functional analog of β_1 in the generalized linear model setting; $\mu = E(Y;\theta,\phi) = b'(\theta)$; and g() is the link function from the generalized linear model. The functional generalized linear model assumes that each predictor can be modeled as a smooth curve from a specific functional family. The model uses natural cubic splines (Silverman [31]; Green and Silverman [32]). The resulting parameterization is

$$X(t) = s(t)^T\gamma, \gamma \sim N(\mu_\gamma, \tau)$$

where $s(t)^T$ is the q-dimensional spline basis at time t; and γ is the q-dimensional spline coefficients for the predictors. μ_γ and τ are the mean and variance of the γ's. With the setting of $g(\mu)$ and $X(t)$, the final link function can be written as

$$g(\mu_j) = \beta_0 + \int \omega_1(t)s(t)^T\gamma_j dt = \beta_0 + \beta_1^T\gamma_j \qquad (21)$$

where $\beta_1^T = \int \omega_1(t)s(t)^T dt$. $j = 1, \ldots, N$; and N corresponds to the response-predictor pairs. From here, the functional generalized linear model can be written as

$$p(y_j; \theta_j, \phi) = \exp\left(\frac{y_j\theta_j - b(\theta_j)}{a(\phi)} + c(y_j, \phi)\right),$$

$$g(\mu_j) = \beta_0 + \beta_1^T\gamma_j, \gamma_j \sim N(\mu_\gamma, \tau), j = 1, \ldots, N$$

where N represents the number of response-predictor pairs; and the other notation and parameter settings can be found in formulas (19), (20), and (21).

When illustrating a situation in which the response is a Bernoulli variable, such as indicators of certain diseases, the probability $E(Y|X) = P(Y = 1|X)$ is modeled using possible link functions such as the probit or complementary log-log, just as with the generalized linear model. Under such situations, the canonical and most commonly applied link function—i.e., the logistic link function—under the functional generalized linear model becomes

$$Y_j = 1 \text{ with probability } \left(1 + \exp\left(-\beta_0 - \beta_1^T\gamma_j\right)\right)^{-1},$$

$$Y_j = 0 \text{ with probability } \left(1 + \exp\left(\beta_0 + \beta_1^T\gamma_j\right)\right)^{-1},\ \gamma_j \sim N(\mu_\gamma, \tau)$$

In general, the new predicted response is 1 if $E(Y_{new}|X_{new}) = P(Y_{new} = 1|X_{new}) > 0.5$; and 0 otherwise. Therefore, the predicted response $Y_{new} = 1$ if

$$\left(1 + \exp\left(-\beta_0 - \beta_1^T\mu_{\gamma|x_{new}}\right)\right)^{-1} > 0.5$$

The functional generalized linear models with a logistic link function can easily be applied to binary classification cases using functional predictors. This model also has the ability to be modified in order to accommodate problems in functional predictors caused by missing data or inconsistently observed time intervals between individuals.

The detailed computation and fitting procedure in the functional generalized linear model are explained in [33].

2.4.4 Classification Using Generalized Kernel Additive Models

Generalized additive models (GAM) (Hastie and Tibshirani [13]) are a typical extension of generalized linear models, when the linear predictor is not restricted as linear in the covariates but rather in the summation form of smoothing functions applied to the covariates. Generalized additive models are good for finding the balance between the flexibility and complexity of the model and can also serve as a great tool for practitioners to decide which model is more adequate.

To extend the generalized linear models to functional data, several previous works have explored using various aspects of approximations. Müller and Yao [34] construct an additive model utilizing a projection of the functional components on the eigen-basis of the covariance operator. Ferraty and Vieu [35] establish a two-step procedure to estimate an additive model of two functional predictors. Fan and James [36] propose a functional additive regression model, which further develops the functional linear regression using ideas from the penalized linear squares optimization approach. In general, the FGLM uses the concept of generalized linear models but replaces the linear combination of the covariates with the inner product in the functional space.

Another way to extend the GLMs for multivariate data is to express the systematic components as the sum of smooth functions, which is known as the general additive model (GAM). The extension using GAMs in functional data analysis has been identified as the generalized kernel additive model (GKAM) [33], and it is based on mixing the iteratively reweighted least squares and back-fitting algorithms adapted to the functional situation. This model establishes an algorithm to estimate several classes of regression models for functional data with responses in the exponential family.

The extension of generalized additive models to a functional context can be expressed by the following model. Suppose we have a functional predictor X and a response Y

$$E(Y|X) = \mu = g^{-1}(\eta_X) = g^{-1}\left(\beta_0 + \sum_{i=1}^{p} f_i(X_i)\right) \quad (22)$$

where $g()$ refers to the link function; f_j are the partial functions that need to be estimated; and X_i are the functional predictors, $i = 1, \ldots, p$. The estimation of the f_i functions is done by using functional kernel estimates of the partial functions and by considering the response as continuous.

The generalized kernel additive model follows the work by Ferraty and Vieu [35], in which the estimation of the f_j functions is constructed utilizing functional kernel estimates of the partial functions while considering the response as continuous. The proposed solution for the estimation of the f_j functions from Ferraty and Vieu [35] is a one-cycle conditional algorithm that considers one step for each functional covariate conditioned upon the previous estimation. The GKAM thus extends to situations where the responses belong to the exponential distribution family. Also, it further develops the algorithm so as to overcome situations in which there is not enough information either about the form of the link function for the response or about the shape of the partial functions. Specifically, the estimation of the partial functions in a nonparametric way makes the algorithm in the GKAM applicable to functional covariates in Banach spaces or in metric spaces.

The algorithm is designed to solve a broader class of models which widens the assumptions regarding the link in generalized additive models. Manuel [33] adapts the techniques proposed in Roca-Pardiñas et al. [37] that allow for a nonparametric estimation of the partial functions f_i and a joint nonparametric estimation of the inverse link $g^{-1} = H$, when the predictors are curves. Also, the algorithm is extended by using a back-fitting algorithm and estimation of the partial functions. The detailed estimation and prediction procedures are explained in the paper, along with additional details about the

selection of the kernel function, solutions for unknown link functions, and computational techniques. For a binary response, the logit link is considered as a typical link function. Therefore, the classification criteria can be expressed as

$$Y = 1 \; if \; g^{-1}\left(\beta_0 + \sum_{i=1}^{p} f_i(X_i)\right) > 0.5$$

$$Y = 0 \; if \; g^{-1}\left(\beta_0 + \sum_{i=1}^{p} f_i(X_i)\right) \leq 0.5$$

Regarding the fitting algorithm in the generalized kernel additive model, some concepts are discussed in the paper as well. When the probability is close to zero or one, the inverse of the weights at each back-fitting step can be arbitrary close to zero. Typically, those data are discarded once the weights are zero. If this situation happens for too many curves, the algorithm may try to estimate the partial functions without sufficient data. The solution proposed by the paper to avoid this is to stop the algorithm when the number of weights obviously far from zero is fewer than the equivalent amount of parameters of the estimator.

The detailed algorithm and estimation loop are described in Manuel [33], which also contains the method for estimating the link function from the exponential family and the unknown link function.

CHAPTER 3
PARAMETRIC AND NON-PARAMETRIC APPROACHES IN CIRCULAR-LINEAR DATA

Our major interest is to classify new observations into one of two populations while the time-dependent measures of these observations have similar patterns within a known time period, which is called circular-linear data. This approach aims for situations

where the difference in the functional readings is due to differences in pattern, while discounting the effect of other confounders. For instance, one may have an interest in the pattern of differences in blood pressure between two disease groups. The prior knowledge of some common confounders for blood pressure may include age, gender, and race, which the researcher would tend to control for. The solutions for such a concern are addressed in later sections, when we extend the curve estimation into its application to real data. Instead of using raw data for classification, we replace the readings for each individual with estimated curves that reduce the noise. From here, the curve estimations for each observation are performed separately under the parametric and non-parametric approaches, coupled with the distinct classification rules for each approach. For the non-parametric approach, periodic cubic smoothing splines under the Bayesian framework is proposed. For the purpose of comparison, a circular-linear regression model is considered as a parametric approach.

In the curve estimation stage, the general setting of the model can be described as follows. Considering a simple situation, suppose we have observations from two populations P_1 and P_2 with functional response measures $\boldsymbol{Y}_i^{(k)}, k = 1,2$. The sets of measures are $\boldsymbol{Y}_i^{(k)} = \left[Y_{i1}^{(k)}, Y_{i2}^{(k)}, \dots, Y_{im}^{(k)}\right]^T$. We define

$$Y_{ij}^{(k)} = f_i^{(k)}(\omega t_j) + \varepsilon_{ij}^{(k)} \tag{23}$$

where $f_i^{(k)}(\omega t_j)$ represents the target circular regression functions, which we estimate by either the parametric or non-parametric approach for populations P_1 and P_2, respectively; and $i = 1, \dots, n^{(k)}, j = 1, \dots, m$. $\omega = \frac{2\pi}{T}\ or \frac{360^o}{T}$ and T is the time period, representing the periodicity of responses, which is assumed to be known. Therefore, ω is also known. t_j is the jth time point up to the m time points that we observed, and both t_j and m are assumed to be known. $f_i^{(k)}(\omega t_j)$ are the circadian functions, which depend only on time t_j; and $\omega t_j = \omega t_j + q2\pi$ for any integer q. $\varepsilon_{ij}^{(k)}$ are the normal error terms, and we assume that $\varepsilon_{ij}^{(k)} \sim N(0, \sigma^{2^{(k)}})$ for the sake of simplicity. The estimated curves contain the

correlation between time points as a whole, which allows us to further assume independency among the residuals $\varepsilon_{ij}^{(k)}$.

For simplicity without losing generality, the model for curve estimation considers the population without any linear covariates or reference time for the proposed methods. For the application proposed in the extension work section, we add other linear independent variables as confounders. Also, the reference times, another essential concept in circular data, are used here to achieve a better comparison between estimated curves in real data. From here, we fit the circadian function $f_i^{(k)}(\omega t_j)$ for $\boldsymbol{Y}_i^{(k)}$ using the parametric or non-parametric approach for all response sets. T is the time period, representing the periodicity of the response, and we need to determine it by observing the behavior of data or our prior knowledge of the period. When we determine the time points $j = 1, \dots, m$, $T = m$ is decided so that the entire time points j have a length of 2π.

3.1 Non-Parametric Bayesian Periodic Cubic Smoothing Splines

The prior work by Graham [15] demonstrates the inappropriateness of fitting natural cubic smoothing splines to circular data and introduces an alternative end condition called the "periodic end condition" to achieve proper fitting of circular data. Here we pursue the Bayesian cubic smoothing splines for circular data by using the periodic end condition under the non-parametric scheme for each population. Under the Bayesian model for smoothing splines, we have all ordered time points $t_1 < t_2 < \cdots < t_m$, which are all used as knot points in Bayesian natural cubic smoothing splines. Writing this in the matrix form, we derive the cubic smoothing splines $\boldsymbol{f}$ by minimizing

$$(\boldsymbol{y} - \boldsymbol{f})^T(\boldsymbol{y} - \boldsymbol{f}) + \alpha \boldsymbol{f}^T \boldsymbol{K} \boldsymbol{f} \tag{24}$$

where $\boldsymbol{y}$ is the matrix of observed data; α is the smoothing penalty; $\boldsymbol{f^T K f}$ is in a quadratic form in the second derivative; and $\boldsymbol{K}$ is an inverse covariance matrix of $\boldsymbol{f}$.

We assign singular normal priors to $\boldsymbol{f}$ that have a mean of zero and a variance or $\alpha^{-1}\sigma^2\boldsymbol{K}^-$. Also, the smoothing parameter $\tau = \frac{\alpha}{\sigma^2}$ has the gamma prior, and the nuisance parameter σ^2 has the inverse gamma prior. For the construction of $\boldsymbol{K}$, we need to consider the continuity condition, which is specified in the cubic smoothing splines, and substitute the natural end condition with the periodic condition for the circular data. For each interval $(t_1, t_2), (t_2, t_3), \dots, (t_{m-1}, t_m)$, f_k represents the twice differentiable function for each interval (t_k, t_{k+1}). The continuity condition in all the interior knots points $t_2, \dots, t_{m-1}$ is given by

$$f_{k-1}(\omega t_k) = f_k(\omega t_k)$$

$$f'_{k-1}(\omega t_k) = f'_k(\omega t_k)$$

$$f''_{k-1}(\omega t_k) = f''_k(\omega t_k)$$

Also, the cubic spline f for interval (t_1, t_m) is periodic when it satisfies the periodic end condition, which is given by

$$f(\omega t_m) = f(\omega t_1)$$

$$f'(\omega t_m) = f'(\omega t_1)$$

$$f''(\omega t_m) = f''(\omega t_1)$$

Note that the original end condition in natural cubic smoothing splines is called the natural end condition, which refers to $f''(\omega t_m) = f''(\omega t_1) = 0$.

