Penalized Regression Methods in Time Series and Functional Data Analysis

by

Li Zhang

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

In this thesis, we study penalized methods in time series and functional data analysis.

In the first part, we introduce regularized periodograms for spectral analysis of unevenly spaced time series. The regularized periodograms, called regularized least squares periodogram and regularized quantile periodogram, are derived from trigonometric least squares and quantile regression with Group Lasso penalty. A simple model provides a theoretical justification for the use of regularized least squares periodogram as a tool for detecting hidden frequency in the time series. We give a data-dependent procedure for selection of the regularization parameter. An extensive simulation studies are conducted to examine whether our periodogram functions have the power to detect frequencies from the unevenly spaced time series with big gaps and outliers.

In the second part, we propose a penalized likelihood approach for the estimation of the spectral density of a stationary time series. The approach involves L_1 penalties, which were shown to be an attractive regularization device for nonparametric regression, image reconstruction, and model selection. We explore the use of penalties based on the total variation of the estimated derivatives of spectral density. An asymptotic analysis of the integrated absolute error between the estimator and the true spectral density is presented and gives a consistency result under certain regularity conditions. We also investigate the convergence of the total variation penalized Whittle likelihood estimator to the true spectral density via simulations.

In the third part, we treat discrete time series data have as functional covariates in functional regression models with a scalar response. We develop an efficient

wavelet-based regularized linear quantile regression framework for coefficient estimation in the functional linear model. The coefficient functions are sparsely represented in the wavelet domain, and we suppose that only a few of them are linked to the response. Subsequently, we derive an estimator for the slope functions through composite quantile regression and sparse Group Lasso penalty. We also establish the rate of convergence of the estimator under mild conditions, showing that this rate is dependent on both the sample size and the number of observed discrete points for the predictive functions. Finally, we conduct a simulation study to figure whether our method can identify relevant functional variables.

We illustrate the empirical performance of all the proposed methods on several real data examples.

Acknowledgements

My deepest and sincere gratitude firstly goes to my supervisor, Dr. Ivan Mizera for his constant encouragement and advice throughout my study. Without his immense knowledge and patience in correcting my writing, this thesis could not be completed.

I would like to express my sincere gratitude to Dr. Linglong Kong for enlightening me and insightful comments on my research. I am also taking this opportunity to thank Dr. Dana Cobzas for her support of my research and offering me the research assistantship for the last year, making my pursuit of Ph.D. continue smoothly. In addition, I am also thankful to all professors who taught me a lot in the classes.

My sincere thanks also go to my committee members Dr. Cristina Anton, Dr. Guodong Li and Dr. Douglas Wiens for serving on my dissertation examination committee.

I thank my fellow colleagues and friends: Yongquan Liu, Sile Tao, Wei Tu, Bo Wang, Dengdeng Yu, and many others; from whom I learned new things, and with whom I gained new experiences, and had fun during my study.

My love and gratitude go to my family, and especially my parents and parents-in-law for their continuous support and constant encouragement throughout my study. Special thanks to my wife for believing in me and always being there for me with great patience and love throughout the last six years.

Table of Contents

1 Introduction and Overview of the Thesis **1**

1.1 General Introduction 1

1.1.1 Spectral Analysis 4

1.1.2 Functional Data Analysis 7

1.2 Contributions and Outline of the Thesis 8

2 Regularized Periodograms **11**

2.1 Introduction 11

2.2 Periodogram 16

2.2.1 Classic Periodogram 16

2.2.2 A Linear Model 18

2.2.3 Regularized Periodograms 19

2.3 Orthonormal Design 21

2.4 Practical Implementation 24

2.4.1 Optimization Algorithm 24

2.4.2 Tuning Parameter Selection 25

2.5 Simulation Study 28

2.5.1 Simulation Study for Single-period Model 29

2.5.2 Simulation Study for Two-period Model 31

2.6 Real Data Application 32
2.6.1 Single SPBs: HD 138764 33
2.6.2 Binary SPBs: HD 123515 34

3 Nonparametric Spectral Density Estimation by Total Variation Regularization 38

3.1 Introduction 38
3.2 Methodology Development 40
3.2.1 The Whittle Likelihood and Definition of the Estimator . . . 41
3.2.2 Computation 43
3.3 Asymptotic Approximation 44
3.4 Numerical Illustration 53
3.4.1 Simulation Study 53
3.4.2 Analysis of the Sunspot Dataset 58
3.4.3 Analysis of the Váh River Dataset 59

4 Sparse Wavelet Quantile Regression with Multiple Predictive Curves 61

4.1 Introduction 61
4.2 Wavelet-based Sparse Group LASSO 65
4.2.1 Some Background on Wavelets 65
4.2.2 Model Estimation 67
4.3 Implementations 69
4.3.1 Algorithm 69
4.3.2 Selection of Tuning Parameters 71
4.4 Consistency of the Wavelet-based Group Lasso Estimator 72
4.5 Numerical Studies 77
4.5.1 Simulations 77

4.5.2 Application to Real Data 82
4.6 Appendix 86

5 Conclusion **94**

List of Figures

2.1 (a) Light curve of HD 123515. (b) Zoom on the rightmost light curve of HD 123515 from Julian Day 50544 to 50579 13

2.2 Threshold function $g_{th}(z)$. 24

2.3 Each curve is the stability path that represents the selected probability for a fixed frequency corresponding to the 50 different $\lambda's$. The red line is the stability path for frequency $f_1 = 0.7944$ and indicates that there might be one frequency in the Geneva data HD 138764. . 35

2.4 Estimated stability path for HD 123515 using RQP. Each curve is the stability path that represents the selected probability for a fixed frequency corresponding to the 60 different $\lambda's$. (a) Red lines are the stability paths of the selected frequencies selected by setting $\pi_{thr} = 0.7$. The interval of interesting frequencies is [0, 1]. (b) As for (a) but the interesting frequencies belong to the interval [0.5, 0.75]. 37

3.1 Spectral density functions of Model (M1)–(M4). 55

3.2 The panels show boxplots of the the L_1-errors from the 1,000 Monte-Carlo Replicates. The Panel (a) shows the result for AR(2) process, (b) for AR(12), (c) for AR(4) and (d) for ARMA(2,2). The letter beneath the boxplots indicate the method; T stands for PTVE, introduced in this chapter, P, S, A, indicate respectively PLE, SPE and ARE methods. 57

3.3 All four spectral estimates of the square root transformed Sunspot Data. 58

3.4 The yearly discharge time series of the Váh, period 1931-2002. . . . 60

3.5 All four spectral estimates of the Váh Data. The letter P stands for periodogram. The PTVE and periodogram reveal a peak at $0.27778 \times 2\pi$. The PLE, ARE and SPE indicate a peak at $\omega = 0.27778 \times 2\pi$, $0.26389 \times 2\pi$ and $0.26389 \times 2\pi$, respectively. 60

4.1 Slope fucntions β_1–β_4. 80

4.2 Boxplot of L_2 norm for each slope function, by using the quantile spare Group Lasso method. 87

4.3 Boxplot of L_2 norm for each slope function, by using the quantile Lasso method. 92

4.4 Boxplot of L_2 norm for each slope function, by using the quantile Group Lasso method. 93

List of Tables

2.1 This table shows the detection rate of the 4 Methods from the 200 Monte Carlo replicates for the model M1 and M2 with different error distributions. 31

2.2 This table shows the detection rate of four Methods from the 200 Monte Carlo replicates for the model M3 with different error distributions. 33

2.3 Comparison of frequencies estimation result for the HD123515. There are just three frequencies detected by Hall and Li [33]. 36

3.1 Mean L_1-error from the 1,000 Monte Carlo Replicates. 56

4.1 Simulation summary of SNR=5. The first column n is the size of training data. The second column is the type of noise. The third column is the method we used, Q for the quantile sparse Group Lasso, L for the quantile Lasso, and G for the quantile Group Lasso. GS means λ was selected by the validation method (gold standard). GIC means λ selected via the GIC criterion. MISE stands for mean integrated errors. PE, GA and VA indicate prediction error, group accuracy and variable accuracy, respectively. 83

4.2 Individual funtional L_2 error of SNR=5. The first column n is the size of training data. The second column is the type of noise. The third column is the method we used. ISE1: $\|\hat{\beta}_1 - \beta_1\|_2^2$; ISE2: $\|\hat{\beta}_2 - \beta_2\|_2^2$; ISE3: $\|\hat{\beta}_3 - \beta_3\|_2^2$; ISE4: $\|\hat{\beta}_4 - \beta_4\|_2^2$. 84

4.3 Selected ROIs for the suggestion 7 regions, R and L indicate the region has been selected from the right brain and left brain, respectively. The symbol × means the brain region has not been chosen. . 86

4.4 Selected ROIs for the ADHD-200 fMRI Dataset. 87

4.5 Simulation summary of SNR=1, as for Table 4.1. 88

4.6 Individual funtional L_2 error when SNR=1, as for Table 4.2. 89

4.7 Simulation summary of SNR=10, as for Table 4.1. 90

4.8 Individual funtional L_2 error when SNR=10, as for Table 4.2. 91

Chapter 1

Introduction and Overview of the Thesis

1.1 General Introduction

Penalized regression methods for statistical inference received lots of attention in recent years. In the background of regularization, penalization was devised by Tikhonov [77] for approximating the solution of a set of unsolvable integral equations. Similar concept is the basis for the ridge regression, which was formally introduced by Hoerl and Kennard [37] about 40 years later. Penalized maximum likelihood methods for density estimation were introduced by Good [31], who suggested using Fisher information for the location parameter of the density as a penalty functional. In 1996, Tibshirani [76] proposed the Lasso technique that involves minimizing the residual sum of squares, subject to a constraint on the sum of absolute value of the regression coefficients. A similar formulation was proposed by Chen, Donoho, and Saunders [15] under the name basis pursuit, for denoising using overcomplete dictionaries. The emergence of the least angle regression

(LARS) algorithm of Efron, Hastie, Johnstone, *et al.* [25] provided an efficient solution to the optimization problem underlying the Lasso. After that, penalization techniques gained significant impetus in statistics, especially in the applications; an enormous amount of work in statistics is dealing with penalization in a broad spectrum of problems. Comprehensive reviews are Bickel and Li [5], Hesterberg, Choi, Meier, *et al.* [36], Fan and Lv [26], Vidaurre, Bielza, and Larrañaga [78], and the references therein.

In high dimensional data analysis, penalized likelihood methods have been extensively applied for the simultaneous selection of important variables and estimation of their effects. In the situation when the design matrix $\boldsymbol{X}$ and the response vector $\boldsymbol{Y}$ are known, the penalized likelihood has a generalized formulation

$$n^{-1}\ell(\boldsymbol{\beta}) - p_\lambda(\boldsymbol{\beta}), \tag{1.1}$$

where $\ell(\boldsymbol{\beta})$ is the log-likelihood function and $p_\lambda(\boldsymbol{\beta})$ is a penalty function indexed by the nonnegative regularization parameter λ. For example, we consider the linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then the Lasso estimation can be defined as

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \tag{1.2}$$

where $\|\cdot\|_2$ and $\|\cdot\|_1$ denote the L_2 norm and L_1 norm, respectively. The major advantage of the lasso is that it offers interpretable, stable models and efficient prediction at a reasonable cost.

For function estimation, the roughness penalties are imposed to obtain a general smoothing estimator. A traditional measure of roughness of a function is by its integrated squared second derivative. For instance, we estimate a curve g from

observations $y_j = g(t_j) + \varepsilon_j$ by minimizing the penalized residual sum of squared

$$\min_g \Big\{ \sum_{j=1}(y_j - g(t_j))^2 + \lambda \|g^{(2)}\|_2^2 \Big\}.$$

The penalized method can be viewed as a way of quantifying the conflict between smoothness and goodness-of-fit to the data, since the term of the residual sum of squares measures how well g fits the data.

There is an enormous amount effort to explore alternatives form of both the fidelity to the data and penalty functions to achieve modified objectives. For linear regression, the function $\ell(\boldsymbol{\beta})$ in (1.1) can be a quasi-likelihood or a loss function, such as least absolute deviation, quantile regression, composite quantile regression. For instance, the β can be estimated by the quantile regression

$$\sum_{i=1}^{n} \rho_\tau(y_i - b_\tau - \boldsymbol{x}_i^T \boldsymbol{\beta}), \tag{1.3}$$

where $\rho_\tau(x) = \tau x_+ + (1-\tau)x_-$ and b_τ is an additive constant. For grouped variables, Yuan and Liu [87] proposed a generalized Lasso, called the Group Lasso, to do variable selection at the group level. Simon, Friedman, Hastie, *et al.* [72] introduced a regularized model for linear regression with L_1 and L_2 penalties for the problems with grouped covariates, which are believed to have sparse effects both on a group and within group level. To estimate the conditional quantile function, Koenker, Ng, and Portnoy [44] suggested a nonparametric approach based on minimizing total variation penalties with quantile fidelity to the data.

