ESTIMATION OF CIRCADIAN PARAMETERS AND INVESTIGATION IN CYANOBACTERIA VIA SEMIPARAMETRIC VARYING COEFFICIENT PERIODIC MODELS

A Dissertation

by

YINGXUE LIU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2007

Major Subject: Statistics

ESTIMATION OF CIRCADIAN PARAMETERS AND INVESTIGATION IN CYANOBACTERIA VIA SEMIPARAMETRIC VARYING COEFFICIENT PERIODIC MODELS

A Dissertation

by

YINGXUE LIU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Naisyin Wang
Committee Members,	G. Donald Allen
	Raymond J. Carroll
	Erning Li
	Michael Longnecker
Head of Department,	Simon J. Sheather

August 2007

Major Subject: Statistics

ABSTRACT

Estimation of Circadian Parameters and Investigation in Cyanobacteria via
Semiparametric Varying Coefficient Periodic Models. (August 2007)
Yingxue Liu, B.S., Peking University, P. R. China;
M.S., Texas A&M University
Chair of Advisory Committee: Dr. Naisyin Wang

This dissertation includes three components. Component 1 provides an estimation procedure for circadian parameters in cyanobacteria. Component 2 explores the relationship between baseline and amplitude by model selection under the framework of smoothing spline. Component 3 investigates properties of hypothesis testing. The following three paragraphs briefly summarize these three components, respectively.

Varying coefficient models are frequently used in statistical modeling. We propose a semiparametric varying coefficient periodic model which is suitable to study periodic patterns. This model has ample applications in the study of the cyanobacteria circadian clock. To achieve the desired flexibility, the model we consider may not be globally identifiable. We propose to perform local approximations by kernel based methods and focus on estimating one solution that is biologically meaningful. Asymptotic properties are developed. Simulations show that the gain by our procedure over the commonly used method is substantial. The methodology is illustrated by an application to a cyanobacteria dataset.

Smoothing spline can be implemented, but a direct application with the penalty selected by the generalized cross-validation often leads to non-convergence outcomes.

We propose an adjusted cross-validation instead, which resolves the difficulties. Biologists believe that the amplitude function of the periodic component is proportional to the baseline function. To verify this belief, we propose a full model without any assumptions regarding such a relationship, and two reduced models with the ratio of baseline and amplitude to be a constant and a quadratic function of time, respectively. We use model selection techniques, Akaike information criterion (AIC) and Schwarz Bayesian information criterion (BIC), to determine the optimal model. Simulations show that AIC and BIC select the correct model with high probabilities. Application to cyanobacteria data shows that the full model is the best model.

To investigate the same problem in component 2 by a formal hypothesis testing procedure, we develop kernel based methods. In order to construct the test statistic, we derive the global degree of freedom for the residual sum of squares. Simulations show that the proposed tests perform well. We apply the proposed procedures to the data and conclude that the baseline and amplitude functions share no linear or quadratic relationship.

To Lian and Shu.

ACKNOWLEDGEMENTS

I feel very lucky to have had Dr. Naisyin Wang as my advisor in this most important stage in my life. She guided me during my research from the very early stages to writing the technical report. All the things I learned from her will benefit my entire career. She also showed me what enthusiastic and successful researchers women can become. She encouraged me when I felt frustrated, and she truly cares about my well-being. She is more than my advisor, she is the person that I can always trust and ask for advice for whatever reason. I wish I could have found better words than "Thank you" to convey my appreciation for her.

A special thanks is owed to Dr. Yuedong Wang who jointly worked with me on the second component of my dissertation. I would like to thank him for his very helpful advice.

My thanks are extended to Dr. Allen for serving on both my M.S. committee and Ph.D. committee. Thank you for being consistently supportive. I wish to say thanks to Dr. Carroll and Dr. Li for serving on my Ph.D. committee.

I also want to thank Dr. Longnecker and Dr. Speed for being my co-advisors for my M.S., and further thank Dr. Longnecker for the many kinds of advice and help he gave me. Our department will not survive without him.

And finally, to my husband Lian and my son Shu.

TABLE OF CONTENTS

Page

ABSTRACT iii

DEDICATION v

ACKNOWLEDGEMENTS vi

TABLE OF CONTENTS vii

LIST OF TABLES ix

LIST OF FIGURES xi

CHAPTER

I INTRODUCTION 1

1.1 What Is a Circadian Program? 1
1.2 The Cyanobacterial Circadian Clock 2
1.3 Structure Overview 4

II ESTIMATION OF CIRCADIAN PARAMETERS IN CYANOBACTERIA VIA SEMIPARAMETRIC VARYING COEFFICIENT PERIODIC MODELS 5

2.1 Introduction 5
2.2 Traditional Approach 8
2.3 Varying Coefficient Periodic Models 11
2.4 Parametric and Semiparametric Estimation Methods . . . 13
2.5 Asymptotic Properties 22
2.6 Numerical Outcomes 23
2.7 Concluding Remarks 36

CHAPTER Page

III MODEL SELECTION USING SMOOTHING SPLINE IN CYANOBACTERIA 39

3.1 Introduction 39
3.2 Models 41
3.3 Spline Estimations 42
3.4 Model Selection 46
3.5 Numeric Outcomes 49
3.6 Concluding Remarks 56

IV HYPOTHESIS TESTING USING LOCAL LINEAR REGRESSION IN CYANOBACTERIA 57

4.1 Introduction 57
4.2 Models 58
4.3 Kernel Estimation Methods 59
4.4 Testing Hypothesis 64
4.5 Degree of Freedom for RSS 65
4.6 Numerical Outcomes 67
4.7 Concluding Remarks 71

V CONCLUSION AND FUTURE RESEARCH 73

REFERENCES 75

APPENDIX A 78

APPENDIX B 85

APPENDIX C 86

VITA 92

LIST OF TABLES

TABLE		Page
1	Monte Carlo biases, SE's and MSE's of all estimators by all five methods when number of curves $m = 1$.	26
2	Monte Carlo biases, SE's and MSE's of all estimators, average lengths and coverage probabilities for confidence intervals of τ and ϕ. Methods used are DFT, SVCC and SVCSc. Number of curves $m = 6$.	28
3	Monte Carlo biases, SE's and MSE's of all estimators, average lengths and coverage probabilities for confidence intervals of τ and ϕ. Methods used are DFT, SVCC and SVCSc. Number of curves $m = 18$.	30
4	Estimates for cyanobacteria circadian datasets by all five methods.	32
5	Bootstrap analysis of cyanobacteria circadian data sets. Monte Carlo biases, SE's, MSE's of $\hat{\theta}$ and $\hat{\Sigma}_\theta$ by all five methods, and average lengths and coverage probabilities of confidence intervals for τ and ϕ by all five methods.	34
6	Average lengths and coverage probabilities of confidence intervals for τ and ϕ by both (2.10) and (2.9).	36
7	Comparison of bootstrap confidence interval and normal confidence interval for τ and ϕ.	37
8	Model selection in Example 3.1	48
9	Inferences for estimators in Example 3.1	49
10	Proportions of selection in Example 3.1	50
11	Model selection in Example 3.2	50
12	Inferences for estimators in Example 3.2	52
13	Proportions of selection in Example 3.2	52

TABLE Page

14 Model selection in Example 3.3 52

15 Inferences for estimators in Example 3.3 54

16 Proportions of selection in Example 3.3 54

17 Model selection in cyanobacteria circadian data 55

LIST OF FIGURES

FIGURE		Page
1	Plot of bioluminescence vs time, data were recorded every 2.116 hours. X-axis, time in hours. Y-axis, bioluminescence. Time is divided by dashed lines to indicate subjective days. The parameters of circadian period and relative phase are indicated.	3
2	Four data curves. Plot of bioluminescence vs time, data were recorded every 2.116 hours. The plots in the top panel have two stages of growth cycle (exponential stage and sustained stage), while the plots in the bottom panel have three stages (exponential stage, sustained stage and senescent stage). The red dashed lines are the underlying baseline functions.	6
3	The bottom-right curve in Figure 2 and the DFT estimated curve. The black solid line is the observed curve, and the red dashed line is the DFT estimated curve.	11
4	DFT estimates of τ, ϕ, β_0 and β_1 vs magnitude of noise s in Example 2.1, from left to right. The dots are our estimated parameter values through DFT for different s; the solid lines provide the true parameter values.	12
5	The bottom-right curve in Figure 2 and the QVC estimated curve. The black solid line is the observed curve, and the red dashed line is the QVC estimated curve.	15
6	The bottom-right curve in Figure 2 and the SVCC estimated curve. The black solid line is the observed curve, and the red dashed line is the SVCC estimated curve.	18
7	The bottom-right curve in Figure 2 and the SVCSc estimated curve. The black solid line is the original curve, and the red dashed line is the SVCSc estimated curve.	19
8	The bottom-right curve in Figure 2 and the SVCSd estimated curve. The black solid line is the original curve, and the red dashed line is the SVCSd estimated curve.	20

FIGURE Page

9 Sensitivity analysis for simulated data ($m = 6$). Plot of within-curve variation $\hat{\Sigma}_\varepsilon$ versus $\hat{\theta}$ by SVCSc. Left graph is for τ and right graph is for ϕ. The solid lines are between-curve variation $\hat{\Sigma}_\theta$, and 'percent's indicate the maximum values of $\hat{\Sigma}_\varepsilon/\hat{\Sigma}_\theta$. 29

10 Sensitivity analysis for simulated data ($m = 18$). Plot of within-curve $\hat{\Sigma}_\varepsilon$ versus $\hat{\theta}$ by SVCSc. Left graph is for τ and right graph is for ϕ. The solid lines are between-curve variation $\hat{\Sigma}_\theta$, and 'percent's indicate the maximum values of $\hat{\Sigma}_\varepsilon/\hat{\Sigma}_\theta$. 31

11 Sensitivity analysis for cyanobacteria circadian data. Plot of within-curve variation $\hat{\Sigma}_\varepsilon$ versus $\hat{\theta}$ by SVCSc. Left graph is for τ and right graph is for ϕ. The solid lines are between-curve variation $\hat{\Sigma}_\theta$, and 'percent's indicate the maximum values of $\hat{\Sigma}_\varepsilon/\hat{\Sigma}_\theta$. 35

12 Histograms of $\hat{\tau}$ and $\hat{\phi}$ in bootstrap analysis of cyanobacteria circadian data. 37

13 Plot of bioluminescence vs time. Data are recorded every 2.116 hours. 40

14 Spline fitted curves for overall, $\beta_0(t)$ and $\beta_1(t)$ in Example 3.1, with smoothing parameter chosen by GCV. No results for model (3.1), $limnla = -0.97$ for model (3.2), and $limnla = -0.70$ for model (3.3). The bold lines are the true curves, the green dashed lines are estimated curves under model (3.2) and the blue dot lines are estimated curves under model (3.3). 45

15 Spline fitted curves for overall, $\beta_0(t)$ and $\beta_1(t)$ in Example 3.1, with smoothing parameter chosen by adjusted CV. $limnla = \log_{10}(n\lambda) = 7.6$ for all three models. The bold lines are the true curves, the red solid lines are the estimated curves under model (3.1), the green dashed lines are the estimated curves under model (3.2), and the blue dot lines are the estimated curves under model (3.3). 47

16 Spline fitted curves for overall, $\beta_0(t)$ and $\beta_1(t)$ in Example 3.2. $limnla = \log_{10}(n\lambda) = 7.4$ for all three models. The bold lines are the true curves, the red solid lines are the estimated curves under model (3.1), the green dashed lines are the estimated curves under model (3.2), and the blue dot lines are the estimated curves under model (3.3). 51

FIGURE Page

17 Spline fitted curves for overall, $\beta_0(t)$ and $\beta_1(t)$ in Example 3.3. $limnla = \log_{10}(n\lambda) = 7.1$ for all three models. The bold lines are the true curves, the red solid lines are the estimated curves under model (3.1), the green dashed lines are the estimated curves under model (3.2), and the blue dot lines are the estimated curves under model (3.3). 53

18 Spline fitted curves for overall, $\beta_0(t)$ and $\beta_1(t)$ in cyanobacteria circadian data. $limnla = 8.2$ for all three models. The red solid lines are the estimated curves under model (3.1), the green dashed lines are the estimated curves under model (3.2), and the blue dot lines are the estimated curves under model (3.3). 55

19 Plot of bioluminescence vs time. 58

20 Plot of the diagonal values of the hat matrix H vs t in Example 4.1. The top plot is for model (4.1), the middle plot is for model (4.2) and the bottom plot is for model (4.3). Interior points ($t > min(t) + h/2$ and $t < max(t) - h/2$) are the red solid points between two dashed lines. 68

CHAPTER I

INTRODUCTION

Semiparametric models use different structures to best accommodate the problems in hand while allow modeling flexibility. A useful class of models is that of varying coefficient models. Varying coefficient models have applications on many aspects and disciplines. This dissertation investigates the application of a useful extension of varying coefficient models, semiparametric varying coefficient periodic models, on analyzing the circadian patterns of cyanobacteria. It contains three components. In chapter II, we propose semiparametric varying coefficient periodic models and provide estimating procedures for the circadian parameters. In chapter III, we investigate further properties that are of biological interest using model selection techniques. This is performed under the framework of smoothing spline. Chapter IV proposes procedures for hypothesis testing under the framework of local linear regression. The rest of this chapter provides the biological background behind chapter II, III and IV. Section 1.1 provides the definition and characteristics of circadian programs. Section 1.2 describes the cyanobacterial circadian clock. The overview structure of this dissertation is provided in Section 1.3.

1.1 What Is a Circadian Program?

Circadian rhythms are endogenous biological programs that time metabolic and/or behavioral events to occur at optimal phases of the daily cycle. They have three diag-

---

The format and style follow that of *Biometrics*.

nostic characteristics. The first is that in constant conditions, the programs free-run with a period that is about 24 hours in duration. The second is that, in an appropriate environmental cycle (usually a light-dark and/or temperature cycle), the rhythm will take on the period of the environmental cycle, that is, circadian rhythms will entrain to the environmental cycle. The final characteristic is that the period of the free-running rhythm is nearly the same at different constant ambient temperatures within the physiological range; that is, circadian rhythms are temperature compensated. It is these three characteristics that define circadian rhythms, not the details of their biochemical mechanisms. Indeed, questions of considerable interest are whether circadian mechanisms have evolved more than once and, if so, whether completely different biochemical processes have been harnessed to the task in different organisms. The fascination of circadian rhythms is how a biochemical mechanism can keep time so precisely over such a long time constant (about 24 hours) at different ambient temperatures (Johnson and Golden, 1999).

1.2 The Cyanobacterial Circadian Clock

Like the circadian clocks of plants, animals and fungi, the cyanobacterial clock generates rhythms of biological processes that exhibit an approximate 24-hour period even in the absence of an environmental cycle, can be synchronized with the environment through light or temperature cues, and maintain a nearly constant period over a range of physiologically relevant temperatures (Golden, 2003).

The revelation that gene expression, as reflected by luciferase reporter fusions, is under circadian control allowed the automated monitoring of oscillations in bioluminescence as a readout from the clock (Figure 1). Cells are incubated in a 12-hour light: 12-hour dark cycle to synchronize to the clock, then released into continuous light (time 0), the so called constant condition, for several days. From Figure 1, we

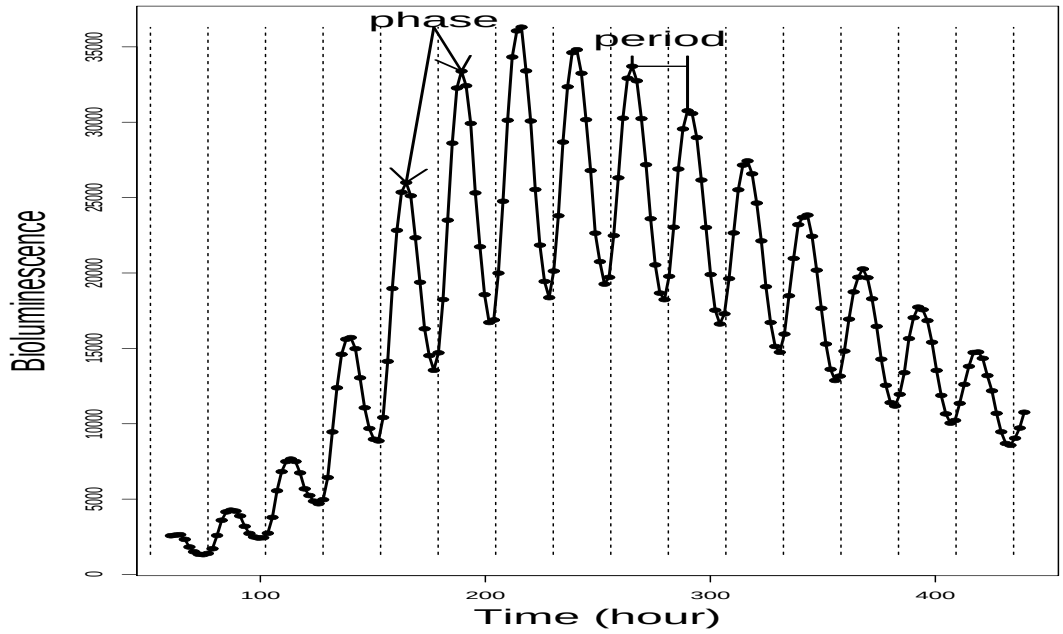

Figure 1: Plot of bioluminescence vs time, data were recorded every 2.116 hours. X-axis, time in hours. Y-axis, bioluminescence. Time is divided by dashed lines to indicate subjective days. The parameters of circadian period and relative phase are indicated.

observe a clear periodic pattern with period and phase, baseline of the whole curve and amplitude of the periodic component. Both baseline and amplitude are changing with time.

The bioluminescence patterns have three stages. In early stage, the luminescence patterns display a clear circadian rhythm that grows in baseline and amplitude exponentially (exponential stage). After the growth rate of the culture slows, the luminescence pattern stabilizes into a circadian pattern of consistent baseline and amplitude (sustained stage). As the culture ages, the rhythm slowly damps (senescent stage). The damping observed in the senescent stage probably results from nutrient depletion (Johnson and Golden, 1999).

These rhythmic colonies exhibited a range of waveforms and amplitudes, and they also showed at least two predominant phase relationships. We defined Class-1

genes as those whose expression peaks at the end of the day (peak at dusk; trough at dawn) and Class-2 genes as those peaking at the end of the night (peak at dawn; trough at dusk) (Johnson and Golden, 1999). Data in Figure 1 can be classified as Class-1 genes.

1.3 Structure Overview

Chapter II proposes a varying coefficient periodic model, and semiparametric estimation procedures. This chapter contains a traditional discrete Fourier transformation method, our proposed model, semiparametric kernel based methods, asymptotic properties of the semiparametric methods and numeric results. Chapter III is devoted to the study of the relationship between the baseline and amplitude functions. We present this as a problem under the general framework of smoothing spline. This chapter focuses on the solution of the identifiability problem for smoothing spline and presents simulations and application to cyanobacteria circadian data to select the optimal model. Chapter IV explores the relationship between the baseline and amplitude functions by hypothesis testing. The main accomplishment of this chapter is the derivation of the global degree of freedom for residual sum of squares. Chapter V concludes our studies and considers a future research topic of testing if the period and phase of cyanobacteria data are constants. Proofs of the theorems and lemmas are detailed in the appendices.

CHAPTER II

ESTIMATION OF CIRCADIAN PARAMETERS IN CYANOBACTERIA VIA SEMIPARAMETRIC VARYING COEFFICIENT PERIODIC MODELS

2.1 Introduction

Circadian rhythms are endogenous biological programs that coordinate a living being's internal/external functions to match the phase of the daily cycle. Cyanobacterium, *Synechococcus elongates PCC 7942*, is the simplest organism, in terms of genome size and unicellular structure, that is known to possess an endogenous circadian clock. The study of circadian patterns in cyanobacteria provides a powerful tool to assist researchers to better understand the circadian input pathways. For cyanobacteria, rhythmic gene expressions can be monitored by bioluminescence produced from luciferase reporter gene(s). The gene expression patterns, measured by luminescence, tend to reflect the stage of growth cycle of cyanobacterium colonies. For example, in early stage ("exponential stage"), the luminescence patterns display an exponential growth in the number of cells in the culture. This growth is reflected by, e.g., the growing amplitude. When the growth rate slows, the luminescence pattern stabilizes into a so called "sustained stage". As the culture ages, the pattern damps into a "senescent stage" which is believed to be a consequence of the stress from nutrient depletion (Johnson and Golden, 1999). Figure 2 displays bioluminescence patterns collected from CikA (circadian input kinase) at four different cultures. CikA is a key component of the circadian clock input pathway of cyanobacteria. A cikA null mutant strain could show shorten period, reduced amplitude and lack of ability in sensing environmental change (Zhang, Dong, and Golden, 2006). The plots in the top panel illustrate patterns that contain the exponential and sustained phases,

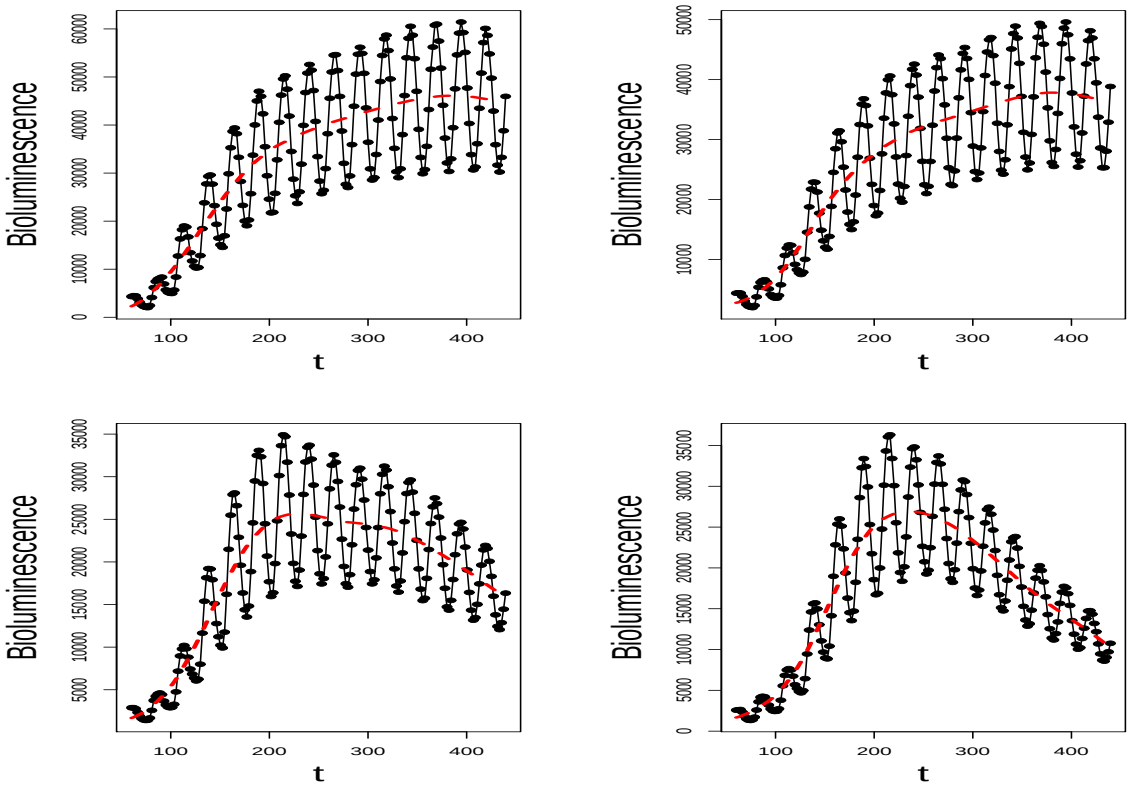

Figure 2: Four data curves. Plot of bioluminescence vs time, data were recorded every 2.116 hours. The plots in the top panel have two stages of growth cycle (exponential stage and sustained stage), while the plots in the bottom panel have three stages (exponential stage, sustained stage and senescent stage). The red dashed lines are the underlying baseline functions.

while the plots in the bottom panel display all three stages. In the past, researchers tend to focus on observations collected from the sustained stage only. This practice has changed though the same method is applied to analyze the data. The use of one parametric model to capture patterns at different stages and from different cultures appears to be challenging.

A fast Fourier transformation nonlinear least square (FFT-NLLS) algorithm is by far the most commonly used approach to analyze circadian rhythm data (e.g. Dodd et al., 2005; Niinuma et al., 2005). However, the performances of this approach are

not always satisfactory. We will illustrate some potential drawbacks with examples in Section 2.2. Our numerical illustrations will also show that our proposed methods could greatly improve the estimation precision of certain key parameters in Section 2.6.

As researchers realize that parametric models might be inadequate to capture certain underlying relationships between response variables and the associated covariates in some practical situations, there has been upsurge of interests and effort in the development of nonparametric and semiparametric models and methods; see, e.g., Hastie and Tibshirani (1990), Green and Silverman (1994), Wand and Jones (1995) and Fan and Gijlels (1996), among others. The most relevant literature to this manuscript is that of varying coefficient models. The modeling techniques were systematically investigated in seminal work of Cleveland, Grosse, and Shyu (1991) and Hastie and Tibshirani (1993). For this kind of models, Fan and Zhang (1999) proposed a two-step procedure to accommodate varying degrees of smoothness among coefficient functions. The varying coefficient partially linear model is a useful extension of varying coefficient model, and has been investigated by Zhang, Lee, and Song (2002) and Li et al. (2002). Their estimation and inference procedures are further systematically studied by Fan and Huang (2005).

In this paper, we extend the varying coefficient partially linear model to a scenario where a periodic component is part of the semiparametric varying coefficient model. One embedded difficulty is that the most natural model has a global identifiability issue such that a direct application of spline based methods can result multiple answers. Our strategy is to use kernel based methods, approximate the model locally and show that one solution, which is close to our initial estimate, is the correct one to use to addresses the biological questions. More details regarding local identifiability and the associated conditions will be provided in Section 2.3 and 2.4.4, respectively.

There are other models and spline-based approaches that have been used to analyze other circadian rhythm data. For example, Wang and Brown (1996) developed a flexible shape-invariant model using a periodic spline function as the common basis curve, and estimate the individual's mean, phase and amplitude. They assume that the mean, phase and amplitude are scalar and do not change with time. Equivalently, Luan and Li (2003) adopted a similar modeling structure when they analyzed microarray time course data. Figure 2 has clearly shown that there may not be a common basis curve across different cultures. That is, the proposed approaches above are not applicable for the more general cases.

The rest of this chapter is organized as follows. Section 2.2 briefly describes a traditional approach of FFT-NLLS. Sections 2.3 and 2.4 provide the model we study and our proposed parametric and semiparametric methods, respectively. In Section 2.5, we investigate the asymptotic properties of the semiparametric methods. The numerical studies of their performances and analysis of circadian dataset are provided in Section 2.6. Section 2.7 contains concluding remarks. All theoretical conditions and proofs are relegated to Appendix A and B.

2.2 Traditional Approach

Traditionally, the period and phase of circadian rhythm data were estimated using the Fourier transformation based methods, which we briefly describe in Section 2.2.1. However, the performances of such approaches are not always satisfactory. For example, when the error variation increases and/or when the coefficient functions vary with time, the variation of the estimates of the rhythm parameters increases. In Section 2.2.2, we illustrate some potential drawbacks with examples.