With the above periodic end condition, we derive the inverse covariance matrix $\boldsymbol{K}$ for the Bayesian periodic cubic smoothing splines, which gives a smooth and closed estimation of curves with a circular feature. The inverse covariance matrix $\boldsymbol{K}$ is now an $(m-1) \times (m-1)$ matrix instead of $m \times m$ in the natural end condition, as we have the additional constraint that the two ends of the curve have to be close. The construction of the $\boldsymbol{K}$ matrix uses the same idea as the suggested method in the Bayesian natural cubic

smoothing splines, with some modification to meet the periodic end condition. The detailed form of the matrix $\boldsymbol{K}$ can be described as

$$\boldsymbol{K} = \begin{bmatrix} k_{1,1} & k_{1,2} & 0 & & & & 0 & k_{1,m-1} \\ k_{2,1} & k_{2,2} & k_{2,3} & \cdots & & & & 0 \\ 0 & k_{3,2} & k_{3,3} & & & & & \\ & \vdots & & \ddots & & & \vdots & \\ & & & & k_{m-3,m-3} & k_{m-3,m-2} & 0 & \\ 0 & & & \cdots & k_{m-2,m-3} & k_{m-2,m-2} & k_{m-2,m-1} & \\ k_{m-1,1} & 0 & & & 0 & k_{m-1,m-2} & k_{m-1,m-1} & \end{bmatrix}$$

The elements in the $\boldsymbol{K}$ matrix having values other than 0 are the elements on diagonal ($k_{1,1},\ldots, k_{m-1,m-1}$), elements asides to diagonal (such as $k_{2,1}, k_{2,1}, \ldots, k_{m-1,m-2}, k_{m-2,m-1}$) and the two elements on the end of antidiagonal ($k_{m-1,1}$ and $k_{1,m-1}$). For the actual value in each element is

$$k_{1,1} = \frac{12}{(t_2 - t_1)^3} \times \left(1 + \frac{12}{(t_2 - t_1)^2}\right) + \frac{12}{(t_m - t_{m-1})^3} \times \left(1 + \frac{12}{(t_m - t_{m-1})^2}\right)$$

for other diagonal elements $k_{j,j}, j = 2, \ldots, m-1$,

$$k_{j,j} = \frac{12}{\left(t_{j+1} - t_j\right)^3} \times \left(1 + \frac{12}{\left(t_{j+1} - t_j\right)^2}\right) + \frac{12}{\left(t_j - t_{j-1}\right)^3} \times \left(1 + \frac{12}{\left(t_j - t_{j-1}\right)^2}\right)$$

and for the elements asides to diagonal,

$$k_{j-1,j} = k_{j,j-1} = -\frac{12}{\left(t_j - t_{j-1}\right)^3} \times \left(1 + \frac{12}{\left(t_j - t_{j-1}\right)^2}\right)$$

For the elements on the end of antidiagonal,

$$k_{m-1,1} = k_{1,m-1} = -\frac{12}{(t_m - t_{m-1})^3} \times \left(1 + \frac{12}{(t_m - t_{m-1})^2}\right)$$

For each individual used in training data and testing, we derive the periodic pattern using the Bayesian periodic cubic smoothing splines (BPCSS). For simpler

notation, we temporarily omit the k group notation for now. The same smoothing method is applied to each individual in both groups. The simplified model with response Y_{ij} , $i = 1, \dots, n$, and time point $t_j, j = 1, \dots, m$ is given by

$$Y_{ij} = g_i\ (\omega t_j) + \varepsilon_{ij} \tag{25}$$

where $g_i\ (\omega t_j)$ is the evaluated functional value of ωt_j for response Y_{ij} ; $\omega = \frac{2\pi}{T}\ or \frac{360^o}{T}$ and $\omega t_j = \omega t_j + q2\pi$ for any integer q; and ε_{ij} is the random error component.

When defining $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{im})^T$, we assume $\boldsymbol{Y}_i$ belongs to a multivariate normal distribution with a mean vector $\boldsymbol{g}_i = [g_i\ (\omega t_1), g_i\ (\omega t_2), \dots, g_i\ (\omega t_m)]$ and a diagonal variance-covariance matrix $\sigma_y^2 \boldsymbol{I}$. Note that m is the number of predetermined time points, which we assume to be a known constant for each case. $\boldsymbol{g}_i$ represents the circular explanatory variable, which can be treated as a function depending only on the discrete time t. $\boldsymbol{g}_i$ is estimated by BPCSS. $\epsilon_{ij} \sim N(0, \sigma_y^2)$ are the errors, which we assume to be uncorrelated between any two readings for one individual; i.e., $\sigma_{ab} = 0\ when\ a \neq b$. The correlation between readings within the same individual is controlled by the smoothing splines $\boldsymbol{g}_i$. The setting of the correlation between time points is explained next. Also, all individuals are also assumed to be independent of each other.

Under the Bayesian framework, we assign prior distributions for all the unknown parameters of interest to us. The prior for σ_y^2 is an inverse gamma distribution, which is denoted by $\sigma_y^2 \sim IG(A_\sigma, B_\sigma)$. The prior for g_i is a singular normal distribution, which is denoted by $\boldsymbol{g_i} \sim SN\left(0, \alpha_i^{-1} \sigma_y^2 \boldsymbol{K}^-\right)$, where α_i is the smoothing penalty and the smoothing parameter $\tau_i = \alpha_i / \sigma_y^2$. The prior for τ_i is given by a gamma distribution, which is denoted by $\tau_i \sim G(A_\tau, B_\tau)$. All hyperparameters are assumed to be known and predetermined so the priors are proper.

The $\boldsymbol{K}$ is an $m-1 \times m-1$ matrix with a rank $m-2$ which satisfies $\int [g_i''(t)]^2 dt$; and $\boldsymbol{K}^-$ is the generalized inverse of $\boldsymbol{K}$. The periodic end condition is predetermined in the $\boldsymbol{K}$ matrix, which yields the smoothed closed curve for estimation.

The likelihood of joint posterior distribution is proportional to

$$\propto \left(\sigma_y^2\right)^{-\frac{m}{2}} \exp\left[-\frac{1}{2\sigma_y^2}(\mathbf{Y}_i - \mathbf{g}_i)^T(\mathbf{Y}_i - \mathbf{g}_i)\right]$$

$$\times \tau_i^{\frac{m-2}{2}} \exp\left\{\frac{-\tau_i}{2}\boldsymbol{g}_i^T\boldsymbol{K}\boldsymbol{g}_i\right\}$$

$$\times \tau_i^{A_\tau - 1} \exp[-(\tau_i B_\tau)]$$

$$\times \left(\sigma_y^2\right)^{-(A_\sigma+1)} \exp(-\frac{B_\sigma}{\sigma_y^2})$$

To derive the full conditional posterior for $\boldsymbol{g}_i$,

$$\boldsymbol{g}_i|\cdot \propto \left(\sigma_y^2\right)^{-\frac{m}{2}} \exp\left[-\frac{1}{2\sigma_y^2}(\mathbf{Y}_i - \mathbf{g}_i)^T(\mathbf{Y}_i - \mathbf{g}_i)\right]$$

$$\times \tau_i^{\frac{m-2}{2}} \exp\left\{\frac{-\tau_i}{2}\boldsymbol{g}_i^T\boldsymbol{K}\boldsymbol{g}_i\right\}$$

$$\propto \exp\left[-\frac{1}{2\sigma_y^2}(\mathbf{g}_i)^T(\mathbf{g}_i) + \frac{-\tau_i}{2}\boldsymbol{g}_i^T\boldsymbol{K}\boldsymbol{g}_i - 2\boldsymbol{g}_i^T\mathbf{Y}_i\right]$$

$$\propto \exp\left[-\frac{1}{2\sigma_y^2}\mathbf{g}_i{}^T(\mathbf{I} + \alpha_i\boldsymbol{K})\mathbf{g}_i - 2\boldsymbol{g}_i^T\mathbf{Y}_i\right]$$

where $\alpha_i = \tau_i\sigma_y^2$ or $\tau_i = \alpha_i/\sigma_y^2$. We therefore have the full conditional posterior for $\boldsymbol{g}_i$ as

$$\boldsymbol{g}_i|\cdot \sim Normal((\boldsymbol{I} + \alpha\boldsymbol{K})^{-1}\boldsymbol{Y}_i, (\boldsymbol{I} + \alpha\boldsymbol{K})^{-1}\sigma_y^2)$$

To derive the full conditional posterior for τ_i,

$$\tau_i|\cdot \propto \tau_i^{\frac{m-2}{2}} \exp\left\{\frac{-\tau_i}{2}\boldsymbol{g}_i^T\boldsymbol{K}\boldsymbol{g}_i\right\} \times \tau_i^{A_\tau - 1} \exp[-(\tau_i B_\tau)]$$

$$\propto \tau_i^{\frac{m-2}{2} + A_\tau - 1} \exp\left\{-\tau_i[\frac{\boldsymbol{g}_i^T\boldsymbol{K}\boldsymbol{g}_i}{2} + B_\tau]\right\}$$

We therefore have the full conditional posterior for τ_i as

$$\tau_i|\cdot \sim Gamma(\frac{m-2}{2} + A_\tau, \frac{\boldsymbol{g}_i^T\boldsymbol{K}\boldsymbol{g}_i}{2} + B_\tau)$$

To derive the full conditional posterior for σ_y^2,

$$\sigma_y^2|\cdot \propto \left(\sigma_y^2\right)^{-\frac{m}{2}} \exp\left[-\frac{1}{2\sigma_y^2}(\mathbf{Y}_i-\mathbf{g}_i)^T(\mathbf{Y}_i-\mathbf{g}_i)\right]$$

$$\times \left(\sigma_y^2\right)^{-(A_\sigma+1)} \exp(-\frac{B_\sigma}{\sigma_y^2})$$

$$\propto \left(\sigma_y^2\right)^{-\frac{m}{2}+(A_\sigma+1)} \exp\left[-\frac{1}{\sigma_y^2}[\frac{(\mathbf{Y}_i-\mathbf{g}_i)^T(\mathbf{Y}_i-\mathbf{g}_i)}{2}+B_\sigma\right]$$

We therefore have the full conditional posterior for σ_y^2 as

$$\sigma_y^2|\cdot \sim Inverse\ Gamma(\frac{m}{2}+A_\sigma, \frac{(\mathbf{Y}_i-\mathbf{g}_i)^T(\mathbf{Y}_i-\mathbf{g}_i)}{2}+B_\sigma)$$

As we have the full conditional posterior distribution available, we use Gibb's sampler to collect the parameters and we make inferences among those parameters by Bayes' estimator. In the later simulation and real data estimation, 1,000 iterations are collected and the burn-in process discards the first 500 iterations. We also expect a huge inverse in the cubic smoothing splines, and the Cholesky decomposition is applied.

3.2 Extension Work to NHANES Data Analysis in Bayesian Periodic Cubic Smoothing Splines

One of the most important features in circular data is the estimation of the reference time. We assume the observations from the same population follow the same pattern in the circular data and the reference time is the time point that represents the maximum reading. For instance, in their daily activities, people can have the same pattern of activity readings but different reference times or peak times because of habit, work shifts, or living area. People can have the same 8 hour job with different starting times and have the same activity and sleeping pattern, but their maximum activity times are shifted by the difference between the starting times. Therefore, the alignment of functional time effects can certainly contribute to the accuracy of the estimation and allow for better inferences about the overall activity pattern in the particular group.