1.1.1 Spectral Analysis

Time series analysis, aimed at extracting meaningful statistics and other characteristics of the data, is widely used in many areas, such as science, economics, medicine, and others. An example of the discipline where the search for frequencies from unevenly spaced time series is an important topic, is astroparticle physics. In the field of neuroscience, functional magnetic resonance imaging of brain-wave time series may be used to study the differences in patterns of brain activation between cases and controls in attention deficit hyperactivity disorder (ADHD) research; see Paloyelis, Mehta, Kuntsi, *et al.* [61].

In general, there are two rather distinct approaches to time series analysis: the frequency domain approach and the time domain approach. The simplicity of visualization of periodicities is the distinct advantage of the frequency approach, whereas the time domain approach focuses on modeling some future value of a time series as a parametric function of the current and past value. In this thesis, we concentrate on the first approach, frequency domain analysis. More specifically, our primary interest is related to periodic and/or systematic sinusoidal variations.

In the frequency domain, the partition of the various kinds of periodic variation in a time series is accomplished by evaluating separately the variance associated with each periodicity of interest. The variance profile over frequency is called the spectral density. Let X_t be a real value stationary time series with mean μ and autocovariance $C_{xx}(k) = E[(X_{t+k} - \mu)(X_t - \mu)]$. By the Wiener-Khintchine theorem, there exists a monotonically increasing function $F(\omega)$ in the frequency domain such that

$$C_{xx}(k) = \int_{-\pi}^{\pi} e^{ik\omega} dF(\omega), \tag{1.4}$$

where the integral is a Riemann-Stieltjes integral, the function F is the spectral distribution function and i is the imaginary unit. From Lebesgue's decomposition theorem, the spectral distribution function can be split into three components, discrete, continuous, and singular. In our applications, we neglect—as is quite common in the literature—the singular part, and assume that the spectral distribution consists of a discrete and a continuous component. Using the Wold decomposition, every stationary process can be represented as

$$X_t = X_t^{(d)} + X_t^{(n)},$$

where $X_t^{(d)}$ and $X_t^{(n)}$ are purely deterministic and purely nondeterministic, respectively, corresponding to the decomposition of a spectral distribution to the discrete and remaining parts. In Chapter 2, we consider the $C_{xx}(k)$ is a sum of sinusoids; that implies that the spectral distribution function is discrete. Technically, the results of Chapter 2 are thus applicable only to discrete spectral distributions, to time series with "point spectrum". In Chapter 3, the autocovariance function, $C_{xx}(k)$, of a stationary process is absolutely summable; that indicates that F is continuous with the spectral density function f as

$$f(\omega) = \frac{dF(\omega)}{d(\omega)};\ 0 < \omega < \pi. \tag{1.5}$$

This function is also known as the power spectral function or the spectrum. The results of Chapter 3 thus fully apply only to time series with continuous spectrum, with the spectral distribution possessing a density. The autocovariance function

then has the representation

$$C_{xx}(k) = \int_{-\pi}^{\pi} e^{ik\omega} f(\omega) d\omega,$$

as the inverse transform of the spectral density, which in turn has the representation

$$f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} C_{xx}(k) e^{-ik\omega} \quad -\pi \leq \omega \leq \pi. \tag{1.6}$$

Let $X_1, \ldots, X_n$ be a finite sample of size n. The sample autocovariance estimator is then defined by $\widehat{C}_{xx}(k) = \frac{1}{n} \sum_{l=1}^{n-k} (X_l - \mu)(X_{l+k} - \mu)$. To estimate the spectral density, the natural step is to plug in the sample autocovariance function in place of an unknown autocovariance function in the formula (1.6). The resulting estimator (up to the factor $1/2\pi$) is called the periodogram

$$I_n(\omega) = \sum_{k=-n}^{n} \widehat{C}_{xx}(k) e^{-ik\omega} = n^{-1} \left| \sum_{j=1}^{n} X_j e^{-i\omega j} \right|^2 \quad -\pi \leq \omega \leq \pi. \tag{1.7}$$

In practice, the periodogram is a very useful tool for describing a time series data set. If a time series has a strong sinusoidal signal for some frequency, then there will be a peak in the periodogram at that frequency. The formula (1.7) also shows why the periodogram is such a useful tool for searching periodicities: indeed, a peak in $I_n(\omega)$ at the frequency $\omega = \omega^*$ indicates a possible periodic phenomenon with the period $p^* = 2\pi/\omega^*$. Note that the periodogram is customarily calculated at the Fourier frequencies $\omega = 2\pi k/n$ for some integer $0 < k \leq \frac{n}{2}$. For more background on time series analysis and its applications, refer to the monograph of Shumway and Stoffer [69].

1.1.2 Functional Data Analysis

Functional data analysis (FDA) is about the analysis of information on curves, images, functions, or more general objects. It has become a major branch of non-parametric statistic and is a fast evolving area as more data has arisen where the observation can be viewed as a function. One of the standard functional linear models relates functional covariates to a scalar response via

$$y = \beta_0 + \int_I x(t)\beta(t)dt + \varepsilon, \tag{1.8}$$

which has been studied extensively, see Wang, Chiou, and Müller [82], Morris [56].

A common method involves representing the covariate $x(t)$ and the coefficient function $\beta(t)$ by a linear combination of known functional basis. Specifically, consider an orthonormal basis ϕ_k, $k \geq 1$, of the function space. Expanding both $x(t)$ and $\beta(t)$ in this basis leads to $x(t) = \sum_{k=1}^{\infty} a_k\phi_k$ and $\beta(t) = \sum_{k=1}^{\infty} \beta_k\phi_k$. Model (1.8) is seen to be equivalent to the linear model of the form

$$y = \beta_0 + \sum_{k=1}^{\infty} \beta_k a_k + \varepsilon^*, \tag{1.9}$$

where in implementation the sum on the right-hand side of the above equation is replaced by a finite sum that is truncated at the first K term, and ε^* is the error term from the truncation and noise.

The simple functional liner model (1.8) can be extended to multiple functional covariates $x_1(t), \ldots, x_m(t)$, also including additional vector covariates $\boldsymbol{u} = (u_1, \ldots, u_q)$, by

$$y = \beta_0 + \sum_{l=1}^{m} \int_{I_l} x_l(t)\beta_l(t)dt + \boldsymbol{u}^T\boldsymbol{\gamma} + \varepsilon, \tag{1.10}$$

where the interval I_l is the domain of $x_l(t)$, ε is the error term. Hereafter, we set $I_j = [0,\ 1]$ to simplify our model. We can expand the coefficient functions $\beta_l(t)$ in term of suitable bases and estimate $\beta_l(t)$ and $\boldsymbol{\gamma}$ simultaneously. Full details of this model are discussed in the context of the particular application in Chapter 4. For more other linear models, see the monograph of Ramsay [64].

1.2 Contributions and Outline of the Thesis

The thesis is arranged into five chapters. The first chapter presents a brief introduction to the penalized methods, the spectral density, and functional regression models.

Many unevenly spaced time series coming from the natural sciences exhibit periodic structures where the series repeat approximately the same pattern over time. Discovering their period and repetitive behavior they exhibit is an important task toward understanding their characteristics. Due to the importance of detecting frequency from the unevenly spaced time series, Chapter 2 makes a contribution to this challenging problem. We devote to developing effective procedures to estimate the frequency from the unevenly spaced time series. We construct two periodogram-like functions, through penalized trigonometric least squares and quantile regression. To explain how the method works, we prove Theorem 2.3.1 to show the effect of the proposed regularization under the orthonormal design. We adapt the stability selection method of Meinshausen and Bühlmann [54] to select the regularization parameter. The application of regularized periodograms significantly improves the detecting rate. Simulation studies with different sample size and noise type are utilized to validate the effectiveness of the proposed two periodograms. The capability of frequency detecting from the unevenly spaced time series data with big gaps is

demonstrated on real data.

Penalized likelihood methods have been applied successfully in nonparametric function estimation and variable selection problems. For instance, L_1 penalty usually enriches the models with variable selection and a reasonable bias-variance trade-off. A number of penalties based on the L_1 penalty have been proposed for adaptation to specific types of problems or improvement of the statistical properties. In Chapter 3, we employ the total variation penalty to identify a local maximum in the spectral density while still maintaining the desired smoothness. Then, logarithmic spectral density estimation is obtained as a result of minimizing our objective function that combines the Whittle likelihood approximation and total variation penalty. We also prove Theorem 3.3.6 to describe the asymptotic rate of convergence of the estimator with the tuning parameter given in Lemma 3.3.5. Finally, this methodology is illustrated with simulated and real data sets. A simulation study with autoregressive and moving average process provides the quantification of the performance characteristic of the approach, in comparison to some well-established methods; the latter include the L_2 penalized likelihood method of Pawitan and O'Sullivan [62], smoothed periodogram, and autoregressive spectral density estimators. The Váh River example shows that our nonparametric spectral density estimator appears to have some distinct advantage to capture local features.

In Chapter 4, we consider the variable selection problem in functional linear quantile regression. The covariates are given as both functional and scalar types while the response is scalar, and the conditional quantile for each fixed quantile index is modeled as a linear function of the covariates. Our approach to estimating the coefficient functions is to use wavelet basis functions expansion. The advantages with respect to the use of wavelet bases are that the resulting representation is sparsity with a few non-zero coefficients and that it is trivial to extend our approach to

higher dimensional predictors. In order to select functional variables each of which is controlled by multiple wavelet coefficients, we treat these coefficients as grouped parameters and then apply the group Lasso penalty. Since the coefficient function is sparsity expressed in the wavelet domain, we employ the Lasso to select a relatively small number of nonzero wavelet coefficients. On the other hand, the Lasso penalty imposes smoothness of the coefficient functions. We reformulate the final optimization problem as the standard second order cone program and solve it by the interior point methods. We also prove Theorem 4.4.1 that gives an asymptotic error bound, the bound on the difference between our estimator and true functional coefficients. This bound is explicitly specified in terms of the sample size, the number of observed discrete point for the predictive functions and the smoothness of slope functions. We give the finite sample performance of our procedure via a simulation study to demonstrate that our estimator has the desired effect of group-wise and within group sparsity. We also analyze the ADHD-200 fMRI dataset with 59 covariates. Our aim of the analysis is to select important brain regions that are related to the ADHD index. The real data results confirm the effectiveness of the proposed method and have been supported by other independent and different functional neuroimaging studies.

We end the thesis with conclusions in Chapter 5.

Chapter 2

Regularized Periodograms

We propose the use of regularized least squares periodogram and regularized quantile periodogram to detect frequency from unevenly spaced time series. To explain how the method works, we prove Theorem 2.3.1 to show the effect of the proposed regularization under an idealized setting. The superiority of regularized periodograms in handing big gaps and outliers in the time series is supported by simulations. Two real-data examples are analyzed to illustrate the empirical performance of the proposed procedure.

2.1 Introduction

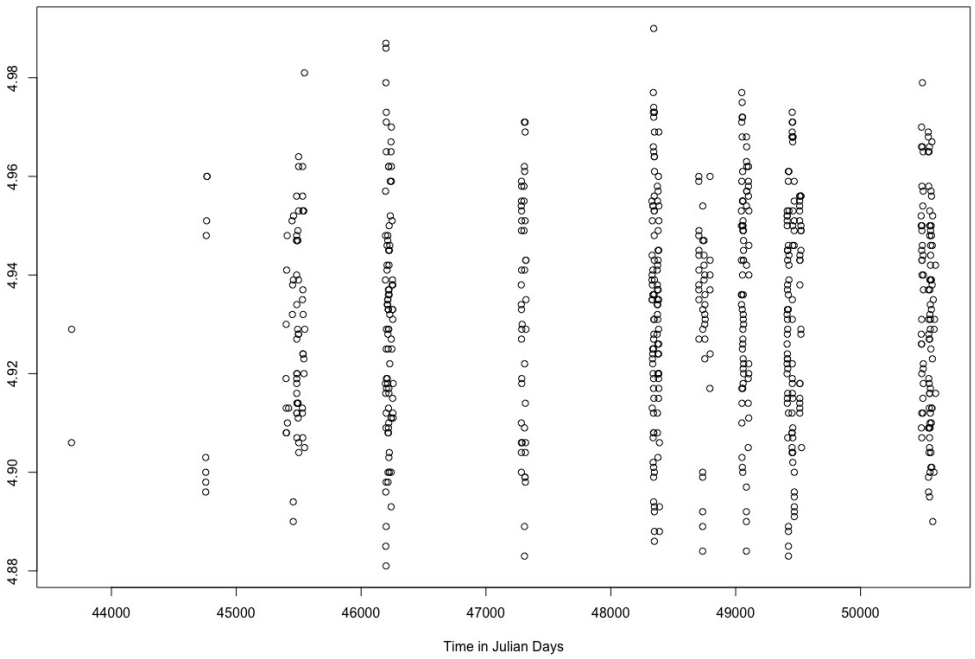
Detection and estimation of periodic patterns from unevenly sampled time series is a frequent problem in various fields of study, like genetics [1], seismology [4], biological rhythm research [65], hematology [28], paleoclimatology [57], and astroparticle physics[52], [66], [75]. In all these articles, periodicity detection in unevenly spaced time series is a common work. The analysis of unevenly spaced time series is an important task in science. Although there is an extensive theory for the

analysis of equally spaced data, very little theory exists specifically for unevenly spaced time series.