2.2.1 Analysis via Periodogram and Fourier Transformation

Let Y and t denote the response and covariate variables, respectively, and let ε be the error term. A plausible model for periodic data with discrete time points, $t = 0, \cdots, n-1$, is

$$y_t = \beta_0 + \beta_1 \cos\{2\pi(t/\tau - \phi)\} + \varepsilon_t, \tag{2.1}$$

where τ denotes the period and ϕ denotes the phase of the periodic component. It is common to assume that $\mathrm{E}(\varepsilon_t) = 0$, and $\mathrm{var}(\varepsilon_t) = \sigma_\varepsilon^2$. Notice by this setup, the range of ϕ should be the interval $[0, 1)$; the growth baseline, β_0, and the amplitude of the periodic component, β_1, are assumed to be time invariant. We can also assume that both β_0 and β_1 non-negative. If $\beta_1 > 0$, then a rhythm exists, and y_t reaches its peak at $t = \tau\phi$. This y_t repeats itself with a period of τ.

The most commonly implemented approach to estimate τ, ϕ, β_0 and β_1 is through discrete Fourier transformation. We let w denote the frequency of the periodic component, that is, $w = 2\pi/\tau$, and let $w_j = 2\pi j/n$ to be the so called jth Fourier frequency, where $j = -n/2, \cdots, -1, 1, n/2$. The discrete Fourier transform (DFT) of y_t is defined to be

$$J(w_j) = (1/n) \sum_{t=0}^{n-1} y_t e^{-iw_j t};$$

while the periodogram of y_t at frequency w_j is $I(w_j) = (n/2\pi)|J(w_j)|^2$.

A peak in the periodogram at a provided frequency indicates a strong harmonic component at that frequency in $\{y_t\}$. Consequently, one can estimate the frequency w by the argument that maximizes the realization of the periodogram:

$$\hat{w} = \{w_j : \max\, I(w_j)\},$$

$$\hat{\tau} = 2\pi/\hat{w},$$

$$\hat{\phi} = \arctan\{-\mathrm{Im}J(\hat{w})/\mathrm{Re}J(\hat{w})\}/2\pi,$$

where $\mathrm{Re}J(\hat{w})$ and $\mathrm{Im}J(\hat{w})$ are the real and imaginary components of $J(\hat{w})$, respectively. By minimizing

$$\sum_{t=0}^{n-1}\Big[y_t-\beta_0-\beta_1\cos\{2\pi(t/\hat{\tau}-\hat{\phi})\}\Big]^2,$$

we obtain

$$\hat{\beta}_0=\frac{1}{n}\sum_{t=0}^{n-1}y_t=J(0),\ \hat{\beta}_1=2\sqrt{\mathrm{Re}J(\hat{w})^2+\mathrm{Im}J(\hat{w})^2}.$$

This estimation algorithm is referred to as "Fast Fourier Transformation – Nonlinear Least Square" (FFT-NLLS) approach in the Chronobiology literature. Note that in model (2.1), t is set to be in $\{0,\cdots,n-1\}$, which is not always true in real applications. A pre-application standardization procedure is often adopted. For more details on DFT, please see Bloomfield (1976) and Brockwell and Davis (1996).

2.2.2 Potential Drawbacks of the Traditional Method

We list two potential drawbacks that we have encountered of the traditional method. First, DFT cannot properly estimate intercept and amplitude when they change with time. If one ignores the fact and apply DFT regardlessly, the outcomes tend to suffer due to the model mis-specification. For the last dataset in Figure 2, we plot the DFT estimated curve in Figure 3. From Figure 3 we observe that the traditional estimates are way off the true values. The estimate of period τ is roughly acceptable, but other estimates behave very badly.

The second drawback is that DFT is sensitive to large-size noises. The effect of magnitude of noises can be seen at the following example.

Example 2.1. Let $y_t=1+3\cos(0.5t-2)+\varepsilon_t$, $t=0,\cdots,99$, with $\varepsilon_t\sim N(0,s^2)$. That is, $n=100$, $w=0.5$, $\tau=4\pi$, $\phi=1/\pi$, $\beta_0=1$, $\beta_1=3$ and $\sigma_\varepsilon^2=s^2$. To know how $\hat{\tau}$, $\hat{\phi}$, $\hat{\beta}_0$ and $\hat{\beta}_1$ perform when s gets larger, we plot $\hat{\tau}$, $\hat{\phi}$, $\hat{\beta}_0$ and $\hat{\beta}_1$ versus s in Figure 4, where the values of s vary from 0 to 8, with each two adjacent points

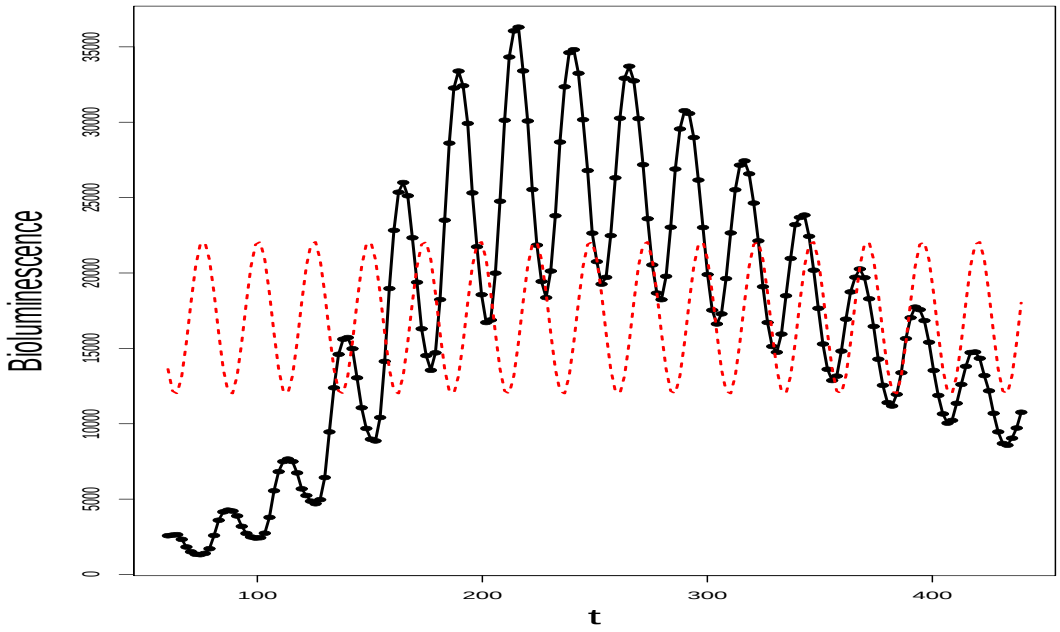

Figure 3: The bottom-right curve in Figure 2 and the DFT estimated curve. The black solid line is the observed curve, and the red dashed line is the DFT estimated curve.

0.08 apart. Figure 4 suggests that the quality of estimation deteriorates for a large s. This phenomenon is especially obvious for $\hat{\beta}_0$ and $\hat{\beta}_1$. That is, FFT-NLLS is quite sensitive towards noises. Compared to $\hat{\beta}_0$ and $\hat{\beta}_1$, $\hat{\tau}$ and $\hat{\phi}$ are relatively stable until s becomes fairly large.

2.3 Varying Coefficient Periodic Models

For cyanobacteria circadian rhythm data, we observe that each item possesses an overall smooth growth trend as well as a periodic pattern; both terms change over time. Let y and t be the response and covariate variable, respectively. We consider the following varying coefficient periodic model:

$$y_{ij} = \beta_{0i}(t_{ij}) + \beta_{1i}(t_{ij}) \cos\{2\pi(t_{ij}/\tau_i - \phi_i)\} + \varepsilon_{ij}, i = 1, \cdots, m; j = 1, \cdots, n_i, \quad (2.2)$$

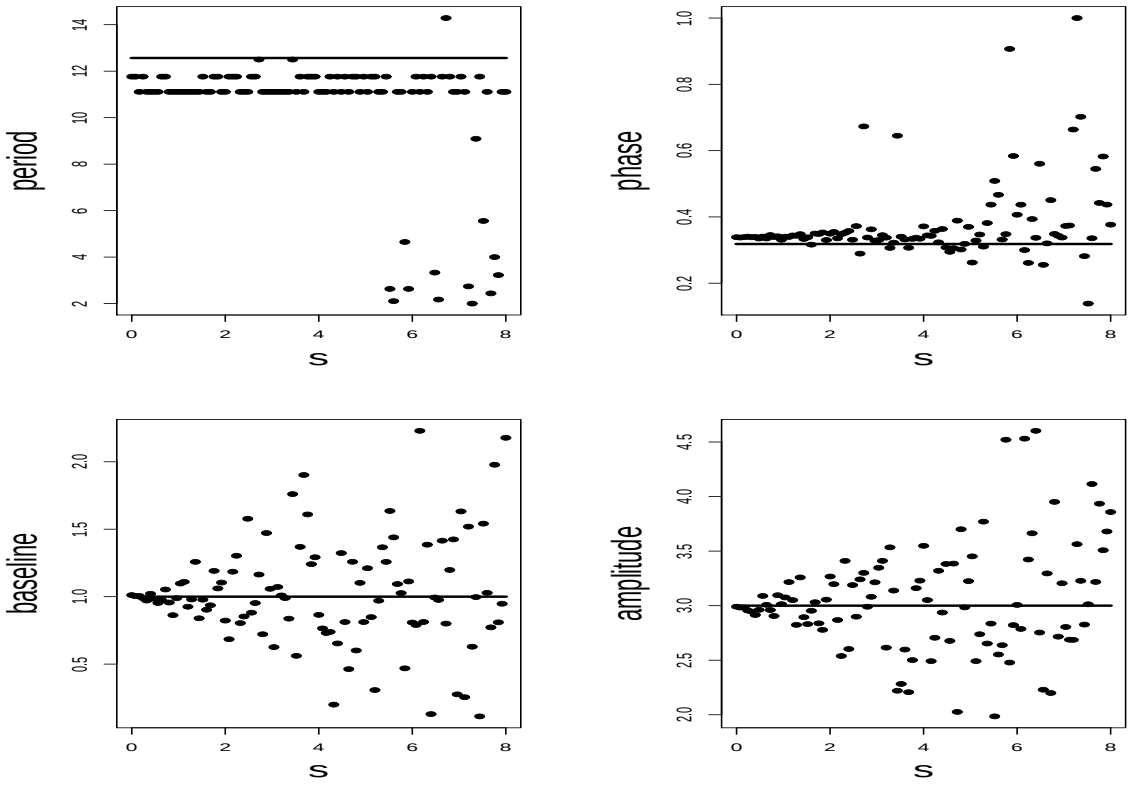

Figure 4: DFT estimates of τ, ϕ, β_0 and β_1 vs magnitude of noise s in Example 2.1, from left to right. The dots are our estimated parameter values through DFT for different s; the solid lines provide the true parameter values.

where y_{ij} is the response of ith individual at the jth time point t_{ij}; $\beta_{0i}(t)$ is the underlying baseline growth function of the ith individual; $\beta_{1i}(t) > 0$ is the amplitude function of the periodic component of the ith individual; τ_i denotes the period of the ith individual; $0 \leq \phi_i < 1$ is the phase of the ith individual; n_i is the total number of observations of subject i; and m is the total number of subjects. We assume that the error terms ε_{ij} are independent and identically distributed with $\mathrm{E}(\varepsilon_{ij}) = 0$, and $\mathrm{var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$. One advantage of this model is that it allows the baseline and amplitude functions to vary with time, which reflect the growth patterns that biologists believe in. It also contains the period and phase parameters τ and ϕ,

so we can conveniently study the property of the periodic component by evaluating these two estimated parameters. Note that, one can replace $\cos(\cdot)$ by another suitable parametric function, and our methods could still be applicable.

Let $\theta_i = (\tau_i, \phi_i)^T$. We assume all θ_i's are random samples from the same distribution with mean θ and covariance Σ_θ. Our ultimate goal is to provide interval estimation of $\theta = (\tau, \phi)^T$, which are the population means of the period and phase in the data. The covariance Σ_θ measures the between-item variation of θ_i and its magnitude is also of interest.

Without loss of generality and slightly abusing the notations, we assume that observations from all items are collected at the same time points and present the general model as

$$Y_i = \sum_{k=0}^{p} \beta_{ki}(t) X_k(t; \theta_i) + \boldsymbol{\varepsilon}_i, i = 1, \cdots, m. \tag{2.3}$$

where $t = (t_1, \cdots, t_n)^T$, $Y_i = (y_{i1}, \cdots, y_{in})^T$, and $\boldsymbol{\varepsilon}_i = (\varepsilon_1, \cdots, \varepsilon_n)^T$. $X_k(t; \theta_i)$ are equivalently defined as Y_i but $X_k(t; \theta_i)$ could contain the unknown parameter θ_i. In our special case, $X_0(t; \theta_i) = (1, \cdots, 1)^T = \mathbf{1}^T$, and $X_1(t; \theta_i) = \cos\{2\pi(t/\tau_i - \phi_i)\}$.

Evaluating model (2.2), we notice an identifiability problem in the model. That is, one can hardly identify $\beta_0(t)$ and $\beta_1(t)$. One extreme case is that we set $\beta_1(t)$ equals to 0, and $E(Y|t) = \beta_0(t)$, so the circadian pattern with be captured by $\beta_0(t)$ alone. Obviously in this case, we can not obtain the desired estimates of the period τ, the phase ϕ, the baseline function $\beta_0(t)$ and the amplitude function $\beta_1(t)$. We will address this issue again after our semiparametric estimation methods are provided in Section 2.4.

2.4 Parametric and Semiparametric Estimation Methods

In this Section, we provide parametric and semiparametric estimation methods to estimate the circadian data. We introduce a parametric method in Section 2.4.1 and

semiparametric methods in Section 2.4.2. Section 2.4.3 contains details, including bandwidth selection and initial estimates of the parameters, that are relevant to the semiparametric procedures. Identifiability problem is addressed in Section 2.4.4.

2.4.1 Quadratic Varying Coefficient Method

In the traditional approaches, the estimates of the baseline, β_0, and amplitude, β_1 can only be constants. To improve this, we assume, for item i, both $\beta_{0i}(t)$ and $\beta_{1i}(t)$ are quadratic functions of time t in model (2.2). Also, from Example 2.1, the estimates of τ and ϕ are relatively stable for noises, so, we use estimates of τ and ϕ from DFT, as our initial estimates to update $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$. We refer to this approach as the quadratic varying coefficient method (QVC).

For item i, the estimation steps can be described as below:

1. Estimate τ_i and ϕ_i by DFT, denote them as $\hat{\tau}_i^{[0]}$ and $\hat{\phi}_i^{[0]}$, and let $\hat{\theta}_i^{[0]} = (\hat{\tau}_i^{[0]}, \hat{\phi}_i^{[0]})^T$.

2. Assume $\beta_{0i}(t)$ and $\beta_{1i}(t)$ are quadratic functions of t, then model (2.3) becomes

$$Y_i \;=\; (b_{00i} + b_{01i}t + b_{02i}t^2) + (b_{10i} + b_{11i}t + b_{12i}t^2)X_1(t;\theta_i^{[0]}) + \boldsymbol{\varepsilon}_i.$$

   Estimate $b_{00i}, b_{01i}, b_{02i}, b_{10i}, b_{11i}, b_{12i}$ by fitting the above linear model, and obtain the estimates of $\beta_{0i}(t)$ and $\beta_{1i}(t)$ by

$$\hat{\beta}_{0i}(t) = \hat{b}_{00i} + \hat{b}_{01i}t + \hat{b}_{02i}t^2, \text{ and } \hat{\beta}_{1i}(t) = \hat{b}_{10i} + \hat{b}_{11i}t + \hat{b}_{12i}t^2.$$

3. Find $\hat{\tau}_i$ and $\hat{\phi}_i$ by minimizing residual sum of squares $\text{RSS} = \sum_{j=1}^{n}(y_{ij} - \hat{y}_{ij})^2$ in the neighborhood of $\hat{\tau}_i^{[0]}$ and $\hat{\phi}_i^{[0]}$, where $\hat{y}_{ij} = \hat{\beta}_{0i}(t_j) + \hat{\beta}_{1i}(t_j)\cos\{2\pi(t/\hat{\tau}_i - \hat{\phi}_i)\}$.

Each θ_i can be estimated by $\hat{\theta}_i = (\hat{\tau}_i, \hat{\phi}_i)^T$. We repeat steps 1-3 for all m items, and obtain $\hat{\theta}_i$, $i = 1, \cdots, m$. We let the final estimates of θ and Σ_θ to be the sample mean and sample variance of $\hat{\theta}_i$'s:

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i, \;\; \hat{\Sigma}_\theta = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \hat{\theta})^{\otimes 2}.$$

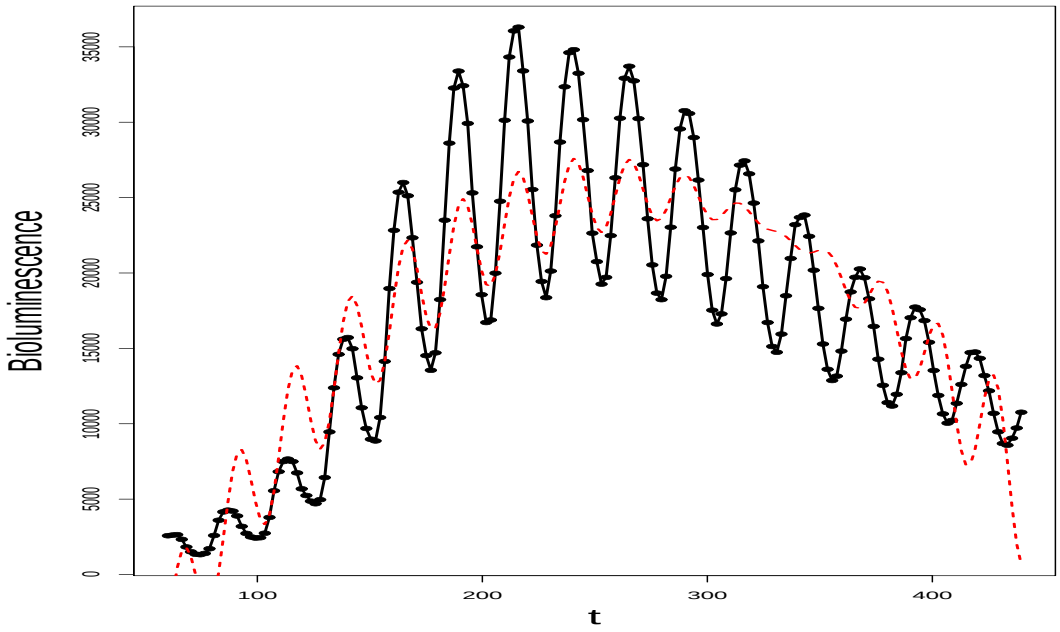

Figure 5: The bottom-right curve in Figure 2 and the QVC estimated curve. The black solid line is the observed curve, and the red dashed line is the QVC estimated curve.

We expect that when $\hat{\beta}_{0i}(t)$ and $\hat{\beta}_{1i}(t)$ are better estimates than what we obtain from DFT, $\hat{\tau}_i$ and $\hat{\phi}_i$ will be more accurate than $\hat{\tau}_i^{[0]}$ and $\hat{\phi}_i^{[0]}$.

For the last data set in Figure 2, the QVC estimated curve is provided in Figure 5. By Figure 5, the estimates of the period τ, phase ϕ and baseline function $\beta_0(t)$ seem better than the traditional method, while the estimate of the amplitude function $\beta_1(t)$ is much worse. This is because the initial estimates of τ and ϕ comes from DFT is not accurate, and estimate of $\beta_1(t)$ depends on these initial estimates greatly. Because of the model set up in (2.2), $\hat{\beta}_0(t)$ is less affected than $\hat{\beta}_1(t)$.

2.4.2 Semiparametric Local Linear Varying Coefficient Methods

2.4.2.1 Component-wise update estimation

Quadratic varying coefficient method improves over the traditional approach. However, if $\beta_0(t)$ and $\beta_1(t)$ are not quadratic functions of t, QVC may result in poor estimates for $\beta_0(t)$ and $\beta_1(t)$, and consequently we obtain poor estimate for θ. When the parametric forms of $\beta_0(t)$ and $\beta_1(t)$ are unknown, it is natural to estimate them nonparametricly. We use the local linear approach to estimate the nonparametric component and use least squares to estimate the parametric component, update the estimates iteratively and component-wise until residual sum of squares (RSS) achieves the minimum. We refer to this method as the semiparametric local linear varying coefficient method with component-wise update estimation (SVCC). The execution of SVCC is as below:

For subject i, let $\hat{\tau}_i^{[0]}$ and $\hat{\phi}_i^{[0]}$ be the initial estimates of τ_i and ϕ_i, $\hat{\beta}_{0i}^{[0]}(t)$ and $\hat{\beta}_{1i}^{[0]}(t)$ be the initial estimates of $\beta_{0i}(t)$ and $\beta_{1i}(t)$. Let $\hat{\tau}_i^{[k-1]}$, $\hat{\phi}_i^{[k-1]}$, $\hat{\beta}_{0i}^{[k-1]}(t)$ and $\hat{\beta}_{1i}^{[k-1]}(t)$ denote the estimates at step $(k-1)$. At the kth step,

1. We find $\hat{\tau}_i^{[k]}$ and $\hat{\phi}_i^{[k]}$ by minimizing RSS $= \sum_{j=1}^{n}(y_{ij} - \hat{y}_{ij})^2$ in the neighborhood of $\hat{\tau}_i^{[k-1]}$ and $\hat{\phi}_i^{[k-1]}$, where $\hat{y}_{ij} = \hat{\beta}_{0i}^{[k-1]}(t_j) + \hat{\beta}_{1i}^{[k-1]}(t_j)\cos\{2\pi(t_j/\hat{\tau}_i - \hat{\phi}_i)\}$.

2. Suppose that τ_i and ϕ_i are known and equal to $\hat{\tau}_i^{[k]}$ and $\hat{\phi}_i^{[k]}$, model (2.3) becomes

$$Y_i = \beta_{0i}(t) + \beta_{1i}(t)X_1(t;\hat{\theta}_i^{[k]}) + \boldsymbol{\varepsilon}_i.$$

   Our model is a common varying coefficient model. We use the local linear approach to update $\hat{\beta}_{0i}(t)$ and $\hat{\beta}_{1i}(t)$ separately with bandwidths h_0 and h_1. The estimates are

$$\hat{\beta}_{0i}^{[k]}(t_0) = (1,0)(\mathbf{X}_0^T W_0 \mathbf{X}_0)^{-1}\mathbf{X}_0^T W_0 Y_i,$$

$$\hat{\beta}_{1i}^{[k]}(t_0) = (1,0)(\mathbf{X}_1^T W_1 \mathbf{X}_1)^{-1}\mathbf{X}_1^T W_1 Y_i, \tag{2.4}$$

where

$$\mathbf{X}_0 = \begin{pmatrix} 1 & t_1 - t_0 \\ \vdots & \vdots \\ 1 & t_n - t_0 \end{pmatrix}, \ \mathbf{X}_1 = \begin{pmatrix} X_1(t_1;\hat{\theta}_i^{[k]}) & X_1(t_1;\hat{\theta}_i^{[k]})(t_1 - t_0) \\ \vdots & \vdots \\ X_1(t_n;\hat{\theta}_i^{[k]}) & X_1(t_n;\hat{\theta}_i^{[k]})(t_n - t_0) \end{pmatrix},$$

$$W_0 = \text{diag}(K_{h_0}(t_1 - t_0), \cdots, K_{h_0}(t_n - t_0)),$$

$$\text{and } W_1 = \text{diag}(K_{h_1}(t_1 - t_0), \cdots, K_{h_1}(t_n - t_0)).$$

Iterate the algorithm until RSS stops decreasing, we obtain $\hat{\theta}_i = (\hat{\tau}_i, \hat{\phi}_i)^T$, $\hat{\beta}_{0i}(t)$, and $\hat{\beta}_{1i}(t)$, for $i = 1, \cdots, m$. And let the final estimates to be

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i, \ \hat{\Sigma}_\theta = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \hat{\theta})^{\otimes 2}. \tag{2.5}$$

For the last dataset in Figure 2, the SVCC estimated curve is plotted in Figure 6. By Figure 6, we see great improvement of estimating data. The estimated curve fit the observed data very well, but it is off when t is larger than 400.

2.4.2.2 Simultaneously update estimation

To have more accurate updates, instead of updating $\beta_{0i}(t)$ and $\beta_{1i}(t)$ component-wise as in SVCC, we update them simultaneously to use all the information. We first use the same bandwidth h for $\beta_0(t)$ and $\beta_1(t)$. In step 2 of the kth step in Section 2.4.2.1, instead of estimating $\beta_0(t)$ and $\beta_1(t)$ using equation (2.4), We estimate them by

$$\hat{\beta}_{0i}^{[k]}(t_0) = (1,0,0,0)(\mathbf{X}^T W \mathbf{X})^{-1}\mathbf{X}^T W Y_i,$$

$$\hat{\beta}_{1i}^{[k]}(t_0) = (0,0,1,0)(\mathbf{X}^T W \mathbf{X})^{-1}\mathbf{X}^T W Y_i, \tag{2.6}$$

where $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1)$, and $W = \text{diag}(K_h(t_1 - t_0), \cdots, K_h(t_n - t_0))$. We refer to this approach as the semiparametric local linear varying coefficient method with simultaneously update estimation using a common bandwidth (SVCSc).

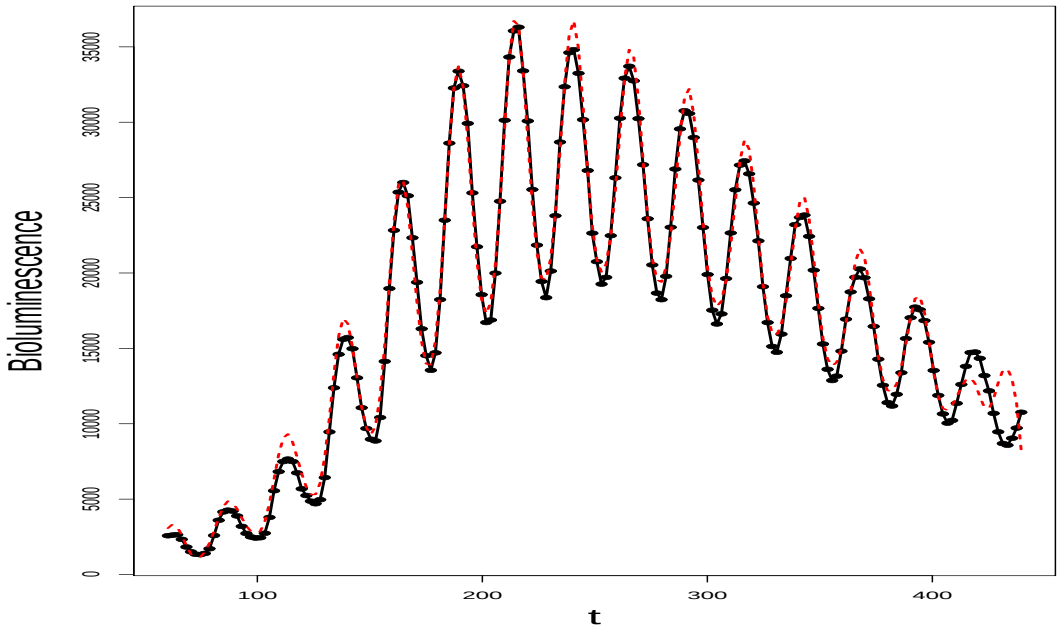

Figure 6: The bottom-right curve in Figure 2 and the SVCC estimated curve. The black solid line is the observed curve, and the red dashed line is the SVCC estimated curve.