Other than the reference time, it is common to consider other effects such as age, gender, and other information that may serve as confounders in real data. These effects have an additive effect on the curves which can rise, change magnitude, or shift the pattern that we observe and may also produce more noise which may or may not change over time. In order to make a more accurate interpretation of the only-time-dependent pattern, it is natural to include those effects in the model. We consider the confounding effects here as linear variables and the only-time-dependent pattern as the circular variable. Our ultimate goal is to classify new observations into one of the populations using only the circular variable part. Therefore, a different group of individuals is modeled separately, and we will make an inference based on the difference in the linear predictors and the time effects. Under the additive regression model setting, each estimated curve is established, which yields an additive model with the linear predictors and the periodic predictor. In real data, one may consider the classification to depend only on, or to make sense only for, the periodic pattern of the time effect and to treat other linear effects as confounders. Moreover, comparisons between populations and new observations are based only on the shape of the patterns, not on shifts among the patterns. The corresponding concept in the circular-linear regression model is to estimate the reference time for each curve as well as the other independent variables. For the application proposed herein, we consider the circular-linear regression model with a response $\boldsymbol{Y}_i^{(k)} = \left[Y_{i1}^{(k)}, \dots, Y_{im}^{(k)}\right], i = 1, \dots, n^{(k)}$; time points $t_j, j = 1, \dots, m$; and linear predictor $\boldsymbol{X}_i^{(k)} = [X_{i1}^{(k)}, \dots, X_{ip}^{(k)}]$ such that

$$Y_{ij}^{(k)} = X_{i1}^{(k)}\beta_1^{(k)} + \cdots + X_{ip}^{(k)}\beta_p^{(k)} + g_i^{(k)}\left(\omega t_j - \omega r_i\right) + \varepsilon_{ij}^{(k)} \tag{26}$$

where $\beta_i, \dots, \beta_p$ are linear coefficients. The superscript (k) represents the model used to estimate each population $k = 1,2$; $g_i^{(k)}(\omega t_j - \omega r_i)$ are the functions of $\omega t_j - \omega r_i$; $\omega = \frac{2\pi}{T}\ or \frac{360^o}{T}$ and $\omega(t_j - r_i) = \omega(t_j - r_i) + q2\pi$ for any integer q; r_i is the reference direction or acrophase of the $g_i^{(k)}$ function; and $\varepsilon_{ij}^{(k)}$ is the random error component. Also, we assume the linear predictor $\boldsymbol{X}_i^{(k)}$ and the circular predictor are mutually independent. More specifically, $0 \le \omega t_i \le 2\pi$ and $-\omega r_i \le \omega\left(t_j - r_i\right) \le \omega\left(t_j - r_i\right) +$

2π. Therefore, the t_j here is the reference number for each time point $[t_1, t_2, \dots, t_m] = [1,2, \dots, m]$. We can thus extract the periodic feature of the responses $Y_{ij}^{(k)}$ and also determine the effects of other linear independent variables. For simplicity, we suppress the superscript in the model description provided later on.

In the non-parametric approach, we consider the Bayesian regression model combined with Bayesian periodic cubic smoothing splines. Therefore, the prior distributions are assigned to both unknown parameters in the linear covariates and periodic cubic smoothing splines. The regression model is constructed under the Bayesian framework. For shorter notation, let $\boldsymbol{g_i}$ represent the functional values $[g_i(\omega t_1 - \omega r_i), \dots, g_i(\omega t_m - \omega r_i)]^T$. The likelihood distribution for $\boldsymbol{Y}_i$ is a multivariate normal distribution, which is noted by $\boldsymbol{Y}_i \sim MVN(\boldsymbol{1X_i\beta} + \boldsymbol{g}_i, \sigma_y^2\boldsymbol{I})$. The $\boldsymbol{1X_i\beta}$ matrix here is an $m \times 1$ matrix with the same elements $\boldsymbol{X_i\beta}$, which affect the response equally at each time point by shifting $\boldsymbol{g}_i$. The prior distribution for σ_y^2 is an inverse gamma distribution, which is noted by $\sigma_y^2 \sim IG(A_\sigma, B_\sigma)$. For the sake of simplicity without losing generality, we use a flat prior for β such as $\boldsymbol{\beta} \sim MVN(0, \sigma_\beta \boldsymbol{I})$, while $\sigma_\beta = 100$ is often used, and $\boldsymbol{\beta}$ is a $P \times 1$ matrix. This setting is commonly used when no prior information is available about the linear predictive covariate in the Bayesian linear regression. For $\boldsymbol{g}_i$, α_i is the smoothing penalty, and the smoothing parameter $\tau_i = \alpha_i/\sigma_y^2$ as before. In BPCSS, $\boldsymbol{g}_i$ has a singular normal prior, which is denoted by $\boldsymbol{g}_i \sim SN(0, \alpha_i^{-1}\sigma_y^2\boldsymbol{K}^-)$. The prior for τ_i is a gamma distribution, which is denoted by $\tau_i \sim G(A_\tau, B_\tau)$. Similarly, $\boldsymbol{K}$ is an $m \times m$ matrix with a rank $m - 2$ which satisfies $\int [g_i''(t)]^2 dt$; and $\boldsymbol{K}^-$ is the generalized inverse of $\boldsymbol{K}$. The periodic end condition is predetermined in the $\boldsymbol{K}$ matrix, which yields the smooth and closed curve for estimation. All hyperparameters are assumed to be known and predetermined, so the priors are proper.

We need to derive the full conditional posterior distribution for Bayes' estimator. The joint posterior distribution is proportional to

$$\propto \prod_1^n \left(\sigma_y^2\right)^{-\frac{m}{2}} \exp\left[-\frac{1}{2\sigma_y^2}(\boldsymbol{Y}_i - \boldsymbol{1X_i\beta} - \boldsymbol{g}_i)^T(\boldsymbol{Y}_i - \boldsymbol{1X_i\beta} - \boldsymbol{g}_i)\right]$$

$$\times \left(\sigma_{\beta}^{2}\right)^{-\frac{P}{2}} \exp\left[-\frac{1}{2\sigma_{\beta}^{2}}(\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{\beta})\right]$$

$$\times \prod_{1}^{n}[\tau_{i}^{\frac{m-2}{2}} \exp\left\{\frac{-\tau_{i}}{2}\boldsymbol{g_i^T K g_i}\right\}$$

$$\times \tau_{i}^{A_{\tau}-1}\exp\{-(\tau_{i}B_{\tau})\}]$$

$$\times \left(\sigma_{y}^{2}\right)^{-(A_{\sigma}+1)} \exp\left(-\frac{B_{\sigma}}{\sigma_{y}^{2}}\right)$$

We next consider the full conditional posterior for each unknown parameter. To derive the full conditional posterior for $\boldsymbol{\beta}$, we use

$$\boldsymbol{\beta}|\cdot \propto \prod_{1}^{n} \exp\left[-\frac{1}{2\sigma_{y}^{2}}(\boldsymbol{Y}_{i}-\boldsymbol{1X_i\beta}-\boldsymbol{g}_{i})^{T}(\boldsymbol{Y}_{i}-\boldsymbol{1X_i\beta}-\boldsymbol{g_i})\right]$$

$$\times \exp\left[-\frac{1}{2\sigma_{\beta}^{2}}(\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{\beta})\right]$$

After rearrangement, we have

$$\boldsymbol{\beta}|\cdot \propto \exp\{-\frac{1}{2}\left[\boldsymbol{\beta}^{T}\left(\frac{1}{\sigma_{\beta}^{2}}\boldsymbol{I}+\frac{1}{\sigma_{y}^{2}}\sum\boldsymbol{X_i^T 1^T 1X_i}\right)\boldsymbol{\beta}-2\boldsymbol{\beta}^{T}[\frac{1}{\sigma_{y}^{2}}\sum\boldsymbol{X_i^T 1^T}(\boldsymbol{Y_i}-\boldsymbol{g_i})]\right]\}$$

We therefore designate the full conditional posterior for β as

$$\boldsymbol{\beta}|\cdot \sim N(\left(\frac{1}{\sigma_{\beta}^{2}}\boldsymbol{I}+\frac{1}{\sigma_{y}^{2}}\sum\boldsymbol{X_i^T 1^T 1X_i}\right)^{-1}[\frac{1}{\sigma_{y}^{2}}\sum\boldsymbol{X_i^T 1^T}(\boldsymbol{Y_i}-\boldsymbol{g_i})],\left(\frac{1}{\sigma_{\beta}^{2}}\boldsymbol{I}+\frac{1}{\sigma_{y}^{2}}\sum\boldsymbol{X_i^T 1^T 1X_i}\right)^{-1})$$

To derive the full conditional posterior for each and every $\boldsymbol{g_i}$, we use

$$\boldsymbol{g_i}|\cdot \propto \prod_{1}^{n} \exp\left[-\frac{1}{2\sigma_{y}^{2}}(\boldsymbol{Y_i}-\boldsymbol{1X_i\beta}-\boldsymbol{g_i})^{T}(\boldsymbol{Y_i}-\boldsymbol{1X_i\beta}-\boldsymbol{g_i})\right]$$

$$\times \prod_{1}^{n}[\tau_{i}^{\frac{m-2}{2}} \exp\left\{\frac{-\tau_{i}}{2}\boldsymbol{g_i^T K g_i}\right\}$$

After rearrangement, we have

$$\boldsymbol{g_i}|\cdot \propto \exp\{-\frac{1}{2}\left[\boldsymbol{g_i^T}\left(\tau_i\boldsymbol{K}+\frac{1}{\sigma_y^2}\boldsymbol{I}\right)\boldsymbol{g}_i\right]-2\boldsymbol{g_i^T}(\frac{1}{\sigma_y^2}(\boldsymbol{Y_i}-\boldsymbol{1X_i\beta}))\}$$

If we pull σ_y^2 out, we would have $\exp -\frac{1}{2\sigma_y^2}\{[\boldsymbol{g_i^T}(\alpha_i\boldsymbol{K}+\boldsymbol{I})\boldsymbol{g_i}]-2\boldsymbol{g_i^T}(\boldsymbol{Y_i}-\boldsymbol{1X_i\beta})\}$, where $\alpha_i=\tau_i\sigma_y^2$.

We therefore have the full conditional posterior for the $\boldsymbol{g}_i$'s as

$$\boldsymbol{g_i}|\cdot \sim N((\alpha_i\boldsymbol{K}+\boldsymbol{I})^{-1}(\boldsymbol{Y_i}-\boldsymbol{1X_i\beta}),(\alpha_i\boldsymbol{K}+\boldsymbol{I})^{-1}\sigma_y^2)$$

for each $\boldsymbol{g}_i$. Note that the τ_i is paired with $\boldsymbol{g}_i$; therefore, it yields a different value for every individual i. However, $\boldsymbol{K}$ would be the same for each individual i, as long as each individual has the same time point and the same time interval. Otherwise, different $\boldsymbol{K}$'s are needed for each individual.

To derive the full conditional posterior for each τ_i corresponding to $\boldsymbol{g_i}$, we use

$$\tau_i|\cdot \propto \tau_i^{\frac{m-2}{2}}\exp\left\{\frac{-\tau_i}{2}\boldsymbol{g_i^T K g_i}\right\}\times\tau_i^{A_\tau-1}\exp\{-(\tau_i B_\tau)\}$$

$$\propto \tau_i^{\frac{m-2}{2}+A_\tau-1}\exp\left\{-\tau_i[\frac{\boldsymbol{g_i^T K g_i}}{2}+B_\tau]\right\}$$

We therefore have the same full conditional posterior as before for τ_i, namely

$$\tau_i|\cdot \sim Gamma(\frac{m-2}{2}+A_\tau,\frac{\boldsymbol{g_i^T K g_i}}{2}+B_\tau)$$

For σ_y^2, we derive the full conditional posterior from full conditional prior as

$$\sigma_y^2|\cdot \propto \prod_1^n\left(\sigma_y^2\right)^{-\frac{m}{2}}\exp\left[-\frac{1}{2\sigma_y^2}(\boldsymbol{Y_i}-\boldsymbol{1X_i\beta}-\boldsymbol{g_i})^T(\boldsymbol{Y_i}-\boldsymbol{1X_i\beta}-\boldsymbol{g_i})\right]$$

$$\times\left(\sigma_y^2\right)^{-(A_\sigma+1)}\exp\left(-\frac{B_\sigma}{\sigma_y^2}\right)$$

After arrangement, we have

$$\sigma_y^2|\cdot \propto \left(\sigma_y^2\right)^{-\frac{nm}{2}+A_\sigma-1} \exp\left[-\frac{1}{\sigma_y^2}\left(\frac{[\sum(\boldsymbol{Y_i}-\boldsymbol{1X_i\beta}-\boldsymbol{g_i})^T(\boldsymbol{Y_i}-\boldsymbol{1X_i\beta}-\boldsymbol{g_i})]}{2}+B_\sigma\right)\right]$$