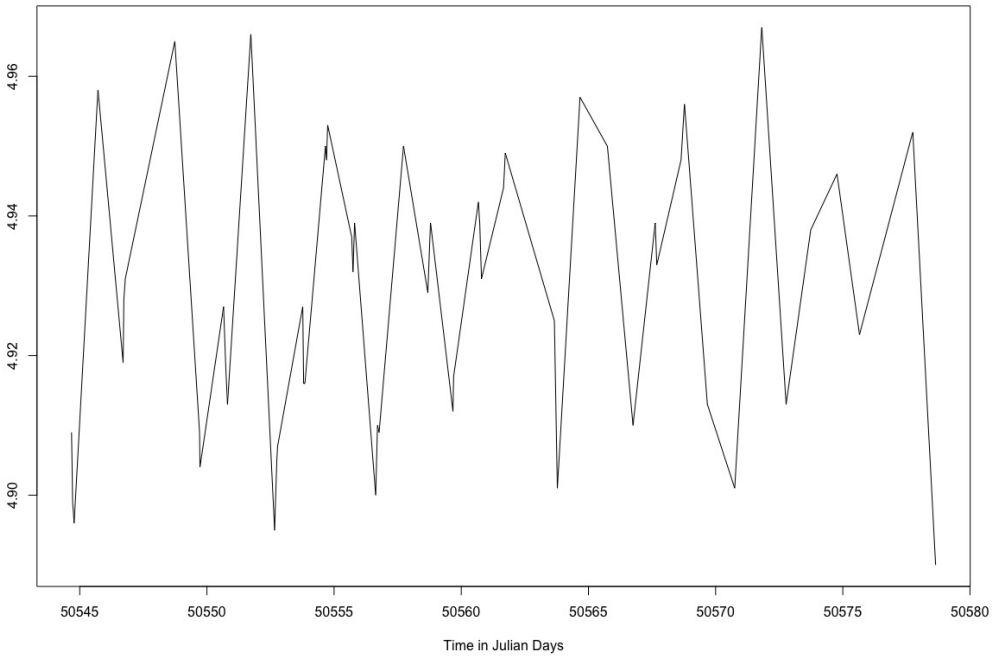
In particular, astronomical data generally suffer from incomplete sampling. A number of factors influence the sampling of time series. For instance, the day-night alternation or bad meteorological conditions may generate gaps in the time series. Long-time observations are generally unevenly sampled, because of the telescope schedule and other reasons. As an example we discuss the light curve depicted in Figure (2.1), showing the observations of the radiation of slowly pulsating B-star HD 123515, see De Cat and Aerts [21], Hall and Yin [34]. Besides of the irregularities of the light curve, a certain additional correlated noise and outliers also are typical for light curves in astroparticle physics.

Classical methods based on the periodogram are widely used by astronomers and statisticians for unevenly spaced data; they can be traced to Lomb [52], Scargle [66], Chen [16]. Recently, instances of its applications are to be found in De Cat and Aerts [21], Hall and Li [33], Lévy-Leduc, Moulines, and Roueff [47]. They are based on fitting a sinusoid of different trial periods to the unevenly sampled time series using least squares regression and taking the respective goodness-of-fit criterion as periodogram threshold. For multi-sine fitting, prewhitening method such as CLEANEST algorithm of [29], widely used in astrophysics, removes the peaks in the periodogram as single frequency components. However, these methods may fail in some cases as pointed out by Bourguignon, Carfantan, and Böhm [6], for example, the false peaks.

Compared to the least squares method, least absolute deviations is robust in that it is resistant to outliers in the data. Li [48] employed the least absolute deviations in the field of time series analysis. He derived a new type of periodogram, called the Laplace periodogram, by changing least absolute deviations from least squares



(a)



(b)

Figure 2.1: (a) Light curve of HD 123515. (b) Zoom on the rightmost light curve of HD 123515 from Julian Day 50544 to 50579

in the harmonic regression procedure that produces the ordinary periodogram of an evenly spaced time series. Moreover, the classical problem of estimating the frequency of a sinusoidal signal from noisy observations where the noise has a heavy-tailed distribution was considered. His simulations showed that the Laplace periodogram provides an efficient alternative to the ordinary periodogram and gives a more accurate estimator under the noise setting. Quantile periodograms are constructed from trigonometric quantile regression in Li [49], which can be viewed as generalizations of the Laplace periodogram. Li [49] gave numerical and theoretical results to demonstrate the capability of the quantile periodograms for detecting hidden periodicity in the quantiles. Inspired by the Laplace periodogram and the success of quantile periodograms in the frequency estimation from evenly sampled time series, we apply not only the least squares but also the quantile regression technique as the data fitting criterion to detect frequencies from unevenly sampled time series.

We jointly estimate periodogram for all our interesting frequencies as an alternative to the usual harmonic regression procedure. Following recent works in the last decade [15], [7], [55], we regard the problem of detecting frequency as an underdetermined inverse problem where the spectrum is discretized in an arbitrarily thin frequency grid. Chen, Donoho, and Saunders [15] developed a principle, called Basis Pursuit, for decomposing a signal into an optimal superposition of dictionary elements, where optimal means having the smallest L_1 norm of coefficients among all such decompositions. An illustrative example of frequency detection by using the Lasso penalty was discussed in Meinshausen and Yu [55]. The results highlight that the true frequencies are with high probability picked up by the Lasso and coefficients of the true frequencies are much larger than the coefficient of the resonance frequency with an appropriate choice of the penalty parameter. Meinshausen

and Yu [55] also suggested that we can employ the Group Lasso, grouping together the sine and cosine part of identical frequencies. The Group Lasso as an extension of the Lasso to do the variable selection on (predefines) groups of variables in regression models has become a popular model selection and shrinkage estimation method in many applications [87]. The attractive property of Group Lasso is enforcing the structural sparsity. In our problem, the explanatory factor frequencies are represented by the two input variables, the sine and cosine part of the frequencies. Therefore, we adapt the Basis Pursuit principle where the optimal means having the smallest L_2 norm, the Group Lasso penalty, of coefficients. This results in two periodogram-like functions, called the regularized least squares periodogram and regularized quantile periodogram by minimizing least squares and quantile loss function with the Group Lasso penalty, respectively.

To detecting frequencies from the unevenly sampled time series, we derived the regularized least squares periodogram and regularized quantile periodogram. A variant of these periodograms is developed in some details, with particular emphasis on the optimization algorithm and on the selection of tuning parameters. Simulation studies are conducted to measure the performance of these periodograms relative to some well-established techniques such as adjusted periodogram in Hall and Li [33], Laplace periodogram in Thieler, Backes, Fried, *et al.* [75] and Li [48]. In addition, we apply these periodograms for detecting frequencies from the real light curve HD 123515 which was also considered in Hall and Li [33]. It is shown that our periodograms can identify the fourth important frequency while the adjusted periodogram cannot.

The rest of this chapter is organized as follows. In Section 2.2, a unified concept from which we construct the methods is generalized based on review classical periodogram methods to detect frequencies. A simple example is given in Section

2.3 to show that the effect of Group Lasso is equivalent to a shrinkage function. The practical implementation, including optimization strategy and tuning parameter selection, is presented in Section 2.4. Section 2.5 provides some simulation studies to compare our regularized periodogram techniques with some other existing methods. To illustrate the power of the new methodology, more comprehensive analyses of the light curve observed for HD 18764 and HD 123515 are given in Section 2.6.

2.2 Periodogram

Most of the popular periodogram methods are based on fitting a model to the folded time series using least squares regression. This section considers the different criterion to fit the observation data and defines the regularized periodogram.

2.2.1 Classic Periodogram

Spectral analysis is a critical approach to describe the fluctuation of time series in term of sinusoidal behavior at various frequencies. It is well known that the periodogram plays a major role in estimating the spectral density. The value of the periodogram for a given frequency ω are obtained via fitting a sinusoidal with that frequency to the analyzed series using the fitted model

$$\alpha(\omega)\cos(\omega t) + \beta(\omega)\sin(\omega t), \tag{2.1}$$

also with the added intercept, if the series is not centered (or considered centered) about zero. For simplicity, we assume that the series is centered.

Given observed time series data $\{y(t_j)\}_{j=1}^n$, we denote $\widehat{\beta}_n(\omega)$ to be the least

squares regression solution

$$\widehat{\boldsymbol{\beta}}_n(\omega) := \arg\min_{\beta \in R^2} \|\boldsymbol{Y} - \boldsymbol{X}(\omega)\beta\|_2^2, \tag{2.2}$$

where $\boldsymbol{Y} = (y(t_1), \ldots, y(t_n))^T$ is the $n \times 1$ vector, $\boldsymbol{X}(\omega)$ is the regression matrix corresponding to the model (2.1). If the time series data $\{y(t_j)\}_{j=1}^n$ is observed at equidistant points $t_k = k\Delta$, where Δ is the sample period, then the ordinary periodogram can be written as

$$I_n(\omega) = \frac{n}{4}\|\widehat{\boldsymbol{\beta}}_n(\omega)\|_2^2, \tag{2.3}$$

at any Fourier frequency; this is the same as the ordinary periodogram defined by the discrete Fourier transform as the equation (1.7). Note that this formula also can be rewritten as $I_n(\omega) = \frac{1}{n}\|\hat{\boldsymbol{y}}(\omega)\|_2^2$ with $\widehat{\boldsymbol{y}}_n(\omega) = \boldsymbol{X}(\omega)\widehat{\boldsymbol{\beta}}_n(\omega)$.

The definition (2.3) given above applies to any time series: not only to equidistantly sampled as above but also possible non-evenly sampled at arbitrary where are ordered observation times $\{t_j\}_{j=1}^n$. Consider the following least squares fitting problem:

$$\widehat{\boldsymbol{\beta}}_n(\omega) := \arg\min_{\boldsymbol{\beta} \in R^2} \|\boldsymbol{Y} - \boldsymbol{A}(\omega)\boldsymbol{\beta}\|_2^2, \tag{2.4}$$

where the jth row of the matrix $\boldsymbol{A}(\omega)$ is $\boldsymbol{A}_j(\omega) = [\cos(\omega t_j - \tilde{\phi}),\ \sin(\omega t_j - \tilde{\phi})]$ and the $\tilde{\phi}$ is subject to the constraint: $\sum_{j=1}^n \sin(\omega(t_j - \tilde{\phi}))\cos(\omega(t_j - \tilde{\phi})) = 0$. Then so-called Lomb-Scargle periodogram that defined by Lomb [52] and Scargle [66]

$$P_n(\omega) = \frac{1}{2}\left\{\frac{[\sum_{j=1}^n y_j \cos\omega(t_j - \tilde{\phi})]^2}{\sum_{j=1}^n \cos^2\omega(t_j - \tilde{\phi})} + \frac{[\sum_{j=1}^n y_j \sin\omega(t_j - \tilde{\phi})]^2}{\sum_{j=1}^n \sin^2\omega(t_j - \tilde{\phi})}\right\}, \tag{2.5}$$

can be rewritten as $P_n(\omega) = \frac{1}{n}\widehat{\boldsymbol{\beta}}_n(\omega)^T(\boldsymbol{A}^T\boldsymbol{A})\widehat{\boldsymbol{\beta}}_n(\omega) = \frac{1}{n}\|\widehat{\boldsymbol{y}}(\omega)\|_2^2$, where ω could be any real number.