For the last dataset in Figure 2, the SVCSc estimated curve is provided in Figure 7. We see that taking all the information into account does improve the estimates in Figure 6 (SVCC), especially when $t > 400$.

Since $\beta_0(t)$ and $\beta_1(t)$ may have different level of smoothness, we are interested in knowing whether updating them simultaneously with different bandwidths h_0 and h_1 could improve the performance. In step 2 of the kth step in Section 2.4.2.1, The estimates for $\beta_0(t)$ and $\beta_1(t)$ are

$$\hat{\beta}_{0i}^{[k]}(t_0) = (1,0,0,0)(\mathbf{X}^T W_0 \mathbf{X})^{-1}\mathbf{X}^T W_0 Y_i,$$

$$\hat{\beta}_{1i}^{[k]}(t_0) = (0,0,1,0)(\mathbf{X}^T W_1 \mathbf{X})^{-1}\mathbf{X}^T W_1 Y_i.$$

We refer to this approach as semiparametric local linear varying coefficient method with simultaneously update estimation using different bandwidths (SVCSd).

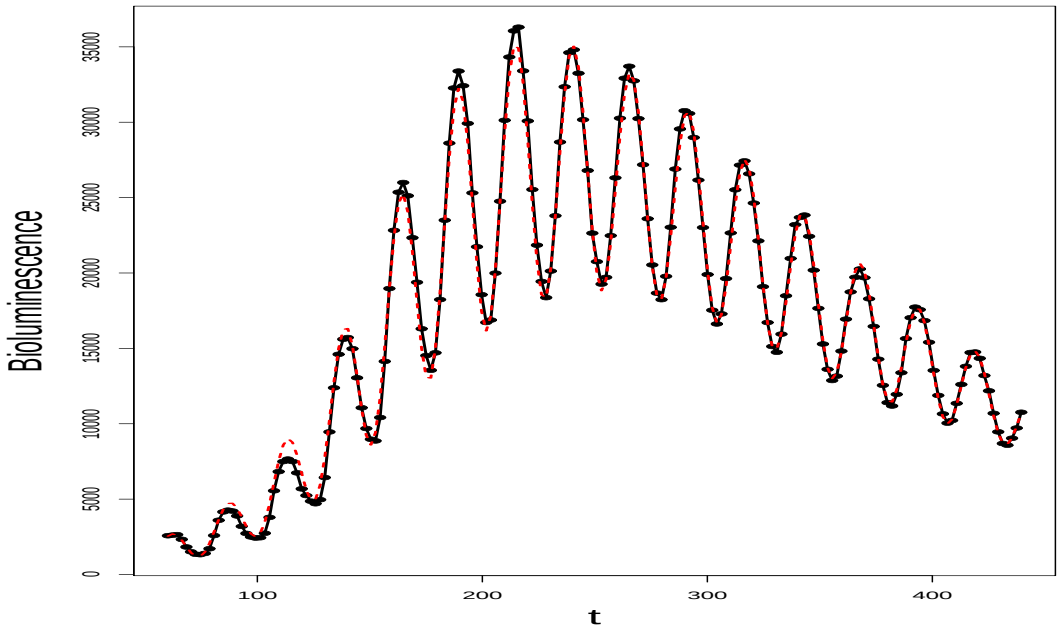

Figure 7: The bottom-right curve in Figure 2 and the SVCSc estimated curve. The black solid line is the original curve, and the red dashed line is the SVCSc estimated curve.

For the last dataset in Figure 2, the SVCSd estimated curve is provided in Figure 8. The estimated curve in Figure 8 is very similar to that in Figure 7, which suggests that $\beta_0(t)$ and $\beta_1(t)$ share similar smoothness.

2.4.3 Bandwidth Selection and Initial Estimates

In our semiparametric methods in Section 2.4.2, we need to select the optimal bandwidths, as well as a set of initial estimates. The simplest and most effective bandwidth selection method could be cross-validation. We perform the cross-validation method as below:

1. Divide time into c intervals, each has k points, then leave the 1 point out in each interval, so we will use the remaining $n - c$ points to produce estimates and make predicts to the left-out c points.

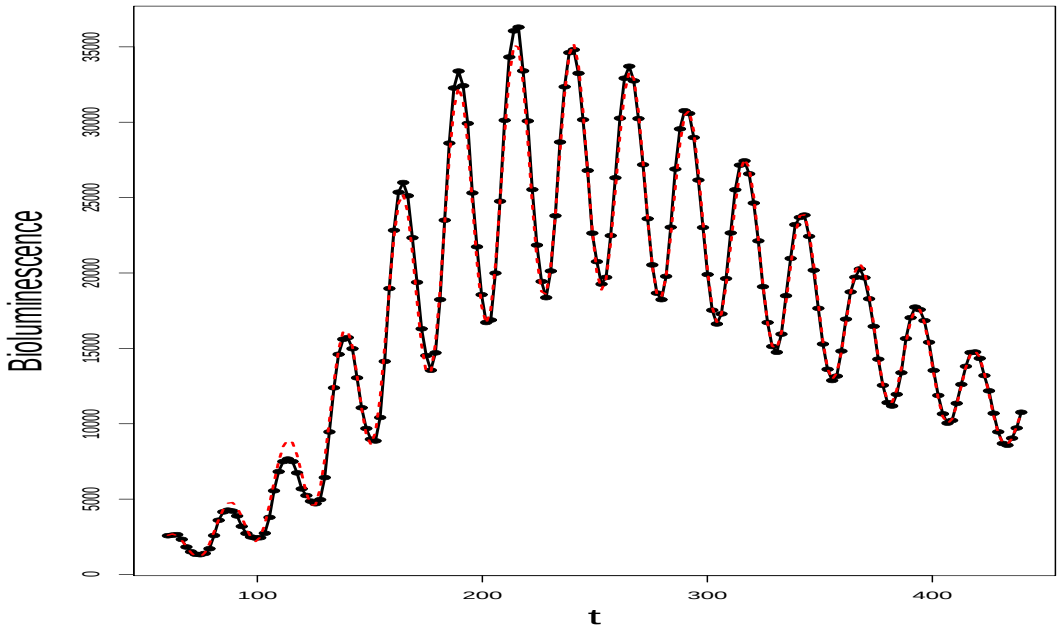

Figure 8: The bottom-right curve in Figure 2 and the SVCSd estimated curve. The black solid line is the original curve, and the red dashed line is the SVCSd estimated curve.

2. Select a h, estimate using these $n-c$ points, obtain $\hat{y}_{ck}^h$, obtain prediction mean square error (MSE)

$$G_{ck}(h) = \{y_{ck} - \hat{y}_{ck}^h\}^2.$$

3. For a provided C, repeat step 1. and 2. for c from 1 to C, and k from 1 to $K = \text{int}(n/c)$, sum the G_{ck} up, we obtain

$$G(h) = \sum_{c=1}^{C}\sum_{k=1}^{K} G_{ck}(h) = \sum_{c=1}^{C}\sum_{k=1}^{K}\{y_{ck} - \hat{y}_{ck}^h\}^2.$$

The optimal bandwidth should be $h = \text{argmin}\, G(h)$. The idea of choosing two bandwidths h_0 and h_1 is similar. Our experience shows that the bandwidth selected by cross-validation provide decent enough results, as shown in Section 2.6.

To estimate the unknown curves, we need to find a set of initial estimates. Here is how we obtain the initial estimates in Section 2.4.4.2: For item i,

1. Select the local minimum and maximum points.

2. Obtain the median of two adjacent minimum and maximum points, consider the medians as our baseline. Use local linear regression to fit $\hat{\beta}_{0i}^{[0]}(t)$ from these points.

3. Subtract minimum from adjacent maximum points, consider these as two times the amplitude. Use local linear regression to fit $\hat{\beta}_{1i}^{[0]}(t)$.

4. The distance from adjacent minimum and maximum points are half the period, average them to obtain the initial estimate of τ, $\hat{\tau}_i^{[0]}$. Using the maximum points as the start point of each cycle, and minimum points as middle point of each cycle, by averaging them we can obtain $\hat{\phi}_i^{[0]}$.

2.4.4 Identifiability Problem

As we mentioned in Section 2.3, our varying coefficient periodic model (2.2) is not global identifiable. To avoid the identifiability problem, we choose the initial estimates that are biologically meaningful and refine our estimates from there. Also, we use kernel based methods, semiparametric local linear varying coefficient methods, to approximate the model locally. In equation (2.6), we need to inverse the matrix $\mathbf{X}^T W \mathbf{X}$, which requires the matrix to be non-singular, so we must have a large enough bandwidth h. Further, if we only use partial data from one cycle, we could just use observations near the peak or near the trough; both would provide misleading outcomes. Consequently, we let the lower bound of the bandwidth to be around the period, 24. A large bandwidth ensure the smoothness of $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$, which leads to the results we desire.

2.5 Asymptotic Properties

We study the asymptotic properties of the estimates we obtain from our semiparametric methods in this Section. We focus on the performance of the estimates $\hat{\theta}$ from the semiparametric local linear varying coefficient method with simultaneously update estimation using a common bandwidth (SVCSc) throughout this Section.

Theorem 1. Under the regularity conditions provided in Appendix A,

$$\sqrt{m}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma), \tag{2.7}$$

where $\Sigma = \Sigma_\theta + \Sigma_\varepsilon$ and $\Sigma_\varepsilon = \frac{1}{n}\mathcal{K}_\theta^T(t)\mathcal{K}_\theta(t)\sigma_\varepsilon^2$.

Theorem 1 tells us that our $\hat{\theta}$ is a good estimate of θ. The variation Σ contains two components: Σ_θ and Σ_ε. Σ_θ is the between-item variation, and remains as a constant; while Σ_ε is the within-item variation, with the order of $1/n$. If n is large enough, Σ_ε is negligible. To prove Theorem 1, we need the following Lemma.

Lemma 1. Suppose θ_i is known for item i, under the regularity conditions provided in Appendix A,

$$\hat{\beta}_{ki}(t_0) - \beta_{ki}(t_0) = \{\mathcal{B}_k(t_0)h^2 + \frac{1}{\sqrt{nh}}\mathcal{K}_{ki}^T(t_0)\boldsymbol{\varepsilon}_i\}(1 + o_p(1)), \text{ for } k = 0, \cdots, p. \tag{2.8}$$

The proof of Theorem 1 and Lemma 1 are provided in Appendix A. Lemma 1 tells us that the MSE's of $\hat{\beta}_{0i}(t)$ and $\hat{\beta}_{1i}(t)$ are $O_P\{h^4 + (nh)^{-1}\}$ when θ_i's are known.

Also, we notice that our $\hat{\theta}$ is obtained from averaging the estimates of θ_i's as in (2.5). If we obtain the estimate of θ from EM algorithm, that is, we find $\hat{\theta}_{\text{EM}}$ that minimizes

$$\int \sum_{j=1}^{n_i} \{y_{ij} - \hat{y}_{ij}\}^2 g(\theta) d\theta_i,$$

where $g(\theta)$ is the marginal density of θ_i. If n_i's are large enough, it is easy to prove that

$$\hat{\theta}_{\text{EM}} - \theta = (\hat{\theta} - \theta)(1 + o_p(1)).$$

Thus, $\sqrt{m}(\hat{\theta}_{\text{EM}} - \theta) \xrightarrow{d} N(0, \Sigma)$ with the same Σ described in Theorem 1 (2.7). So our $\hat{\theta} = \dfrac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i$ performs as good as $\hat{\theta}_{\text{EM}}$ from EM algorithm if n is large enough.

2.6 Numerical Outcomes

Using the asymptotic property of θ provided in Theorem 1, we introduce some inference procedures for θ in Section 2.6.1. We study the numerical performance of the proposed approaches by simulation studies in Section 2.6.2. The analysis of the circadian dataset is provided in Section 2.6.3. Throughout this Section, we use the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$.

2.6.1 Inference Procedures for θ

According to Theorem 1 (2.7), the variation of $\hat{\theta}$ can be divided into a between-item variation Σ_θ and a within-item variation Σ_ε. We already propose several methods to estimate Σ_θ in Section 2.4. To study the relationship of Σ_θ and Σ_ε and obtain the estimate of the total variation Σ, we use bootstrap approach to obtain the estimate of the within-item variation Σ_ε of a given dataset. Then we can compute an approximate confidence interval for θ.

Given an original dataset with m subjects, the steps of estimating Σ_ε are as below: For the original dataset, we obtain $\hat{\theta}_i$, $\hat{\beta}_{0i}(t)$, $\hat{\beta}_{1i}(t)$ and $\hat{\Sigma}_\theta$, $i = 1, \cdots, m$. The fitted value of Y_i are $\hat{Y}_i = \hat{\beta}_{0i}(t) + \hat{\beta}_{1i}(t)\cos\{2\pi(t/\hat{\tau}_i - \hat{\phi}_i)\}$. For item i,

1. Generate B items by

$$Y^*_{ib} = \hat{\beta}_{0i}(t) + \hat{\beta}_{1i}(t)\cos\{2\pi(t/\hat{\tau}_i - \hat{\phi}_i)\} + \boldsymbol{\varepsilon}^*_{ib},\ b = 1, \cdots, B,$$

   where $\boldsymbol{\varepsilon}^*_{ib}$ is the bootstrap sample of the estimating error $\hat{\boldsymbol{\varepsilon}}_i = Y_i - \hat{Y}_i$. Notice that the between-item variation Σ_θ does not play a role here, because we use the same $\hat{\tau}_i$ and $\hat{\phi}_i$ for all B items.

2. Obtain $\hat{\theta}^*_{ib} = (\hat{\tau}^*_{ib}, \hat{\phi}^*_{ib})^T$, $b = 1, \cdots, B$, for the bootstrap data. Our estimate of Σ_ε for item i is

$$\hat{\Sigma}_{\varepsilon i} = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*_{ib} - \bar{\hat{\theta}}^*_{ib})^{\otimes 2},$$

where $\bar{\hat{\theta}}^*_{ib} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_{ib}$.

Repeat 1 and 2 for every item, we obtain $\hat{\Sigma}_{\varepsilon i}$, $i = 1, \cdots, m$. Our estimate of the within-item variation is

$$\hat{\Sigma}_\varepsilon = \frac{1}{m}\sum_{i=1}^{m}\hat{\Sigma}_{\varepsilon i}.$$

By (2.7), the estimate of the total variation Σ is

$$\hat{\Sigma} = \hat{\Sigma}_\theta + \hat{\Sigma}_\varepsilon.$$

According to Theorem 1, we can compute the approximate $100(1-\alpha)\%$ confidence intervals of θ by

$$\hat{\theta} \pm t_{\alpha/2, m-2}\sqrt{\hat{\Sigma}/m}. \tag{2.9}$$

If n is large enough, $\hat{\Sigma}_\varepsilon$ is negligible. We can simplify the equation to

$$\hat{\theta} \pm t_{\alpha/2, m-2}\sqrt{\hat{\Sigma}_\theta/m}. \tag{2.10}$$

This will save us a lot of computational time. We recall that equation (2.10) requires two assumptions, one is that Σ_ε is negligible, the other is that $\hat{\theta}$ achieves normality.

2.6.2 Simulation Study

We study the performances of all methods with one item ($m = 1$) first, then extend to the multiple-item cases. We explore the methods both for a small study ($m = 6$) and for a large study ($m = 18$).

First, we conduct a small simulation to evaluate the numerical performances of the parametric and nonparametric approaches introduced in Section 2.4 as well

as the traditional approach in Section 2.2. When there is one single item, that is, $m = 1$, we generate $n = 100$ observations, with $t \in \{60, 63, \cdots, 357\}$ under model (2.2). We set $\theta = (24, 0.5)$ and the (1,1), (1,2) and (2,2) elements of Σ_θ 0.1, 0 and 0.05, respectively. We generate θ_i, $i = 1$, from the distribution $N(\theta, \Sigma_\theta)$, and set $\beta_0(t) = \exp(-0.1 + 0.034t - 0.00007t^2)$, $\beta_1(t) = \exp(0.8 + 0.024t - 0.00005t^2)$, and $\varepsilon \sim N(0, 1^2)$.

We carry out the estimation of model (2.2) by all five methods: (1) the traditional method (DFT); (2) the quadratic varying coefficient method (QVC); (3) the semiparametric local linear varying coefficient method with component-wise update estimation (SVCC); (4) the semiparametric local linear varying coefficient method with simultaneously update estimation using a common bandwidth (SVCSc); and (5) the semiparametric local linear varying coefficient method with simultaneously update estimation using different bandwidths (SVCSd). By the bandwidth selection method provided in Section 2.4.3, we obtain $h_0 = 40$, $h_1 = 80$ for SVCC; a common bandwidth $h = 40$ for SVCSc; and $h_0 = 40$, $h_1 = 50$ for SVCSd. The data generation and the estimation processes are repeated 500 times. For $\hat{\tau}$ and $\hat{\phi}$, we compute the Monte Carlo biases, standard errors and mean square errors (MSE's). For $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$, we compute the Monte Carlo biases, standard errors and MSE's for every t_j, $j = 1, \cdots, 100$, then we average them over all t_j's. All the results are reported in Table 1.

In Table 1, by comparing the Monte Carlo MSE's, we notice that DFT resulted in poor estimates of all parameters. QVC didn't perform better than DFT for the estimation of θ, which is not surprising since we use the wrong parametric models for $\beta_0(t)$ and $\beta_1(t)$, and bad estimates of $\beta_0(t)$ and $\beta_1(t)$ result in bad estimates of θ. MSE of $\hat{\beta}_1(t)$ for QVC is huge because we allow negative values by assuming $\beta_1(t)$ a quadratic function of t while $\beta_1(t)$ should always be positive as the amplitude

Table 1: Monte Carlo biases, SE's and MSE's of all estimators by all five methods when number of curves $m = 1$.

estimates	Method	bias	SE	MSE
$\hat{\tau}$ (24)	DFT	-0.8804	1.2519	2.3424
	QVC	-0.9985	1.1999	2.4368
	SVCC	-0.0364	0.3200	0.1037
	SVCSc	-0.0348	0.3175	0.1020
	SVCSd	-0.0346	0.3177	0.1021
$\hat{\phi}$ (0.5)	DFT	-0.0141	0.3096	0.0961
	QVC	0.0125	0.3077	0.0948
	SVCC	0.0230	0.2165	0.0474
	SVCSc	0.0221	0.2157	0.0470
	SVCSd	0.0221	0.2158	0.0471
$\hat{\beta}_0(t)$	DFT	13.9822	0.2506	258.3662
	QVC	3.7465	0.6140	19.7979
	SVCC	0.5761	0.6731	1.7437
	SVCSc	0.5420	0.2565	0.5269
	SVCSd	0.5420	0.2566	0.5269
$\hat{\beta}_1(t)$	DFT	9.3368	1.8360	116.8756
	QVC	24.2889	14.1053	837.991
	SVCC	1.0634	0.9389	4.5503
	SVCSc	0.2926	0.3480	0.2664
	SVCSd	0.4469	0.3238	0.4277

function. Semiparametric methods produce much smaller MSE's. SVCC performed satisfactorily, but the methods SVCSc and SVCSd provided the best results, especially for estimating $\beta_0(t)$ and $\beta_1(t)$. SVCSd doesn't outperform SVCSc in this situation, which is because the true functions of $\beta_0(t)$ and $\beta_1(t)$ share similar smoothness. Also, for the semiparametric methods, MSE's of $\hat{\tau}$ are close to 0.1, MSE's of $\hat{\phi}$ are close to 0.05, which are the diagonal components in the true Σ_θ. This is also an indicator that semiparametric methods perform well.

Second, we evaluate the numerical performances of estimating the population mean θ and the between-item variation Σ_θ when there are multiple items ($m = 6$). We generate our data with $m = 6$ items and each item has $n = 100$ observations

under model (2.2). $t \in \{60, 63, \cdots, 357\}$, for all $i = 1, \cdots, 6$. We generate $\theta_i = (\tau_i, \phi_i)$, $i = 1, \cdots, 6$, from $N(\theta, \Sigma_\theta)$, where $\theta = (24, 0.5)$ and the (1,1), (1,2), (2,2) elements in Σ_θ are 0.1,0 and 0.05, respectively. We used the best quadratic functions in the analysis of a pilot data set as the basis of constructing $\beta_0(t)$ and $\beta_1(t)$. The actual functions of $\beta_0(t)$ and $\beta_1(t)$ we use are provided in Appendix B. The errors are independently generated from $N(0, 1^2)$.

In Table 1, the performances of DFT and QVC are similar, and the performances of SVCSc and SVCSd are very close when estimating θ. Therefore, we concentrate only on DFT, SVCC and SVCSc in this case. Since the data generation is the same, we use the same bandwidths as we obtained when $m = 1$, and calculate the estimates of θ and Σ_θ for the generated datasets. Also we compute the 95% confidence intervals of θ for each generated dataset by (2.10). The data generation and the estimation processes are repeated 500 times. For $\hat{\theta}$, we computed the Monte Carlo biases, standard errors and MSE's. We also calculate the average lengths and coverage probabilities of the confidence intervals calculated by (2.10). For $\hat{\Sigma}_\theta$, we computed the Monte Carlo biases, standard errors and MSE's, and we report (1,1), (1,2) and (2,2) elements of $\hat{\Sigma}_\theta$ separately, which are $\hat{\Sigma}_{\theta 11}$, $\hat{\Sigma}_{\theta 12}$ and $\hat{\Sigma}_{\theta 22}$ respectively. The results are summarized in Table 2.

Table 2 shows that SVCSc can provide us decent point estimates and intervals estimates for θ. As in Table 1, SVCC and SVCSc outperform DFT significantly with small Monte Carlo biases, standard errors and MSE's. The coverage probabilities of the confidence intervals are very close to 95% for both τ and ϕ, and the average lengths are very small.

In Table 2, we calculate the confidence intervals for θ by (2.10), which assume that Σ_ε is negligible for one of our simulated data when $n = 100$. We raise the question that if this assumption is true. To find out, we perform a sensitivity analysis

Table 2: Monte Carlo biases, SE's and MSE's of all estimators, average lengths and coverage probabilities for confidence intervals of τ and ϕ. Methods used are DFT, SVCC and SVCSc. Number of curves $m = 6$.

For $\theta = (\tau, \phi)^T$

					Confidence Intervals	
estimates	Method	bias	SE	MSE	avg length	cov prob
$\hat{\tau}$ (24)	DFT	-0.8379	0.4997	0.9518	2.6464	0.79
	SVCC	-0.0216	0.1364	0.0191	0.7279	0.95
	SVCSc	-0.0120	0.1327	0.0177	0.7059	0.96
$\hat{\phi}$ (0.5)	DFT	0.0217	0.1293	0.0172	0.6695	0.94
	SVCC	0.0058	0.0840	0.0071	0.4436	0.97
	SVCSc	0.0020	0.0843	0.0071	0.4480	0.97

For Σ_θ

estimates	Method	bias	SE	MSE
$\hat{\Sigma}_{\theta 11}$ (0.1)	DFT	1.3399	0.6159	2.1747
	SVCC	0.0139	0.0730	0.0055
	SVCSc	0.0072	0.0678	0.0046
$\hat{\Sigma}_{\theta 12}$ (0)	DFT	0.0236	0.1636	0.0273
	SVCC	-5e-04	0.0309	0.0010
	SVCSc	-1e-04	0.0314	0.0010
$\hat{\Sigma}_{\theta 22}$ (0.05)	DFT	0.0419	0.0378	0.0032
	SVCC	-0.0083	0.0239	6e-04
	SVCSc	-0.0074	0.0247	7e-04

for our simulated data. Based on the study of Table 1 and Table 2, we select SVCSc as our interested method. We use one set of generated datasets with $m = 6$ and the same setup as above, then perform bootstrapping method described in Section 2.6.1 to obtain $\hat{\Sigma}_{\varepsilon i}$. We select $B = 10$ here, and we can easily show that 10 is large enough for estimating the within-item variation. In Figure 9, we plot $\hat{\Sigma}_{\varepsilon 11}$ versus $\hat{\tau}$, and $\hat{\Sigma}_{\varepsilon 22}$ versus $\hat{\phi}$, where $\hat{\Sigma}_{\varepsilon 11}$ and $\hat{\Sigma}_{\varepsilon 22}$ are the diagonal elements of $\hat{\Sigma}_\varepsilon$. From Figure 9 we see that the "percent", which is the maximum value of $\hat{\Sigma}_\varepsilon/\hat{\Sigma}_\theta$, is less than 3%. We conclude that Σ_ε is negligible. Therefore $\hat{\Sigma}_\theta$ can be a good estimator of Σ, and the confidence intervals calculated by (2.10) is accurate enough.

Third, we evaluate the numerical performances of estimating θ and Σ_θ when we

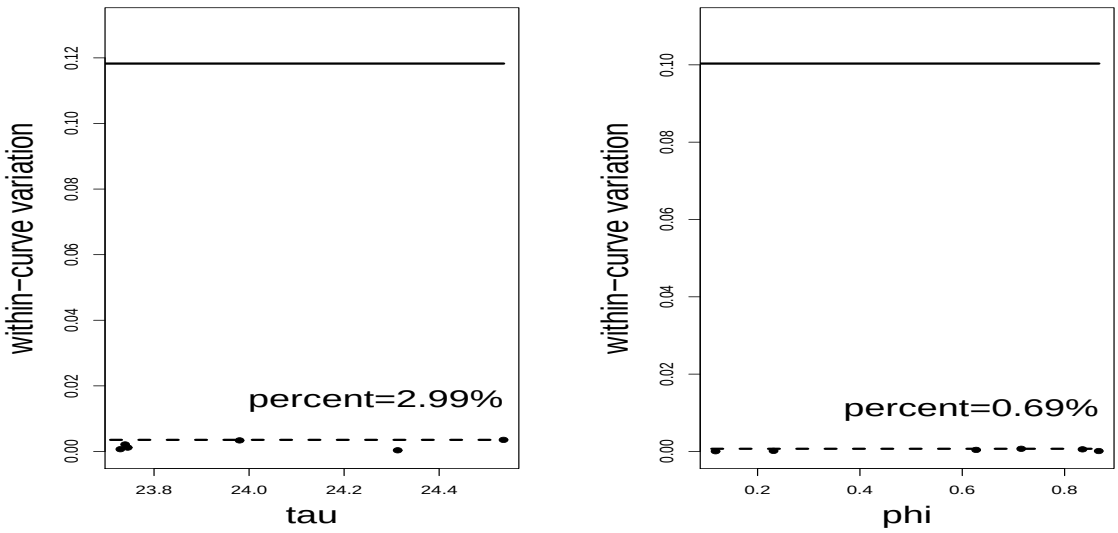

Figure 9: Sensitivity analysis for simulated data ($m = 6$). Plot of within-curve variation $\hat{\Sigma}_\varepsilon$ versus $\hat{\theta}$ by SVCSc. Left graph is for τ and right graph is for ϕ. The solid lines are between-curve variation $\hat{\Sigma}_\theta$, and 'percent's indicate the maximum values of $\hat{\Sigma}_\varepsilon/\hat{\Sigma}_\theta$.

have a large study ($m = 18$). The data generation is similar as $m = 6$ case and we use $\beta_{0i}(t)$'s and $\beta_{1i}(t)$'s, $i = 1, \cdots, 6$, from the $m = 6$ case 3 times. We concentrate only on DFT, SVCC and SVCSc. Since the data generation process is the same, we use the same bandwidths as obtained when $m = 1$, and calculate the estimates of θ and Σ_θ for the generated datasets. Also we compute the 95% confidence intervals of θ for each generated dataset by (2.10). The data generation and the estimation procedures are repeated 500 times. For θ, we compute the Monte Carlo biases, standard errors, MSE's, the average lengths and coverage probabilities of the confidence intervals. For Σ_θ, we compute the Monte Carlo biases, standard errors and MSE's. The results are summarized in Table 3. Table 3 provides us the same conclusion as Table 2: SVCSc overperform all other methods in estimating θ and Σ_θ. Notice that MSE's are smaller in Table 3 than in Table 2 because of the larger m.