We therefore have the full conditional posterior for σ_y^2 as

$$\sigma_y^2|\cdot \sim IG\left(\frac{nm}{2}+A_\sigma, \frac{[\sum(\boldsymbol{Y_i}-\boldsymbol{1X_i\beta}-\boldsymbol{g_i})^T(\boldsymbol{Y_i}-\boldsymbol{1X_i\beta}-\boldsymbol{g_i})]}{2}+B_\sigma\right)$$

After deriving all the full conditional posterior distributions, we use Gibb's sampler to collect MCMC samples and estimate the parameters using Bayes' estimators. After taking a closer look at the algorithm, one may notice that the procedures for collecting MCMC samples are very similar to the back-fitting procedure in the multivariate regression approach. Indeed, they are almost the same in terms of the algorithm, except that our algorithm is performed under a full Bayesian framework. The back fittings are iterated within the algorithm. Instead of finding the convergence of the unknown parameters, the Bayesian approach utilizes the Bayesian estimator and thus provides a different concept of the variation estimation. In both the simulation and the NHANES data analysis, we consider the non-informative priors as

$$\boldsymbol{g}_i \sim SN(0, \alpha_i^{-1}\sigma_y^2\boldsymbol{K}^-).$$

$$\sigma_y^2 \sim IG(1,5)$$

$$\tau_i \sim G(1, 0.001)$$

$$\boldsymbol{\beta} \sim MVN(0, 100\boldsymbol{I})$$

By finding the estimated curve for each individual and the point estimation for the overall linear covariates, we can now proceed to determine the essential feature in our circular-linear data analysis, the reference time. To estimate the reference time r_i for an individual i, we consider following argument minimum:

$$r_i = argmin_{r_i}\left[\int_i \left(g_1(\omega t_j\ ) - g_i(\omega t_j\ - \omega r_i)\right)^2 dt\right] \qquad (27)$$

When the time points have condensed enough, the integration (27) above can be approximated by summing up the area between two curves within each interval

$[t_1, t_2], .., [t_{m-1}, t_m], [t_m, t_1]$. The argument minimum is achieved by searching through all possible reference time points, such as $r_i = t_1, r_i = t_2, \dots, r_i = t_m$. The reference time r_i for the i-th individual is determined as one of the observed time points t_z, when $r_i = t_z$ minimizes the argument.

Therefore, when r_i is derived for each individual, all curves are aligned to the curve of the first observation by changing the starting time point t_1 to r_i, and such manipulation is called alignment. For example, the original time point starting at t_1, $[t_1, t_2, \dots, t_m]$, is changed to $[t_z, t_{z+1}, \dots, t_m, t_1, t_2, \dots, t_{z-1}]$ when $r_i = t_z$. The responses corresponding to the time points are also changed in the same manner. The concept of alignment is used to accommodate the idea of analysis of the pattern of response. For individuals who have the same response pattern with different reference times, such individuals can provide totally different information for analysis of the pattern when alignments are not performed in advance. For example, two individuals are in the analysis who have exactly the same activity pattern during a day except that one works a day shift and the other works a night shift. Those two individuals should provide the same information for the analysis of the pattern. However, without alignment, they would provide completely opposite information.

Under the Bayesian framework, we also propose an additional classification method that utilizes the MCMC samples we collected in the BPCSS. This classification method is based on the area difference between two curves. Therefore, the fundamental idea of this method is to determine which learning curve the new curve is closer to and then to decide the membership of the new curve or the individual. This closeness can be determined differently depending on various methods. The method we consider uses the area between two curves to determine the closeness, which is straight forward and reasonable. Once we estimate the curves for each individual, we use the estimated curves to estimate the group mean curves. For each subject, Bayesian smoothing splines provide a vector of functional values of the fitted curve, which are evaluated at given time points through MCMC samples. We estimate the individually fitted curve by connecting the components of the vector with linear interpolation. The group mean curve of that population is derived after the alignment is made. The classification rules in the non-

parametric approach can be directly applied to the group mean curves for each population and the aligned curves of new observations. Also, when we interested in classification solely based on the periodic features and treat the linear independent variables as confounders, we discount the effects of the linear independent variables as well as the reference time shifting before we proceed to the classification problem. Therefore, the testing individuals use the estimated coefficients for linear independent variables in the comparisons with the learning mean curve model.

For the group mean curves, we first take the mean of the functional evaluations of the universal time points throughout all individuals, such that

$$\bar{g}^{(k)}(t) = \left[\frac{1}{n^{(k)}} \sum_{i=1}^{n^{(k)}} g_i^{(k)}(\omega t_1), \frac{1}{n^{(k)}} \sum_{i=1}^{n^{(k)}} g_i^{(k)}(\omega t_2), \dots, \frac{1}{n^{(k)}} \sum_{i=1}^{n^{(k)}} g_i^{(k)}(\omega t_m) \right]^T,$$

We use the linear interpolations, which we denote it as $\bar{g}^{(k)}(t)$, to estimate the group mean curve. When we have the condensed time points, the linear interpolations are the approximation of a smooth interpolation. To obtain the MCMC samples for the group mean curves, we collect the MCMC samples by choosing one sample among the MCMC samples for each individual curve and calculate the averages for the collected samples at each time point. We use the group mean curves and their corresponding MCMC samples for the investigation of the group differences in a Bayesian approach. For instance, we collect 1,000 MCMC samples among 300 individuals for each population by using Bayesian periodic cubic smoothing splines. By finding the MCMC samples among the 300 individuals, we obtain 1,000 iterations of group mean curves for each population.

The classification rule, which uses the area between the curves, is based on the functional classification rules. When a new observation y_0 comes in, we use the same approach to get the fitted curve $g_0(t)$, then proceed to classification. We use the MCMC samples from the group mean curves to achieve the probability of correct classification. Using the MCMC samples we collected, the classification rule applies to each iteration and is given by

Classify y_0 to G_1 if

$$\int \left(\bar{g}^{(1)}(t) - g_0(t)\right)^2 dt < \int \left(\bar{g}^{(2)}(t) - g_0(t)\right)^2 dt$$

Classify y_0 to G_2 otherwise.

where $\bar{g}^{(1)}(t)$ and $\bar{g}^{(2)}(t)$ are the group mean curves for the one MCMC sample, derived from populations G_1 and G_2. y_0 is the new observation, with a fitted curve of $g_0(t)$. The integration can be approximated by summing up the area between the two curves.

This approach is relatively straightforward, but we can certainly consider a proper weight λ on the right-hand side of the equation, such as $\lambda\int \left(\bar{g}^{(2)}(t) - g_0(t)\right)^2 dt$. For instance, a lot of classification methods consider the cost of misclassification. The λ here can be a function of the misclassification cost intended to optimize the classification result. We do not consider the misclassification cost at this point, but it can certainly be further developed when the misclassification cost is a concern.

In the non-parametric approach under the Bayesian framework, the probability of correct classification can be derived from MCMC samples from both the population and the new observation. Among the MCMC iteration samples that we collected with the periodic cubic smoothing splines method, one mean curve for each population can be derived from samples of one iteration within that population. For each iteration, we apply the classification rule above once and determine the membership of the new individual. Out of all the membership decisions for each iteration, we can calculate the probability of correct classification by using the proportion of classifications among all iterations. For example, we have 1,000 iterations of mean curves from each population and one mean curve from the new observation. The classification rule provides one decision for each iteration. Supposing we have 900 iterations out of the 1,000 that classify y_0 to G_1, we can say that the probability of classifying y_0 to G_1 is 0.9 and the probability of classifying y_0 to G_2 is 0.1. We can then decide the membership of y_0 based on the arbitrary probability. The threshold of the classification probability will generally be 0.5 if no other prior information or cost of misclassification is given.

The method described here of using the area between curves gives a general classification approach that can be applied under a Bayesian framework. Through the MCMC samples we collect, the further developed classification methods can be considered to replace area calculations such as the voting feature classification. When there is an ideally customized classification rule that only gives straight membership instead of probability, such a classification method can always be replicated and validated under the Bayesian setting to produce a reasonable probability of classification.

3.3 Parametric Approach Using Circular-Linear Regression Model

For fitting the curves under a parametric setting, we consider the multivariate circular-linear regression model presented in SenGupta and Ugwuowo [6]. We assume the response at the jth time point in the k population $Y_{ij}^{(k)}(t)|t = j$ is normally distributed, with a mean $\mu_{t|t=j}$ and a variance σ_{μ}^2 for all $t_j,\ i = 1, \dots, n^{(k)},\ j = 1, \dots, m$. If $Y_{ij}^{(k)}(t)$ is not normally distributed, a transformation such as log transformation is considered before proceeding to the regression problem. For each stratified population k, the circular-linear regression is applied for the population, and the polynomial circadian regression function is given by

$$Y_{ij}^{(k)} = A_i + B_i\, cos(\omega t_j) + C_i\, sin(\omega t_j) + \epsilon_{ij} \qquad (28)$$

where A_i, B_i, and C_i are the regression coefficients; $\omega = \frac{2\pi}{T}\ or \frac{360^o}{T}$; ϵ_{ij} is the random error component; and $T = m$ is decided in order to make the total of the time points j has a length of 2π.

We consider least square estimation (LSE) to estimate the regression coefficients. Under the normality assumption, least square estimators are equivalent to maximum likelihood estimators (MLEs), and we can use this advantage to make inferences about the coefficient estimators, if needed. Several standard diagnosis tools are available to

assess the goodness-of-fit for the model. In our simulation, we use the model introduced above, which is appropriate for a close curve with one peak. However, in NHANES data analysis, we can further consider constructing a polynomial regression function in order to obtain a better fit of the data using the suggested function introduced in section 2.1.1. Details of the extension of the model used are explained in section 3.4.

In the parametric approach, we estimate the functions for each individual using the circular-linear regression model and treating universal time points as the predictor variable. Unlike the ordinary approach used in the linear regression to find one model for one population k, we find one model for each individual i and find the best fit with respect to the observations from that individual. The predicted values of $Y_{ij}^{(k)}$, $\hat{Y}_{ij}^{(k)}$, are obtained from the regression model and used as the inputs for the classification procedures later on. The major purpose of the regression fitting is to reduce the noise and thus improve the performance of the classifications.

Under the supervised classification setting, we select individuals from each group, the normal group and the insomnia group, for which we are already aware of their membership in the population; we refer to those individuals as the learning samples. In contrast, the testing samples are treated as new individuals who must be classified into one of the populations. For both the learning and testing samples, we consider the regression model (27) for each individual and obtain the predicted values $\hat{Y}_{ij}^{(k)}$ for each time point t_j as the input information for the classification methods.

Further development of the methodology for this extension of the circular-linear regression model is considered in the work for the NHANES data analysis. This extension includes a change in the circadian function of the regression model based on the suggestion in SenGupta and Ugwuowo [6]. A detailed approach is presented in the next section.

3.4 Extension Work to NHANES Data Analysis in Circular-Linear Regression Model

In real data analysis, one may consider the classification to depend only on, or to make sense only for, the periodic patterns of the time effect. In the extension work for the parametric approach, we focus on the model fitting procedure in order to find the predicted values and yield a better performance in the classification procedures. By observing the patterns of the curves in the NHANES daily activity monitor data, it appears that more than one peak exists in more than half of the individuals. Also, the patterns are mostly symmetric, so we would not consider a model that is especially designed for skewed oscillations. Therefore, we would choose to use a model similar to the model (7) we used for the simulation study but that allows more than one peak in close curves.

Suppose the response at the jth time point in the k population $Y_{ij}^{(k)}(t)|t = j$ is normally distributed, with a mean $\mu_{t|t=j}$ and a variance σ_{μ}^2 for all $t_j,\ i = 1, \dots, n^{(k)},\ j = 1, \dots, m$. We consider the following circular-linear regression model:

$$Y_{ij}^{(k)} = A_i + B_i\, cos(\omega t_j) + C_i\, sin(\omega t_j) + D_i\, cos(2\omega t_j) + E_i\, sin(2\omega t_j) + \epsilon_{ij} \qquad (29)$$

where A_i, B_i, C_i, D_i, and E_i are the regression coefficients; $\omega = \frac{2\pi}{T}\ or \frac{360^o}{T}$; and ϵ_{ij} is the random error component. $T = m$ is decided to ensure that the total of the time points j has a length of 2π.