The Lomb-Scargle periodogram provides an analytic solution and is, therefore, both convenient to be used and efficient. Moreover, for evenly sampled data and Fourier frequencies, the Lomb-Scargle periodogram (2.5) coincides with the ordinary one, as defined by (2.3). Hence, this periodogram is a common and useful tool in the frequency analysis of unevenly spaced data. However, due to the Lomb-Scargle periodogram equivalent to least squares fitting of a sine wave, some robust methods can be helpful if the observation data contain outliers. Schimmel [67] provided some examples to illustrate that non-sinusoidal signals or outlying large amplitude features can make more challenging the interpretation of Lomb-Scargle periodogram and eventually lead to a misleading conclusion. This motivates us to extend the capability of least squares techniques and then use of robust regression for periodogram.

2.2.2 A Linear Model

As extension of (2.1), we consider the model as linear combination of an arbitrarily large number N of sine waves with frequencies $\omega_k = 2\pi f_{max}\frac{k}{N}$ for $k = 1, \ldots, N$:

$$\sum_{k=1}^{N} \left[\beta_k \cos(\omega_k t_i) + \beta_{N+k} \sin(\omega_k t_i)\right], \tag{2.6}$$

where β_k is unknown coefficient. When $N = [\frac{n}{2}]$ and $f_{\max} = 0.5$, we can obtain ordinary periodogram by the solution from (2.2) with the regression matrix corresponding to the above model. Model (2.6) is linear and consider jointly a high number N of potential frequencies, where we just need to estimate the $(2N)\times 1$ vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{2N})^T$. In the next subsection, we propose two methods to estimate

these parameters, then use them to define our regularized periodograms.

There are two fundamental differences between the model (2.6) and the classic multi-sine fitting approaches. In our model, the frequencies are fixed, and we just need to estimate the coefficients. To detect the hidden periodicity, we also consider a sparse solution, in which the whole spectrum can be estimated jointly.

2.2.3 Regularized Periodograms

The frequency resolution in model (2.6) is intrinsically limited by the discretization step of the frequency grid $\frac{f_{max}}{N}$. To avoid missing frequencies in the true model, we make the mesh of the frequencies becomes finer and finer, for example, N is as large as we need. Thus, the size of unknown coefficients, $2N + 1$, must be vast to yield a resolution comparable to that of classic prewhitening methods, even larger than the amount of data. In this case, an infinite number of unknown coefficients β_k supremely fit the data with model (2.6). Among all possible solutions, our aim is to obtain the sparsest one that has the fewest non-zero elements in the vector $(\sqrt{\beta_1^2 + \beta_{N+1}^2}, \ldots, \sqrt{\beta_N^2 + \beta_{2N}^2})^T$. Given observed time series data $\{y(t_j)\}_{j=1}^n$, we generate the design matrix $\boldsymbol{X}$ corresponding to the model (2.6) and consider the objective function

$$L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}) + \lambda \sum_{i=1}^{N} \sqrt{\beta_i^2 + \beta_{i+N}^2}, \tag{2.7}$$

where $L(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta})$ is a general loss function, the tuning parameter $\lambda > 0$ balances between data fidelity and the sparsity. The typical loss function used for parameter estimation is the least squares loss function. In [6] and [7], a penalized least squares method has been studied. Based on these, we define the regularized least squares

periodogram (RLSP) througth the penalized linear regression solution

$$\widehat{\boldsymbol{\beta}_{n_{LS}}} := \arg\min_{\boldsymbol{\beta},\, \beta_0} n^{-1}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{i=1}^{N} \sqrt{\beta_i^2 + \beta_{i+N}^2}, \tag{2.8}$$

and the value of regularized least squares periodogram can be defined via the solution of the above problem as the form (2.3).

The least squares loss function is very sensitive to outliers in the observation data. Therefore, instead of fitting the models mentioned above by least squares regression, we may apply the quantile regression. Many robust loss functions have been taken into account in [75], such as least absolute deviations, Tukey and Huber loss functions. Li [48] derived the Laplace periodogram by using least absolute deviation and showed an asymptotic distribution of the Laplace periodogram. He also demonstrated that this type periodogram is effectiveness in dealing with contamination data. After that, he generalized the Laplace periodogram to the quantile periodograms by using quantile regression and demonstrated that the quantile periodograms not only share the property of the ordinary periodogram but also offer a much richer view than the one provided by the ordinary periodogram [49]. More specifically, Li [49] showed that the quantile periodogram could detect hidden periodicity in the quantiles. Based on these, we substitute the objective function of quantile regression for the least squares loss function in the model (2.8). The regularized quantile periodogram version is obtained via the penalized linear quantile regression solution

$$\widehat{\boldsymbol{\beta}_{n_Q}} := \arg\min_{\boldsymbol{\beta},\, \beta_0} \sum_{t} \rho_\tau(Y_t - \beta_0 - \boldsymbol{X}_t\boldsymbol{\beta}) + \lambda \sum_{i=1}^{N} \sqrt{\beta_i^2 + \beta_{i+N}^2}. \tag{2.9}$$

The periodogram value can be defined through $\widehat{\beta_{n_Q}}$ as the form (2.3). Denote

$I_{n,RLSP}(\cdot)$ and $I_{n,RQP}(\cdot)$ be the regularized least squares periodogram function and regularized quantile periodogram function, respectively.

2.3 Orthonormal Design

To see clearly the effect of Group Lasso, we consider the particular case of an orthonormal design with $\hat{\Sigma} = n^{-1}X^TX = I_{n\times n}$, where X is a $n \times n$ design matrix. We have the following theorem analogous to Lemma 2.1 in Bühlmann and Van De Geer [10], which gives an explicit solution of the optimization problem (2.8) under this case.

Theorem 2.3.1. *Denote the vector $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)^T$ be the ordinary least squares initial estimator that $\boldsymbol{Z} = n^{-1}(\boldsymbol{X}^T\boldsymbol{Y})$. Then the solution of the problem (2.8) equals*

$$\hat{\beta}_j(\lambda) = Z_j(1 - \lambda/a_j)_+ \quad \textit{and} \quad \hat{\beta}_{j+\frac{n}{2}}(\lambda) = Z_{j+\frac{n}{2}}(1 - \lambda/a_j)_+,$$

where $(u)_+ = \max(u, 0)$ denotes the positive part of u and $a_j = 2\sqrt{Z_j^2 + Z_{j+\frac{n}{2}}^2}$ for $j = 1, \ldots, n/2$.

Proof. To simplify, we denote the objective function by

$$Q_\lambda(\boldsymbol{\beta}) = n^{-1}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_{1,2}.$$

For a minimizer $\hat{\boldsymbol{\beta}}(\lambda)$ of Problem (2.8), it is necessary and sufficient that the subdifferential at $\hat{\boldsymbol{\beta}}(\lambda)$ is zero. If the jth group component $\hat{\beta}_j^2(\lambda) + \hat{\beta}_{j+\frac{n}{2}}^2(\lambda) \neq 0$, then the

ordinary first derivatives at $\hat{\boldsymbol{\beta}}(\lambda)$ have to be zero:

$$\begin{aligned}
\frac{\partial Q_\lambda(\boldsymbol{\beta})}{\partial \beta_j}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\lambda)} &= 2n^{-1}\boldsymbol{X}_j^T(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Y}) + \lambda\frac{\beta_j}{\sqrt{{\beta_j}^2+\beta_{j+\frac{n}{2}}^2}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\lambda)} = 0, \\
\frac{\partial Q_\lambda(\boldsymbol{\beta})}{\partial \beta_{j+\frac{n}{2}}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\lambda)} &= 2n^{-1}\boldsymbol{X}_{j+\frac{n}{2}}^T(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Y}) + \lambda\frac{\beta_{j+\frac{n}{2}}}{\sqrt{{\beta_j}^2+\beta_{j+\frac{n}{2}}^2}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\lambda)} = 0,
\end{aligned}$$

where $\boldsymbol{X}_j$ is the jth column of the regression matrix $\boldsymbol{X}$. Since the design matrix $\boldsymbol{X}$ is an orthogonal matrix, the above two equations can be rewritten as

$$\begin{aligned}
\beta_j(2+\frac{\lambda}{\sqrt{{\beta_j}^2+\beta_{j+\frac{n}{2}}^2}})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\lambda)} &= 2Z_j, \\
\beta_{j+\frac{n}{2}}(2+\frac{\lambda}{\sqrt{{\beta_j}^2+\beta_{j+\frac{n}{2}}^2}})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\lambda)} &= 2Z_{j+\frac{n}{2}}. \qquad (2.10)
\end{aligned}$$

We have either $Z_j \neq 0$ or $Z_{j+\frac{n}{2}} \neq 0$ because of the non-equality constraints ${\hat{\beta}_j}^2(\lambda) + \hat{\beta}_{j+\frac{n}{2}}^2(\lambda) \neq 0$. Suppose that $Z_j \neq 0$. Then the solution of above equations can be expressed in the form of

$$\hat{\beta}_j = \mathrm{sign}(Z_j)(|Z_j| - \frac{\lambda}{2\sqrt{1+r_j^2}}), \quad \hat{\beta}_{j+\frac{n}{2}} = \mathrm{sign}(Z_{j+\frac{n}{2}})(|Z_{j+\frac{n}{2}}| - \frac{\lambda|r_j|}{2\sqrt{1+r_j^2}}), \qquad (2.11)$$

where r_j is the ratio of $Z_{j+\frac{n}{2}}$ to Z_j, such as $r_j = Z_{j+\frac{n}{2}}/Z_j$.

On the other hand, if the jth component satisfies the condition $\hat{\beta}_j^2(\lambda)+\hat{\beta}_{j+\frac{n}{2}}^2(\lambda) = 0$, then the subdifferential at $\hat{\boldsymbol{\beta}}(\lambda)$ has to include the zero element. Therefore, we have two real numbers g_j and $g_{j+\frac{n}{2}}$ with the constraint condition $g_j^2 + g_{j+\frac{n}{2}}^2 \leq 1$ such that

$$G_j(\hat{\beta}(\lambda)) + \lambda g_j = 0 \quad \text{and} \quad G_{j+\frac{n}{2}}(\hat{\beta}(\lambda)) + \lambda g_{j+\frac{n}{2}} = 0,$$

where $G(\beta) = -2n^{-1}X^T(Y - X\beta)$ is the gradient of $n^{-1}\|Y - X\beta\|_2^2$. This fact implies that

$$Z_j^2 + Z_{j+\frac{n}{2}}^2 \leq \lambda^2/4 \quad \text{if } \hat{\beta}_j^{\,2}(\lambda) + \hat{\beta}_{j+\frac{n}{2}}^2(\lambda) = 0.$$

Note that if $\hat{\beta}_j^{\,2}(\lambda) + \hat{\beta}_{j+\frac{n}{2}}^2(\lambda) \neq 0$, we have

$$Z_j^2 + Z_{j+\frac{n}{2}}^2 = \lambda^2/4 + (\beta_j^2 + \beta_{j+\frac{n}{2}}^2) + \lambda\sqrt{\beta_j^2 + \beta_{j+\frac{n}{2}}^2} > \lambda^2/4.$$

We can now demonstrate the conclusion in this theorem by contradiction. If $Z_j^2 + Z_{j+\frac{n}{2}}^2 \leq \lambda^2/4$, then we have $\hat{\beta}_j^{\,2}(\lambda) + \hat{\beta}_{j+\frac{n}{2}}^2(\lambda) = 0$. Otherwise either $\hat{\beta}_j(\lambda) \neq 0$ or $\hat{\beta}_{j+\frac{n}{2}}^2(\lambda) \neq 0$ holds that implies $Z_j^2 + Z_{j+\frac{n}{2}}^2 > \lambda^2$ and leads to a contraction. Analogously, if $Z_j^2 + Z_{j+\frac{n}{2}}^2 > \lambda^2/4$, then the inequality $\hat{\beta}_j^{\,2}(\lambda) + \hat{\beta}_{j+\frac{n}{2}}^2(\lambda) > 0$ holds. The exact form of $\hat{\beta}_j$ in this theorem is coming from the system of equations (2.10) solution. □

Consider an evenly sample time series with $n = 100$ and $\lambda = 0.2$. The regression matrix is defined by the Fourier frequencies. By using Theorem 2.3.1, the value of regularized least squares periodogram can be clearly rewritten as a function of the ordinary periodogram,

$$I_{n,RLSP}(\omega_k) = g_{th}(I_n(\omega_k)),$$

where $g_{th}(z) = \text{sign}(z)(|z| - \sqrt{|z|})_+$ is a thresholding function depicted in Figure 2.2. There, we show that the regularized least squares periodogram involves shrinkage, either to zero or to a value which is smaller than the ordinary periodogram. Thus, regularized least squares periodogram function yields a substantially sparser than the ordinary periodogram in some sense that there are a few frequencies ω_k such that $I_{n,RLSP}(\omega_k) > 0$. Meanwhile, regularized least squares periodogram function

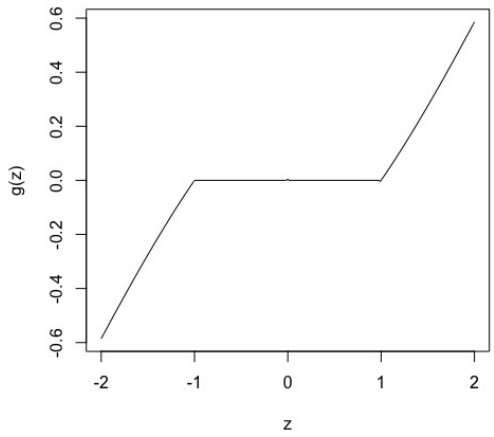


Figure 2.2: Threshold function $g_{th}(z)$.

keeps the major characteristics of the ordinary periodogram

2.4 Practical Implementation

The estimations in optimization problems (2.8) and (2.9) require two essential points to be examined with precision. First, an efficient optimization algorithm is needed in order to solve these two problems. Moreover, a practical rule is necessary for the tuning parameter selection.