Table 3: Monte Carlo biases, SE's and MSE's of all estimators, average lengths and coverage probabilities for confidence intervals of τ and ϕ. Methods used are DFT, SVCC and SVCSc. Number of curves $m = 18$.

For $\theta = (\tau, \phi)^T$

					Confidence Intervals	
estimates	Method	bias	SE	MSE	avg length	cov prob
$\hat{\tau}$ (24)	DFT	-0.8636	0.2819	0.8252	1.1950	0.20
	SVCC	-0.0149	0.0779	0.0063	0.3295	0.97
	SVCSc	-0.0119	0.0756	0.0059	0.3147	0.94
$\hat{\phi}$ (0.5)	DFT	0.0140	0.0731	0.0055	0.3158	0.96
	SVCC	-3e-04	0.0529	0.0028	0.2062	0.91
	SVCSc	2e-04	0.0497	0.0025	0.2120	0.95

For Σ_θ

estimates	Method	bias	SE	MSE
$\hat{\Sigma}_{\theta 11}$ (0.1)	DFT	1.3476	0.3159	1.9157
	SVCC	0.0120	0.0388	0.0016
	SVCSc	0.0018	0.0328	0.0011
$\hat{\Sigma}_{\theta 12}$ (0)	DFT	0.0360	0.0898	0.0094
	SVCC	-3e-04	0.0175	3e-04
	SVCSc	0.0024	0.0169	3e-04
$\hat{\Sigma}_{\theta 22}$ (0.05)	DFT	0.0508	0.0193	0.003
	SVCC	-0.0063	0.0135	2e-04
	SVCSc	-0.0040	0.0136	2e-04

To test if Σ_ε is negligible, we carry on a sensitivity analysis for our simulated data using SVCSc as our interested method. We use one generated dataset with $m = 18$ and the same setup as above, then perform bootstrapping method in Section 2.6.1 to obtain $\hat{\Sigma}_{\varepsilon i}$, $i = 1, \cdots, 18$. Set $B = 10$. In Figure 10, we plot $\hat{\Sigma}_{\varepsilon 11}$ versus $\hat{\tau}$, and $\hat{\Sigma}_{\varepsilon 22}$ versus $\hat{\phi}$, where $\hat{\Sigma}_{\varepsilon 11}$ and $\hat{\Sigma}_{\varepsilon 22}$ are the diagonal elements of $\hat{\Sigma}_\varepsilon$. From Figure 10 we observe that the 'percent' is stable for different θ's, and $\hat{\Sigma}_\varepsilon$ is less than 5% of $\hat{\Sigma}_\theta$. We conclude that Σ_ε is negligible. Therefore $\hat{\Sigma}_\theta$ can be a good estimator of the overall variation Σ.

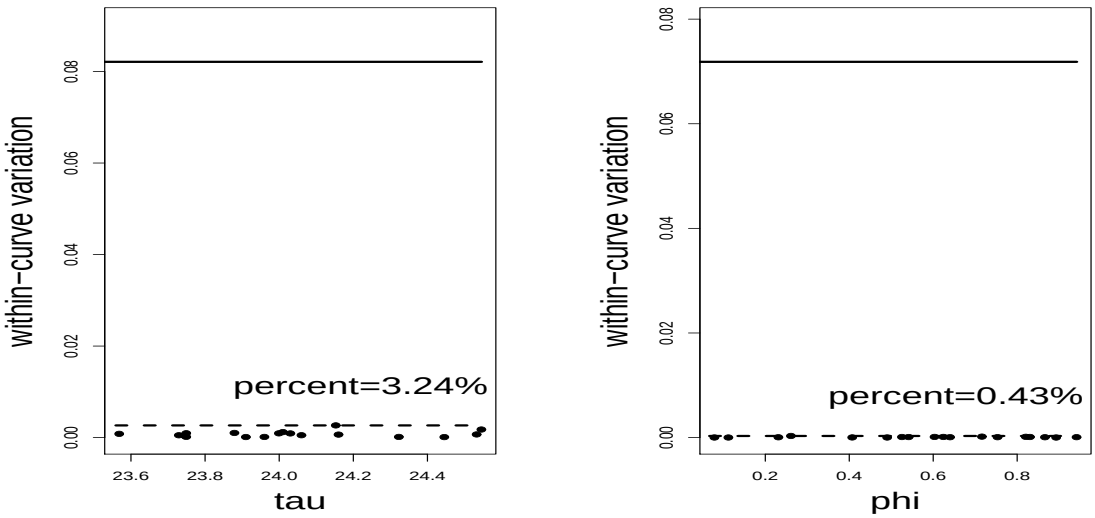

Figure 10: Sensitivity analysis for simulated data ($m = 18$). Plot of within-curve $\hat{\Sigma}_\varepsilon$ versus $\hat{\theta}$ by SVCSc. Left graph is for τ and right graph is for ϕ. The solid lines are between-curve variation $\hat{\Sigma}_\theta$, and 'percent's indicate the maximum values of $\hat{\Sigma}_\varepsilon/\hat{\Sigma}_\theta$.

2.6.3 Analysis of Circadian Dataset

In this Subsection, we provide estimates of our circadian example using all the methods first. Then we study the performances of these methods by a bootstrap procedure. We also check the two assumptions we made for calculating confidence intervals (2.10) and provide alternative solutions if the assumptions are violated.

For our circadian dataset, $m = 13$ and n_i's are 214 for 7 items and 172 for the remaining 6 items. We perform the estimation of these five methods: DFT, QVC, SVCC, SVCSc and SVCSd. As people traditionally do, we drop the first 60 hours because the bacteria are adjusting to the enviornment and are relatively unstable. That is, we use $t > 60$. Thus, n_i's reduced to 186 for 7 items and 145 for the remaining 6 items. The optimal bandwidth chosen for SVCC by cross-validation as in Section 2.4.3 is $h_0 = 40$, $h_1 = 40$. $h = 40$ for SVCSc, and $h_0 = 40$, $h_1 = 50$ for

Table 4: Estimates for cyanobacteria circadian datasets by all five methods.

For θ

	Method	Estimates	confidence interval by (2.10)	confidence interval (2.9)
$\hat{\tau}$	DFT	25.5388	(24.7492 , 26.3285)	(23.5406 , 27.5371)
	QVC	25.4688	(24.7482 , 26.1895)	(24.4386 , 26.4991)
	SVCC	25.4015	(25.3026 , 25.5004)	(24.9616 , 25.8415)
	SVCSc	25.4369	(25.3355 , 25.5383)	(25.3312 , 25.5427)
	SVCSd	25.4369	(25.3355 , 25.5383)	(25.3315 , 25.5424)
$\hat{\phi}$	DFT	0.3930	(0.1243 , 0.6616)	(0.0923 , 0.6936)
	QVC	0.6193	(0.3583 , 0.8802)	(0.2997 , 0.9388)
	SVCC	0.4887	(0.4531 , 0.5243)	(0.4457 , 0.5317)
	SVCSc	0.4738	(0.4350 , 0.5127)	(0.4332 , 0.5145)
	SVCSd	0.4738	(0.4350 , 0.5127)	(0.4334 , 0.5143)

For Σ_θ and Σ_ε

	DFT	QVC	SVCC	SVCSc	SVCSd
$\hat{\Sigma}_{\theta 11}$	1.6732	1.3937	0.0262	0.0276	0.0276
$\hat{\Sigma}_{\theta 12}$	0.3560	0.3203	-0.0082	-0.0092	-0.0092
$\hat{\Sigma}_{\theta 22}$	0.1937	0.1827	0.0034	0.0040	0.0040
$\hat{\Sigma}_{\varepsilon 11}$	9.0424	1.4548	0.4931	0.0024	0.0023
$\hat{\Sigma}_{\varepsilon 12}$	-0.2258	0.0335	0.0155	-0.0010	-9e-04
$\hat{\Sigma}_{\varepsilon 22}$	0.0489	0.0913	0.0016	4e-04	4e-04

SVCSd. We report the estimates of θ, Σ_θ and the confidence interval by (2.10) and (2.9) from the five methods in Table 4.

In Table 4, the estimates of τ and ϕ in the five methods are all around 25 and 0.5. But the estimates of Σ_θ and Σ_ε are quite different between the semiparametric methods (SVCC, SVCSc and SVCSd) and the basic ones (DFT and QVC). We conclude that the semiparametric methods have much less variations. Consequently, the lengths of the confidence intervals are shorter for semiparametric methods. The confidence interval for ϕ by DFT is around $(0.1, 0.7)$, and the range of ϕ is $(0, 1)$, so this confidence interval provides little information about ϕ. Because of this poor estimation accuracy, analysis about phase ϕ is an under-explored area in microbiology. While the confidence interval for ϕ by SVCSc is $(0.43, 0.51)$, which is narrow enough

for us to conclude that ϕ is around 0.5. Based on the classification of genes described in Section 1.2, we know that our data can be classified as Class-1 genes. For SVCSc and SVCSd, $\hat{\Sigma}_\varepsilon$ are significantly smaller than $\hat{\Sigma}_\theta$, indicating smaller estimating errors. Also, SVCSc and SVCSd have almost the same results, this shows that $\beta_0(t)$ and $\beta_1(t)$ share similar smoothness.

To study the performances of our estimates, we carry out a bootstrap procedure. From our original circadian dataset, we can obtain $\hat{\beta}_{0i}(t)$'s, $\hat{\beta}_{1i}(t)$'s and $\hat{\varepsilon}_{ij}$'s in addition to $\hat{\theta}$ and $\hat{\Sigma}_\theta$. Based on the simulation study in Section 2.6.2, we select the estimates of the circadian dataset from SVCSc. We generate one dataset with 13 items under model (2.2), where t_{ij}'s are the same as our circadian data. $\theta_i = (\tau_i, \phi_i)$, $i = 1, \cdots, 13$ are generated from $N(\hat{\theta}, \hat{\Sigma}_\theta)$, where $\hat{\theta} = (25.4369, 0.4738)^T$ and the (1,1), (1,2) and (2,2) elements of $\hat{\Sigma}_\theta$ are 0.0276, -0.0092 and 0.0040, respectively. So for the generated dataset, θ and Σ_θ are known. Use $\hat{\beta}_{0i}(t)$ and $\hat{\beta}_{1i}(t)$, $i = 1, \cdots, 13$, from SVCSc as our $\beta(t)$'s. For the error term, obtain the empirical cumulative distribution function (CDF) of the $\hat{\varepsilon}_{ij}$'s, and draw from this CDF. We perform the five methods on the generated dataset, and obtain $\hat{\theta}$, $\hat{\Sigma}_\theta$ and the normal 95% confidence interval for θ by (2.10). The data generation and estimation precesses are repeated 500 times. As before, we compute the Monte Carlo biases, standard errors and MSE's as well as the average lengths and coverage probabilities of the confidence intervals for θ. And we calculated the Monte Carlo biases, standard errors and MSE's for $\hat{\Sigma}_\theta$. The results are summarized in Table 5.

From Table 5, the estimates from the semiparametric methods (SVCC, SVCSc and SVCSd) have much smaller Monte Carlo MSE's than the basic methods (DFT and QVC). The coverage probabilities are closer to 0.95 for the semiparametric methods, though they have much smaller average lengths. In sum, semiparametric methods overperform the basic methods when estimating θ and Σ_θ. Furthermore, SVCSc and

Table 5: Bootstrap analysis of cyanobacteria circadian data sets. Monte Carlo biases, SE's, MSE's of $\hat{\theta}$ and $\hat{\Sigma}_\theta$ by all five methods, and average lengths and coverage probabilities of confidence intervals for τ and ϕ by all five methods.

For θ

					Confidence Intervals	
estimates	Method	bias	SE	MSE	avg length	cov prob
$\hat{\tau}$ (25.43)	DFT	-0.2691	0.2771	0.1492	1.3318	0.77
	QVC	-0.3839	0.2726	0.2217	1.1964	0.71
	SVCC	-0.0399	0.0575	0.0049	0.2536	0.96
	SVCSc	-0.0136	0.0472	0.0024	0.2042	0.93
	SVCSd	-0.0136	0.0472	0.0024	0.2041	0.93
$\hat{\phi}$ (0.47)	DFT	-0.0397	0.1123	0.0142	0.4909	0.92
	QVC	0.0343	0.1018	0.0115	0.4800	0.96
	SVCC	0.0156	0.0262	9e-04	0.1143	0.95
	SVCSc	0.0049	0.0198	4e-04	0.0860	0.95
	SVCSd	0.0049	0.0197	4e-04	0.0860	0.95

For Σ_θ

estimates	Method	bias	SE	MSE
$\hat{\Sigma}_{\theta 11}$ (0.0276)	DFT	1.2466	0.6380	1.9612
	QVC	1.0309	0.6161	1.4424
	SVCC	0.0186	0.0286	0.0012
	SVCSc	0.0023	0.0208	4e-04
	SVCSd	0.0023	0.0206	4e-04
$\hat{\Sigma}_{\theta 12}$ (-0.0092)	DFT	0.1176	0.1368	0.0325
	QVC	0.0984	0.1130	0.0224
	SVCC	-4e-04	0.0070	5e-05
	SVCSc	0.0077	0.0035	7.1e-05
	SVCSd	0.0077	0.0035	7.1e-05
$\hat{\Sigma}_{\theta 22}$ (0.0040)	DFT	0.1592	0.0290	0.0262
	QVC	0.1513	0.0208	0.0233
	SVCC	0.0050	0.0034	4e-05
	SVCSc	0.0011	0.0019	5e-06
	SVCSd	0.0011	0.0019	5e-06

SVCSd provide better estimates than SVCC, though they are not very different from each other.

In Table 5, we calculate the confidence intervals by (2.10). This requires two assumptions. One is that Σ_ε is negligible, the other that is $\hat{\theta}$ achieves normality when

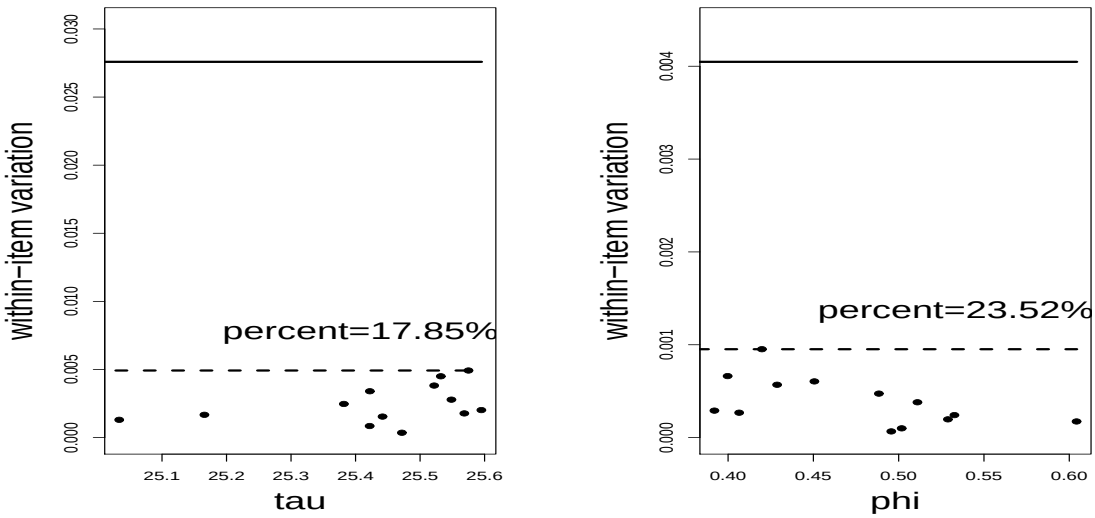

Figure 11: Sensitivity analysis for cyanobacteria circadian data. Plot of within-curve variation $\hat{\Sigma}_\varepsilon$ versus $\hat{\theta}$ by SVCSc. Left graph is for τ and right graph is for ϕ. The solid lines are between-curve variation $\hat{\Sigma}_\theta$, and 'percent's indicate the maximum values of $\hat{\Sigma}_\varepsilon/\hat{\Sigma}_\theta$.

$m = 13$, $n_i = 186$ for 7 items $n_i = 145$ for the other 6 items. We check these two assumptions in the following paragraphs.

We conduct a sensitivity analysis for our circadian data using SVCSc to check the first assumption. We perform bootstrapping method described in Section 2.6.1 to obtain the estimates of the within-item variation for each item i, $\hat{\Sigma}_{\varepsilon i}$. The number of bootstrap B is set to be 10. In Figure 11, we plot $\hat{\Sigma}_{\varepsilon 11}$ versus $\hat{\tau}$, and $\hat{\Sigma}_{\varepsilon 22}$ versus $\hat{\phi}$, where $\hat{\Sigma}_{\varepsilon 11}$ and $\hat{\Sigma}_{\varepsilon 22}$ are the diagonal elements of $\hat{\Sigma}_\varepsilon$. From Figure 11, some $\hat{\Sigma}_{\varepsilon i}$'s are almost 20% of $\hat{\Sigma}_\theta$. We conclude that Σ_ε is not negligible comparing to Σ_θ for our circadian data. Therefore the assumption of Σ_ε is not correct, $\hat{\Sigma}_\theta$ can not be a good estimate of Σ. The confidence intervals calculated in Table 5 are not accurate enough.

Since the first assumption is violated, we modify our confidence intervals by cal-

Table 6: Average lengths and coverage probabilities of confidence intervals for τ and ϕ by both (2.10) and (2.9).

		Confidence Intervals by (2.10)		Confidence Intervals by (2.9)	
estimates	Method	avg length	cov prob	avg length	cov prob
$\hat{\tau}$	SVCSc	0.2111	0.94	0.2365	0.95
$\hat{\phi}$	SVCSc	0.0863	0.94	0.0908	0.95

culating the confidence intervals for θ using (2.9) for the first 100 of the 500 bootstrap datasets. Report the average lengths and coverage probabilities of the confidence intervals by both (2.10) and (2.9) in Table 6. Notice that we have bigger average lengths of the confidence intervals by (2.9) than those by (2.10), because we use larger variations. Consequently, the coverage probabilities obtain a little higher than those in Table 5 and closer to 0.95. However, by setting $B = 10$ when calculating the $\hat{\Sigma}_\varepsilon$'s, we need 10 times more computational time for (2.9) than for (2.10).

To check the second assumption, we drew the histogram of our 500 $\hat{\tau}$'s and $\hat{\phi}$'s in Figure 12. Figure 12 shows that $\hat{\tau}$ and $\hat{\phi}$ are pretty normally distributed for $m = 13$, $n_i = 186$ for 7 items $n_i = 145$ for the other 6 items.

If normality is not reached, there is an alternative way to obtain more robust confidence intervals. Based on our 500 $\hat{\tau}$'s and $\hat{\phi}$'s, we can obtain one bootstrap 95% confidence interval for each of τ and ϕ. The comparison of the bootstrap confidence interval and the normal confidence interval (in Table 4) are reported in Table 7. The lengths are pretty close, which also indicates our estimations of variations are accurate enough.

2.7 Concluding Remarks

In this chapter we proposed an intuitively appealing semiparametric varying coefficient model for the analysis of cyanobacteria circadian rhythm data. We proposed

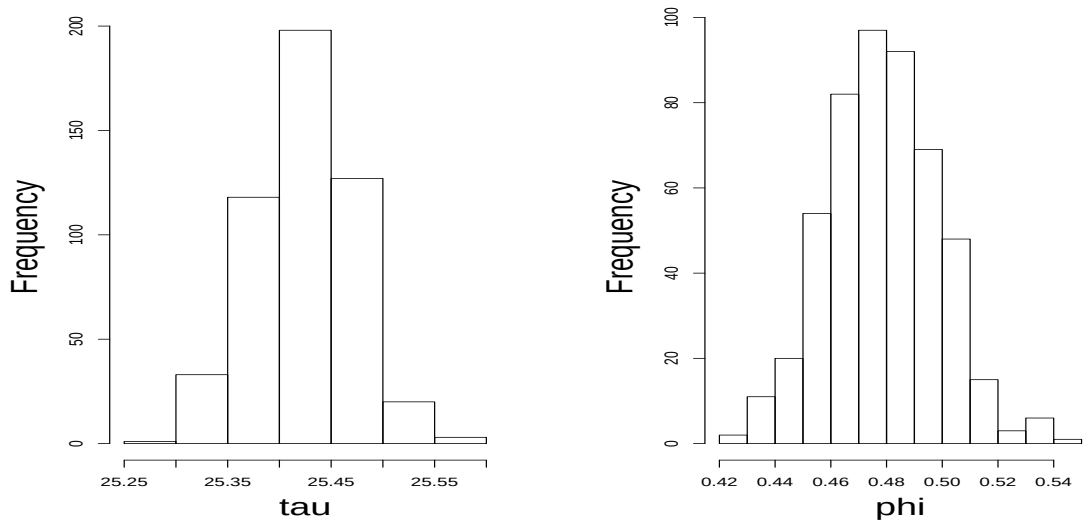

Figure 12: Histograms of $\hat{\tau}$ and $\hat{\phi}$ in bootstrap analysis of cyanobacteria circadian data.

Table 7: Comparison of bootstrap confidence interval and normal confidence interval for τ and ϕ.

	bootstrap CI	length1	normal CI by (2.10)	length2	length1/length2
τ	(25.3298, 25.5114)	0.1816	(25.3355 , 25.5383)	0.2028	0.8955
ϕ	(0.4411, 0.5171)	0.0760	(0.4350 , 0.5127)	0.0777	0.9781

semiparametric local linear approaches when multiple sets of data are collected. Although the varying coefficient model is not globally identifiable, our local linear regression solves the problem by selecting good initial estimates and sufficiently large bandwidths so that the resulting answer provides a useful tool for biologists. One advantage of semiparametric approach is its good quality under reasonable regularity conditions. Asymptotic properties are developed. By simulations and the application to a circadian dataset, we find that significant reduction of the Monte Carlo mean-squared errors obtained by the traditional FFT-NLLS method can be achieved by our procedure. This reduction of variation makes accurately estimating phase ϕ pos-

sible, which is an under-explored area in microbiology because of the poor estimation accuracy.

CHAPTER III

MODEL SELECTION USING SMOOTHING SPLINE IN CYANOBACTERIA

3.1 Introduction

Many hormones and other physiological measurements in living beings vary in a circadian pattern. For example, it is well-known that body temperature is lowest during sleep in the early morning and rises after awakening. Circadian rhythms has been found to exist in humans, plants, animals and lower organisms. The existence of an endogenous circadian oscillator in cyanobacteria has been recognized for less than twenty years. The cyanobacterial clock generates rhythms of biological processes that exhibit an approximate 24-hour period even in the absence of an environment cue, through light or temperature, and maintain a nearly constant period over a range of physiologically relevant environmental conditions (Golden, 2003).

Our data were collected at Dr. Susan Golden's laboratory, Department of Biology, Texas A&M University. In Dr. Golden's lab, cells are incubated in a 12-hour light-dark cycle to synchronize the clock, then are released into continuous light for several days . Bioluminescence levels measure the gene expression level around the clock (Figure 13). From Figure 13 we can see the circadian pattern in the data. This periodic pattern can be fitted by a function of cosine wave. Also, there is a baseline in the data changing with time, which is the growth curve of the bacteria. We notice that the amplitude of the periodic component also change with time. Furthermore, from Figure 13, we observe that the baseline and amplitude functions seem to be positively correlated.

In chapter II we estimated circadian parameters using semiparametric local linear regression. Besides kernel based methods, smoothing spline is another commonly

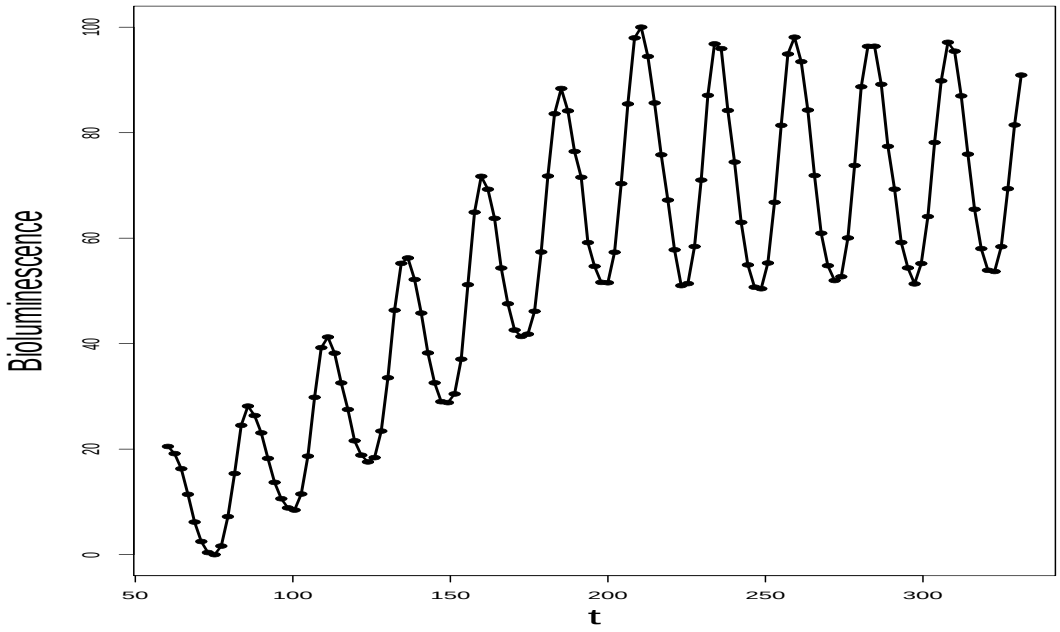

Figure 13: Plot of bioluminescence vs time. Data are recorded every 2.116 hours.

used nonparametric technique. This chapter we use smoothing spline to estimate our data and further investigate interesting properties of the data. Wang and Ke (2002) developed an R package ASSIST for fitting semiparametric models using smoothing spline. Because of the identifiability problem we mentioned in chapter II, a default application in ASSIST often lead to non-convergence outcomes or answers which are not biologically plausible. We solve the identifiability problem by proposing a new smoothing parameter selection method: adjusted cross-validation (adjusted CV).

For our cyanobacteria data, the baseline function $\beta_0(t)$ reflects a measurement of the amount of bacteria, and the amplitude function $\beta_1(t)$ is a measurement of the corresponding circadian component. It is reasonable to expect that $\beta_1(t)/\beta_0(t)$ is less than 1. It is also believed that the ratio of the amplitude and baseline functions is a constant of time. The goal of our study is to investigate the relationship between the baseline and amplitude functions. To execute this, we use a general model which

makes no assumptions about the relationship between the baseline and amplitude functions, and two reduced models which assume specific relationship of baseline and amplitude. We perform model selection to choose the optimal model.

Our proposed models and the potential identifiability problem are described in Section 3.2. Section 3.3 provides smoothing spline estimation methods for the models and a solution to the identifiability problem. Model selection methods we use are in Section 3.4. Numeric outcomes for both simulated data and cyanobacteria circadian data are provided in Section 3.5. Section 3.6 contains the concluding remarks.