Adding the terms $cos(2\omega t_j)$ and $sin(2\omega t_j)$ creates smaller periods $T/2$ in addition to the overall period $T = 2\pi/\omega$. Thus, the model allows more than one peak to occur within the close curves and thus is certainly more feasible for the real data analysis. Next, we consider the model selection procedure during the model fitting for each individual. Backward model selections are used based on the p-value with an alpha level of 0.05 as the criteria for variables dropping out from the individual model. The variables here refer to terms such as $B_i\, cos(\omega t_j)$ and $C_i\, sin(\omega t_j)$ in the model, and the predicted values are obtained from the finalized model after the backward model selection is done.

The procedure here allows each individual to have a different model, while ensuring that the models are also valid under the significance level criteria for the parametric approach concept.

To make the parametric approach comparable to the Bayesian periodic cubic smoothing splines in the non-parametric approach, some pre-process data manipulation must be done before the circular-linear regression model with backward selection can be applied to each individual in both groups. First, in order to reduce the potential multicollinearity problem that may arise from using all 288 activity readings on corresponding time points as predictors, we reduce the activity readings on time points to 48 by summing up every six activity readings. For example, activity readings that were made every five minutes during the day are cumulated to represent readings every thirty minutes during the day. Furthermore, a log transformation is applied to the activity readings to alleviate potential problems coming from the large difference in deviations in the activity readings between individuals. Also, a log transformation is suggested when proposing the parametric approach where a measurement may not be normally distributed. In addition, the concept of alignment to each individual curve is also considered, in order to enable a comparison between the parametric approach and the non-parametric approach. To afford a fair comparison between the two approaches, the same alignment must be applied to the curves in non-parametric approach so that every curve is in the same direction. This alignment procedure is critical, as the purpose of our study is to observe and make inferences about the circular patterns that are not impacted by the different direction between curves. The practical use of alignment may be a considered as a supplementary procedure when analyzing real data as well. Finally, the overall effect of linear predictor variables such as age is considered in the same way as in the non-parametric approach. However, the analysis from the non-parametric approach in the NHANES activity reading data already shows that the age effect is homogeneous across the two populations. Therefore, the age effect should not be included in the circular-linear regression model under the parametric setting.

The predicted values of the responses from each individual in all populations are used as the inputs in the classification methods given in the NHANES data analysis.

Details of the setting of the simulation study and the real data analysis are discussed in chapters 4 and 5.

CHAPTER 4

SIMULATION

In the simulation, we consider a supervised classification problem using parametric and non-parametric fitting to the curves, coupled with several classification methods. For comparison, we also directly apply the classification method to raw data, which may show the effect of noise reduction using estimation before classification. Several classification methods were considered for this simulation. Principal component analysis (PCA) is one of the most popular methods for dealing with high-dimensional data and has also been extensively used in such classification problems. The support vector machines (SVM) approach is another famous dimension reduction method, which utilizes the hyperplanes concept to achieve good classification results. Classification using a functional generalized linear model (FGLM) is also very common in functional data analysis. Circular-linear data can be considered as a special type of longitudinal data; therefore, functional data analysis is certainly an appropriate classification approach. One extension of the functional generalized linear model that releases the restriction to linear covariates is the generalized kernel additive model (GKAM). It provides more flexibility comparing to FGLM and may be preferable when dealing with data containing limited information about the effect of the predictive variables. We demonstrate the classification performance in our simulation study between combinations of data fitting methods and classification methods under three different circular-linear functions and two different levels of artificial noise. Therefore, we consider six different scenarios, and we also evaluate the performance of classifications by comparing the sensitivity (SEN), specificity (SPE), false discovery rate (FDR), false omission rate (FOR), and correct classification rate. Sensitivity and specificity are almost always desirable when

conducting research about the effectiveness of prediction or examination. False discovery rate and its opposite, false omission rate, provide important information about the correct prediction rate among the positively predicted and negatively predicted observations. Finally, the correct classification rate, which is also known as the concordance rate, is the most representative value for determining the performance among the classification combinations.

Under each of the scenarios, 25 learning samples are generated from each of the two different functions, which demonstrate two different populations or groups. Another 25 testing samples are generated for each group in the same fashion, which are considered as new samples and are used to examine the performance of the classifications. The predicted values of those learning and testing samples from either the parametric approach or the non-parametric approach are used as the learning and testing inputs for the classification methods. Also, the performance of the classification methods when directly applying to the raw samples is also presented for comparison. Under this supervised classification setting, all classification methods are trained using 50 learning samples, and then the performance is measured using testing samples. The simulation study is repeated 100 times for each scenario and each combination. The final results are represented by the mean and standard deviation of the sensitivity, specificity, false discovery rate, false omission rate, and correct classification rate among all 100 replicated simulations studied. All the random samples generated are log transformed, and all sample readings less than or equal to 0 are set as 0.01.

We generate sequences of readings from two different circadian functions along with their membership labels as a demonstration of the linear response variable with circular predictors. With predetermined coefficients, the observations are generated from two different populations, denoted by G_1 and G_2. We denote the reading from the ith individual at the jth time in the kth population by $Y_{ij}^{(k)}$, and we denote the sequences of readings from the ith individual as $Y_{ij}^{(k)} = [Y_{i1}^{(k)}, Y_{i2}^{(k)}, \dots, Y_{i288}^{(k)}]$. The corresponding circular predictors t_j are also determined, $t_j = [1, \dots, 288]$. We first consider the sine-cosine function under the given set of true coefficients $(A^{(k)}, B^{(k)}, C^{(k)})$, which is given by

$$Y_{ij}^{(k)} = A^{(k)} + B^{(k)}cos(\omega t_j) + C^{(k)}sin(\omega t_j) + \varepsilon_{ij}$$

where $\omega = \frac{2\pi}{T}$; and $T = 288$ is the known time period. We consider Gaussian random errors $\varepsilon_{ij} \sim Normal(0,\ {\sigma^{(k)}}^2)$, with a constant variance ${\sigma^{(k)}}^2$. Using the samples from the sine-cosine function should be advantageous for the parametric approach, as we use the circular-linear regression model under the same function. We set the time points number to be as large as 288 for examining the classification, under the condition that the dimensions of the predictors must be much larger than the number of the individual. Also, if we consider that the measures or readings are collected every five minutes, we can at most have 288 repeated measurements during a single day, which is assumed to be the known circadian pattern for these measurements. We consider the following four scenarios for the coefficients and the variances of random errors to generate readings for the populations k, $k = 1,2$.

- Scenario 1: $\left(A^{(1)}, B^{(1)}, C^{(1)}, \sigma^{(1)}\right) = (10, 1, 7, 2.5)$ and $\left(A^{(2)}, B^{(2)}, C^{(2)}, \sigma^{(2)}\right) = (10, 1, 6, 2.5)$
- Scenario 2: $\left(A^{(1)}, B^{(1)}, C^{(1)}, \sigma^{(1)}\right) = (10.5, 1, 7, 2.5)$ and $\left(A^{(2)}, B^{(2)}, C^{(2)}, \sigma^{(2)}\right) = (10, 1, 7, 2.5)$
- Scenario 3: $\left(A^{(1)}, B^{(1)}, C^{(1)}, \sigma^{(1)}\right) = (10, 1, 7, 3.5)$ and $\left(A^{(2)}, B^{(2)}, C^{(2)}, \sigma^{(2)}\right) = (10, 1, 6, 3.5)$
- Scenario 4: $\left(A^{(1)}, B^{(1)}, C^{(1)}, \sigma^{(1)}\right) = (10.5, 1, 7,3.5)$ and $\left(A^{(2)}, B^{(2)}, C^{(2)}, \sigma^{(2)}\right) = (10, 1, 7,3.5)$

We further consider generating samples from normal functions in order to explore distinctions from the sine-cosine functions. As the parametric circular-linear regression naturally uses the sine and cosine function in the model, we expect the parametric model to have a better fit than the non-parametric model. To compare the two estimation methods fairly, the samples from non-sine-cosine functions should not favor either the

parametric approach or the non-parametric approach. We consider the fifth and sixth scenario as the following function, with responses $Y_{ij}^{(k)}$ and circular predictors t_j:

$$Y_{ij}^{(k)} = 1 + \frac{D^{(k)}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\omega t_j - \pi\right)^2\right) + \varepsilon_{ij}$$

where $\omega = \frac{2\pi}{T}$; $T = 288$ is the known time period, $t_j = 1, \ldots, 288$;

$\varepsilon_{ij} \sim Normal(0,\ {\sigma^{(k)}}^2)$, with a constant variance ${\sigma^{(k)}}^2$, and 1 is added as an intercept to prevent most of the random sample from falling below 0.

- Scenario 5: $\left(D^{(1)}, \sigma^{(1)}\right) = (4, 0.6)$ and $\left(D^{(2)}, \sigma^{(2)}\right) = (3.5, 0.6)$
- Scenario 6: $\left(D^{(1)}, \sigma^{(1)}\right) = (4, 0.8)$ and $\left(D^{(2)}, \sigma^{(2)}\right) = (3.5, 0.8)$

The normal function scenario here is used to approximate the circadian function without the sine and cosine function. With the proper parameter setting and a small enough variance, the two ends of the normal function beyond $[0,2\pi]$ are almost parallel and connected. We only consider one period of time points, which makes this approximation viable.

The choices for the coefficients are based on the pattern of the circadian function that we defined. The values we selected for the coefficients are closely related to each other. For instance, the coefficients $C^{(k)}$ for the sine functions differ by two, which coincides with the choice of the standard deviation $\sigma^{(k)}$ to ensure that two true circadian functions are crossed over, as we desired. Therefore, the choices for the coefficients are arbitrary but sensitive to the desired level challenge for the classification problem. When two circadian functions are distributed too close to each other, the difference between two functions becomes unobservable, which also makes the classifications meaningless. Conversely, there is no challenge to the classification problems if the two functions are distributed far apart from each other.

Illustrations of one random individual sample from both populations in each scenario are provided here to demonstrate the readings generated from the circadian functions as well as the shape of the true function:\

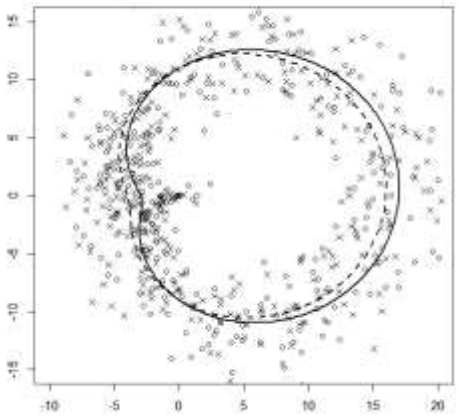
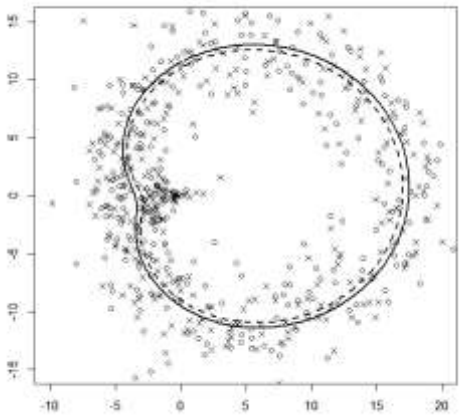
Figure 1. One individual sample from each population in scenario 1	Figure 2. One individual sample from each population in scenario 2
	