2.4.1 Optimization Algorithm

A variety of methods have been proposed for computing $\widehat{\beta_{n_{LS}}}$ in 2.8, including the group LARS approach of Yuan and Liu [87]. Following the recent work in Koenker and Mizera [43], Group Lasso can be implemented as a special instance of second-order cone program (SOCP), minimizing the square of quadratic norm of the residuals while either bounding the penalty or adding it multiplied by a Lagrange multiplier λ to the objective function. These so-called SOCP include linear

constraints and the second-order cone constraints; the objective function is linear, see details in Koenker and Mizera [43]. We utilize the R code in [43] to solve our optimization problem (2.8).

For the quantile Group Lasso problem 2.9, we transform our original optimization problem to the standard form of the SOCP. The following SOCP is equivalent to (2.9):

$$\begin{aligned} \arg\min \quad & \sum_{i=1}^{n} \tau u_i^+ + (1-\tau) u_i^- + \lambda \sum_{j=1}^{N} l_j \\ \text{Subject to} \quad & u_i^- \le Y_{t_i} - \beta_0 - X_t\beta \le u_i^+ \\ & \sqrt{\beta_j^2 + \beta_{j+N}^2} \le l_j \\ & u_i^-, u_i^+, I_j \ge 0 \; for \; i = 1, \; \ldots \; n, \; j = 1, \; \ldots \; N. \end{aligned} \tag{2.12}$$

In the problem (2.12), we introduce new variables, u_i^+ and u_i^- is the positive and negative part of the $Y_{t_i} - \beta_0 - X_t\beta$, respectively. The I_j is an upper bound of $\sqrt{\beta_j^2 + \beta_{j+N}^2}$.

2.4.2 Tuning Parameter Selection

Like any other penalized regression procedure, the performance of the regularized periodograms critically depends on properly tuning parameter λ in (2.7). A general criterion to choose the λ is the k-fold cross validation, which has been widely applied to various regression problem and usually gives a competitive performance. Other common approaches depend on the AIC and BIC scores that trade off the goodness of fit with model complexity. However, we concentrate on the frequency estimation rather than prediction. In fact, we can take the detecting frequency as the structure learning problem, estimation of model structure from data. Recently, Meinshausen and Bühlmann [54] introduced a new method called stability selection

whose goal is to provide an algorithm for performing model selection in a structure learning problem and controlling the number of false discoveries. There are two main advantages over competing approaches: works in high-dimensional data and provides control on the family-wise error rate in the finite sample setting other than an asymptotic guarantee.

In the following, we adapt the stability selection algorithm for our case. Assume that we fix the N and f_{max} in the model (2.6), then we generate our regression matrix $\boldsymbol{X}$ corresponding to this model.

1. Define a candidate set of tuning parameters Λ and a subsample number M

2. For each value of $\lambda \in \Lambda$, do:

   (a) Start with the full data set $(\boldsymbol{Y}, \boldsymbol{X})$.

   (b) For each s in 1,... M, do:

      i. Subsample from the full data set without replacement to generate a smaller dataset, given $(\boldsymbol{Y}_{(s)}, \boldsymbol{X}_{(s)})$.

      ii. Run the problem (2.8) or (2.9) on dataset $(\boldsymbol{Y}_{(s)}, \boldsymbol{X}_{(s)})$ with parameter λ to get the coefficient $\boldsymbol{\beta}$, compute the periodogram at each different frequencies, then we obtain a frequency selection set $\hat{S}_i^{\lambda}$, the detail related to defining the selection set is given later.

   (c) Given the selection sets from the M times subsample, calculate the empirical selection probability for each frequency:

$$\hat{\Pi}_k^{\lambda} = \mathbb{P}(f_k \in \hat{S}^{\lambda}) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}\{k \in \hat{S}_i^{\lambda}\}.$$

3. Given the selection probabilities for each frequency with different tuning parameters, respectively, then construct the final stable frequencies set accord-

ing to the following definition:

$$\hat{S}^{stable} = \{f_k : \max_{\lambda \in \Lambda} \hat{\Pi}_k^{\lambda} \geq \pi_{thr}\}.$$

We suggest two ways to define the thresholding value, and then the selection set $\hat{S}_i^{\lambda}$ just includes the frequencies with periodogram greater than this thresholding value. First, we set a thresholding level by taking a small percentage, for example, 1%, 5% or 10%, of the maximum periodogram value. Second, the thresholding value is the kth largest periodogram value. For instance, we set k=10, then each time just 10 frequencies with top 10 maximum periodogram values enter into the selection set. Another thresholding parameter π_{thr} needs to consider. Meinshausen and Bühlmann [54] suggested the threshold values $\pi_{thr} \in (0.6\ 0.9)$. Conservatively, we can set the interval as (0.5 0.8). Our real data examples show that the results vary little for this threshold parameter and are not sensitive to the choices of λ. For instance, Figure 2.3 indicates that we can detect the true frequency with high probability when λ belongs to a wide subinterval of Λ. These results consist with the conclusion in Meinshausen and Bühlmann [54].

Our procedure identifies a set of stable frequencies that are selected with high probability rather than simply finds the best value of $\lambda \in \Lambda$ and then uses it in the problem 2.8 or 2.9. With the stability selection, we do not simply select one model. Instead, we choose all frequencies that occur in a large fraction of the resulting selection sets. We keep frequencies with a high selection probability and disregard those with low selection probabilities. In practice, we can plot the stability path for each frequency which is potentially very useful for improved frequency estimation. For every frequency $f_1, \ldots, f_N$, the stability path is given by the selection probabilities $\hat{\Pi}_k^{\lambda}, \lambda \in \Lambda$. It can be seen in Figure 2.3 that there is a frequency that stands out

clearly from other frequencies.

2.5 Simulation Study

In this section, we present the empirical performance of the regularized periodograms by using simulated data. For brevity in figures and tables, we use capital abbreviation RLSP standing for regularized least squares periodogram, RQP standing for regularized quantile periodogram, AP and LP indicating respectively, adjusted periodogram of Hall and Li [33], Laplace periodogram of Li [48] and Thieler, Backes, Fried, *et al.* [75],

Suppose that data (t_i, y_i) are generated by the regression model,

$$y_i = h(t_i) + \sigma\varepsilon_i, \tag{2.13}$$

where $h(\cdot)$ is a continuous function and ε_i is the error term. The value of σ depends on the noise sample variance of ε_i. Following the data setting in [48], the noise is a normalized sequence error, normalized for each realization so that the sample variance of the noise equals 0.5 and thus the SNR equal 1. In Subsections 2.5.1, we consider that $h(\cdot)$ is a single periodic function. $h(\cdot)$ is extended to the case of multi-periodic function in Subsection 2.5.2. For each data setting, the errors ε_i in (2.13) are taken to be one of the following:

I. Standard normal distribution: $N(0,\ 1)$,

II. Mixed variance normal distribution: $0.9N(0,\ 1) + 0.1N(0,\ 10)$,

III. Standard Cauchy distribution; $C(0,\ 1)$.

The last two distributions may represent the case that outliers are presented.

For all simulations, the tuning parameter λ is selected from the candidate set Λ, including twenty-five grid points in a logarithmic sequence from e^{-3} to e^{2}, by the proposed procedure in Subsection 2.4.2. For each $\lambda \in \Lambda$, the probability for every frequency to be selected was from the 50 times randomly subsampling without replacement at a subsample size $n/2$. Specifically, We define the thresholding values by the 5% of the maximum value of RLSP and RQP in the step b(ii) of the selection algorithm in Subsection 2.4.2. Finally, we keep the frequencies with a high selection probability and disregard those with low selection probabilities by setting the exact cutoff $\pi_{thr} = 0.7$. We set $N = 500$ and $f_{max} = 1$ for all different settings. In quantile regression, we consider $\tau = 0.5$.

2.5.1 Simulation Study for Single-period Model

Two models are that at (2.13), with h given by M1 and M2, respectively,

$$\begin{aligned} \text{M1: } h(t) &= \cos(0.152 \times 2\pi t), \\ \text{M2: } h(t) &= g(t/\sqrt{2}), \end{aligned}$$

where $g(t)$ defined as $1 - \cos(2\pi t^2)$ on the interval $[0, 1]$, extended to the real line by periodicity. Model M1 is a sinusoidal signal, which has been investigated by Li [48] with successive equally spaced points in time and the standard Cauchy error setting. Hall and Li [33] studied the non-sinusoidal periodic function M2 with the standard Gaussian noise. They took the design points t_i to be the uniformly distribution on the interval $[0, n]$.

We study the performance of RLSP and RQR in the following three sampling strategies. Strategy 1, we take 200 design points t_i that follow the Uniform distribution on the interval $[0, 200]$ and sort these points. Gaps are introduced by dismissing

time points from strategy 1, see [29]. In strategy 2, we generate 200 available time points by strategy 1, then eliminating $t_{36*k+27}, \ldots, t_{36*(k+1)}$ for $k = 0, \ldots, 4$. Reorder the rest 150 time points as $t_1, \ldots t_{150}$ and remove t_{3k} for $k = 1, \ldots, 50$. This reduces the amount of available data from 200 to 100. These gaps can be explained to simulate annual obscuration by the Sun and monthly obscuration by the moon. In strategy 3, we have a gap 20 every 36 continuous time points via taking away $t_{36*k+17}, \ldots, t_{36*(k+1)}$ for $k = 0, \ldots, 4$ and a gap of 1 every 3 time points, reducing the number of data from 200 to 66. Therefore, sample size is $n = 200, 100$, or 66. Overall, there are nine settings considering all the factors for each model.

The performance of our proposed estimation procedures was compared to other two periodogram estimation methods: the adjusted periodogram in Hall and Li [33], Laplace periodogram in Li [48] and Thieler, Backes, Fried, *et al.* [75]. For the adjusted periodogram and Laplace periodogram, we also apply the 50 times subsampling with sample size $n/2$. In each time, we get the frequency selection set based on the estimating value of adjusted periodogram and Laplace periodogram of the subsampling dataset. Then, we calculate the empirical selection probability for each frequency from 50 frequency selection sets. Finally, the set of stable frequencies can be obtained from the empirical selection probability. We repeat our simulation 200 times for each different data setting and calculate the fraction of correctly found periods, referred to as the detection rate.

To compare performance of the four periodograms, namely RQP, RLSP, LP, and AP, we list the detection rate for the periodic functions in M1 and M2. As showed in Table 2.1, most of the cases, the regularized periodograms provide more accurate frequency estimation for all data setting in respect of detection rate, especially in small sample size. Apparently, regularized least squares periodogram method gives the best detection rate when errors follow the standard normal distribution. How-

ever, the detection rate of regularized least squares periodogram method is about 2% and 6% less than regularized quantile periodogram method with mixed normal distribution and sample size 66 in, respectively M1 and M2. Comparably, when errors are Cauchy distributed, the performance of the four methods has a more noticeable difference. To be specific, in the M2 with sample size 66, the detection rate in RQP method is about 13%, 19%, 41%, more than the detection rate from the RLSP, LP, and the AP method, respectively.