3.2 Models

For our data, we observe that the curve possesses an overall growth as well as a periodic pattern as functions of time. The periodic component can be fitted by a cosine wave and the baseline and amplitude functions can be fitted by nonparametric functions. Let y be the response variable and t be its associated covariate. We consider the following varying coefficient periodic model:

$$y_j = \beta_0(t_j) + \beta_1(t_j)\cos\{2\pi(t_j/\tau - \phi)\} + \varepsilon_j,\ j = 1, \cdots, n,$$

where y_j is the response at the jth time point t_j; $\beta_0(t)$ (bounded away from 0) is the underlying baseline growth function; $\beta_1(t)$ (bounded away from 0) is the amplitude function of the periodic component; τ denotes the period and $0 \leq \phi < 1$ is the phase; n is the total number of observations. We assume that the error terms ε_j's are independent and identically distributed with $\mathrm{E}(\varepsilon_j) = 0$ and $\mathrm{var}(\varepsilon_j) = \sigma_\varepsilon^2$.

The general form of the model is

$$Y = \beta_0(t) + \beta_1(t)\cos\{2\pi(t/\tau - \phi)\} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $t = (t_1, \cdots, t_n)^T$, $Y = (y_1, \cdots, y_n)^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$. We refer to model (3.1) as the general model.

In model (3.1), we do not make any assumptions about the relationship between $\beta_0(t)$ and $\beta_1(t)$. Biologists believe that the periodic component $\beta_1(t)$ is proportional to the growth component $\beta_0(t)$. That is, $\beta_1(t) = c\beta_0(t)$ with a positive constant c. To relax this assumption, we replace c by a quadratic function of t. Define $\gamma(t) = \log\beta_1(t) - \log\beta_0(t)$, therefore $\beta_1(t)/\beta_0(t) = \exp\{\gamma(t)\}$. Consequently, we propose the following model:

$$Y = \beta_0(t)\left[1 + \exp\{\gamma(t)\}\cos\{2\pi(t/\tau - \phi)\}\right] + \boldsymbol{\varepsilon}, \tag{3.2}$$

where $\gamma(t) = a_0 + a_1 t + a_2 t^2$. Model (3.2) implies that the relationship between $\beta_0(t)$ and $\beta_1(t)$ can be expressed as a quadratic function. That is, $\gamma(t)$ is a quadratic function.

The traditional belief of the ratio between $\beta_0(t)$ and $\beta_1(t)$ being a constant implies the following model:

$$Y = \beta_0(t)\left[1 + \exp(\gamma)\cos\{2\pi(t/\tau - \phi)\}\right] + \boldsymbol{\varepsilon}, \tag{3.3}$$

Model (3.3) is a special case of model (3.2) when $\gamma(t)$ is a constant.

Evaluating model (3.1), we notice a potential identifiability problem in the model. That is, one can not identify $\beta_0(t)$ and $\beta_1(t)$ in model (3.1). One extreme case is that $\beta_1(t) = 0$, and $E(Y|t) = \beta_0(t)$, so the circadian pattern will be captured by $\beta_0(t)$ alone. We will address this problem more in Section 3.3.

3.3 Spline Estimations

To estimate the parameters and $\beta(t)$'s in models (3.1), (3.2) and (3.3), we use the software package ASSIST developed by Wang and Ke (2002). ASSIST is a suite of R functions for fitting various nonparametric or semiparametric nonlinear models using smoothing spline. We use the function 'snr' in the package to fit our models.

Section 3.3.1 provides the initial estimates we use in ASSIST; the results of a direct application of ASSIST using the generalized cross-validation are given in Section 3.3.2. Our proposed adjusted cross-validation and its corresponding results are given in Section 3.3.3.

3.3.1 Initial Estimates

We denote the initial estimates in model (3.1) as $\hat{\tau}^{[0]}$, $\hat{\phi}^{[0]}$, $\hat{\beta}_0^{[0]}(t)$ and $\hat{\beta}_1^{[0]}(t)$, respectively. Here is how we obtain the initial estimates:

1. Select local minimum and maximum points.

2. Obtain the median of two adjacent minimum and maximum points, consider the medians as our baseline. Use least squares method to fit $\hat{\beta}_0^{[0]}(t)$, which is an exp(quadratic) function.

3. Subtract minimum from adjacent maximum points, consider these as two times amplitude. Use least squares method to fit $\hat{\beta}_1^{[0]}(t)$, which is an exp(quadratic) function.

4. The distance from adjacent minimum and maximum points are half the period, average them to obtain the initial estimate of τ, $\hat{\tau}^{[0]}$. Equivalently, average individual phases to obtain $\hat{\phi}^{[0]}$.

For model (3.2), the initial estimates of a_0, a_1 and a_2 are calculated by fitting the exp(quadratic) model $\hat{\beta}_1(t)/\hat{\beta}_0(t) = \exp(a_0 + a_1 t + a_2 t^2)$. While for model (3.3), we fit $\hat{\beta}_1(t)/\hat{\beta}_0(t) = \exp(\gamma)$ and obtain the initial estimate of γ.

3.3.2 When ASSIST Is Used Directly

We use an example to demonstrate the identifiability problem when using the generalized cross-validation (GCV) to choose the smoothing parameter.

Example 3.1 In this example, we use the dataset in Figure 13 to generate datasets. For our circadian dataset, $n = 156$. We only use $t > 60$. Thus n reduces to 129, and $t \in \{60.35, 62.46, \cdots, 331.20\}$. We estimate the curve first to obtain $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$, then use 'approxfun' in R to obtain the 'true' $\beta_0(t)$ and $\beta_1(t)$. We then generate datasets from model (3.1) by setting $\tau = 24$, $\phi = 0.5$, $\varepsilon \sim N(0, 1^2)$, and $\beta_0(t)$, $\beta_1(t)$ as above.

We use generalized cross-validation (GCV) to choose the smoothing parameter (λ). In ASSIST, a new parameter limnla, which equals to $\log_{10}(n\lambda)$, is considered as the smoothing parameter. We can not obtain a convergence outcome for model (3.1), and the results under model (3.2) and model (3.3) are plotted in Figure 14. In Figure 14, although the overall fitted curves are very close to the true curve, $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$ are far away from the true $\beta_0(t)$ and $\beta_1(t)$. However, in chapter II, local linear regression approach seems successfully avoid the identifiability problem. Using the bandwidth selected by cross-validation, we obtain the $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$ that are close to the truth. By comparing the results by local linear regression and spline, we find that local linear regression turns to favor much smoother results than spline. To solve this problem for spline, we propose an adjusted cross-validation procedure to find our smoothing parameter in Section 3.3.3.

3.3.3 Smoothing Parameter Selection: Adjusted Cross-Validation

In this Subsection, we provides a method to choose a smoothing parameter in ASSIST that would lead to the desirable outcome. We first subtract the initial estimate of $\beta_0(t)$ from y, then perform the cross-validation method as below:

1. Divide time into c intervals, each has k points, then leave the 1 point out in each interval, so we will use the remaining $n - c$ points to produce estimates and make predicts to the left-out c points.

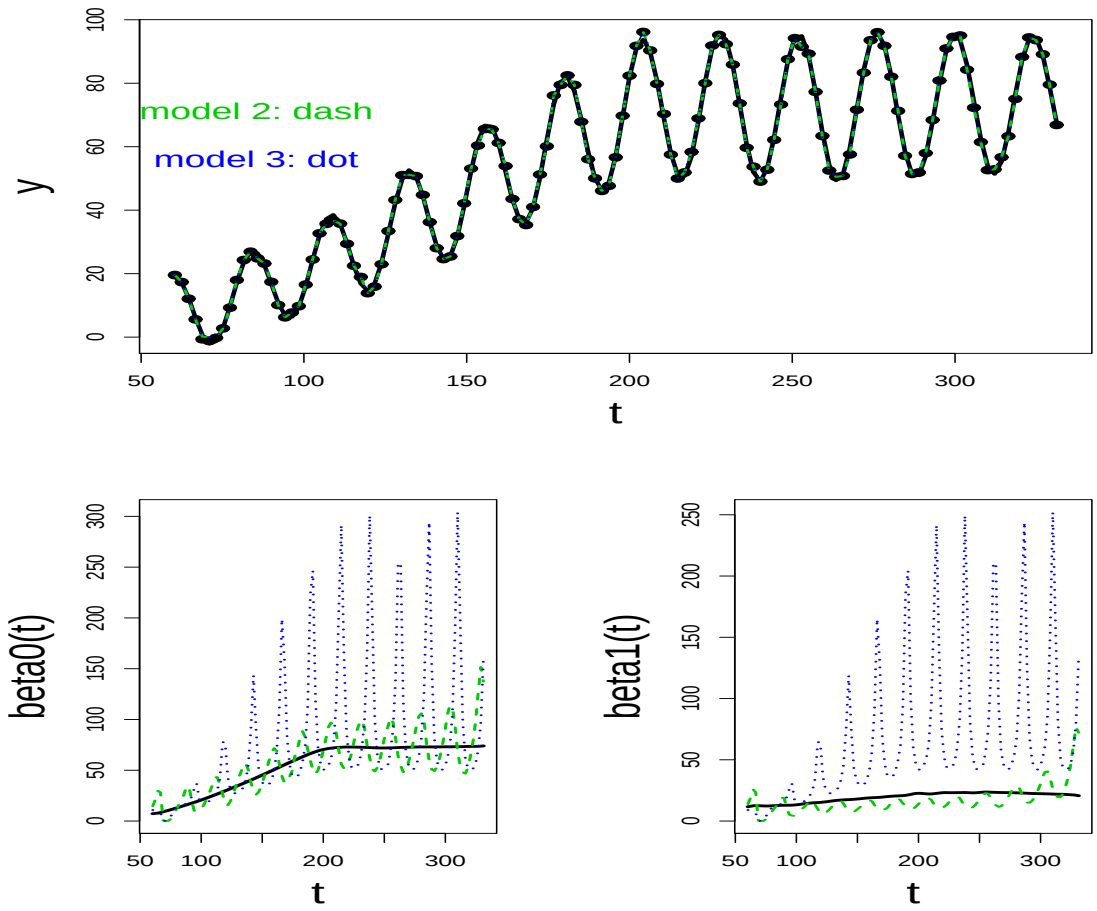

Figure 14: Spline fitted curves for overall, $\beta_0(t)$ and $\beta_1(t)$ in Example 3.1, with smoothing parameter chosen by GCV. No results for model (3.1), $limnla = -0.97$ for model (3.2), and $limnla = -0.70$ for model (3.3). The bold lines are the true curves, the green dashed lines are estimated curves under model (3.2) and the blue dot lines are estimated curves under model (3.3).

2. Select a smoothing parameter λ, and conduct estimation procedures using the $n - c$ points, obtain $\hat{y}$ and corresponding prediction MSE

$$G_{ck}(\lambda) = \{y_{ck} - \hat{y}_{ck}^{\lambda}\}^2.$$

3. For a given C, repeat steps 1 and 2, for c from 1 to C, and k from 1 to

$K = \text{int}(n/c)$, and obtain

$$G(\lambda) = \sum_{c=1}^{C}\sum_{k=1}^{K} G_{ck}(\lambda) = \sum_{c=1}^{C}\sum_{k=1}^{K}\{y_{ck} - \hat{y}_{ck}^{\lambda}\}^2.$$

The optimal bandwidth is $\lambda = \text{argmin } G(\lambda)$. In ASSIST package, there is an option to set "limnla", which equals to $\log_{10}(n\lambda)$. We simply use this option to obtain optimal λ.

Note that for all 3 models, we use a common smoothing parameter. For Example 3.1, we select $limnla = 7.6$ by adjusted cross-validation. The fitted curves under all three models are presented in Figure 15. Figure 15 shows satisfactory results for all three models, which indicates that adjusted CV provides a practical solution to our problems.

3.4 Model Selection

After we obtain estimated parameters in each of the three models, we can select the optimal model using Akaike Information Criterion (AIC) and Schwarz Bayesian Information Criterion (BIC).

The Akaike information criterion (AIC), developed by Hirotsugu Akaike in 1971 and proposed in Akaike (1974), is a measure of the goodness of fit of an estimated statistical model. It is grounded in the entropy concept. The AIC is an operational way of trading off the complexity of an estimated model against how well the model fits the data. In the general case, the AIC is

$$AIC = -2\log(L) + 2p,$$

where L is the likelihood function, p is the number of parameters used.

Assume that the errors in models (3.1) to (3.3) are independent and normally distributed. Let n be the number of observations and RSS be the residual sum of

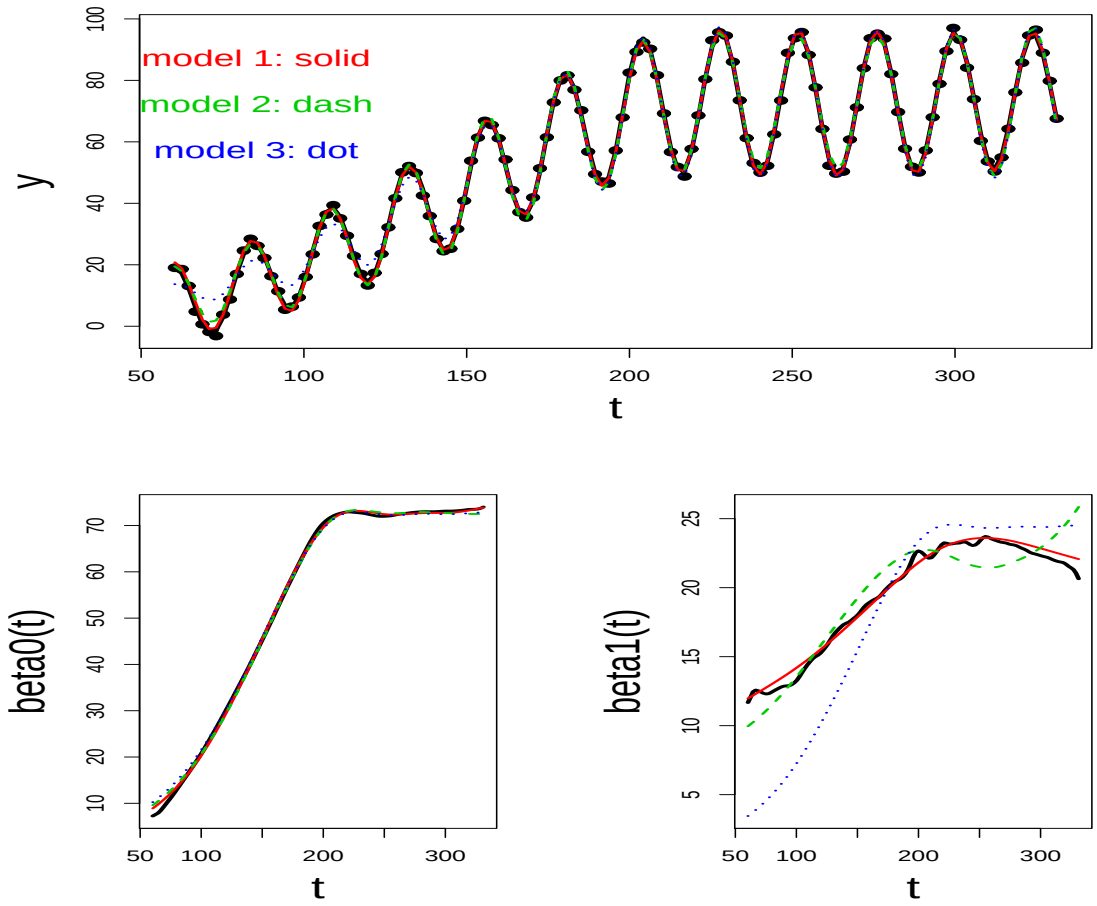

Figure 15: Spline fitted curves for overall, $\beta_0(t)$ and $\beta_1(t)$ in Example 3.1, with smoothing parameter chosen by adjusted CV. $limnla = \log_{10}(n\lambda) = 7.6$ for all three models. The bold lines are the true curves, the red solid lines are the estimated curves under model (3.1), the green dashed lines are the estimated curves under model (3.2), and the blue dot lines are the estimated curves under model (3.3).

squares. Then AIC becomes

$$AIC = n\log(2\pi\hat{\sigma}_e^2) + RSS/\hat{\sigma}_e^2 + 2p.$$

Increasing the number of free parameters to be estimated improves the goodness of fit. This phenomenon is independent of the number of free parameters in the data generating process. Hence AIC not only measures the level of goodness of fit, but also includes a penalty term which increases with the number of estimated parameters.

Table 8: Model selection in Example 3.1

model	AIC	BIC	RSS	MSE	df
(3.1)	399.3249	441.5515	129.6104	51.5387	114.2345
(3.2)	499.1393	537.0184	290.3463	189.9354	115.7547
(3.3)	674.9702	706.9828	1174.8094	1044.6025	117.8061

This penalty term discourages overfitting. The preferred model is the one with the lowest AIC value. The AIC methodology intends to find the model that best explains the data with a minimum number of free parameters.

The Schwarz Bayesian information criterion (BIC) is an alternative statistical criterion for model selection. It is named after Schwarz (1978) which provides a Bayesian argument to support its use. The formula for BIC is

$$BIC = -2\log(L) + p\log(n).$$

Under the assumption that the errors are normally distributed, this expression becomes:

$$BIC = n\log(2\pi\hat{\sigma}_e^2) + RSS/\hat{\sigma}_e^2 + p\log(n).$$

Given any m models, the model that has the lowest value of BIC is the one to be preferred. The BIC is a decreasing function of RSS, the measurement of goodness of fit, and an increasing function of p. The BIC penalizes free parameters more severe than does the Akaike information criterion.

For Example 3.1, the values of AIC and BIC are reported in Table 8, in which $RSS = \sum(\hat{y} - y)^2$, $MSE = \sum(\hat{y} - \text{true } y)^2$ and the degree of freedom $df = n - p$. In Table 8, model (3.1) minimizes both AIC and BIC, so we select model (3.1) as our best model. By the data generation algorithms described in Section 3.3.2, model (3.1) is acturally the true model. That is, AIC and BIC select the correct model.

Table 9: Inferences for estimators in Example 3.1

estimates	model	bias	SE	MSE
$\hat{\tau}$ (24)	(3.1)	0.0056	0.0081	1e-04
	(3.2)	0.0074	0.0081	1e-04
	(3.3)	0.0045	0.0083	1e-04
$\hat{\phi}$ (0.5)	(3.1)	-0.0025	0.0034	2e-05
	(3.2)	-0.003	0.0034	2e-05
	(3.3)	-0.0018	0.0035	2e-05
$\hat{\beta}_0(t)$	(3.1)	0.3012	0.2105	0.2752
	(3.2)	0.4658	0.2066	0.5375
	(3.3)	0.6676	0.2059	0.9415
$\hat{\beta}_1(t)$	(3.1)	0.217	0.229	0.1445
	(3.2)	1.189	0.2186	2.0981
	(3.3)	2.9092	0.1269	14.5637

3.5 Numeric Outcomes

This Section provides the numeric results for both simulated and real data. In Section 3.5.1 to 3.5.3, we provide 3 examples for model selection; precisely, data are generated under three different models. We apply model selection techniques on a cyanobacteria dataset in Section 3.5.4.

3.5.1 Example 3.1. Data Generated Under Model (3.1)

Example 3.1. As described in Section 3.3.2, data were generated under model (3.1).

We repeat the data generation and estimation procedure 100 times with the smoothing parameter chosen by adjusted CV ($limnla = 7.6$). Table 9 reports the Monte Carlo biases, standard errors and MSE's. We can see that smoothing spline provides satisfactory estimates for both models (3.1) and (3.2). We have the smallest MSE's under model (3.1). The MSE differs the most for different models when the goal is to estimate $\beta_0(t)$ and $\beta_1(t)$.

For these 100 datasets, we performed model selection; count the number of se-

Table 10: Proportions of selection in Example 3.1

method	selection	model (3.1)	model (3.2)	model (3.3)
Spline	AIC	100%	0%	0%
	BIC	100%	0%	0%

Table 11: Model selection in Example 3.2

model	AIC	BIC	RSS	MSE	df
(3.1)	369.8029	414.0228	101.7302	16.3498	113.5375
(3.2)	367.4249	407.7773	102.9347	13.7781	114.8898
(3.3)	526.4069	560.9916	365.3745	272.4258	116.9067

lections of each model, and report the proportions of the selections in Table 10. For all datasets, both AIC and BIC select model (3.1) 100%.

3.5.2 Example 3.2. Data Generated Under Model (3.2)

Example 3.2. We generate one dataset under model (3.2). Each item has $n = 129$ observations, with t the same as in Example 3.1. We set $\tau = 24$ and $\phi = 0.5$, $\beta_0(t) = \exp(2 + 0.0165t - 0.00003t^2)$, $\gamma(t) = -0.5 - 0.0065t + 0.00001t^2$ and $\boldsymbol{\varepsilon} \sim N(0, 1^2)$.

We conducted the estimations in ASSIST. Using the adjusted cross-validation described in Section 3.3.3, we selected $limnla = 7.4$ and plot the estimated curves in Figure 16. We observe that similar outcomes were obtained when we assume the true model was either (3.1) or (3.2). The differences between these outcomes and those obtained assuming model (3.3) were sizable.

The AIC and BIC values are reported in Table 11. Both AIC and BIC select model (3.2) as the best model, which is the correct selection.

We repeated the data generation and estimation procedures 100 times, and reported the Monte Carlo biases, standard errors and MSE's in Table 12. We have very similar outcomes for model (3.1) and model (3.2) in Table 12, and worse outcomes for model (3.3).

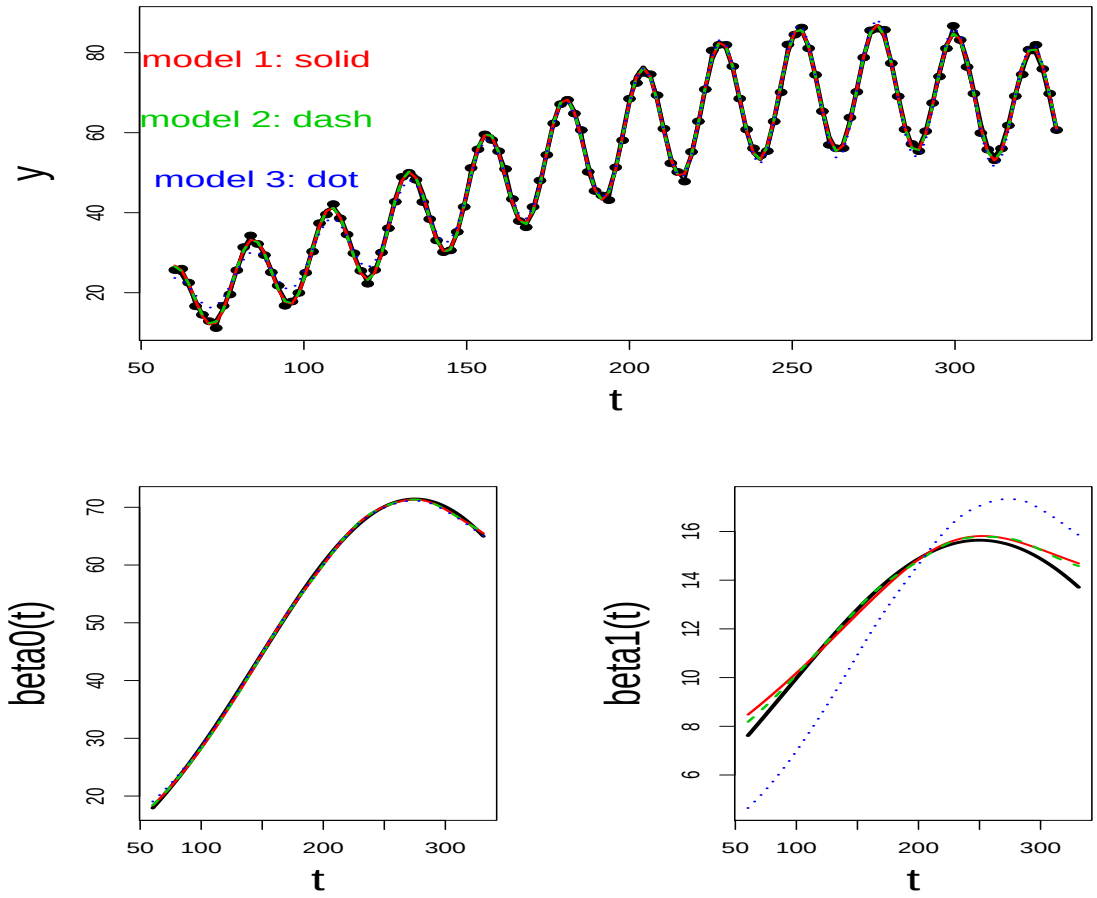

Figure 16: Spline fitted curves for overall, $\beta_0(t)$ and $\beta_1(t)$ in Example 3.2. $limnla = \log_{10}(n\lambda) = 7.4$ for all three models. The bold lines are the true curves, the red solid lines are the estimated curves under model (3.1), the green dashed lines are the estimated curves under model (3.2), and the blue dot lines are the estimated curves under model (3.3).

We ran the model selection procedure for these 100 datasets. In Table 13, AIC selects model (3.2) 94% of the times and model (3.1) 6% of the times. BIC selects model (3.2) 99% of the times and model (3.1) 1% of the times. Since model (3.2) is the correct model, the use of BIC provided a more accurate selection than the use of AIC.

Table 12: Inferences for estimators in Example 3.2

estimates	model	bias	SE	MSE
$\hat{\tau}$ (24)	(3.1)	-0.0008	0.0118	1e-04
	(3.2)	-0.0013	0.0118	1e-04
	(3.3)	-0.0002	0.0117	1e-04
$\hat{\phi}$ (0.5)	(3.1)	1e-05	0.0049	2e-05
	(3.2)	0.0002	0.0049	2e-05
	(3.3)	-0.0001	0.0048	2e-05
$\hat{\beta}_0(t)$	(3.1)	0.0572	0.2234	0.0597
	(3.2)	0.0576	0.2218	0.0598
	(3.3)	0.248	0.2202	0.145
$\hat{\beta}_1(t)$	(3.1)	0.0599	0.2228	0.0607
	(3.2)	0.0375	0.2117	0.0481
	(3.3)	1.7846	0.121	4.0367

Table 13: Proportions of selection in Example 3.2

method	selection	model (3.1)	model (3.2)	model (3.3)
Spline	AIC	6%	94%	0%
	BIC	1%	99%	0%

Table 14: Model selection in Example 3.3

model	AIC	BIC	RSS	MSE	df
(3.1)	377.3668	433.9193	99.017	17.6599	109.2251
(3.2)	369.3669	414.3292	101.3816	14.5043	113.2779
(3.3)	367.6993	406.9426	103.6046	12.1563	115.2777

3.5.3 Example 3.3. Data Generated Under Model (3.3)

Example 3.3. We generate one dataset under model (3.3). Each item has $n = 129$ observations, with t as in Example 3.1. We set $\tau = 24$ and $\phi = 0.5$, $\beta_0(t) = \exp(1 + 0.02t - 0.00004t^2)$, $\gamma = -0.5$ and $\boldsymbol{\varepsilon} \sim N(0, 1^2)$.

Again, using adjusted CV in Section 3.3.3, we obtained $limnla = 7.1$. Estimated curves are provided in Figure 17, where all three models share similar outcomes.

The values of AIC and BIC were reported in Table 14. According to AIC and BIC, we select model (3.3) as the best model.