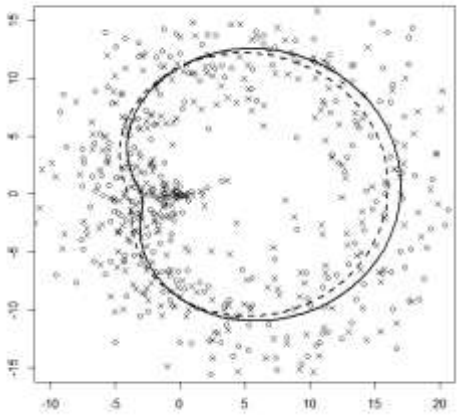
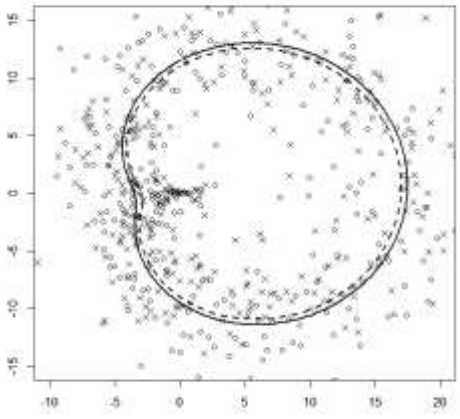
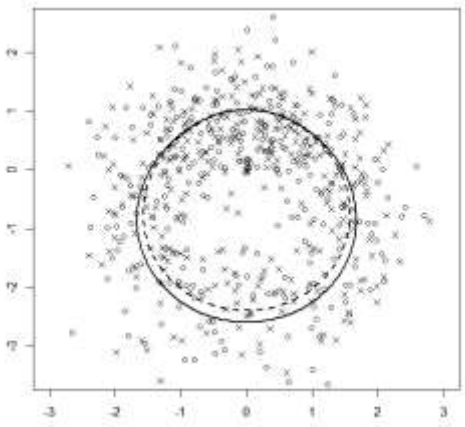
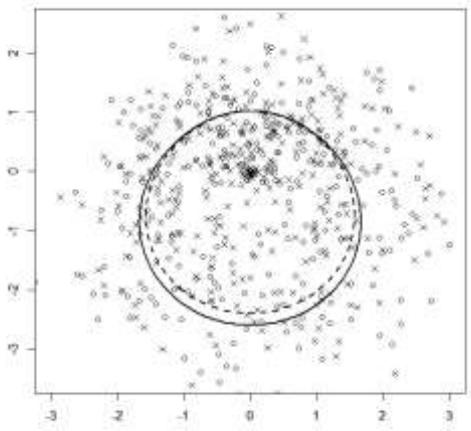
Figure 3. One individual sample from each population in scenario 3	Figure 4. One individual sample from each population in scenario 4
	

Figure 5. One individual sample from each population in scenario 5	Figure 6. One individual sample from each population in scenario 6
	

The symbols here represent the readings generated for one sample from each population along with their respective circadian functions and predetermined coefficients. The small circles with a solid line represent the readings and true circadian functions from population 1, while the crosses and dashed line represent the readings and true circadian functions from population 2.

In scenarios 1 and 3, we illustrate a situation in which two population functions cross each other with different variances. These scenarios examine whether the classification methods have the ability to determine which region gives more strength to the correct classification. They also test the precision of the data fitting from the circular-linear regression model and the BPCSS. In scenarios 2 and 4, we illustrate a situation in which two population functions have no crossover with respect to each other. More particularly, one population function has higher values than the other at all times in the true functions. However, even though the two population functions are apart from each other, we assign a relatively small difference between the two functions in order to heighten the intensity of the classification challenge. The choices for the variance here are 2.5 and 3.5, to represent moderate and huge noise in the readings. Because we use data fitting methods to eliminate most of the noise, we would expect the variation challenge here to be decent. For the normal function scenarios 5 and 6, we

consider different shapes of functions which reflect a different concept than those in the sine-cosine functions. The differences between the two groups are smaller than those in the previous scenarios; therefore, the choice of variance is also relevant to the difference. One may visualize the variations through the graphs provided and expect that the classification challenge is no easier here than with the sine-cosine function scenarios. Later on, for the real data, we apply the log transformation to the samples generated, which also reduces the effect of the variation and the potentially large number.

As the classifications are performed under a simulated setting, we are able to identify the true membership of the samples in the testing group. Therefore, we have the ability to determine the proportion of correct classifications, and we present the result in the following table for each data fitting approach in each scenario. By denoting classification to population G_1 as positive and population G_2 as negative, we have the following table:

Table 1. Table for classification result in population G_1 and G_2

	CLASSIFIED AS G_1	**CLASSIFIED AS PG_2**	**TOTAL NUMBER**
TRUE G_1	# of True positive	# of False negative	25
TRUE G_2	# of False positive	# of True negative	25
TOTAL NUMBER	# of All positive	# of All negative	50

The evaluations of the classification performances are based on the ability to yield correct classifications for 50 testing samples in the simulation study. We consider sensitivity and specificity as part of the performance evaluation, which can be calculated as $\frac{\#\ of\ True\ positive}{25}$ and $\frac{\#\ of\ True\ negative}{25}$, respectively. We also consider the false discover rate, the false omission rate, and the concordance rate as the other measures of the classification performance. The false discover rate (FDR) can be calculated as $\frac{\#\ of\ false\ positive}{\#\ of\ all\ positive}$, and the false omission rate (FOR) can be calculated as $\frac{\#\ of\ false\ negative}{\#\ of\ all\ negative}$.

The correct classification rate, or concordance rate, is calculated as $\frac{\#\text{ of True negative}+\#\text{ of True positive}}{50}$.

Moreover, because the true population functions are known in the simulation, calculations of the mean integrated square error (MISE) are possible and can provide information about how close the estimated curves were to the true functions. The MISE can be calculated by

$$E\int \left(f_n(t) - f(t)\right)^2 dt$$

where $f_n(t)$'s are n samples of estimated curves from the target population; and $f(t)$ is the known true function of the population. The MISE is also known as the L^2 risk function, but we use it here mainly for evaluation of the closeness between the estimated curves and the true functions. The integration can be approximated by summing up the area between the two curves when the time points are condensed.

Table 2. mean of MISE of each scenario in population G_1 and G_2

MISE	S1_$\boldsymbol{G_1}$	S1_$\boldsymbol{G_2}$	S2_$\boldsymbol{G_1}$	S2_$\boldsymbol{G_2}$	S3_$\boldsymbol{G_1}$	S3_$\boldsymbol{G_2}$	S4_$\boldsymbol{G_1}$	S4_$\boldsymbol{G_2}$
BPCSS	28.863	12.653	18.707	28.692	73.283	44.343	56.103	73.613
CLR	27.016	11.997	17.726	26.89	70.101	42.708	53.083	70.492
RAW	152.359	89.62	113.517	152.062	315.159	233.999	263.122	315.709

MISE	S5_$\boldsymbol{G_1}$	S5_$\boldsymbol{G_2}$	S6_$\boldsymbol{G_1}$	S6_$\boldsymbol{G_2}$
BPCSS	4.481	3.076	5.121	4.758
CLR	5.286	4.287	6.167	5.157
RAW	58.762	58.964	104.096	104.429

We note here that the calculations of MISE are performed for every replicated simulation study, and the numbers shown in the table above are averages for the MISE among all 100 replicated studies and all 25 learning samples in the respective population. The MISE for raw samples is also calculated, which uses the original samples generated

from the population functions. Our intention in showing the MISE for raw samples is to represent the effect of noise and also to demonstrate the value of noise reduction by the data fitting methods. The goodness-of-fit here is demonstrated by MISE instead of the commonly used mean square error (MSE) for the following reason. MSE can not yield a fair comparison of goodness-of-fit between parametric and non-parametric methods, because the elements considered in the two methods are quite different. When we consider the least squares estimation fit in the parametric models, non-parametric smoothing splines can always produce zero MSE, once we omit the smoothing penalty and make the splines become interpolating lines between readings. That is certainly not the desired outcome; therefore, comparisons based on MSE are not ideal.

In scenarios 1 through 4, we expect the circular-linear regression (CLR) to have a better fit in general than BPCSS, because the circular-linear regression model naturally uses sine and cosine terms in the model. The MISE in the table does reflect the expected phenomenon, with parametric fittings that are closer to the true function. In scenarios 5 and 6, the BPCSS turns out to have a better fit than circular-linear regression, because the true functions are no longer a combination of the sine and cosine functions. The artificial noise is significantly reduced by the estimation processes, which improves the performance of the classification models.

The performance for each combination of data fitting method and classification method is represented in tables 3 through 6. We observe that the BPCSS combined with the SVM yields the best performance in general out of any other combination in terms of correct classification rate (CCR). This statement is also true when comparing the other results, including sensitivity, specificity, FDR, and FOR. Conditioned on data fitting using circular-linear regression, the performances among the different classification methods are about the same, but SVM is the best choice most of time when combined with circular-linear regression. The classification performance when directly using raw data is obviously poorer than the performance with the use of data fitting procedures, except for classification using the functional generalized linear model. The FGLM is quite competitive in classification performance even without the data fitting procedure among all combinations.

For scenarios 5 and 6, the BPCSS combined with the generalized kernel additive model (GKAM) gives the best performance out of all combinations. However, the difference in performance is not as obvious as in scenarios 1 through 4 when compared with other classification methods conditioned on BPCSS data fitting. Also, the area between curves (ABC) using MCMC samples in BPCSS has quite similar and competitive performance to the other best combinations in scenarios 5 and 6. In circular-linear regression estimation, GKAM is also the best choice out of the four different classification method considered. Among the classification methods using raw data, FGLM constantly demonstrates good performance, as in scenarios 1 through 4, and based on the simulation results would still be the best choice for use without a data fitting procedure.

In summary, Bayesian periodic cubic smoothing splines combined with a good choice of classification method can always produce a slightly better classification performance when compared to the parametric approach. BPCSS also have a more consistent and robust result when information about the true functions is unknown. To improve classification performance, data fitting procedures that pre-process the raw data can certainly reduce the noise and thus achieve a better classification outcome. These additional procedures would be desirable when it is believed that large variations in the measures exist due to unknown or uncontrollable sources. When no data fitting procedures are carried out before the use of classification methods, the FGLM is shown to provide the best classification ability in the simulation study. The appropriate classification method to couple with the estimation method should thus be investigated according to the data.

Table 3. Simulation result for scenario 1

Fitting Method	Classification	SEN	SEN_SD	SPE	SPE_SD	FDR	FDR_SD	FOR	FOR_SD	CCR	CCR_SD
BPCSS	GKAM	0.874	0.085	0.919	0.051	0.082	0.049	0.115	0.067	0.897	0.047
	GLM	0.955	0.056	0.957	0.047	0.041	0.043	0.042	0.049	0.956	0.035
	PCA	0.888	0.089	0.892	0.079	0.103	0.067	0.105	0.07	0.89	0.056
	SVM	0.974	0.034	0.97	0.038	0.029	0.036	0.025	0.033	0.972	0.026
	ABC	0.856	0.09	0.947	0.046	0.058	0.052	0.128	0.072	0.902	0.059
CLR	GKAM	0.908	0.073	0.916	0.056	0.082	0.051	0.087	0.062	0.912	0.046
	GLM	0.931	0.075	0.92	0.061	0.076	0.053	0.064	0.064	0.926	0.043
	PCA	0.927	0.067	0.913	0.069	0.081	0.058	0.07	0.058	0.92	0.042
	SVM	0.939	0.053	0.94	0.049	0.058	0.046	0.058	0.049	0.94	0.037
Raw Data	GKAM	0.62	0.127	0.92	0.064	0.11	0.078	0.286	0.07	0.77	0.068
	GLM	0.958	0.049	0.953	0.057	0.044	0.051	0.04	0.044	0.955	0.036
	PCA	0.869	0.091	0.842	0.084	0.149	0.066	0.127	0.077	0.856	0.057
	SVM	0.955	0.044	0.768	0.086	0.191	0.059	0.053	0.049	0.861	0.046

Table 4. Simulation result for scenario 2

Fitting Method	Classification	SEN	SEN_SD	SPE	SPE_SD	FDR	FDR_SD	FOR	FOR_SD	CCR	CCR_SD
BPCSS	GKAM	0.808	0.078	0.757	0.083	0.227	0.06	0.197	0.067	0.782	0.048
	GLM	0.886	0.08	0.885	0.076	0.11	0.06	0.109	0.067	0.885	0.049
	PCA	0.836	0.089	0.826	0.086	0.168	0.069	0.16	0.074	0.831	0.059
	SVM	0.892	0.069	0.914	0.045	0.086	0.042	0.101	0.058	0.923	0.038
	ABC	0.886	0.065	0.754	0.1	0.213	0.076	0.131	0.073	0.82	0.071
CLR	GKAM	0.848	0.091	0.823	0.086	0.168	0.068	0.15	0.077	0.835	0.058
	GLM	0.894	0.067	0.905	0.064	0.093	0.057	0.102	0.059	0.899	0.048
	PCA	0.884	0.073	0.904	0.063	0.094	0.057	0.109	0.062	0.894	0.043
	SVM	0.907	0.067	0.903	0.054	0.094	0.049	0.089	0.058	0.905	0.041
Raw Data	GKAM	0.758	0.102	0.511	0.116	0.389	0.053	0.313	0.097	0.635	0.061
	GLM	0.896	0.068	0.907	0.062	0.091	0.055	0.099	0.058	0.902	0.046
	PCA	0.748	0.103	0.749	0.089	0.248	0.066	0.245	0.076	0.749	0.062
	SVM	0.629	0.11	0.91	0.057	0.122	0.073	0.285	0.06	0.769	0.058