	Mode	M1				M2			
Dist	n	RLSP	RQP	LP	AP	RLSP	RQP	LP	AP
I	200	1.000	1.000	1.000	1.000	0.938	0.880	0.625	0.627
	100	1.000	1.000	1.000	0.993	0.922	0.895	0.610	0.665
	66	0.988	0.983	0.965	0.950	0.760	0.703	0.517	0.550
II	200	1.000	1.000	1.000	1.000	0.927	0.902	0.585	0.637
	100	1.000	1.000	0.990	0.988	0.868	0.915	0.620	0.623
	66	0.973	0.990	0.985	0.958	0.688	0.740	0.578	0.547
III	200	0.998	1.000	1.000	0.995	0.887	0.970	0.785	0.573
	100	0.995	1.000	1.000	0.965	0.880	0.943	0.703	0.647
	66	0.865	1.000	1.000	0.877	0.665	0.890	0.608	0.487

Table 2.1: This table shows the detection rate of the 4 Methods from the 200 Monte Carlo replicates for the model M1 and M2 with different error distributions.

2.5.2 Simulation Study for Two-period Model

A similar regression function used here is identical to that treated by Hall and Yin [34] and Hall and Li [33]. In particular, the function h in the model (2.13) can be represented as

$$\text{M3: } h(t) = 2.5 + \sin(2\pi t/\theta_1 + \phi_1) + \sin(2\pi t/\theta_2 + \phi_2), \tag{2.14}$$

where $\theta_1 = 3$, $\theta_2 = \sqrt{2}$, $\phi_1 = \pi/4$ and $\phi_2 = \pi/12$. Two phases, ϕ_1 and ϕ_2, are introduced to bring a little difference between our function $h(\cdot)$ and the regression function in Hall and Li [33]. The errors ε_i are also independently taken from three different distributions. Sample size is $n = 200,\ 100$, or 66.

Table 2.2 contains the result of the simulation study in the detection rate for estimating the two unknown frequencies. We observe that regularized periodograms largely improve the frequencies estimation of the four methods. The detection rate in regularized periodograms is about 20% higher than the unpenalized methods with the normal distribution errors. Moreover, when errors follow mixed normal distribution, regularized quantile periodogram still performs better than other three methods. In particular, for sample size 66 with standard normal distribution and mixed normal distribution, the four periodograms fail to estimate the two frequencies, meaning that they almost lose their estimation ability. In addition, regularized quantile periodogram delivers the highest detection rate and provides a reliable estimation under the Cauchy error setting. In summary, our proposed regularized periodograms are more efficient in frequencies estimation from the data including outliers and big gaps.

2.6 Real Data Application

In this section, we focus on estimating the intrinsic frequencies on the 'slowly pulsating B stars' (SPBs) which are B-type variables pulsating in high-radial-order g-modes with periods of the order of days, refer to [79]. This property indicates that we can set $f_{max} = 1$ in this section. There is a series of papers study the frequencies for the SPBs, see [79], [21]. We implement our method on two real examples. One is HD 138764 with single period SPBs; another one is a multi-periodicity SPBs,

		θ_1				θ_2			
Dist	n	RLSP	RQP	LP	AP	RLSP	RQP	LP	AP
I	200	0.792	0.760	0.505	0.590	0.863	0.805	0.565	0.618
	100	0.575	0.540	0.282	0.405	0.605	0.550	0.310	0.395
	66	0.228	0.172	0.107	0.185	0.225	0.198	0.107	0.207
II	200	0.820	0.853	0.598	0.650	0.860	0.885	0.557	0.568
	100	0.545	0.618	0.417	0.465	0.627	0.670	0.370	0.415
	66	0.210	0.240	0.168	0.210	0.172	0.237	0.175	0.170
III	200	0.782	1.000	0.565	0.618	0.818	0.968	0.590	0.595
	100	0.545	0.950	0.527	0.450	0.497	0.925	0.495	0.380
	66	0.250	0.723	0.315	0.235	0.260	0.675	0.362	0.242

Table 2.2: This table shows the detection rate of four Methods from the 200 Monte Carlo replicates for the model M3 with different error distributions.

called HD 123515. The two dataset were gathered by the Geneva P7 photometer of the Geneva Observatory. All real dataset studied in this paper can be downloaded from:

`http://www.ster.kuleuven.be/~roy/helas/`

We thank Professor Conny Aerts for giving us above address.

2.6.1 Single SPBs: HD 138764

We now analyze the Geneva data for the star HD 138764. Thanks to the photometric measurements of the HIPPARCOS mission, photometric variability of HD 138764 is beyond any doubt now. There are 89 time points with large gaps in the Geneva data with the time recorded in Heliocentric Julian Date form. We implement our regularized quantile periodogram to detect hidden frequency from this dataset. Consider $\tau = 0.5$ and $N = 5000$. The candidate set of tuning parameters Λ includes 50 grid points of equally spaced on the log-scale over $[e^{-1},\ e^{3}]$. For each tuning parameter λ, we run the subsampling procedure 100 times. In each time, random pick 60 time points from the 89 ones and estimate the selection set based on the value of

the regularized quantile periodogram from the subsample dataset. Especially, the selection set includes the frequency whose regularized quantile periodogram value is greater than the 1% of the maximum one. After 100 times, we compute the empirical selection probability for each frequency under the fixed tuning parameter. Finally, for each frequency, we obtain 50 empirical selection probabilities corresponding to the 50 $\lambda's$. We plot these probabilities as the function of λ that is the stability path, see (2.3). Setting $\pi_{thr} = 0.8$, we find the important intrinsic frequency at $f_1 = 0.7944$. This result confirms the conclusion of [21] that suggests a single strong intrinsic frequency $f_1 = 0.7944$ in the Geneva data. If we set $\pi_{thr} = 0.6$, then we may get two more frequencies, $f_2 = 0.7934$ and $f_3 = 0.7930$ that can be seen as the aliases. If we set $\pi_{thr} = 0.45$, then a new frequency $f_4 = 0.0326$ enter into the selection set. Here we are not concerned with astronomical explanations of this frequency, but focus on the frequency estimation from astronomical data.

2.6.2 Binary SPBs: HD 123515

The data are treated by De Cat and Aerts [21], Hall and Yin [34], [33]. Waelkens [79] conducted an analysis using the first 209 observations in the dataset. Later, De Cat and Aerts [21] and Hall and Yin [34] also employed a periodogram-based approach to treating a version of Waelkens' dataset expanded to 630 observations. The dataset consists of a few old measurements besides new ones gathered during, and the time of measurements is separated by several years.

We first consider the regularized quantile periodogram. Let $N = 8000$, $f_{\max} = 1$, and $M = 100$. The set Λ contains a grid of 60 values of λ equally spaced on the log-scale over $[e^{-2}, e^{4}]$. We take 400 observation points from the 630 samples to generate a subsample each time. Figure 2.4a shows the result of the above process.

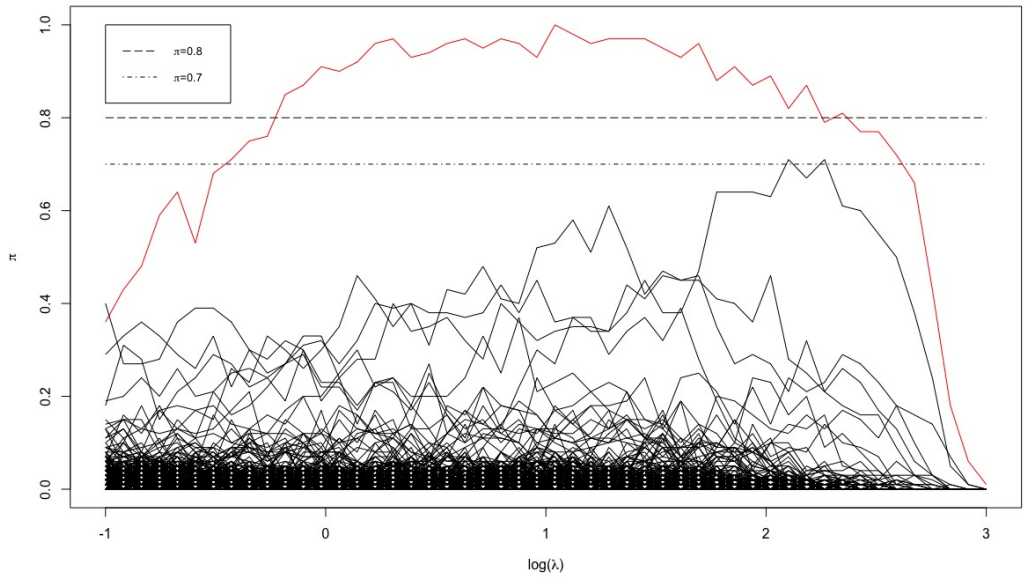
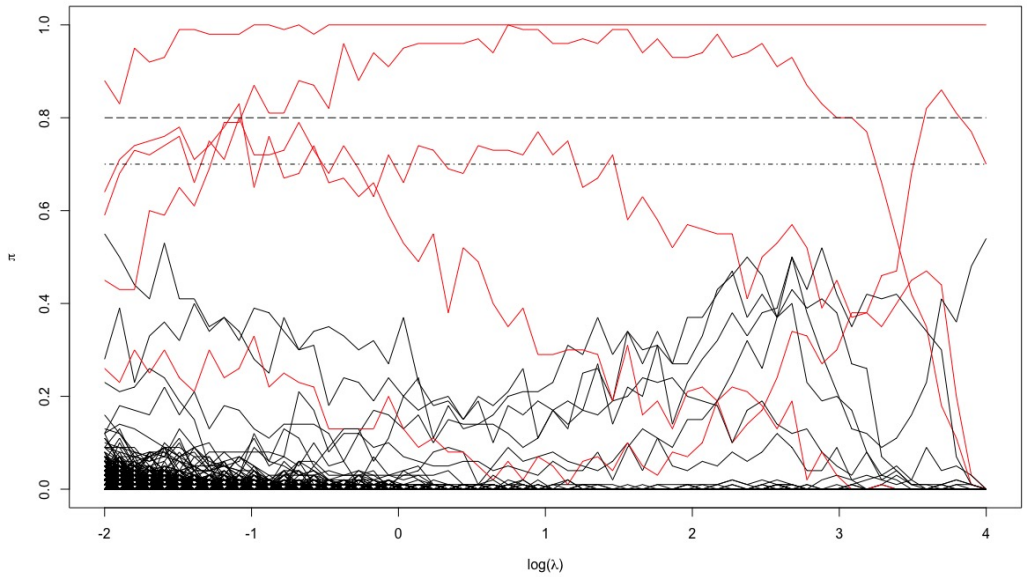


Figure 2.3: Each curve is the stability path that represents the selected probability for a fixed frequency corresponding to the 50 different $\lambda's$. The red line is the stability path for frequency $f_1 = 0.7944$ and indicates that there might be one frequency in the Geneva data HD 138764.

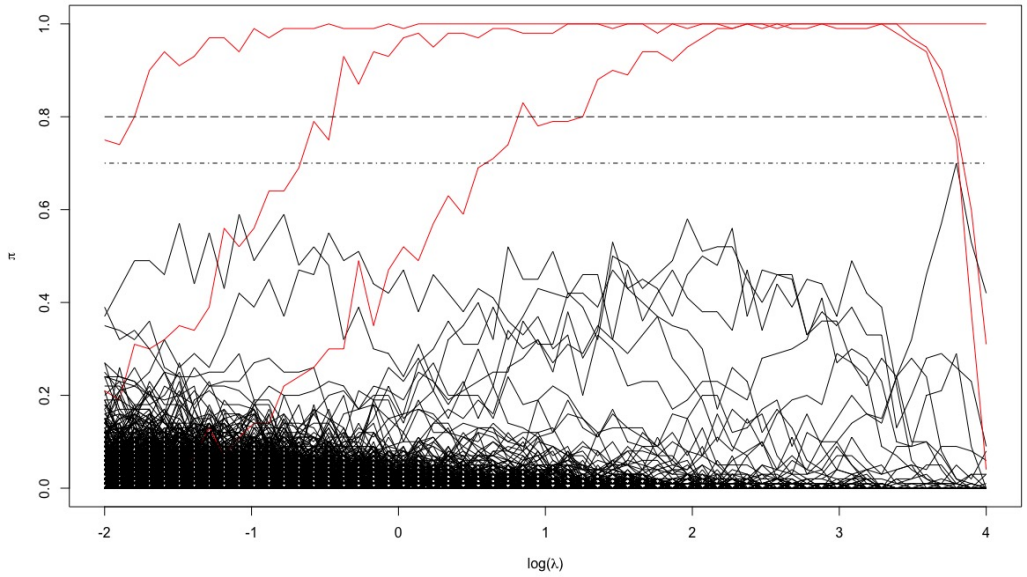
	f_1	f_2	f_3	f_4
Waelkens [79]	0.72861	0.68521	0.65928	0.45834
De Cat and Aerts [21]	0.72585	0.68528	0.65929	0.55198
Hall and Li [33]	0.72585	0.68531	0.65933	
RQP	0.72584	0.68528	0.65928	0.55200

Table 2.3: Comparison of frequencies estimation result for the HD123515. There are just three frequencies detected by Hall and Li [33].