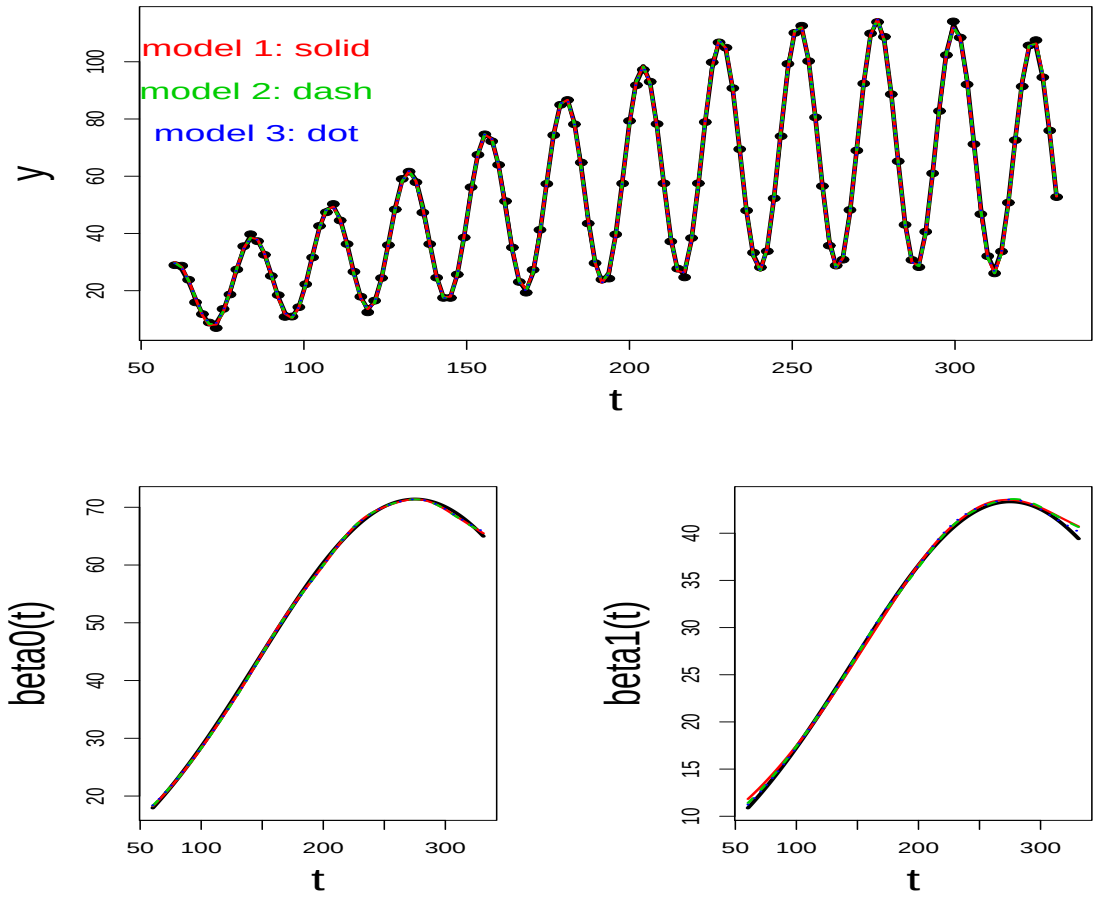

Figure 17: Spline fitted curves for overall, $\beta_0(t)$ and $\beta_1(t)$ in Example 3.3. $limnla = \log_{10}(n\lambda) = 7.1$ for all three models. The bold lines are the true curves, the red solid lines are the estimated curves under model (3.1), the green dashed lines are the estimated curves under model (3.2), and the blue dot lines are the estimated curves under model (3.3).

We repeat the data generation and estimation procedure 100 times, and reported the Monte Carlo biases, standard errors and MSE's in Table 15. All three models provide similar Monte Carlo biases, standard errors and MSE's. Model (3.3) is even slightly better than model (3.1) and model (3.2).

We ran the model selection for these 100 datasets, and reported the proportions of selecting different models in Table 16. AIC selects model (3.3) 89% of the times

Table 15: Inferences for estimators in Example 3.3

estimates	model	bias	SE	MSE
$\hat{\tau}$ (24)	(3.1)	-0.0006	0.0053	3e-05
	(3.2)	0.0013	0.0197	4e-04
	(3.3)	-0.0006	0.0053	3e-05
$\hat{\phi}$ (0.5)	(3.1)	0.0001	0.0023	1e-05
	(3.2)	-0.0004	0.0063	4e-05
	(3.3)	0.0002	0.0023	1e-05
$\hat{\beta}_0(t)$	(3.1)	0.0457	0.2409	0.0647
	(3.2)	0.0397	0.2372	0.0632
	(3.3)	0.0416	0.2256	0.0561
$\hat{\beta}_1(t)$	(3.1)	0.0497	0.2776	0.0873
	(3.2)	0.0239	0.2425	0.0622
	(3.3)	0.0237	0.1692	0.0312

Table 16: Proportions of selection in Example 3.3

method	selection	model (3.1)	model (3.2)	model (3.3)
Spline	AIC	0%	11%	89%
	BIC	0%	0%	100%

and model (3.2) 11% of the times. BIC selects model (3.3) 100% of the times. The use of BIC provided a more accurate selection than the use of AIC, since model (3.3) is the correct model.

3.5.4 Application to a Cyanobacteria Circadian Rhythm Dataset

Based on the simulation results, AIC and BIC choose the correct model with high probabilities. Now we apply AIC and BIC to our real data in Figure 13. We obtained the estimates of parameters using ASSIST. By adjusted cross-validation in Section 3.3.3, we obtain $limnla = 8.2$. The estimated curves are provided in Figure 18. Though the true curve is not known in this case, we can see that the estimations are pretty close to the observations.

AIC and BIC that we calculate are reported in Table 17. Both AIC and BIC

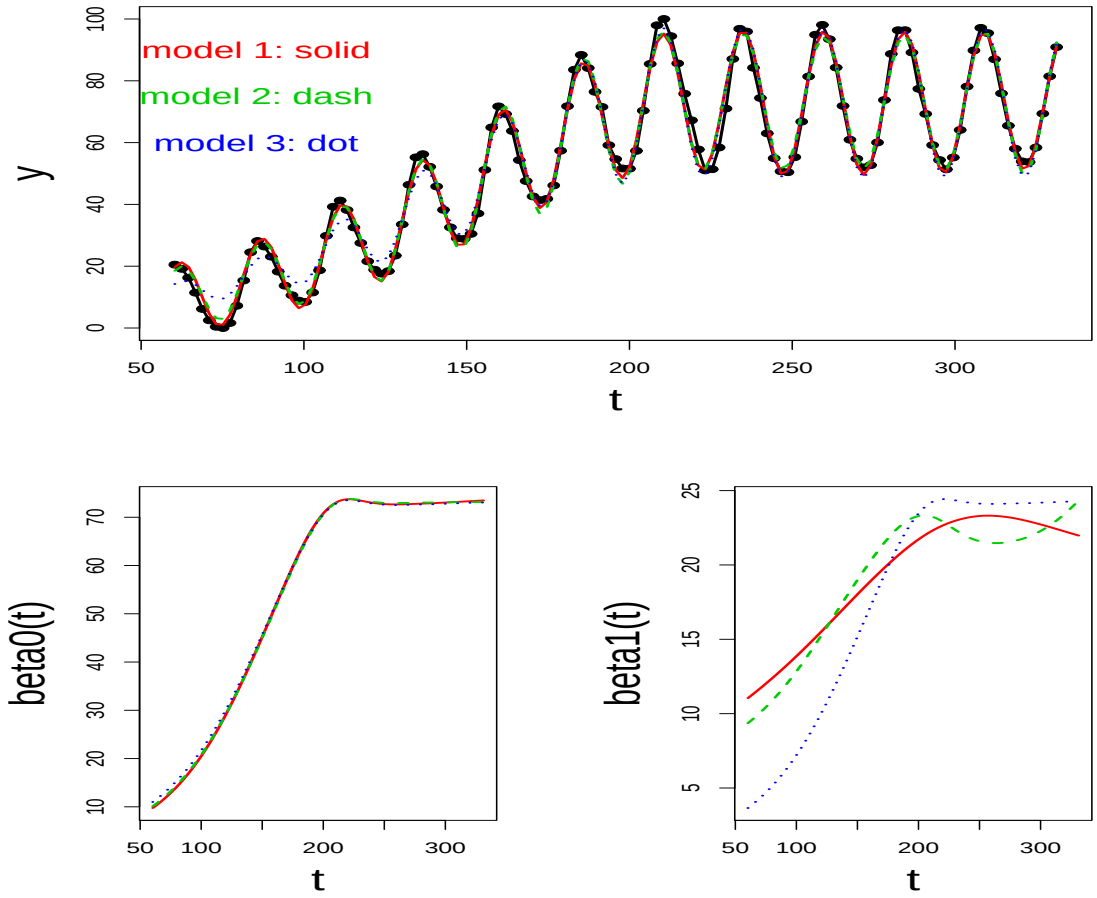

Figure 18: Spline fitted curves for overall, $\beta_0(t)$ and $\beta_1(t)$ in cyanobacteria circadian data. $limnla = 8.2$ for all three models. The red solid lines are the estimated curves under model (3.1), the green dashed lines are the estimated curves under model (3.2), and the blue dot lines are the estimated curves under model (3.3).

Table 17: Model selection in cyanobacteria circadian data

model	AIC	BIC	RSS	df
(3.1)	635.8879	669.0265	860.5469	117.4123
(3.2)	652.0414	683.9436	986.3708	117.8447
(3.3)	724.7136	750.7937	1792.2142	119.8805

select model (3.1) as the best model. This suggests that $\gamma(t)$ is not a constant, nor a quadratic function of time t.

3.6 Concluding Remarks

To investigate the relationship between the baseline and amplitude functions in cyanobacteria circadian data, we propose three models. One is the general model, while the other two are reduced models with assumption that the ratio of baseline and amplitude is a quadratic function of time or a constant. We use an R package ASSIST to estimate the data using smoothing spline. ASSIST is very flexible to use for semiparametric models, but with the identifiability problem, we can not obtain the estimates we desire by directly using ASSIST. The adjusted cross-validation we propose provides a practical solution to the identifiability problem. Using model selection techniques, we conclude that the ratio of the baseline and amplitude functions is neither a constant nor a quadratic function of time.

CHAPTER IV

HYPOTHESIS TESTING USING LOCAL LINEAR REGRESSION IN CYANOBACTERIA

4.1 Introduction

In chapter III, we estimate the cyanobacteria circadian data using smoothing spline with the R package ASSIST, and we investigate the relationship between the baseline function and the amplitude function by model selection techniques. In this chapter, we use testing hypothesis approaches to perform an equivalent investigation. The emphasis of the two approaches, model selection versus hypothesis testing, is slightly different. In former, the "best" model is selected while the complexity of the model is taken into account. In latter, the goal is to check whether a simple, "qualified" null model is acceptable in the sense that it is not significantly different from a more complex general model. Because of the heavy computational intensity using ASSIST, we perform estimation using local linear regression. To perform the hypothesis tests, we need to derive or approximate the degrees of freedom for residual sum of squares (RSS) in local linear regression. Fan, Yao, and Cai (2003) proposed an estimator of the local degree of freedom for the local linear regression procedure. We extend this approach and the original principle of Hastie and Tibshirani (1990) to derive the global degree of freedom.

The general model and the two reduced models we propose are given in Section 4.2. Kernel based estimation methods for all models are provided in Section 4.3. Section 4.4 describes the hypothesis testing procedures, and Section 4.5 contains three theorems regarding the properties of the degrees of freedom for RSS. Numerical outcomes are provided in Section 4.6. Section 4.7 contains the concluding remarks.

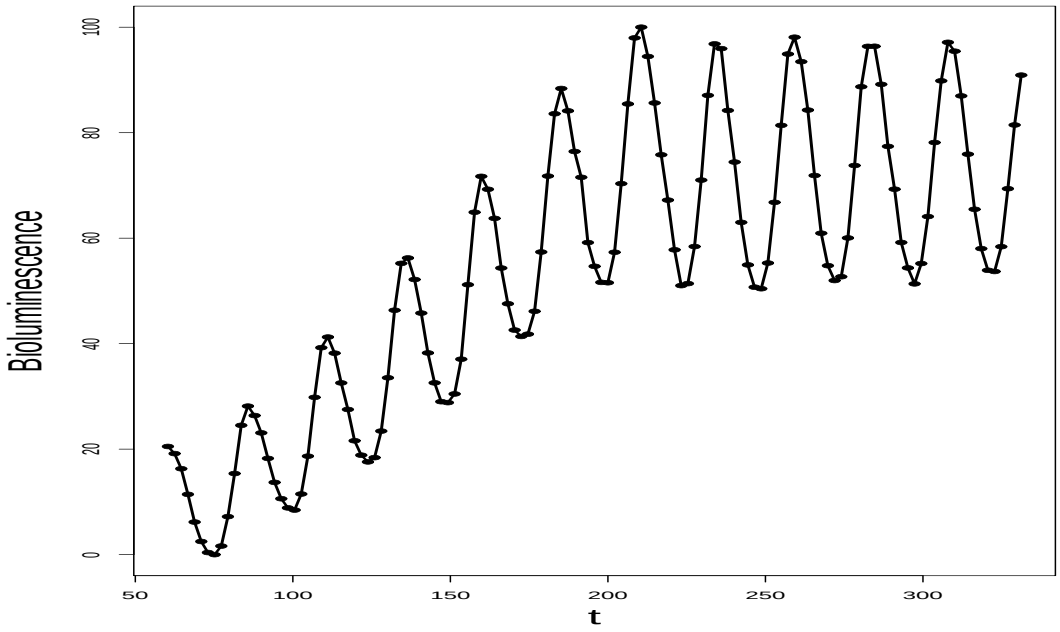

Figure 19: Plot of bioluminescence vs time.

4.2 Models

For data in Figure 19, the periodic component can be fitted by a cosine wave and the baseline and amplitude functions can be fitted by nonparametric procedures. Let y be the response variable and t be its associated covariate. We consider the following varying coefficient periodic model:

$$Y = \beta_0(t) + \beta_1(t)\cos\{2\pi(t/\tau - \phi)\} + \boldsymbol{\varepsilon}, \tag{4.1}$$

where $t = (t_1, \cdots, t_n)^T$, $Y = (y_1, \cdots, y_n)^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$. And, $\beta_0(t)$ (bounded away from 0) is the underlying baseline growth function; $\beta_1(t)$ (bounded away from 0) is the amplitude function of the periodic component; τ denotes the period and $0 \leq \phi < 1$ is the phase; n is the total number of observations. We assume that ε_j's are independent and identically distributed with $\mathrm{E}(\varepsilon_j) = 0$ and $\mathrm{var}(\varepsilon_j) = \sigma_\varepsilon^2$, $j = 1, \cdots, n$. We refer to model (4.1) as the general model.

In model (4.1), we do not make any assumptions about the relationship between $\beta_0(t)$ and $\beta_1(t)$. Biologists believe that the periodic component $\beta_1(t)$ is proportional to the growth component $\beta_0(t)$. That is, $\beta_1(t) = c\beta_0(t)$ with a positive constant c. To relax this assumption, we replace c by a quadratic function of t. Define $\gamma(t) = \log\beta_1(t) - \log\beta_0(t)$, therefore $\beta_1(t)/\beta_0(t) = \exp\{\gamma(t)\}$. Consequently, we propose the following model:

$$Y = \beta_0(t)\left[1 + \exp\{\gamma(t)\}\cos\{2\pi(t/\tau - \phi)\}\right] + \boldsymbol{\varepsilon}, \tag{4.2}$$

where $\gamma(t) = a_0 + a_1 t + a_2 t^2$. Model (4.2) implies that the relationship between $\beta_0(t)$ and $\beta_1(t)$ can be expressed as a quadratic function of time t. That is, $\gamma(t)$ is a quadratic function.

The traditional belief of the ratio between $\beta_0(t)$ and $\beta_1(t)$ being a constant implies the following model:

$$Y = \beta_0(t)\left[1 + \exp(\gamma)\cos\{2\pi(t/\tau - \phi)\}\right] + \boldsymbol{\varepsilon}, \tag{4.3}$$

Model (4.3) is a special case of model (4.2) when $\gamma(t)$ is a constant.

4.3 Kernel Estimation Methods

Since the parametric structures are unknown of the baseline function $\beta_0(t)$ and amplitude function $\beta_1(t)$, we will use nonparametric methods to estimate them. Further, the circadian component is parametric, it is natural to conduct an investigation using semiparametric techniques.

Section 4.3.1 to 4.3.3 provide the semiparametric methods for model (4.1) to (4.3), respectively. Bandwidth selection and initial estimations are provided in Section 4.3.4.

4.3.1 Semiparametric Local Linear Estimation Method for Model (4.1)

For model (4.1), we use the local linear approach to estimate the nonparametric component ($\beta_0(t)$ and $\beta_1(t)$) and use least squares to estimate the parametric component (τ and ϕ), update the estimates iteratively and component-wise until residual sum of squares (RSS) achieves the minimum. The execution is as below:

Let $\hat{\tau}^{[0]}$ and $\hat{\phi}^{[0]}$ be the initial estimates of τ and ϕ, $\hat{\beta}_0^{[0]}(t)$ and $\hat{\beta}_1^{[0]}(t)$ be the initial estimates of $\beta_0(t)$ and $\beta_1(t)$. Let $\hat{\tau}^{[k-1]}$, $\hat{\phi}^{[k-1]}$, $\hat{\beta}_0^{[k-1]}(t)$ and $\hat{\beta}_1^{[k-1]}(t)$ denote the estimates at step $(k-1)$. At the kth step,

1. We find $\hat{\tau}^{[k]}$ and $\hat{\phi}^{[k]}$ by minimizing RSS $= \sum_{j=1}^{n}(y_j - \hat{y}_j)^2$ in the neighborhood of $\hat{\tau}^{[k-1]}$ and $\hat{\phi}^{[k-1]}$, where $\hat{y}_j = \hat{\beta}_0^{[k-1]}(t_j) + \hat{\beta}_1^{[k-1]}(t_j)\cos\{2\pi(t_j/\hat{\tau} - \hat{\phi})\}$.

2. Suppose that τ and ϕ are known and equal to $\hat{\tau}^{[k]}$ and $\hat{\phi}^{[k]}$. Now the model (4.1) is

$$Y = \beta_0(t) + \beta_1(t)\cos\{2\pi(t/\hat{\tau}^{[k]} - \hat{\phi}^{[k]})\} + \boldsymbol{\varepsilon}.$$

   Our model is a common varying coefficient model. We use the local linear approach to update $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$ simultaneously with the same bandwidth h. At the time point t_0, the estimates are

$$\begin{aligned}\hat{\beta}_0^{[k]}(t_0) &= (1,0,0,0)(\mathbf{X}^T W\mathbf{X})^{-1}\mathbf{X}^T WY,\\ \hat{\beta}_1^{[k]}(t_0) &= (0,0,1,0)(\mathbf{X}^T W\mathbf{X})^{-1}\mathbf{X}^T WY,\end{aligned} \tag{4.4}$$

   where

$$\mathbf{X} = \begin{pmatrix} 1 & t_1 - t_0 & z_1 & z_1(t_1 - t_0) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_n - t_0 & z_n & z_n(t_n - t_0) \end{pmatrix},$$

$$z_j = \cos\{2\pi(t_j/\hat{\tau}^{[k]} - \hat{\phi}^{[k]})\},$$

$$\text{and } W = \text{diag}(K_h(t_1 - t_0), \cdots, K_h(t_n - t_0)).$$

Iterate the algorithm until the RSS stops decreasing, we obtain $\hat{\tau}$,$\hat{\phi}$, $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$.

4.3.2 Semiparametric Local Linear Estimation Method for Model (4.2)

For model (4.2), let $\hat{\tau}^{[0]}$ and $\hat{\phi}^{[0]}$ be the initial estimates of τ and ϕ, $\hat{\beta}_0^{[0]}(t)$ be the initial estimates of $\beta_0(t)$, and $\hat{\gamma}^{[0]}(t)$ be the initial estimate of $\gamma(t)$. Let $\hat{\tau}^{[k-1]}$, $\hat{\phi}^{[k-1]}$, $\hat{\beta}_0^{[k-1]}(t)$ and $\hat{\gamma}^{[k-1]}(t)$ denote the estimates at step $(k-1)$. At the kth step,

1. We find the kth step estimates of all parameters ($\hat{\tau}^{[k]}$, $\hat{\phi}^{[k]}$, $\hat{\gamma}^{[k]}(t) = \hat{a}_0^{[k]} + \hat{a}_1^{[k]}t + \hat{a}_2^{[k]}t^2$) by minimizing RSS $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ in the neighborhood of $\hat{\tau}^{[k-1]}$, $\hat{\phi}^{[k-1]}$ and $\hat{\gamma}^{[k-1]}(t)$.

2. Suppose that τ, ϕ and $\gamma(t)$ are known and equal to $\hat{\tau}^{[k]}$, $\hat{\phi}^{[k]}$ and $\hat{\gamma}^{[k]}(t)$. Now the model (4.2) is

$$Y = \beta_0(t)\left[1 + \exp(\hat{\gamma}^{[k]}(t))\cos\{2\pi(t/\hat{\tau}^{[k]} - \hat{\phi}^{[k]})\}\right] + \boldsymbol{\varepsilon}.$$

   Our model is a common varying coefficient model. We use the local linear approach to update $\hat{\beta}_0(t)$ with bandwidth h. The estimates are

$$\hat{\beta}_0^{[k]}(t_0) = (1,0)(\mathbf{X}_2^T W \mathbf{X}_2)^{-1}\mathbf{X}_2^T W Y, \tag{4.5}$$

   where

$$\mathbf{X}_2 = \begin{pmatrix} 1 + \exp(\hat{\gamma}_1^{[k]})z_1, & \{1 + \exp(\hat{\gamma}_1^{[k]})z_1\}(t_1 - t_0) \\ \vdots & \vdots \\ 1 + \exp(\hat{\gamma}_n^{[k]})z_n, & \{1 + \exp(\hat{\gamma}_n^{[k]})z_n\}(t_n - t_0) \end{pmatrix},$$

$$\hat{\gamma}_j^{[k]} = \hat{\gamma}^{[k]}(t_j).$$

Iterate the algorithm until the RSS stops decreasing, we will obtain $\hat{\tau}$, $\hat{\phi}$, $\hat{\beta}_0(t)$ and $\hat{\gamma}(t)$.

4.3.3 Semiparametric Local Linear Estimation Method for Model (4.3)

In model (4.3), the estimation procedure is the same as that for model (4.2), except that $\gamma(t) = \gamma$. Let $\hat{\tau}^{[0]}$ and $\hat{\phi}^{[0]}$ be the initial estimates of τ and ϕ, $\hat{\beta}_0^{[0]}(t)$ be the initial estimates of $\beta_0(t)$, and $\hat{\gamma}^{[0]}$ be the initial estimate of γ. Let $\hat{\tau}^{[k-1]}$, $\hat{\phi}^{[k-1]}$, $\hat{\beta}_0^{[k-1]}(t)$ and $\hat{\gamma}^{[k-1]}$ denote the estimates at step $(k-1)$. At the kth step,

1. We find the kth step estimates of all parameters ($\hat{\tau}^{[k]}$, $\hat{\phi}^{[k]}$, $\hat{\gamma}^{[k]}$) by minimizing RSS $= \sum_{j=1}^{n}(y_j - \hat{y}_j)^2$ in the neighborhood of $\hat{\tau}^{[k-1]}$, $\hat{\phi}^{[k-1]}$ and $\hat{\gamma}^{[k-1]}$.

2. Suppose that τ, ϕ and γ are known and equal to $\hat{\tau}^{[k]}$, $\hat{\phi}^{[k]}$ and $\hat{\gamma}^{[k]}$. Now the model (4.3) is

$$Y = \beta_0(t)[1 + \exp(\hat{\gamma}^{[k]})\cos\{2\pi(t/\hat{\tau}^{[k]} - \hat{\phi}^{[k]})\}] + \boldsymbol{\varepsilon}.$$

   Our model is a common varying coefficient model. We use the local linear approach to update $\hat{\beta}_0(t)$ with bandwidth h. The estimates are

$$\hat{\beta}_0^{[k]}(t_0) = (1,0)(\mathbf{X}_3^T W \mathbf{X}_3)^{-1}\mathbf{X}_3^T W Y, \tag{4.6}$$

   where

$$\mathbf{X}_3 = \begin{pmatrix} 1 + \exp(\hat{\gamma}^{[k]})z_1, & \{1 + \exp(\hat{\gamma}^{[k]})z_1\}(t_1 - t_0) \\ \vdots & \vdots \\ 1 + \exp(\hat{\gamma}^{[k]})z_n, & \{1 + \exp(\hat{\gamma}^{[k]})z_n\}(t_n - t_0) \end{pmatrix},$$

Iterate the algorithm until the RSS stops decreasing, we will obtain $\hat{\tau}$, $\hat{\phi}$, $\hat{\beta}_0(t)$ and $\hat{\gamma}$.

4.3.4 Bandwidth Selection and Initial Estimates

In Section 4.3.1 to 4.3.3, we need to select an optimal bandwidth h. We use common bandwidth for all three models. The simplest and most effective bandwidth selection

method could be cross-validation. For our model, we perform the cross-validation method as below:

1. Divide time into c intervals, each has k points, then leave the 1 point out in each interval, so we will use the remaining $n - c$ points to produce estimates and make predicts to the left-out c points.

2. Select a h, and conduct estimation procedures using the $n - c$ points to obtain $\hat{y}_{ck}^h$ and corresponding prediction MSE

$$G_{ck}(h) = \{y_{ck} - \hat{y}_{ck}^h\}^2.$$

3. For a given C, repeat steps 1 and 2, for c from 1 to C, and k from 1 to $K = \text{int}(n/c)$, and obtain

$$G(h) = \sum_{c=1}^{C} \sum_{k=1}^{K} G_{ck}(h) = \sum_{c=1}^{C} \sum_{k=1}^{K} \{y_{ck} - \hat{y}_{ck}^h\}^2.$$

The optimal bandwidth is $h = \text{argmin } G(h)$.

To estimate the unknown curves, we need to find a set of initial estimates. Here is how we obtain the initial estimates under model (4.1):

1. Select local minimum and maximum points.

2. Obtain the median of two adjacent minimum and maximum points, consider the medians as our baseline. Use local linear method to obtain $\hat{\beta}_0^{[0]}(t)$ from these points.

3. Subtract minimum from adjacent maximum points, consider these as two times the amplitude. Use local linear method to fit $\hat{\beta}_1^{[0]}(t)$.

4. The distance from adjacent minimum and maximum points are half the period, average them to obtain the initial estimate of τ, $\hat{\tau}^{[0]}$. Using the maximum points as the start point of each cycle, and minimum points as middle point of each cycle, by averaging them we can obtain $\hat{\phi}^{[0]}$.

For model (4.2), we obtain $\hat{\beta}_0^{[0]}(t)$, $\hat{\tau}^{[0]}$ and $\hat{\phi}^{[0]}$ the same way as we did under model (4.1). We fit a exp(quadratic) model on

$$\frac{Y - \hat{\beta}_0^{[0]}(t)}{\hat{\beta}_0^{[0]}(t)\cos\{2\pi(t/\hat{\tau}^{[0]} - \hat{\phi}^{[0]})\}}$$

and t to obtain $\hat{\gamma}^{[0]}(t)$. For model (4.3), obtain the mean of

$$\frac{Y - \hat{\beta}_0^{[0]}(t)}{\hat{\beta}_0^{[0]}(t)\cos\{2\pi(t/\hat{\tau}^{[0]} - \hat{\phi}^{[0]})\}}$$

and set that to be $\hat{\gamma}^{[0]}$.