Table 5. Simulation result for scenario 3

Fitting Method	Classification	SEN	SEN_SD	SPE	SPE_SD	FDR	FDR_SD	FOR	FOR_SD	CCR	CCR_SD
BPCSS	GKAM	0.796	0.102	0.858	0.069	0.146	0.06	0.184	0.072	0.827	0.051
	GLM	0.888	0.084	0.904	0.067	0.094	0.059	0.104	0.071	0.896	0.05
	PCA	0.79	0.097	0.814	0.096	0.184	0.076	0.198	0.071	0.802	0.058
	SVM	0.92	0.061	0.905	0.057	0.091	0.05	0.078	0.056	0.912	0.04
	ABC	0.782	0.099	0.848	0.08	0.162	0.082	0.201	0.082	0.815	0.079
CLR	GKAM	0.819	0.1	0.862	0.078	0.138	0.067	0.166	0.075	0.841	0.054
	GLM	0.87	0.088	0.858	0.085	0.133	0.067	0.124	0.07	0.864	0.05
	PCA	0.877	0.091	0.83	0.077	0.157	0.057	0.121	0.074	0.854	0.047
	SVM	0.851	0.085	0.882	0.074	0.116	0.065	0.139	0.069	0.866	0.05
Raw Data	GKAM	0.583	0.115	0.834	0.09	0.214	0.096	0.329	0.065	0.709	0.068
	GLM	0.888	0.083	0.91	0.066	0.088	0.058	0.104	0.068	0.899	0.047
	PCA	0.762	0.1	0.796	0.095	0.205	0.077	0.224	0.075	0.779	0.063
	SVM	0.834	0.106	0.674	0.118	0.274	0.071	0.186	0.09	0.754	0.062

Table 6. Simulation result for scenario 4

Fitting Method	Classification	SEN	SEN_SD	SPE	SPE_SD	FDR	FDR_SD	FOR	FOR_SD	CCR	CCR_SD
BPCSS	GKAM	0.753	0.104	0.7	0.101	0.28	0.071	0.254	0.085	0.727	0.067
	GLM	0.836	0.092	0.8	0.093	0.187	0.071	0.164	0.076	0.818	0.06
	PCA	0.748	0.109	0.732	0.104	0.258	0.076	0.249	0.083	0.74	0.068
	SVM	0.822	0.082	0.806	0.1	0.184	0.076	0.176	0.071	0.834	0.06
	ABC	0.817	0.098	0.684	0.107	0.276	0.08	0.207	0.1	0.75	0.083
CLR	GKAM	0.77	0.101	0.721	0.098	0.262	0.072	0.235	0.083	0.745	0.067
	GLM	0.826	0.089	0.82	0.09	0.173	0.075	0.17	0.074	0.823	0.061
	PCA	0.801	0.099	0.832	0.085	0.168	0.076	0.187	0.076	0.816	0.062
	SVM	0.828	0.091	0.794	0.092	0.194	0.073	0.172	0.076	0.811	0.061
Raw Data	GKAM	0.674	0.113	0.492	0.122	0.427	0.06	0.394	0.094	0.583	0.067
	GLM	0.846	0.084	0.829	0.092	0.162	0.073	0.152	0.07	0.818	0.058
	PCA	0.683	0.107	0.665	0.097	0.326	0.068	0.317	0.076	0.674	0.065
	SVM	0.512	0.121	0.816	0.082	0.259	0.087	0.37	0.057	0.664	0.063

Table 7. Simulation result for scenario 5

Fitting Method	Classification	SEN	SEN_SD	SPE	SPE_SD	FDR	FDR_SD	FOR	FOR_SD	CCR	CCR_SD
BPCSS	GKAM	0.926	0.05	0.933	0.057	0.065	0.053	0.072	0.047	0.929	0.041
	GLM	0.913	0.076	0.91	0.066	0.087	0.058	0.082	0.065	0.911	0.049
	PCA	0.893	0.077	0.895	0.066	0.102	0.059	0.102	0.067	0.894	0.05
	SVM	0.917	0.064	0.908	0.069	0.087	0.061	0.08	0.058	0.913	0.047
	ABC	0.926	0.055	0.927	0.059	0.071	0.054	0.072	0.051	0.926	0.043
CLR	GKAM	0.922	0.06	0.929	0.052	0.069	0.049	0.074	0.053	0.926	0.039
	GLM	0.9	0.077	0.898	0.075	0.097	0.065	0.095	0.068	0.899	0.051
	PCA	0.9	0.076	0.898	0.07	0.097	0.061	0.095	0.065	0.899	0.047
	SVM	0.915	0.064	0.918	0.055	0.079	0.051	0.081	0.057	0.917	0.042
Raw Data	GKAM	0.786	0.094	0.784	0.094	0.21	0.075	0.209	0.074	0.785	0.062
	GLM	0.908	0.077	0.91	0.063	0.087	0.056	0.087	0.066	0.909	0.049
	PCA	0.888	0.081	0.886	0.092	0.107	0.075	0.107	0.069	0.887	0.058
	SVM	0.785	0.098	0.774	0.097	0.218	0.073	0.212	0.076	0.78	0.064

Table 8. Simulation result for scenario 6

Fitting Method	Classification	SEN	SEN_SD	SPE	SPE_SD	FDR	FDR_SD	FOR	FOR_SD	CCR	CCR_SD
BPCSS	GKAM	0.845	0.094	0.87	0.078	0.128	0.066	0.144	0.076	0.858	0.056
	GLM	0.841	0.09	0.852	0.088	0.144	0.071	0.152	0.075	0.847	0.06
	PCA	0.829	0.092	0.837	0.088	0.158	0.07	0.163	0.073	0.833	0.057
	SVM	0.842	0.092	0.843	0.082	0.151	0.066	0.151	0.072	0.842	0.053
	ABC	0.846	0.091	0.87	0.073	0.127	0.062	0.144	0.072	0.857	0.054
CLR	GKAM	0.85	0.092	0.863	0.085	0.133	0.071	0.142	0.074	0.856	0.056
	GLM	0.842	0.094	0.845	0.089	0.151	0.075	0.152	0.079	0.843	0.065
	PCA	0.838	0.093	0.842	0.091	0.152	0.074	0.154	0.075	0.84	0.056
	SVM	0.838	0.095	0.849	0.09	0.146	0.076	0.154	0.075	0.843	0.059
Raw Data	GKAM	0.697	0.11	0.672	0.11	0.315	0.073	0.304	0.075	0.684	0.063
	GLM	0.836	0.092	0.855	0.085	0.142	0.072	0.155	0.077	0.846	0.062
	PCA	0.799	0.103	0.816	0.1	0.18	0.077	0.19	0.079	0.807	0.064
	SVM	0.687	0.122	0.667	0.121	0.318	0.079	0.31	0.082	0.677	0.065

CHAPTER 5

ANALYSIS OF NHANES DATA

USING THE PROPOSED METHODS

The National Health and Nutrition Examination Survey (NHANES) is a program of studies evaluating the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. We consider herein the classification problem for insomnia determination using NHANES data. In order to make a more precise inference about sleeping duration, sleeping quality, and daytime activity, we take advantage of physical activity monitor data reported in NHANES 2005-2006. The physical activity monitor data and demographic data used herein can be found on the official website of NHANES.

It is difficult to accurately measure activity patterns when studying free-living populations. Questionnaires taken from self-reported interviews are usually of limited value, because participants' perceptions of their activity intensity are subjective and because patterns of physical activity are difficult for participants to recall and quantify. Therefore, records obtained from monitors are believed to be more complete and objective than self-reported information. In NHANES 2005-2006, the activity intensity was measured by an electronic device called the ActiGraph AM-7164, manufactured by ActiGraph of Ft. Walton Beach, FL. This device collects the intensity and duration of an activity such as walking or jogging, and the accumulation of activity intensity is recorded within each minute. Activity intensity readings are collected for 7 consecutive days, thus yielding up to 10,800 readings per participant. Among the participants, demographic information is available that includes age, gender, race, marital status, and so on. In addition, the questionnaires contained information on whether the participant had been diagnosed by a doctor as having a sleeping disorder, and if so, what kind of sleeping disorder had been diagnosed.

We assume the daily activity among the participants to have a similar pattern every day for the 7 days collected. The circular-linear regression model utilizes the periodic natural daily activity to make inferences about the activity for people with insomnia and normal people. Furthermore, other information such as demographic information or disease status can play a role in customizing the inference of activity for individuals with certain characteristics.

In this section, we will compare the differences in features of the fitted curves generated by two different approaches. Any major divergence in patterns between two curves can be used to make suggestions related to the diagnosis or early determination of sleeping disorders, by observing the difference in sleeping duration, sleeping quality, and daytime activity duration and intensity. The results can also serve as an objective measure for determining the severity of insomnia or for the prevention of sleeping disorders by comparing the activity of patients with the estimated curves. To verify the performance of the classification for both the parametric approach and the non-parametric approach using NHANES data, we can consider separating the data into training samples and testing samples randomly with a ratio of approximately 1:1. The number of patients in the normal group and the insomnia group are also in a 1:1 ratio. We do not use all patients in the data sets, because the number of patients in the normal group is about fifteen times greater than that in the insomnia group, and the classification method may not work properly when the patient numbers are so far from equal.

For the daily activity monitor data, the activity readings used for each patient are collected every five minutes. Therefore, there are 288 readings within a day, and we assume the known period of the repeating pattern is twenty-four hours for each patient. 320 insomnia patients and 320 normal patients are equally and randomly distributed to the learning group and the testing group. Data cleaning is performed before the patients' readings are used for estimation and classification, because we notice that there are quite a few unreasonable records. During the data cleaning procedure, we exclude patients who 1) constantly have the same reading along the time period; 2) have unreasonably high readings (thirty times the overall average) for at least 75% of the time; or 3) have zero readings for at least 75% of the time. After data cleaning, we end up with 157 patients in

the insomnia learning group, 149 patients in the insomnia testing group, 150 patients in the normal learning group, and 150 patients in the normal testing group. The same procedure is performed as in the simulation study for both the learning and the testing groups. We first use the parametric or non-parametric model to find predicted values for the activity readings. Then, the values are used as inputs to couple with the classification methods. All activity readings are pre-processed in the following manner before being used in the data fitting procedures. All the activity readings are integers ranging from 0 to approximately 5,000, and we make log transformations of the readings to reduce the number variation among patients. Zero readings are set to 0.01 before log transformation to prevent generation of a negative infinity. All the samples are aligned to the same direction using the reference times found according to the alignment calculation in (27). In the data fitting procedure, as in the simulation study, we consider the predicted values from the parametric circular-linear regression model, the non-parametric BPCSS, and using raw data as inputs for the classification method. To demonstrate the need in the reality, we consider age effect as the overall linear covariate in both data fitting methods and discount the age effects as needed to demonstrate the desired classification problems depending on circadian pattern only.

In the classification step, we consider the same SVM, PCA, FGLM, functional GKAM, and area between curves, which is only applicable in BPCSS, as was used in the simulation study. One may refer to sections 2.2.1 through 2.2.4 for the detailed classification procedures. A similar idea to the leave-k-out method of cross-validations is considered in the NHANES study. Each group of the patients is randomly separated into 10 subgroups for the learning and testing groups in both the normal and insomnia groups. We exclude the first subgroup out of each group—normal learning, normal testing, insomnia learning, and insomnia testing—in the first run and perform all the data fitting and classification as described to access our final result. Therefore, we perform 10 runs by excluding one different subgroup in each group at each run. The 10 results from each combination of data fitting methods and classification methods are collected, and the mean and standard deviation of each combination are reported in table 9.

In data fitting using BPCSS, the coefficient estimation of the overall age effect among learning group in normal patients is -0.000282 with a standard deviation of 0.000533, and it was 0.000871 in the insomnia learning group with a standard deviation of 0.000793. The age range from both groups is from 16 to 78. The estimated amount of the age effect is deducted in the testing group before proceeding to the classification. In data fitting using the parametric approach, the age effect is homogeneous across all the groups; therefore, the age effect is omitted.