There is no longer a clearly gap between the important frequencies and unimportant ones. When $\pi_{thr} = 0.7$, then five frequencies enter into the selection set corresponding $f_1 = 0.725875$, $f_2 = 0.688125$, $f_3 = 0.68525$, $f_4 = 0.6835$ and $f_5 = 0.65925$. The order is not sorting by the value of periodograms. Setting the $\pi_{thr} = 0.5$, we obtain 11 significant frequencies concentrated in the interval $[0.5, 0.75]$. Borrowing the idea of prewhitening, we consider the frequencies in the interval $[0.5,\ 0.75]$ and run above process again. The final result has been shown on the Figure 2.4b. Three strong frequencies, $f_1 = 0.7258438$, $f_2 = 0.6852812$ and $f_3 = 0.6592813$, present in it by setting $\pi_{thr} \in [0.7,\ 0.8]$. This result consists with the highest three frequencies obtained in all of the earlier studies [21], [34], [33]. [79], see Table 2.3. We conservatively select the fourth frequency at $f_4 = 0.552$ by setting $\pi_{thr} = 0.5$. At the same time, there are many aliases frequencies of respectively, f_1, f_2 and f_3 has been found. Implementing the same procedure for the regularized least squares periodogram, we can also find the four relevant frequencies. Table 2.3 demonstrates that our result meets the conclusion of most specialized astronomy literature, which finds the important four frequencies.



(a)



(b)

Figure 2.4: Estimated stability path for HD 123515 using RQP. Each curve is the stability path that represents the selected probability for a fixed frequency corresponding to the 60 different $\lambda's$. (a) Red lines are the stability paths of the selected frequencies selected by setting $\pi_{thr} = 0.7$. The interval of interesting frequencies is $[0,\ 1]$. (b) As for (a) but the interesting frequencies belong to the interval $[0.5,\ 0.75]$.

Chapter 3

Nonparametric Spectral Density Estimation by Total Variation Regularization

We propose the total variation penalized Whittle likelihood approach to the nonparametric estimation of the spectral density of Gaussian processes. Some asymptotic rates of convergence are established. A simulation study with autoregressive and moving average process is conducted to compare the proposed method with some other existing methods including L_2 penalized likelihood method. Two real-data sets are analyzed to illustrate the application of our method.

3.1 Introduction

Spectral analysis is a useful tool to analyze the frequency content in the time series. For instance, the frequency content can be characterized by the spectral density of a stationary time series. One could be interested in finding a few dominant fre-

quencies or frequency regions, which correspond to multimodality in the spectral density. Inference methods for multimodal spectral densities have been considered by Davies and Kovac [20]. They used the taut string method to directly control the number of peaks in the estimate spectral density. Anevski and Soulier [2] considered unimodal spectral density estimation for known mode that the spectral density is a decreasing function on $[0, \pi]$. In this chapter, our objective is to develop the application of the penalized likelihood method to estimate multimodal spectral density.

The penalized likelihood method has been established for nonparametric estimation of function parameters in a variety of setting including regression, density estimation, and time series analysis. Chow and Grenander [18] proposed a L_2 regularized approach, called the sieve method, to obtain a nonparametric maximum likelihood estimator (MLE) based on the Whittle likelihood. Pawitan and O'Sullivan [62] defined a penalized MLE as the maximizer of the Whittle likelihood with roughness. These two methods can be viewed as the L_2 type penalized Whittle likelihood estimation since the measure of the roughness of a function is by its integrated squared mth derivative.

Naturally, we would like to replace the L_2 penalties with L_1 penalties in the penalized Whittle likelihood method to obtain a new estimation principle. The L_1 penalties interpreted as the roughness of the candidate function measured by the total variation of their derivatives for function estimation. Koenker and Mizera [39] developed a unified approach to total variation penalized, L_1 penalty, density estimation offering methods that are capable of identifying qualitative features like sharp peaks. An advantage of using L_1 type penalties is known to be capable of capturing sharp changes in the target function while still maintaining a general smoothing objective. For example, we always calculate an appropriate spectral den-

sity function and identify the hidden periodicities in the data with the peak which is a sharp change in its derivative. This L_1 type penalty contributions inherently exaggerates the contribution to the penalty of jumps and sharp bends in the density; indeed, density jumps and piecewise linear bends are impossible in the L_2 framework since the penalty evaluates them as infinitely rough. But total variation penalties are happy to tolerate such jumps and bends. Therefore they are better suited to identifying discrete jumps in densities or in their derivatives, for example identifying the dominant frequencies in the spectral density. Consequently, our penalty approach has the potential to capture local features in the density more efficiently than do more global approximations method.

The goals of the present chapter are to derive a nonparametric L_1 penalized Whittle likelihood method for the estimation of the spectral density and to study it in some detail. A particular emphasis is on establishing rates of convergence results. We also provide quantification of the performance characteristics of the approach relative to some well-established techniques, including the smoothed periodogram, autoregressive spectral density estimator, by a simulation study with autoregressive and moving average processes.

This chapter is organized as follows. The basic methodology, including the algorithm to solve our optimization problem, is introduced Section 3.2. A result describing the asymptotic rates of convergence of the estimator is presented in Section 3.3. Some illustrations with simulated and real data are given in Section 3.4.

3.2 Methodology Development

We propose an estimator for the spectral density function, a penalized MLE with the total variation penalties. An efficient algorithm is also given in this section.

3.2.1 The Whittle Likelihood and Definition of the Estimator

Consider a real stationary Gaussian time series $\{X_t; t = 0, \pm 1 \dots\}$ with autocovariance function $C_{xx}(k)$ in (1.4). Then, the second order properties of the time series are completely described by the spectral density function f in (1.6). We shall study the estimation f from a finite sample $\{X_0,\ X_1,\ \dots,\ X_n\}$.

Most of the nonparametric estimation procedures are based on the periodogram $I_n(\cdot)$. For a fixed $\omega \in (0,\ \pi)$, $I_n(\omega)$ has an asymptotic exponential distribution with mean $f(\omega)$. Moreover, if two frequencies ω_1 and ω_2 are at least $2\pi/n$ apart, then the covariance between $I_n(\omega_1)$ and $I_n(\omega_2)$ is of the order n^{-1}. Let $I_{n,k}$ and $f_{n,k}$ be the periodogram and spectral density evaluated at the Fourier frequency $\omega_{n,k} = 2\pi k/n$, $k = 0, \pm 1, \dots, \pm v$, where $v = v_n = [(n-1)/2]$ is the greatest integer less than or equal $(n-1)/2$. Then the joint distribution of $(I_{n,1}, \dots, I_{n,v})$ may be approximate by the joint distribution of v independent exponential random variables with mean $f_{n,j}$ for the jth component. Whittle [84], [85] proposed a quasi-likelihood,

$$L_n(f|X_1, \dots, X_n) = \prod_{k=1}^{v} e^{-I_{n,k}/f_{n,k}}/f_{n,k},$$

known as the Whittle likelihood in the literature. By using $I_{n,-k} = I_{n,k}$ and $f_{n,-k} = f_{n,k}$, an approximate negative log-likelihood of $(I_{n,-v}, \dots, I_{n,v})$ given the f is

$$L_n(f) = \frac{1}{n} \sum_{k=-v}^{v} \Big\{ \log f_{n,k} + I_{n,k}/f_{n,k} \Big\}, \tag{3.1}$$

where we ignore the slight difference in the asymptotically distribution at $k = 0$, as [62].

Replacing f with a fixed function g in equation (3.1) as $n \to \infty$, we have that

$$L_n(g) \to L(g) = \int_{-\pi}^{\pi} \Big\{ \log g(\omega) + f(\omega)/g(\omega) \Big\} d\omega.$$

By Lemma 3.3.2, we have $L(g)$ is minimized at $g = f$. Hence $L_n(g)$ should be a reasonable objective function for identification of the spectrum which is also a key property of the Whittle likelihood. But this is not enough, since if we just solve the optimization problem for any function g, then we get the estimate $f(\omega) = I_n(\omega)$. It is well known that this is not even consistent and cannot be accepted.

Our method is inspired by the penalized likelihood approach in [18], [62] and [39]. Analogously, we propose to minimize objective function

$$L_{n,\lambda}(\theta) = L_n(e^{\theta}) + \lambda J(\theta),$$

involving roughness penalties of the logarithmic spectral density. The first term $L_n(e^{\theta})$ should serve as the data fit criterion and $J(\theta)$ is a penalty function. The benefit of the transformation to the logarithmic spectral density function is that it obviates any worries about the non-negativity of the spectral density function. Also we always view spectral densities on a logarithmic scale in practice. The parameter $\lambda > 0$ controls the amount of smoothing: $\lambda = 0$ corresponds to the unpenalized estimator or $\hat{\theta} = \log I_n$ and $\lambda = +\infty$ corresponds to the smoothest model.

Motivated by Koenker and Mizera [39], we consider $J(\theta)$ based on total variation of θ and its derivatives. Recall that the total variation of a function $\theta : \Omega \to R$ is defined as

$$\bigvee_{\Omega} \theta = \sup \sum_{i=1}^{m} |\theta(u_i) - \theta(u_{i-1})|,$$

where the supreme is taken over all partitions, $u_0 < \ \ldots \ < u_m$ of Ω, when θ is

absolutely continuous, we can write

$$\bigvee_{\Omega} \theta = \int_{\Omega} |\theta'(x)| dx.$$

We will focus on penalizing the total variation of $\theta^{(m-1)}$, the $m-1$ derivative of the θ. In fact, we restrict our attention to $m = 2$, rather then other possibilities, like $m = 1$, which would be quite natural in the context of spectral density estimation, or $m = 3$, as suggested for probability density estimation by Silverman [71]. Our choice of $m = 2$ is the same as the choice prevailing in Pawitan and O'Sullivan [62], and also conforms to the total-variation penalty, used in the regression setting, of Koenker, Ng, and Portnoy [45]; see also Koenker and Mizera [39].

The penalty function $J(\theta)$ can be rewritten as

$$J(\theta) = \bigvee_{(-\pi,\ \pi)} \theta^{(m-1)} = \int_{-\pi}^{\pi} |\theta^{(m)}(\omega)| d\omega,$$

under the assumption that $\theta^{(m-1)}$ is absolutely continuous.

3.2.2 Computation

From the computational perspective, total variation based penalties fit comfortably into modern convex optimization setting. The idea is based on the algorithm of Koenker and Mizera [39]. Firstly, note that the first part of our objective function is a convex function with respect to $\theta(\omega_{n,k})$, see Lemma 3.3.2. Here, we just focus our attention on penalizing derivatives of $\theta(\omega) = \log(f(\omega))$, in fact, other convex transformations can be easily accommodated. Our preliminary experimentation with penalization of $\theta^m(\omega)$ with $m = 2$. Restricting attention to θ's for which θ is piecewise linear on $(-\pi,\ \pi)$, we can write $J(\theta)$ as an l_1 norm of the second weighted differ-

ences of $\theta(\omega)$ evaluated at the Fourier frequency $\omega_{n,k}$. Because $I_n(\omega)$ are symmetric about 0, and the resulting spectral density estimate $\theta(\omega)$ is guaranteed symmetric about 0, we just need to consider the interval $(0,\ \pi)$. More explicitly, if $\theta(\omega)$ is a piecewise linear on the partition $\omega_{n,0} < \omega_{n,1} \ldots,\ \omega_{n,v}$, so that

$$\theta(\omega) = \alpha_i + \beta_i \omega \quad \omega \in [\omega_{n,i},\ \omega_{n,i+1}],$$

then the penalty part can be rewritten as following:

$$J(\theta) = \bigvee_{(0,\ \pi)} (\theta)' = \sum_{i=0}^{v} |\beta_i - \beta_{i-1}| = \frac{2\pi}{n} \sum_{i=0}^{v} |\theta(\omega_{n,i-1}) - 2\theta(\omega_{n,i}) + \theta(\omega_{n,i+1})|,$$

where we have imposed continuity of f in the last step and ignore. We can thus parameterize function θ by the discrete function values $\theta_k = \theta(\omega_{n,k})$, this enables us to write our problem as a convex program,

$$\min \left\{ \sum_{i=0}^{v} \left(\theta_{n,i} + I_{n,i} e^{-\theta_{n,i}} \right) + \lambda \sum_{i=0}^{v} \left(u_i^+ + u_i^- \right) \middle| \boldsymbol{D\theta} - \boldsymbol{u}^+ + \boldsymbol{u}^- = 0,\ (\boldsymbol{\theta}, \boldsymbol{u}^+, \boldsymbol{u}^-) \in \mathbb{R}^{v+1} \times \mathbb{R}_+^{2(v+1)} \right\},$$

where $\boldsymbol{D}$ denotes a tridiagonal matrix containing the $2\pi/n$ factors for the penalty contribution, $\boldsymbol{\theta} = (\theta_{n,0}, \ldots, \theta_{n,v})^T$ is a $(v+1) \times 1$ vector, and $\boldsymbol{u}^+$ and $\boldsymbol{u}^-$ represent the positive and negative parts of the vector $\boldsymbol{D\theta}$, respectively.