4.4 Testing Hypothesis

We are interested in three tests. One is the goodness of fit test for model (4.2), when model (4.1) is the full model:

H_0: $\gamma(t)$ is a quadratic function,

H_1: $\gamma(t)$ is general as given in model (4.1).

We refer to this test as Test I.

Another test of interest is the test against the traditional belief:

H_0: $\gamma(t)$ is a constant,

H_1: $\gamma(t)$ is a quadratic function.

We refer to this test as Test II.

The third test of interest is the goodness of fit test for model (4.3), when model (4.1) is the full model:

H_0: $\gamma(t)$ is a constant,

H_1: $\gamma(t)$ is general as given in model (4.1).

We refer to this test as Test III.

For the original data, using the "extra sum of squares" principle, we define the test statistic to be

$$F = \frac{(RSS_{H_0} - RSS_{H_1})/(df_{H_0} - df_{H_1})}{RSS_{H_1}/df_{H_1}}, \tag{4.7}$$

where RSS_{H_1} and RSS_{H_0} mean the residual sums of squares under H_1 and H_0, df_{H_1} and df_{H_0} the degrees of freedom for residual sum of squares under H_1 and H_0, respectively. Since the numerator and denominator may not be independent in (4.7), F may not have a F distribution. We use a bootstrap procedure to calculate the null distribution of F and obtain the p-value of the test. That is, for $b = 1, \cdots, B$,

1. Generate one dataset under H_0, that is, $Y_b^* = \hat{Y}_{H_0} + \boldsymbol{\varepsilon}_{H_0,b}^*$, where $\boldsymbol{\varepsilon}_{H_0,b}^*$ is a bootstrap sample of $\hat{\boldsymbol{\varepsilon}}_{H_0} = Y - \hat{Y}_{H_0}$.

2. Fit the curves under both H_0 and H_1. Calculate F_b^* by

$$F_b^* = \frac{(RSS_{H_0,b} - RSS_{H_1,b})/(df_{H_0,b} - df_{H_1,b})}{RSS_{H_1,b}/df_{H_1,b}}.$$

To construct B F_b^*'s, compute the percentage of F_b^* bigger than F, and set that proportion as p-value.

4.5 Degree of Freedom for RSS

To perform the hypothesis tests described in Section 4.4, we need the degree of freedom for residual sum of squares in the semiparametric methods. We derive the degrees of freedom for RSS's in all 3 models in this Section.

Theorem 2. For model (4.1), the degree of freedom for $RSS = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2$ is

$$n - p = n - \sum_{k=1}^{n} H_{kk} - 2.$$

where

$$H_{kk} = K_h(0)(1, 0, z_k, 0)\{\mathbf{X}^T(t_k)W(t_k)\mathbf{X}(t_k)\}^{-1}(1, 0, z_k, 0)^T,$$

$$\mathbf{X}(t_0) = \begin{pmatrix} 1 & t_1 - t_0 & z_1 & z_1(t_1 - t_0) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_n - t_0 & z_n & z_n(t_n - t_0) \end{pmatrix},$$

and $W(t_0) = \text{diag}(K_h(t_1 - t_0), \cdots, K_h(t_n - t_0))$.

Theorem 3. For model (4.2), the degree of freedom for $RSS = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2$ is

$$n - p = n - \sum_{k=1}^{n} H_{kk} - 5,$$

where

$$H_{kk} = K_h(0)(1 + \exp(\gamma_k)z_k, 0)\{\mathbf{X}_2^T(t_k)W(t_k)\mathbf{X}_2(t_k)\}^{-1}(1 + \exp(\gamma_k)z_k, 0)^T,$$

$$\mathbf{X}_2(t_0) = \begin{pmatrix} 1 + \exp(\gamma_1)z_1, & \{1 + \exp(\gamma_1)z_1\}(t_1 - t_0) \\ \vdots & \vdots \\ 1 + \exp(\gamma_n)z_n, & \{1 + \exp(\gamma_n)z_n\}(t_n - t_0) \end{pmatrix}.$$

Theorem 4. For model (4.3), the degree of freedom for $RSS = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2$ is

$$n - p = n - \sum_{k=1}^{n} H_{kk} - 3,$$

where

$$H_{kk} = K_h(0)(1 + \exp(\gamma)z_k, 0)\{\mathbf{X}_3^T(t_k)W(t_k)\mathbf{X}_3(t_k)\}^{-1}(1 + \exp(\gamma)z_k, 0)^T,$$

$$\mathbf{X}_3(t_0) = \begin{pmatrix} 1 + \exp(\gamma)z_1, & [1 + \exp(\gamma)z_1](t_1 - t_0) \\ \vdots & \vdots \\ 1 + \exp(\gamma)z_n, & [1 + \exp(\gamma)z_n](t_n - t_0) \end{pmatrix}.$$

The proofs of Theorems 2-4 are provided in Appendix C.

After obtaining the degree of freedom for RSS, we can calculate F test statistic according to procedures given in Section 4.4. Note that in practice, local linear

regression will have boundary effect, that is, H_{kk} tend to be much larger on the boundary. We will demonstrate this in the following example.

Example 4.1. Use the data set in Figure 19. H_{kk}'s for all three models are plotted in Figure 20. From Figure 20, we observe a clear boundary effect. So, we use only interior points ($t > \min(t) + h/2$ and $t < \max(t) - h/2$) to calculate the degrees of freedom. We do so for all three models. In the Figure 20 the interior points are the red solid points between two dashed lines.

4.6 Numerical Outcomes

In this Section, we evaluate the performance of the hypothesis tests by simulations, then we apply the tests to cyanobacteria circadian data. Through out this Section, we use cross-validation to select bandwidth $h = 40$. Also, the level of significance for our tests is $\alpha = 0.05$ in this Section.

4.6.1 Simulation Studies for Test I

To evaluate the property of Test I, we first generate data under H_0 (model (4.2)). We set $n = 129$, $t \in \{60.35, 62.46, \cdots, 331.20\}$, $\tau = 24$ and $\phi = 0.5$, $\beta_0(t) = \exp(2 + 0.0165t - 0.00003t^2)$, $\gamma(t) = -0.5 - 0.0065t + 0.00001t^2$ and $\boldsymbol{\varepsilon} \sim N(0, 1^2)$. For the dataset we generate, we let B in Section 4.4 to be 1000. After performing the hypothesis test, we obtain

$$F = 1.5392,\ p - value = 0.828.$$

Since p-value is bigger than $\alpha = 0.05$, we fail to reject H_0. Therefore, we conclude that $\gamma(t)$ is a quadratic function.

We repeat the data generation and hypothesis testing procedure 100 times, let

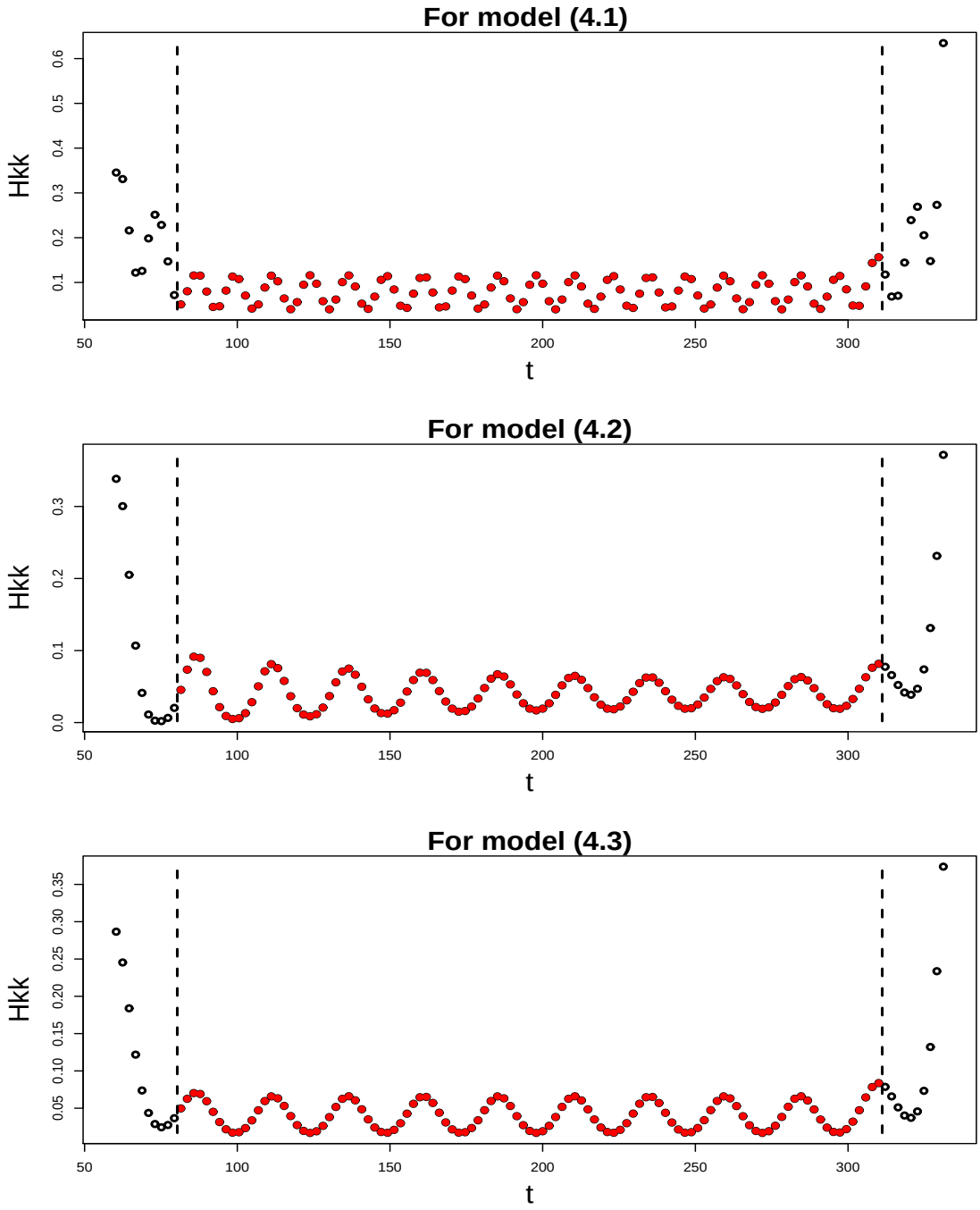

Figure 20: Plot of the diagonal values of the hat matrix H vs t in Example 4.1. The top plot is for model (4.1), the middle plot is for model (4.2) and the bottom plot is for model (4.3). Interior points ($t > min(t) + h/2$ and $t < max(t) - h/2$) are the red solid points between two dashed lines.

$B = 100$, and record all the p-values. We than estimate the size of Test I by

$$\widehat{\text{size}} = \frac{\#\ of\ p - values <= 0.05}{100} = 0.03.$$

Then we generate data under H_1, that is, model (4.1). we use the dataset in Figure 19. For our circadian dataset, $n = 156$. We only use $t > 60$. Thus n reduces to 129, and $t \in \{60.35, 62.46, \cdots, 331.20\}$. We estimated the curve first to obtain $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$, then use 'approxfun' in R to obtain the 'true' $\beta_0(t)$ and $\beta_1(t)$. We generated one dataset by model (4.1) using $\tau = 24$, $\phi = 0.5$, $\varepsilon \sim N(0, 1^2)$, and $\beta_0(t)$, $\beta_1(t)$ as above. For the dataset we generated, we set $B = 1000$ and obtained

$$F = 68.3696,\ p - value = 0.$$

Since p-value is less than 0.05, we reject H_0. We conclude that $\gamma(t)$ is general as given in model (4.1).

We repeat the data generation and hypothesis testing procedure 100 times, let $B = 100$ and record all p-values, we estimate the power of our Test I by

$$\widehat{\text{power}} = \frac{\#\ of\ p - values <= 0.05}{100} = 1.$$

4.6.2 Simulation Studies for Test II

To evaluate the numerical properties of Test II, we first generate data under H_0 (model (4.3)). $n = 129$ and $t \in \{60.35, 62.46, \cdots, 331.20\}$. We set $\tau = 24$ and $\phi = 0.5$, $\beta_0(t) = \exp(1 + 0.02t - 0.00004t^2)$, $\gamma = -0.5$ and $\boldsymbol{\varepsilon} \sim N(0, 1^2)$. For the dataset we generate, we let $B = 1000$, our calculated test statistic and p-value are

$$F = 0.3331,\ p - value = 0.349.$$

p-value is bigger than 0.05, we fail to reject H_0. Thus our conclusion is $\gamma(t)$ is a constant.

We repeat the data generation and hypothesis testing procedure 100 times with $B = 100$, and estimate the size of Test II by

$$\widehat{\text{size}} = \frac{\#\ of\ p-values <= 0.05}{100} = 0.05.$$

Then we generate data under H_1, one dataset under model (4.2). $n = 129$, $t \in \{60.35, 62.46, \cdots, 331.20\}$. We set $\tau = 24$ and $\phi = 0.5$, $\beta_0(t) = \exp(2 + 0.0165t - 0.00003t^2)$, $\gamma(t) = -0.5 - 0.0065t + 0.00001t^2$ and $\boldsymbol{\varepsilon} \sim N(0, 1^2)$. For the dataset we generate, $B = 1000$, we have

$$F = 127.3973,\ p-value = 0.$$

p-value is less than 0.05, reject H_0. We conclude that $\gamma(t)$ is a quadratic function.

We repeat the data generation and hypothesis testing procedure 100 times with $B = 100$, the estimated power of Test II is

$$\widehat{\text{power}} = \frac{\#\ of\ p-values <= 0.05}{100} = 1.$$

4.6.3 Simulation Studies for Test III

We first generate data under H_0, model (4.3). $n = 129$ and $t \in \{60.35, 62.46, \cdots, 331.20\}$. We set $\tau = 24$ and $\phi = 0.5$, $\beta_0(t) = \exp(1 + 0.02t - 0.00004t^2)$, $\gamma = -0.5$ and $\boldsymbol{\varepsilon} \sim N(0, 1^2)$. For the dataset we generate, set $B = 1000$ we obtain

$$F = 1.7293,\ p-value = 0.625.$$

p-value is bigger than 0.05, we fail to reject H_0. We conclude that $\gamma(t)$ is a constant.

Repeat the data generation and hypothesis tests 100 times with $B = 100$, estimate size by

$$\widehat{\text{size}} = \frac{\#\ of\ p-values <= 0.05}{100} = 0.02.$$

Then we generate data under H_1, model (4.1). we use the dataset in Figure 19 to generate our $n = 129$, and $t \in \{60.35, 62.46, \cdots, 331.20\}$. Estimate the curve first to obtain $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$, then use 'approxfun' in R to obtain the 'true' $\beta_0(t)$ and $\beta_1(t)$. Generate one dataset by model (4.1) using $\tau = 24$, $\phi = 0.5$, $\varepsilon \sim N(0, 1^2)$, and $\beta_0(t)$, $\beta_1(t)$ as above. For the dataset we generate, $B = 1000$,

$$F = 185.2273,\ p - value = 0.004.$$

We reject H_0 and conclude that $\gamma(t)$ is general.

Repeat the data generation and hypothesis tests 100 times, estimated power of Test III is

$$\widehat{\text{power}} = \frac{\#\ of\ p - values <= 0.05}{100} = 1.$$

4.6.4 Application to Cyanobacteria Circadian Data

From the simulation studies in Section 4.6.1 to 4.6.3, all three tests perform very well. We apply the tests to our cyanobacteria circadian data. For the dataset in Figure 19, we set $B = 1000$, and perform all three tests, we obtain

$$\begin{aligned} &\text{Test I:} && F = 13.8405, && p - value = 0.001 \\ &\text{Test II:} && F = 44.5338, && p - value = 0.025. \\ &\text{Test III:} && F = 34.2013, && p - value = 0.016. \end{aligned}$$

Since all the p-values are less than 0.05, we reject Test I, II and III. Therefore, we conclude that $\gamma(t)$ is general as given in model (4.1).

4.7 Concluding Remarks

To investigate the relationship between the baseline and amplitude functions of cyanobacteria circadian data, we proposed three models. One is the general model, while the other two are reduced models that satisfy assumptions that the ratio of the baseline

and amplitude functions is a quadratic function of time or a constant. We use semiparametric local linear estimation methods to fit the data under all three models, and we use hypothesis testing procedure to identify plausible reduced models. To perform these tests, we propose a procedure to calculate the degrees of freedom for RSS in the semiparametric methods. By applying the tests to our data, we conclude that the ratio of the baseline and amplitude functions is not a constant or a quadratic function of time.

CHAPTER V

CONCLUSION AND FUTURE RESEARCH

The study of patterns in circadian rhythm provides researchers a better understanding of the circadian input pathways. It also enables research on the interactions of circadian genes and other metabolic or cell signaling pathways to be conducted. In this dissertation, we focus on the study of circadian patterns of cyanobacteria. In chapter II, we proposed a varying coefficient periodic model that contains nonparametric baseline and amplitude functions and a parametric periodic component. This model allows us to easily investigate properties of key circadian parameters such as period and phase. This investigation is the main focus of chapter II. In the front of statistical methodology development, we have provided semiparametric kernel based estimation procedures and investigate their theoretical and numerical properties respectively. In the front of using our proposed methods to enhance research in another discipline, our approaches allow biologists to obtain a sensible confidence range for the parameter phase. This is not achievable before simply due to the large variation induced by the traditional FFT-NLLS procedures. We clearly illustrate this finding in the data example.

The proposed flexible models also allow us to further address a question that is of interest to biologists, namely "Does the circadian component remain invariant across different growth stages?" This question can be addressed via a study of the ratio of $\beta_1(t)$ over $\beta_0(t)$. It has been commonly believed that this ratio should remain a constant across time. However, this belief could be contributed to the past empirical findings when investigations were conducted during the stable sustained stage only. We investigate this problem using model selection and hypothesis testing procedures

under the smoothing spline and kernel local linear framework, respectively. Chapter III reports our developments and findings using smoothing spline. One difficulty is that when the smoothing parameter is chosen by the traditional general cross-validation method, the existing software package such as ASSIST tends to lead to outcomes which either fail to converge or absorb the periodic component into the two nonparametric functions. We overcome this difficulty by introducing a new method of smoothing parameter selection, adjusted cross-validation. Based on the outcomes of model selection techniques, our conclusion is that the log-ratio of the baseline function and amplitude function is not a constant, nor a quadratic function.

An alternative way of investigating the baseline-amplitude relationship is to conduct a hypothesis testing procedure. In chapter IV, we perform three tests under the platform of kernel local linear models. The same conclusions as those in chapter III are obtained. Namely, we can not further reduce the complexity of the proposed semiparametric models without hurting the goodness of fit property. A theoretical contribution in this chapter is to show how to derive the global model degrees of freedom for the kernel based semiparametric methods.

In the future study, we intend to investigate other topics that are of interest to biologists. One of them is whether the parameters period and phase of the circadian component remain constants across time, or do they change after a certain time period after the bacteria are under a constant condition. This question addresses the issue whether the circadian patterns change with the aging process and consequently have effects on other growth related pathways.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Bloomfield, P. (1976). *Fourier Analysis of Time Series: An Introduction*. New York: Wiley.

Brockwell, P. J. and Davis, R. A. (1996). *Time Series: Theory and Methods*. New York: Springer.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). *Local regression models. In Chambers, J. M. and Hastie, T. J. (eds.), Statistical Models in S, pp. 309-376*. Pacific Grove, CA: Wadsworth.

Dodd, A. N., Salathia, N., Hall, A., Kevei, E., Toth, R., Nagy, F., Hibberd, J. M., Millar, A. J., and Webb, A. A. R. (2005). Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science* **309**, 630–633.

Fan, J. and Gijlels, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11**, 1031–1057.

Fan, J., Yao, Q., and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society - Series B: Statistical Methodology* **65**, 57–80.

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

Golden, S. S. (2003). Timekeeping in bacteria: the cyanobacterial circadian clock. *Current Opinion in Microbiology* **6**, 535–540.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman & Hall.

Hastie, T. J. and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall.

Hastie, T. J. and Tibshirani, R. J. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society - Series B: Statistical Methodology* **55**, 757–796.

Johnson, C. H. and Golden, S. S. (1999). Circadian programs in cyanobacteria: Adaptiveness and mechanism. *Annual Review of Microbiology* **53**, 389–409.

Li, Q., Huang, C. J., Li, D., and Fu, T. T. (2002). Semiparametric smooth coefficient models. *Journal of Business and Economic Statistics* **3**, 412–422.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics* **19**, 474–482.

Niinuma, K., Someya, N., Kimura, M., Yamaguchi, I., and Hamamoto, H. (2005). Circadian rhythm of circumnutation in inflorescence stems of arabidopsis. *Plant and Cell Physiology* **46**, 1423–1427.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman & Hall.

Wang, Y. and Brown, M. B. (1996). A flexible model for human circadian rhythms. *Biometrics* **52**, 588–596.

Wang, Y. and Ke, C. (2002). Assist: A suite of s-plus functions implementing spline smoothing techniques. *Manual for the ASSIST Package*, Available at http://www.pstat.ucsb.edu/faculty/yuedong/software.

Zhang, W., Lee, S. Y., and Song, X. (2002). Local polynomial fitting in semivarying coefficient models. *Journal of Multivariate Analysis* **82**, 166–188.

Zhang, X., Dong, G., and Golden, S. S. (2006). The pseudo-receiver domain of cika regulates the cyanobacterial circadian input pathway. *Molecular Microbiology* **60**, 658–668.

APPENDIX A

PROOFS OF THEOREM 1 IN CHAPTER II, SECTION 2.5

Definition of Terms and Regularity Conditions in Theorem 1.

The following notations will be used in the proof of Theorem 1. Let

$$K_h(t-t_0) = \frac{1}{h}K\left(\frac{t-t_0}{h}\right),$$

$$\mu_i = \int t^i K(t)dt,$$

$$\text{and } \nu_i = \int t^i K^2(t)dt.$$

Set $r_{ij}(t) = E(X_iX_j|T=t), r_{ij} = E\{X_i(t)X_j(t)|t_0\}$, for $i,j = 0,\cdots,p$. Put

$$\Psi = \text{diag}(\sigma_\varepsilon^2,\cdots,\sigma_\varepsilon^2),$$

$$\alpha_k(t) = (r_{0k}(t),\cdots,r_{pk}(t))^T,$$

$$\alpha_k = \alpha_k(t_0) \text{ for } k = 0,\cdots,p,$$

and

$$\Omega_k(t) = E\{(X_1,\cdots,X_k)^T(X_1,\cdots,X_k)|T=t\},$$

$$\Omega_k = \Omega_k(t_0), \text{ for } k = 0,\cdots,p.$$

Let

$$S = \Omega_p \otimes \begin{pmatrix} \mu_0 & 0 \\ 0 & \mu_2 \end{pmatrix} \text{ and } A = I_p \otimes \begin{pmatrix} 1 & 0 \\ 0 & h \end{pmatrix}, \tag{A.1}$$

where $\otimes$ denotes the Kronecker product.

Denote the marginal density of t by $f(t)$. Let $\boldsymbol{\varepsilon}_\theta = \theta_i - \theta$ be the error term of the distribution $\theta_i \sim (\theta, \Sigma_\theta)$, we have $\text{E}(\boldsymbol{\varepsilon}_\theta) = 0$ and $\text{var}(\boldsymbol{\varepsilon}_\theta) = \Sigma_\theta$.

We impose the following regularity conditions:

s.1 $\mathrm{E}X_k^{2s} < \infty$, for some $s > 2, k = 0, \cdots, p$.

s.2 $\beta''_{ki}(\cdot)$ is continuous in a neighborhood of t, for $k = 0, \cdots, p$. Assume $\beta''_{ki}(t) \neq 0$, for $k = 0, \cdots, p$.

s.3 $r''_{ij}(\cdot)$ is continuous in a neighborhood of t_0 and $r''_{ij}(t_0) \neq 0$, for $i, j = 0, \cdots, p$.

s.4 The marginal density of t has a continuous second derivative in some neighborhood of t_0 and $f(t_0)$ is bounded away from zero.

s.5 $K(t)$ is a symmetric density function with a compact support.

s.6 $h \to 0$ in such a way that $nh \to \infty$ and $\sqrt{m}h^2 \to 0$.

Proof of Theorem 1.

First, we calculate the asymptotic bias and variance of $\hat{\theta}_i$. For the ith subject $\hat{\theta}_i$, minimizes

$$\sum_{j=1}^{n} \{y_{ij} - \hat{\beta}_{0i}(t_j) - \hat{\beta}_{1i}(t_j) X_1(t_j; \hat{\theta}_i)\}^2.$$

We tentatively drop the subindex i to simplify the notation. We note that

$$\begin{aligned}
& \sum_{j=1}^{n} \{y_j - \hat{\beta}_0(t_j) - \hat{\beta}_1(t_j) X_1(t_j; \hat{\theta})\}^2 \\
= & \left(\sum_{j=1}^{n} \{y_j - \hat{\beta}_0(t_j) - \hat{\beta}_1(t_j) X_1(t_j; \theta) - \hat{\beta}_1(t_j) \frac{\partial}{\partial \theta} X_1(t_j; \theta)(\hat{\theta} - \theta)\}^2 \right) (1 + o_p(\hat{\theta} - \theta)^2) \\
= & \sum_{j=1}^{n} \left\{ \varepsilon_j - [\hat{\beta}_0(t_j) - \beta_0(t_j)] - [\hat{\beta}_1(t_j) - \beta_1(t_j)] X_1(t_j; \theta) - \beta_1(t_j) \frac{\partial}{\partial \theta} X_1(t_j; \theta)(\hat{\theta} - \theta) \right\}^2 \\
& (1 + o_p(1)).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\hat{\theta} - \theta &= (\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^T \boldsymbol{\varepsilon} - (\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1} \hat{\mathbf{X}}^T [\hat{\beta}_0(t) - \beta_0(t)] - (\hat{\mathbf{X}}^T \hat{\mathbf{X}})^{-1} \underline{\hat{\mathbf{X}}}^T [\hat{\beta}_1(t) - \beta_1(t)] \\
&= I_1 + I_2 + I_3, \qquad \text{(A.2)}
\end{aligned}$$

where $\hat{X}_j = \beta_1(t_j)\frac{\partial}{\partial\theta}X_1(t_j;\theta) = \beta_1(t_j)\begin{pmatrix} \frac{\partial}{\partial\tau}X_1(t_j;\theta) \\ \frac{\partial}{\partial\phi}X_1(t_j;\theta) \end{pmatrix}$, $\hat{\mathbf{X}} = (\hat{X}_1, \cdots, \hat{X}_n)^T$, and $\underline{\hat{\mathbf{X}}} = (\hat{X}_1 X_1(t_1), \cdots, \hat{X}_n X_1(t_n))^T$.