In the NHANES data analysis, the best concordance rate out of all the combinations is achieved by BPCSS combined with SVM. GKAM and the area between curves (ABC) coupled with BPCSS also provide a competitive concordance rate. However, GKAM favors the negative group more, while ABC favors the positive group which are reflected on sensitivity and specificity; thus, neither may be desirable unless we have prior information about the cost of misclassification. Without such prior information, we must set the misclassification cost equally for both groups, and the desired classification method must have almost the same sensitivity and specificity—which means that the classification method must not particularly favor either group. Among the classification methods coupled with the parametric data fitting procedure, PCA and FGLM offer about the same level of good performance in classification. Conditioned on using raw data, FGLM was seen to offer the concordance rate, which is the same conclusion that was found in the simulation study. For this data, the concordance rate stays low, because the variety of the activity readings between patients were quite huge. During the data cleaning process, we observed a huge variation in mean activity readings between patients, for which the mean readings could be as low as 20 or as high as 1,500. The activity patterns also varied between patients from having one to multiple peaks during the day. All these variations among patients may have made it very difficult for the classification method to separate the patients into their correct groups. We consider an equal sample size for normal and insomnia groups to be essential for reasonable classification learning. Otherwise, the learning result will highly favor the group having significantly more samples and will classify almost every observation into that group in order to reach highest concordance rate, which is definitely not reasonable or desirable.

Table 9. Classification result in NHANES data

Fitting Method	Classification	SEN	SEN_SD	SPE	SPE_SD	FDR	FDR_SD	FOR	FOR_SD	CCR	CCR_SD
BPCSS	GKAM	0.427	0.043	0.635	0.050	0.458	0.017	0.476	0.008	0.531	0.011
	GLM	0.573	0.025	0.465	0.022	0.481	0.014	0.48	0.018	0.519	0.016
	PCA	0.532	0.032	0.496	0.022	0.485	0.010	0.487	0.011	0.514	0.011
	SVM	0.552	0.031	0.578	0.035	0.451	0.013	0.469	0.010	0.555	0.011
	ABC	0.744	0.045	0.313	0.044	0.478	0.010	0.450	0.024	0.530	0.013
CLR	GKAM	0.379	0.191	0.693	0.161	0.399	0.142	0.471	0.019	0.535	0.018
	GLM	0.596	0.026	0.495	0.033	0.457	0.013	0.451	0.015	0.546	0.014
	PCA	0.589	0.035	0.509	0.017	0.453	0.018	0.447	0.023	0.549	0.021
	SVM	0.628	0.041	0.452	0.024	0.465	0.014	0.452	0.023	0.540	0.017
Raw Data	GKAM	0.087	0.106	0.943	0.083	0.181	0.201	0.493	0.01	0.513	0.016
	GLM	0.544	0.034	0.526	0.031	0.464	0.019	0.466	0.02	0.535	0.019
	PCA	0.472	0.033	0.539	0.016	0.493	0.019	0.496	0.016	0.505	0.017
	SVM	0.453	0.027	0.578	0.032	0.48	0.015	0.488	0.012	0.515	0.013

CHAPTER 6
CONCLUSIONS AND DISCUSSION

From the data fitting results, we can conclude that the proposed Bayesian periodic cubic smoothing splines (BPCSS) approach generally produces a good fit and naturally has more flexibility than the parametric approach to circular-linear data. Compared to the existing parametric circular-linear model, the BPCSS offers equivalent or better performance which is shown by comparing mean integrated square error in the simulation study. When combined with classification methods for supervised classification problems, data fitting using BPCSS is a superior alternative. When comparing the performance of parametric and nonparametric fitting methods using the same classification method, BPCSS consistently perform slightly better on sensitivity, false discovery rate and concordance rate. We also show herein that BPCSS can be flexibly used to calculate reference times, as well as to account for linear predictors in assessing the overall effects for each group. Such desired features are well utilized in the NHANES data analysis, and the results from the real data analysis are consistent with the simulation results. We can certainly see potential uses for the non-parametric approach in circular-linear data with useful features that include the model's flexibility, consistency, and ability to customize. Data fitting procedures reduce the noise in the original data and are shown to be effective in improving the classification performance in terms of sensitivity, specificity, false discovery rate, false omission rate, and concordance rate.

The primary limitation of the BPCSS is its computational expense compared to circular-linear regression. However, this is a common difficulty with most computational Bayesian approaches because of the iterative computations. Nevertheless, the method has the advantage of solving complex problems that are proven difficult under classical approaches. It can be also time-consuming in the parametric approach to decide which model to use, when a complex model such as the polynomial additive model is necessary to provide better estimation.

The parametric circular-regression model can also be modified to possibly achieve a better performance. The symmetric circular-linear model has been considered here as a comparative study to the non-parametric approach. It is possible that using an asymmetric model can offer better performance than the simpler symmetric parametric approach.

The classification of the NHANES data turned out to be considerably harder than the simulation study. The huge variation in the data set that exists both within individual measurements as well as between individuals, resulting in violations of distributional and other assumptions, could be the primary reason for not getting efficient results for the real data. The further investigation of an appropriate transformation may improve the result. Furthermore, the individuals' membership into one of the two groups are self-reported and likely to have inaccuracies. We also lack information whether the individuals with insomnia are still under treatment or if they have recovered from this condition. All these factors may contribute to unsatisfactory classification performance.

There are several directions in which the BPCSS method can be further extended. The BPCSS is an extension of cubic smoothing splines that accommodates circular-linear data under a Bayesian framework and can be considered as a starting point for a Bayesian non-parametric approach to circular-linear data. Therefore, we may certainly consider a similar framework to BPCSS that uses other kinds of smoothing splines. Also, modeling with time-dependent or other kind of covariates could also be customized under the Bayesian framework. Time-dependent covariates have been considered in the parametric circular-linear regression, and can surely be adopted in the BPCSS. Furthermore, we can also consider a more general class of response other than the one-to-one time-dependent response in BPCSS or in circular-linear regression. One may consider that only one response, corresponding to a series of time-dependent predictors, is desirable. Using a similar concept in the functional generalized linear model, it is possible to expand the smoothing function predictors to include circular-linear smoothing function predictors and combining this with the idea of the generalized linear model to adapt the model for any response from the exponential family. Also, missing values in the data have not been considered in this approach. The smoothing splines approach may be used to resolve

missing reading by changing the knot points. Such an approach can also offer a solution to the non-universal time point problem, which is not considered in this paper. Furthermore, classification methods used for multivariate analysis such as PCA and SVM can be modified to accommodate circular-linear data, which considers the correlation between time points using the circular-linear concept. Also, the use of MCMC samples collected from BPCSS can be greatly extended. The classification performance under the simple calculation of the area between curves using MCMC samples may be generalized by using classification rules based on statistics other than area between curves.

Reference

1. Bhattacharya, S. and A. SenGupta, *Bayesian inference for circular distributions with unknown normalising constants.* Journal of statistical planning and inference, 2009. **139**(12): p. 4179-4192.
2. Fisher, N.I., *Statistical analysis of circular data*. 1995: Cambridge University Press.
3. Jammalamadaka, S.R. and A. SenGupta, *Topics in circular statistics*. Vol. 5. 2001: World Scientific.
4. Mardia, K.V. and P.E. Jupp, *Directional statistics*. Vol. 494. 2009: John Wiley & Sons.
5. Roth, T., *Insomnia: definition, prevalence, etiology, and consequences.* Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine, 2007. **3**(5 Suppl): p. S7.
6. SenGupta, A. and F.I. Ugwuowo, *Asymmetric circular-linear multivariate regression models with applications to environmental data.* Environmental and Ecological Statistics, 2006. **13**(3): p. 299-309.
7. Bhattacharya, S. and A. SenGupta, *Bayesian analysis of semiparametric linear-circular models.* Journal of agricultural, biological, and environmental statistics, 2009. **14**(1): p. 33-65.
8. Laird, N.M. and J.H. Ware, *Random-effects models for longitudinal data.* Biometrics, 1982: p. 963-974.
9. Zeger, S.L. and P.J. Diggle, *Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters.* Biometrics, 1994: p. 689-699.
10. Li, Y., X. Lin, and P. Müller, *Bayesian inference in semiparametric mixed models for longitudinal data.* Biometrics, 2010. **66**(1): p. 70-78.
11. De la Cruz, R., et al., *Bayesian Regression Analysis of Data with Random Effects Covariates from Nonlinear Longitudinal Measurements.* arXiv preprint arXiv:1310.8176, 2013.
12. Ryu, D., et al., *Longitudinal studies with outcome-dependent follow-up.* Journal of the American Statistical Association, 2007. **102**(479).
13. Hastie, T.J. and R.J. Tibshirani, *Generalized additive models*. Vol. 43. 1990: CRC Press.
14. De Boor, C., *A practical guide to splines.* Mathematics of Computation, 1978.
15. Graham, N., *Smoothing with periodic cubic splines.* Bell System Technical Journal, 1983. **62**(1): p. 101-110.
16. Richardson, M., *Principal component analysis.* URL: http://people. maths. ox. ac. uk/richardsonm/SignalProcPCA. pdf (last access: 3.5. 2013). Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik@ ntf. uni-lj. si, 2009.
17. Shlens, J., *A tutorial on principal component analysis.* arXiv preprint arXiv:1404.1100, 2014.
18. Friedman, J., T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Vol. 1. 2001: Springer series in statistics Springer, Berlin.
19. Bair, E., et al., *Prediction by supervised principal components.* Journal of the American Statistical Association, 2012.
20. Howley, T., et al., *The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data.* Knowledge-Based Systems, 2006. **19**(5): p. 363-370.

21. Vapnik, V., *The nature of statistical learning theory*. 2013: Springer Science & Business Media.
22. Gunn, S.R., M. Brown, and K.M. Bossley, *Network performance assessment for neurofuzzy data modelling*, in *Advances in intelligent data analysis reasoning about data*. 1997, Springer. p. 313-323.
23. Vapnik, V., S.E. Golowich, and A. Smola. *Support vector method for function approximation, regression estimation, and signal processing*. in *Advances in neural information processing systems 9*. 1996. Citeseer.
24. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine learning, 1995. **20**(3): p. 273-297.
25. Girosi, F., *An equivalence between sparse approximation and support vector machines.* Neural computation, 1998. **10**(6): p. 1455-1480.
26. Smola, A.J. and B. Schölkopf, *On a kernel-based method for pattern recognition, regression, approximation, and operator inversion.* Algorithmica, 1998. **22**(1-2): p. 211-231.
27. Blanz, V., et al., *Comparison of view-based object recognition algorithms using realistic 3D models*, in *Artificial Neural Networks—ICANN 96*. 1996, Springer. p. 251-256.
28. McCullagh, P. and J.A. Nelder, *Generalized linear models*. Vol. 37. 1989: CRC press.
29. Moyeed, R. and P.J. Diggle, *Rates of Convergence in Semi-Parametric Modeling of Longitudinal Data.* Australian Journal of Statistics, 1994. **36**(1): p. 75-93.
30. Diggle, P.J., K. Liang, and S.L. Zeger, *Analysis of longitudinal data.* Journal of the Royal Statistical Society-Series A Statistics in Society, 1995. **158**(2): p. 339.
31. Silverman, B.W., *Some aspects of the spline smoothing approach to non-parametric regression curve fitting.* Journal of the Royal Statistical Society. Series B (Methodological), 1985: p. 1-52.
32. Green, P.J. and B.W. Silverman, *Nonparametric regression and generalized linear models: a roughness penalty approach*. 1993: CRC Press.
33. Febrero-Bande, M. and W. González-Manteiga, *Generalized additive models for functional data.* Test, 2013. **22**(2): p. 278-292.
34. Müller, H.-G. and F. Yao, *Functional additive models.* Journal of the American Statistical Association, 2012.
35. Ferraty, F. and P. Vieu, *Additive prediction and boosting for functional data.* Computational Statistics & Data Analysis, 2009. **53**(4): p. 1400-1413.
36. Fan, Y., G.M. James, and P. Radchenko, *Functional additive regression.* The Annals of Statistics, 2015. **43**(5): p. 2296-2325.
37. Roca-Pardiñas, J., et al., *Predicting binary time series of SO2 using generalized additive models with unknown link function.* Environmetrics, 2004. **15**(7): p. 729-742.