I_1 is the influence function when both $\beta_0(t)$ and $\beta_1(t)$ are known, I_2 is the part influence by unknown $\beta_0(t)$, and I_3 is the part influenced by unknown $\beta_1(t)$. By Lemma 1,

$$\begin{aligned} I_2 &= -(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T[\hat{\beta}_0(t) - \beta_0(t)] \\ &= -(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T[\mathcal{B}_0(t)h^2 + \frac{1}{\sqrt{nh}}\mathcal{K}_0^T(t)\boldsymbol{\varepsilon}](1 + o_p(1)), \end{aligned}$$

and

$$\begin{aligned} I_3 &= -(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\underline{\hat{\mathbf{X}}}^T[\hat{\beta}_1(t) - \beta_1(t)] \\ &= -(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\underline{\hat{\mathbf{X}}}^T[\mathcal{B}_1(t)h^2 + \frac{1}{\sqrt{nh}}\mathcal{K}_1^T(t)\boldsymbol{\varepsilon}](1 + o_p(1)). \end{aligned}$$

Put I_2 and I_3 into (A.2),

$$\hat{\theta} - \theta = \{\mathcal{B}_\theta(t)h^2 + \frac{1}{\sqrt{n}}\mathcal{K}_\theta^T(t)\boldsymbol{\varepsilon}\}(1 + o_p(1)),$$

where

$$\mathcal{B}_\theta(t) = -(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}[\hat{\mathbf{X}}^T\mathcal{B}_0(t) + \underline{\hat{\mathbf{X}}}^T\mathcal{B}_1(t)],$$

$$\mathcal{K}_\theta^T(t) = \sqrt{n}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}[\hat{\mathbf{X}}^T - \frac{1}{\sqrt{nh}}\hat{\mathbf{X}}^T\mathcal{K}_0^T(t) - \frac{1}{\sqrt{nh}}\underline{\hat{\mathbf{X}}}^T\mathcal{K}_1^T(t)].$$

Some calculations show that

$$\begin{aligned} \mathrm{var}(I_1) &= (\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T\Psi\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1} = (\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\sigma_\varepsilon^2 \\ &= (\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\sigma_\varepsilon^2 = \left(\sum_{j=1}^n \beta_1^2(t_j)\frac{\partial}{\partial\theta}X_1(t_j;\theta)\frac{\partial}{\partial\theta}X_1^T(t_j;\theta)\right)^{-1}\sigma_\varepsilon^2. \end{aligned}$$

Consequently, $\mathrm{var}(I_1)$ is of order $O_P(1/n)$. Further,

$$\mathrm{var}(I_2) = (\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T\mathrm{var}(\hat{\beta}_0(t) - \beta_0(t))\hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1} = O\left(\frac{1}{nh}\right)(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1},$$

which is of order $O_P(1/n^2h)$;

$$\begin{aligned}\mathrm{var}(I_3) &= (\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\underline{\hat{\mathbf{X}}}^T\mathrm{var}(\hat{\beta}_1(t)-\beta_1(t))\underline{\hat{\mathbf{X}}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\\ &= O\left(\frac{1}{nh}\right)(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\underline{\hat{\mathbf{X}}}^T\underline{\hat{\mathbf{X}}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1},\end{aligned}$$

which is of order $O_P(1/n^2h)$ as well. Because $nh\to\infty$, $\mathrm{var}(\hat{\theta}-\theta)$ is of order $O_P(1/n)$. Thus, $\mathcal{K}_\theta^T(t)$ is of order 1. Place the subindex i back in, in sum,

$$\hat{\theta}_i-\theta_i=\{\mathcal{B}_{\theta_i}(t)h^2+\frac{1}{\sqrt{n}}\mathcal{K}_{\theta_i}^T(t)\boldsymbol{\varepsilon}_i\}(1+o_p(1)). \tag{A.3}$$

Next, we calculate the asymptotic bias and variance of $\hat{\theta}$. Since $\hat{\theta}=\frac{1}{m}\sum_{i=1}^m\hat{\theta}_i$, we have

$$\hat{\theta}-\theta=\frac{1}{m}\sum_{i=1}^m\hat{\theta}_i-\theta=\frac{1}{m}\sum_{i=1}^m(\hat{\theta}_i-\theta_i)+\frac{1}{m}\sum_{i=1}^m(\theta_i-\theta).$$

By (A.3),

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m(\hat{\theta}_i-\theta_i) &= \frac{1}{m}\sum_{i=1}^m\{\mathcal{B}_{\theta_i}(t)h^2+\frac{1}{\sqrt{n}}\mathcal{K}_{\theta_i}^T(t)\boldsymbol{\varepsilon}_i\}(1+o_p(1))\\ &= \left\{\frac{1}{m}\sum_{i=1}^m\mathcal{B}_{\theta_i}(t)h^2+\frac{1}{m}\frac{1}{\sqrt{n}}(\mathcal{K}_{\theta_1}^T(t),\cdots,\mathcal{K}_{\theta_m}^T(t))\boldsymbol{\varepsilon}\right\}(1+o_p(1))\\ &= \{\mathcal{B}_\theta(t)h^2+\frac{1}{\sqrt{mn}}\mathcal{K}_\theta^T(t)\boldsymbol{\varepsilon}\}(1+o_p(1)),\end{aligned}$$

where $\mathcal{B}_\theta(t)=\frac{1}{m}\sum_{i=1}^m\mathcal{B}_{\theta_i}(t)$, $\mathcal{K}_\theta^T(t)=\sqrt{m}(\mathcal{K}_{\theta_1}^T(t),\cdots,\mathcal{K}_{\theta_m}^T(t))$.

Since $\theta_i\sim N(\theta,\Sigma_\theta)$,

$$\hat{\theta}-\theta = \{\mathcal{B}_\theta(t)h^2+\frac{1}{\sqrt{mn}}\mathcal{K}_\theta^T(t)\boldsymbol{\varepsilon}+\frac{1}{\sqrt{m}}\boldsymbol{\varepsilon}_\theta\}(1+o_p(1)),$$

where $\boldsymbol{\varepsilon}_\theta$ is defined at the beginning of Appendix A.

$$\sqrt{m}(\hat{\theta}-\theta)=\{\mathcal{B}_\theta(t)\sqrt{m}h^2+\frac{1}{\sqrt{n}}\mathcal{K}_\theta^T(t)\boldsymbol{\varepsilon}+\boldsymbol{\varepsilon}_\theta\}(1+o_p(1)).$$

By condition s.6, the bias goes to zero, and the variance is

$$\text{var}\left(\sqrt{m}(\hat{\theta}-\theta)\right) = \Sigma_\theta + \frac{1}{n}\mathcal{K}_\theta^T(t)\mathcal{K}_\theta(t)\sigma_\varepsilon^2.$$

That is,

$$\sqrt{m}(\hat{\theta}-\theta) \to N(0,\Sigma).$$

where $\Sigma = \Sigma_\theta + \Sigma_\varepsilon$, $\Sigma_\varepsilon = \frac{1}{n}\mathcal{K}_\theta^T(t)\mathcal{K}_\theta(t)\sigma_\varepsilon^2$.

Proof of Lemma 1.

We will tentatively assume θ_i is known and prove Lemma (2.8). For item i, model (2.3) is a common varying coefficient model:

$$Y_i = \sum_{k=0}^{p} \beta_{ki}(t)X_k(t) + \boldsymbol{\varepsilon}_i.$$

For simplicity, we tentatively drop the subindex i. One simple approach to estimate the coefficient functions $\beta_k(t)$ is to use local linear modeling. For each given point t_0, approximate the functions locally as

$$\beta_k(t) \approx a_k + b_k(t-t_0), k = 0, \cdots, p,$$

for t in a neighborhood of t_0. This leads to the following local least-squares problem that we minimize

$$\sum_{j=1}^{n}\left[y_j - \sum_{k=0}^{p}\{a_k + b_k(t_j-t_0)\}X_k(t_j)\right]^2 K_h(t_j-t_0)$$

for a given kernel function K and bandwidth h. We use the same bandwidth for $\beta_0(t)$ and $\beta_1(t)$ as in Section 2.4.2.2. The solutions to this problem can be easily obtained, and they can be expressed as

$$\hat{\beta}_k(t_0) = e_{2k+1,2p+2}^T(\mathbf{X}^T W\mathbf{X})^{-1}\mathbf{X}^T WY, \tag{A.4}$$

where $\mathbf{X}$ and W are defined in (2.6).

By Taylor's expansion, we have

$$Y = \mathbf{X}(\beta_0(t_0), \beta_0'(t_0), \cdots, \beta_p(t_0), \beta_p'(t_0))^T + \frac{1}{2}\sum_{k=0}^{p} \begin{pmatrix} \beta_k''(\xi_{1k})(t_1 - t_0)^2 X_k(t_1) \\ \vdots \\ \beta_k''(\xi_{nk})(t_n - t_0)^2 X_k(t_n) \end{pmatrix} + \boldsymbol{\varepsilon}, \tag{A.5}$$

where ξ_{jk} is between t_j and t_0 for $j = 1, \cdots, n$, $k = 0, \cdots, p$. Put (A.5) into (A.4),

$$\begin{aligned} \hat{\beta}_k(t_0) - \beta_k(t_0) &= \frac{1}{2}\sum_{k=0}^{p} e_{2k+1,2p+2}^T (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W \begin{pmatrix} \beta_k''(\xi_{1k})(t_1 - t_0)^2 X_k(t_1) \\ \vdots \\ \beta_k''(\xi_{nk})(t_n - t_0)^2 X_k(t_n) \end{pmatrix} \\ &\quad + e_{2k+1,2p+2}^T (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W \boldsymbol{\varepsilon} \\ &= I_4 + I_5. \end{aligned} \tag{A.6}$$

By calculating the mean and variance, one can easily get

$$\begin{aligned} \mathbf{X}^T W \mathbf{X} &= nf(t_0) \left[\begin{pmatrix} r_{00} & \cdots & r_{0p} \\ \vdots & \ddots & \vdots \\ r_{01} & \cdots & r_{11} \end{pmatrix} \otimes \begin{pmatrix} \mu_0 & 0 \\ 0 & h^2\mu_2 \end{pmatrix} \right] (1 + o_p(1)) \\ &= nf(t_0) A S A (1 + o_p(1)), \end{aligned}$$

where A and S are defined in (A.1). Similarly, we have

$$\begin{aligned} \mathbf{X}^T W \begin{pmatrix} \beta_k''(\xi_{1k})(t_1 - t_0)^2 X_k(t_1) \\ \vdots \\ \beta_k''(\xi_{nk})(t_n - t_0)^2 X_k(t_n) \end{pmatrix} &= nf(t_0) h^2 \beta_k''(t_0) \left[\begin{pmatrix} r_{0k} \\ \vdots \\ r_{pk} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right] \mu_2 \\ &\quad (1 + o_p(1)) \\ &= nf(t_0) h^2 \beta_k''(t_0) A [\alpha_k \otimes (1,0)^T] \mu_2 (1 + o_p(1)). \end{aligned}$$

Therefore,

$$I_4 = \frac{1}{2} h^2 \sum_{k=0}^{p} \beta_k''(t_0) e_{2k+1,2p+2}^T S^{-1} [\alpha_k \otimes (1,0)^T] (1 + o_p(1)) = \mathcal{B}_k(t_0) h^2 (1 + o_p(1)),$$

where $\mathcal{B}_k(t_0) = \frac{1}{2}\sum_{k=0}^{p}\beta_k''(t_0)e_{2k+1,2p+2}^T S^{-1}[\alpha_k \otimes (1,0)^T] = O(1)$.

Also,

$$\begin{aligned} I_5 &= e_{2k+1,2p+2}^T(\mathbf{X}^T W\mathbf{X})^{-1}\mathbf{X}^T W\boldsymbol{\varepsilon} \\ &= \frac{1}{nf(t_0)}e_{2k+1,2p+2}^T S^{-1}A^{-1}X^T W\boldsymbol{\varepsilon}(1+o_p(1)) \\ &= \frac{1}{\sqrt{nh}}\mathcal{K}_k^T(t_0)\boldsymbol{\varepsilon}(1+o_p(1)), \end{aligned}$$

where $\mathcal{K}_k^T(t_0) = \frac{\sqrt{nh}}{nf(t_0)}e_{2k+1,2p+2}^T S^{-1}A^{-1}X^T W = O(1)$. Put the values of I_4 and I_5 back into (A.6),

$$\hat{\beta}_k(t_0) - \beta_k(t_0) = \{\mathcal{B}_k(t_0)h^2 + \frac{1}{\sqrt{nh}}\mathcal{K}_k^T(t_0)\boldsymbol{\varepsilon}\}(1+o_p(1)).$$

Since

$$\mathrm{var}(I_5) = e_{2k+1,2p+2}^T(\mathbf{X}^T W\mathbf{X})^{-1}\mathbf{X}^T W\Psi W\mathbf{X}(\mathbf{X}^T W\mathbf{X})^{-1}e_{2k+1,2p+2},$$

and

$$\mathbf{X}^T W\Psi W\mathbf{X} = nf(t_0)\sigma_\varepsilon^2 h^{-1}\left[\begin{pmatrix} r_{00} & r_{01} \\ r_{01} & r_{11} \end{pmatrix} \otimes \begin{pmatrix} \nu_0 & 0 \\ 0 & h^2\nu_2 \end{pmatrix}\right](1+o_p(1)),$$

$\mathrm{var}(I_5)$ is of order $1/nh$, consequently $\mathcal{K}_k^T(t_0)$ is of order 1. Adding the subindex i back in, our conclusion is

$$\hat{\beta}_{ki}(t_0) - \beta_{ki}(t_0) = \{\mathcal{B}_{ki}(t_0)h^2 + \frac{1}{\sqrt{nh}}\mathcal{K}_{ki}^T(t_0)\boldsymbol{\varepsilon}_i\}(1+o_p(1)).$$

APPENDIX B

$\beta_0(T)$'S AND $\beta_1(T)$'S USED IN SECTION 2.6.2 WHEN $M = 6$

When $m = 6$, we used the best quadratic functions in the analysis of a pilot data set as the basis of constructing $\beta_0(t)$ and $\beta_1(t)$. Precisely, we set

$$\beta_{01}(t) = \exp(-0.1 + 0.034t - 0.00007t^2),$$
$$\beta_{02}(t) = \exp(0.1 + 0.032t - 0.00006t^2),$$
$$\beta_{03}(t) = \exp(-0.4 + 0.033t - 0.00006t^2),$$
$$\beta_{04}(t) = \exp(0.3 + 0.028t - 0.00005t^2),$$
$$\beta_{05}(t) = \exp(0.5 + 0.028t - 0.00005t^2),$$
$$\beta_{06}(t) = \exp(1 + 0.025t - 0.00005t^2);$$

and

$$\beta_{01}(t) = \exp(0.8 + 0.024t - 0.00005t^2),$$
$$\beta_{02}(t) = \exp(1 + 0.022t - 0.00005t^2),$$
$$\beta_{03}(t) = \exp(0.4 + 0.024t - 0.00004t^2),$$
$$\beta_{04}(t) = \exp(1.4 + 0.015t - 0.00003t^2),$$
$$\beta_{05}(t) = \exp(1.4 + 0.016t - 0.00003t^2),$$
$$\beta_{06}(t) = \exp(1.9 + 0.016t - 0.00004t^2).$$

APPENDIX C

DERIVATION OF THEOREM 2 TO 4 IN CHAPTER IV, SECTION 4.5

Proof of Theorem 2:

For model (4.1), first assume τ and ϕ are given since they have better convergence, then in the neighborhood of t_0,

$$\hat{\beta}(t_0) = (\hat{\beta}_0(t_0), \hat{\beta}_0'(t_0), \hat{\beta}_1(t_0), \hat{\beta}_1'(t_0))^T = \{\mathbf{X}^T(t_0)W(t_0)\mathbf{X}(t_0)\}^{-1}\mathbf{X}^T(t_0)W(t_0)Y,$$

where

$$\mathbf{X}(t_0) = \begin{pmatrix} 1 & t_1 - t_0 & z_1 & z_1(t_1 - t_0) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_n - t_0 & z_n & z_n(t_n - t_0) \end{pmatrix},$$

$$z_j = \cos\{2\pi(t_j/\hat{\tau} - \hat{\phi})\},$$

$$W = \mathrm{diag}(K_h(t_1 - t_0), \cdots, K_h(t_n - t_0)).$$

Then

$$
\begin{aligned}
\hat{Y}_k &= \hat{\beta}_0(t_k) + \hat{\beta}_1(t_k) z_k \\
&= (1, 0, z_k, 0)\hat{\beta}(t_k) \\
&= (1, 0, z_k, 0)\{\mathbf{X}^T(t_k) W(t_k) \mathbf{X}(t_k)\}^{-1} \mathbf{X}^T(t_k) W(t_k) Y \\
&= (1, 0, z_k, 0)\{\mathbf{X}^T(t_k) W(t_k) \mathbf{X}(t_k)\}^{-1} \\
&\quad \begin{pmatrix} 1 & \cdots & 1 \\ t_1 - t_k & \cdots & t_n - t_k \\ z_1 & \cdots & z_n \\ z_1(t_1 - t_k) & \cdots & z_n(t_n - t_k) \end{pmatrix} \begin{pmatrix} K_h(t_1 - t_k) Y_1 \\ \vdots \\ K_h(t_n - t_k) Y_n \end{pmatrix} \\
&= (1, 0, z_k, 0)\{\mathbf{X}^T(t_k) W(t_k) \mathbf{X}(t_k)\}^{-1} \begin{pmatrix} \sum_{j=1}^n Y_j K_h(t_j - t_k) \\ \sum_{j=1}^n Y_j K_h(t_j - t_k)(t_j - t_k) \\ \sum_{j=1}^n Y_j K_h(t_j - t_k) z_j \\ \sum_{j=1}^n Y_j K_h(t_j - t_k) z_j (t_j - t_k) \end{pmatrix} \\
&= \sum_{j=1}^n Y_j K_h(t_j - t_k)(1, 0, z_k, 0)\{\mathbf{X}^T(t_k) W(t_k) \mathbf{X}(t_k)\}^{-1} \begin{pmatrix} 1 \\ t_j - t_k \\ z_j \\ z_j(t_j - t_k) \end{pmatrix} \qquad \text{(C.1)}
\end{aligned}
$$

It is easy to see that all those predicted values are linear combinations of $Y = (Y_1, \cdots, Y_n)^T$ with coefficients depending on $\{z_k\}$ only. Namely, we may write

$$(\hat{Y}_1, \cdots, \hat{Y}_n)^T = \mathbf{H} Y,$$

where $\mathbf{H}$ is the $n \times n$ hat matrix, independent of Y. To calculate $\mathrm{tr}\{\mathbf{H}\}$, we note by (C.1) that the (k, j) th element of $\mathbf{H}$ is

$$H_{kj} = K_h(t_j - t_k)(1, 0, z_k, 0)\{\mathbf{X}^T(t_k) W(t_k) \mathbf{X}(t_k)\}^{-1}(1, t_j - t_k, z_j, z_j(t_j - t_k))^T$$

and the kth diagonal element of $\mathbf{H}$ is

$$H_{kk} = K_h(0)(1, 0, z_k, 0)\{\mathbf{X}^T(t_k)W(t_k)\mathbf{X}(t_k)\}^{-1}(1, 0, z_k, 0)^T.$$

Now, we have that $\text{tr}\{\mathbf{H}\} = \sum_{k=1}^{n} H_{kk}$.

$\text{tr}\{\mathbf{H}\}$ can be viewed as the number of parameters used, $p = \text{tr}\{\mathbf{H}\}$.Therefore, the degree of freedom for the residual sum of squares $RSS = (Y - \hat{Y})^T(Y - \hat{Y})$ is

$$n - p = n - \sum_{k=1}^{n} H_{kk}.$$

Now, if τ and ϕ are unknown, the number of parameters will be $p = \text{tr}\{\mathbf{H}\} + 2$, and the degree of freedom for RSS in model (4.1) is

$$n - p = n - \sum_{k=1}^{n} H_{kk} - 2.$$

Proof of Theorem 3:

The proof of Theorem 3 is very similar to the proof of Theorem 2. For model (4.2), assume the parametric components (τ, ϕ and $\gamma(t)$) known,

$$\hat{\beta}(t_0) = (\hat{\beta}_0(t_0), \hat{\beta}_0'(t_0)) = \{\mathbf{X}_2(t_0)^T W(t_0)\mathbf{X}_2(t_0)\}^{-1}\mathbf{X}_2^T(t_0)W(t_0)Y$$

where

$$\mathbf{X}_2(t_0) = \begin{pmatrix} 1 + \exp(\gamma_1)z_1, & \{1 + \exp(\gamma_1)z_1\}(t_1 - t_0) \\ \vdots & \vdots \\ 1 + \exp(\gamma_n)z_n, & \{1 + \exp(\gamma_n)z_n\}(t_n - t_0) \end{pmatrix},$$

$$\gamma_j = \gamma(t_j) = a_0 + a_1 t_j + a_2 t_j^2.$$

So,

$$
\begin{aligned}
\hat{Y}_k &= \hat{\beta}_0(t_k)\{1+\exp(\gamma_k)z_k\} \\
&= (1+\exp(\gamma_k)z_k, 0)\hat{\beta}(t_k) \\
&= (1+\exp(\gamma_k)z_k, 0)\{\mathbf{X}_2(t_k)^T W(t_k)\mathbf{X}_2(t_k)\}^{-1} \\
&\quad \begin{pmatrix} 1+\exp(\gamma_1)z_1 & \cdots & 1+\exp(\gamma_n)z_n \\ \{1+\exp(\gamma_1)z_1\}(t_1-t_k) & \cdots & \{1+\exp(\gamma_n)z_n\}(t_n-t_k) \end{pmatrix} \\
&\quad \begin{pmatrix} K_h(t_1-t_k)Y_1 \\ \vdots \\ K_h(t_n-t_k)Y_n \end{pmatrix} \\
&= (1+\exp(\gamma_k)z_k, 0)\{\mathbf{X}_2(t_k)^T W(t_k)\mathbf{X}_2(t_k)\}^{-1} \\
&\quad \begin{pmatrix} \sum_{j=1}^n Y_j K_h(t_j-t_k)\{1+\exp(\gamma_j)z_j\} \\ \sum_{j=1}^n Y_j K_h(t_j-t_k)\{1+\exp(\gamma_j)z_j\}(t_j-t_k) \end{pmatrix} \\
&= \sum_{j=1}^n Y_j K_h(t_j-t_k)(1+\exp(\gamma_k)z_k, 0)\{\mathbf{X}_2(t_k)^T W(t_k)\mathbf{X}_2(t_k)\}^{-1} \\
&\quad \begin{pmatrix} 1+\exp(\gamma_j)z_j \\ \{1+\exp(\gamma_j)z_j\}(t_j-t_k) \end{pmatrix} \qquad \text{(C.2)}
\end{aligned}
$$

We can write

$$\hat{Y} = (\hat{Y}_1, \cdots, \hat{Y}_n)^T = \mathbf{H}Y,$$

where $\mathbf{H}$ is the hat matrix. By (C.2), the kth diagonal element of $\mathbf{H}$ is

$$H_{kk} = K_h(0)(1+\exp(\gamma_k)z_k, 0)\{\mathbf{X}_2(t_k)^T W(t_k)\mathbf{X}_2(t_k)\}^{-1}(1+\exp(\gamma_k)z_k, 0)^T.$$

The trace of the hat matrix is $\mathrm{tr}\{\mathbf{H}\} = \sum_{k=1}^n H_{kk}$.

Since τ, ϕ and $\gamma(t) = a_0 + a_1 t + a_2 t^2$ are unknown, the number of parameters will be $p = \mathrm{tr}\{\mathbf{H}\} + 5$, and the degree of freedom for RSS under model (4.2) is

$$n - p = n - \sum_{k=1}^n H_{kk} - 5.$$

Proof of Theorem 4:

The proof of Theorem 4 is very similar to the proof of Theorem 3.

For model (4.3), assume the parametric components (τ, ϕ and γ) known,

$$\hat{\beta}(t_0) = (\hat{\beta}_0(t_0), \hat{\beta}_0'(t_0)) = \{\mathbf{X}_3(t_0)^T W(t_0)\mathbf{X}_3(t_0)\}^{-1}\mathbf{X}_3^T(t_0)W(t_0)Y$$

where

$$\mathbf{X}_3(t_0) = \begin{pmatrix} 1+\exp(\gamma)z_1, & [1+\exp(\gamma)z_1](t_1-t_0) \\ \vdots & \vdots \\ 1+\exp(\gamma)z_n, & [1+\exp(\gamma)z_n](t_n-t_0) \end{pmatrix}.$$

Thus,

$$\begin{aligned}
\hat{Y}_k &= \hat{\beta}_0(t_k)\{1+\exp(\gamma)z_k\} \\
&= (1+\exp(\gamma)z_k, 0)\hat{\beta}(t_k) \\
&= (1+\exp(\gamma)z_k, 0)\{\mathbf{X}_2(t_k)^T W(t_k)\mathbf{X}_2(t_k)\}^{-1} \\
&\quad \begin{pmatrix} 1+\exp(\gamma_1)z_1 & \cdots & 1+\exp(\gamma_n)z_n \\ \{1+\exp(\gamma_1)z_1\}(t_1-t_k) & \cdots & \{1+\exp(\gamma_n)z_n\}(t_n-t_k) \end{pmatrix} \\
&\quad \begin{pmatrix} K_h(t_1-t_k)Y_1 \\ \vdots \\ K_h(t_n-t_k)Y_n \end{pmatrix} \\
&= (1+\exp(\gamma)z_k, 0)\{\mathbf{X}_2(t_k)^T W(t_k)\mathbf{X}_2(t_k)\}^{-1} \\
&\quad \begin{pmatrix} \sum_{j=1}^n Y_j K_h(t_j-t_k)\{1+\exp(\gamma)z_j\} \\ \sum_{j=1}^n Y_j K_h(t_j-t_k)\{1+\exp(\gamma)z_j\}(t_j-t_k) \end{pmatrix} \\
&= \sum_{j=1}^n Y_j K_h(t_j-t_k)(1+\exp(\gamma)z_k, 0)\{\mathbf{X}_2(t_k)^T W(t_k)\mathbf{X}_2(t_k)\}^{-1} \\
&\quad \begin{pmatrix} 1+\exp(\gamma)z_j \\ \{1+\exp(\gamma)z_j\}(t_j-t_k) \end{pmatrix}
\end{aligned} \tag{C.3}$$

Write

$$\hat{Y} = (\hat{Y}_1, \cdots, \hat{Y}_n)^T = \mathbf{H}Y,$$

where $\mathbf{H}$ is the hat matrix. By (C.3), the kth diagonal element of $\mathbf{H}$ is

$$H_{kk} = K_h(0)(1 + \exp(\gamma)z_k, 0)\{\mathbf{X}_2(t_k)^T W(t_k)\mathbf{X}_2(t_k)\}^{-1}(1 + \exp(\gamma)z_k, 0)^T.$$

The trace of the hat matrix is $\mathrm{tr}\{\mathbf{H}\} = \sum_{k=1}^{n} H_{kk}$.

Since τ, ϕ and γ are unknown, the number of parameters will be $p = \mathrm{tr}\{\mathbf{H}\} + 3$, and the degree of freedom for RSS under model (4.3) is

$$n - p = n - \sum_{k=1}^{n} H_{kk} - 3.$$

VITA

Yingxue Liu received a Bachelor of Science degree in applied mathematics from Peking University in Beijing, China in July 2002, and a Master of Science degree in statistics from Texas A&M University in College Station, Texas, under the direction of Dr. Michael Longnecker and Dr. F. Michael Speed in May 2004. She continued her studies under the direction of Dr. Naisyin Wang, and received a Doctor of Philosophy degree in statistics from Texas A&M University in August 2007. Her permanent address is:

11-1-2 Aojing Yuan, Huale Jie, Zhongshan District

Dalian, Liaoning, China, 116000.