3.3 Asymptotic Approximation

In this section, we investigate the asymptotic rate of convergence of the estimator under certain regularity conditions on the true spectral density and estimate of the

log spectral density $\theta_{n,\lambda}$. Our objective function can be write as

$$L_{n,\lambda}(\theta) = \int_{-\pi}^{\pi} [\theta(\omega) + I_n(\omega)e^{-\theta(\omega)}]d\omega + \lambda \bigvee_{(-\pi,\ \pi)} \theta^{(m-1)}, \tag{3.2}$$

where $m \geq 2,\ \lambda > 0$ are constants. We can get the estimator by minimizing the objective function over the set

$$A = \left\{\theta :\ \theta^{(i)} \in L^1(-\pi,\ \pi),\ 0 \leq i \leq m\right\}.$$

We apply the theory of [18] and [62] to approximate the asymptotic convergence characteristics of the penalized likelihood estimator. For $r > 0$, let $\mathbb{W}_2^r$ to be the periodic Sobolev space on $[-\pi, \pi]$. The norm on $\mathbb{W}_2^r$ is denoted $\|\cdot\|_r$ and is given by

$$\|g\|_r^2 = \sum_k (1 + |k|^2)^r |\hat{g}_k|^2, \tag{3.3}$$

where $\hat{g}_k$ are the Fourier coefficients of g. A convergence result is obtained under the following conditions.

C1: $(x_t : t \in \mathbb{Z})$ is a real stationary Gaussian process with the true spectral density f_0 that is bounded away from 0. Moreover, there exist a constant m, M such that $f_0(\omega) \in [m,\ M]$ for all $\omega \in [-\pi,\ \pi]$.

C2: The estimate of the log spectral density $\theta_{n,\lambda}$ is the minimizer of $L_{n,\lambda}(\theta)$ in $\mathbb{W}_2^{m+\frac{1}{2}}$.

C3: $\theta_0 = \log f_0 \in \mathbb{W}_2^m$.

Condition C1 guarantees that the log spectral density is well defined. By Sobolev Embedding Theorem in [27], we have the conclusion, for any function $f \in \mathbb{W}_2^{m+\frac{1}{2}}$,

f is to be of differentiability class space $\mathbb{C}^m$ which means the derivatives $f^{(1)} \dots f^{(m)}$ exist and are continuous. Hence, we know our estimator $\theta_{n,\lambda} \in \mathbb{C}^k$. Condition C3 is a smoothness condition. According to [62], the condition can be interpreted as a covariance summability condition of order $l < m - \frac{1}{2}$, for example $\sum_k (1 + |k|^l) C_{xx}(k) < \infty$. Note that Condition C3 also implies $f_0 \in \mathbb{W}_2^m$.

Under these conditions, the following result already shown on [35], [73] and [74].

Lemma 3.3.1. *Let $h(\omega)$ be a continuous symmetric function on $[-\pi, \pi]$. Then with probability 1,*

$$\int_{-\pi}^{\pi} h(\omega)(I_n(\omega) - f(\omega))d\omega \quad \to_n \quad 0, \tag{3.4}$$

$$\int_{-\pi}^{\pi} h(\omega)(I_n^2(\omega) - 2f^2(\omega))d\omega \quad \to_n \quad 0, \tag{3.5}$$

$$\int_{-\pi}^{\pi} h(\omega)(\log(I_n(\omega)e^{\gamma} - \log f(\omega))d\omega \quad \to_n \quad 0, \tag{3.6}$$

where γ is the Euler's constant. Moreover, $\sqrt{n}\int_{-\pi}^{\pi} h(\omega)(I_n(\omega) - f(\omega))d\omega$ converges weakly to the normal distribution $N(0,\ 2\int_{-\pi}^{\pi} h^2(\omega)f^2(\omega)d\omega)$.

The equation (3.5) and (3.6) imply that the sequence $\int_{-\pi}^{\pi} I_n^2(\omega)d\omega$ and $\int_{-\pi}^{\pi} \log I_n(\omega)d\omega$ is bounded with probability 1. The loss function in the right-hand side of (3.2) can be rewritten as

$$L_{n,\lambda}(\theta) = \int_{-\pi}^{\pi} [I_n(\omega)e^{-\theta(\omega)} - \log(I_n(\omega)e^{-\theta(\omega)})]d\omega + \lambda \bigvee_{[-\pi,\ \pi]} (\theta)^{m-1} + \int_{-\pi}^{\pi} \log I_n(\omega)d\omega.$$

The following lemma is elementary, see [18]; we omit the proof.

Lemma 3.3.2. *The function $y(x) = x - \log x$ is strictly convex on $(0, \infty)$ and attains it unique minimum at x=1.*

By this Lemma, $L_{n,\lambda}(\theta)$ is bounded below with probability 1. Since $I_n(\omega)$ converges weakly to $f(\omega)$, the following loss function is related to (3.2),

$$L_\lambda(\theta) = \int_{-\pi}^{\pi} [\theta(\omega) + f(\omega)e^{-\theta(\omega)}]d\omega + \lambda \bigvee_{(-\pi,\ \pi)} (\theta)^{m-1}, \tag{3.7}$$

where $m \geq 2,\ \lambda \geq 0$ are constants. We minimize it and the minimum is taken over A if $\lambda > 0$ and over

$$A' = \{\theta :\ \theta^{(i)} \in L^1(-\pi,\ \pi)\},$$

in case $\lambda = 0$. Similarity to Condition $C2$, we also need a similar condition for the minimizer of (3.7), for example

C4 There is a log spectral density $\theta_\lambda \in \mathbb{W}_2^{(m+\frac{1}{2})}$ which is the minimizer of $L_\lambda(\theta)$ in A.

In fact, the solution θ_λ is unique by the strictly convex properties in the Lemma3.3.2. Using this lemma again, we have

Lemma 3.3.3. *For the $\lambda = 0$, the problem (3.7) has a unique solution $\theta_0(\omega) = \log(f(\omega))$ and any approximating sequence $\theta_k(\omega)$ to the $\theta_0(\omega)$ satisfies*

$$\lim_k \int_{-\pi}^{\pi} \left| e^{-\theta_k(\omega)} f(\omega) - 1 \right| d\omega = 0.$$

Proof. In general, we have that $\lim_k \int_{-\pi}^{\pi} (h_k(\omega) - \log h_k(\omega) - 1)d\omega = 0$ implies $\lim_k \int_{-\pi}^{\pi} \left| h_k(\omega) - 1 \right| = 0$. The proof of the lemma then follows by setting $h_k(\omega) = e^{-\theta_k(\omega)} f(\omega)$. □

For each $\theta(\omega) \in A$, we have $L_\lambda(\theta_\lambda) \leq L_\lambda(\theta) \to_\lambda L_0(\theta)$; therefore

$$\limsup_\lambda L_0(\theta_\lambda) \leq \limsup_\lambda L_\lambda(\theta_\lambda) \leq \min_{A'} L_0(\theta). \tag{3.8}$$

Since $L_{n,\lambda}(\theta_{n,\lambda})$ may be close to the $L_\lambda(\theta_\lambda)$, it is theoretically possible to choose $\lambda(n)$ such that

$$P(\lim_n L_0(\theta_{n,\lambda(n)}) = \min_{A'} L_0(\theta)) = 1. \tag{3.9}$$

In order to verify (3.9), let us introduce a quantity

$$D_{n,\lambda} = \max |L_{n,\lambda}(\theta) - L_\lambda(\theta)|,$$

where the maximum is taken over the set

$$B_\lambda = \left\{\theta : \ \theta \ symmetric, \ \int_{-\pi}^{\pi} (\theta^{(1)}(\omega))^2 d\omega \le \frac{C_1^2}{\lambda^2} \ and \ \|\theta\|_\infty \le \frac{6C_1}{\lambda}\right\},$$

where $C_1 = 2\pi(M - \log(m)/e^\gamma) + 1$. Next, we will show that $\theta_{n,\lambda} \in B_\lambda$ by Lemma 3.3.4, and under λ_n given in 3.3.5, the distance D_{n,λ_n} will go to 0 with probability 1.

Lemma 3.3.4. *For each $\lambda > 0$ and small, we have $P(\theta_{n,\lambda} \in B_\lambda \ for \ large \ n) = 1$.*

Proof. The inequality

$$L_{n,\lambda}(\theta_{n,\lambda}) \le L_{n,\lambda}(0), \tag{3.10}$$

implies that

$$\lambda \int_{-\pi}^{\pi} |\theta_{n,\lambda}^m(\omega)| d\omega \le \int_{-\pi}^{\pi} I_n(\omega) d\omega - \int_{-\pi}^{\pi} \log I_n(\omega) d\omega.$$

From the equation (3.4) and (3.6), it is not hard to show that when n is large enough, we have

$$\int_{-\pi}^{\pi} I_n(\omega) d\omega - \int_{-\pi}^{\pi} \log I_n(\omega) d\omega \le 2\pi(M - \log(m)/e^\gamma) + 1 = C_1.$$

Therefore, we have $\lambda \int_{-\pi}^{\pi} |\theta_{n,\lambda}^m(\omega)| d\omega < C_1$ for the large n. Note that from condi-

tion **C2**, we have $\theta_{n,\lambda} \in \mathbb{C}^m$. Using the Fourier series expansion $\theta_{n,\lambda}(\omega) = \sum a_k e^{(ik\omega)}$, we get

$$|a_k| \leq \frac{\int_{-\pi}^{\pi} |\theta_{n,\lambda}^m(\omega)| d\omega}{2\pi |k|^m} < \frac{C_1}{2\lambda\pi |k|^m},$$

with $m \geq 2$ and $k \geq 1$. By the Parseval's identity,

$$\int_{-\pi}^{\pi} (\theta'_{n,\lambda}(\omega))^2 d\omega = 2\pi \sum |k a_k|^2 \leq 2\pi \sum_{k\neq 0} (\frac{C_1}{2\lambda\pi k^{m-1}})^2 < \frac{C_1^2}{\lambda^2}.$$

We now verify that $|a_0|$ is bounded by a constant. By using the definition of a_0 and (3.10) we have

$$2\pi a_0 + \int_{\pi}^{-\pi} I_n(\omega) e^{-a_0} e^{-\sum_{k\neq 0} a_k e^{(ik\omega)}} d\omega \leq \int_{-\pi}^{\pi} I_n(\omega) d\omega \leq (\int_{-\pi}^{\pi} f(\omega) d\omega + 1) < 2\pi(M+1),$$

for the large enough n.

Hence, we know a_0 has a upper bound $M + 1$. If a_0 has a lower bound, say $a_0 > -4\frac{C_1}{\lambda} - 2\max(0, \log\frac{4}{m})$, then $|a_0| < \max\{M + 1, 4\frac{C_1}{\lambda} + 2\max(0, \log\frac{4}{m})\}$. Otherwise, suppose that $a_0 < -4\frac{C_1}{\lambda} - 2\max(0, \log\frac{4}{m})$. We have the following inequality by the fact that $|a_k| \leq C_1/2\lambda\pi|k|^m$

$$2\pi(M+1) - 2\pi a_0 > \int_{\pi}^{-\pi} I_n(\omega) e^{-a_0} e^{-\sum_{k\neq 0} a_k e^{(ik\omega)}} d\omega \geq e^{-a_0} e^{-2C_1/\lambda} \int_{\pi}^{-\pi} f(\omega)/2 d\omega,$$

where the last step follows from (3.5) with large n. Apply the Taylor expansion to obtain

$$\int_{\pi}^{-\pi} f(\omega)/2 d\omega \geq \pi m \geq 4\pi e^{a_0/2 + 2C_1/\lambda} \geq \pi e^{a_0} e^{2C_1/\lambda} (a_0^2/2 - 2a_0).$$

The above two inequalities imply $|a_0| < 2\sqrt{M+1}$. Finally, we obtain that $\|\theta\|_\infty \leq$

$\sum |a_k| \le |a_0| + \frac{C_1}{\lambda} \le \max\{M+1+\frac{C_1}{\lambda},\ 2\sqrt{M+1}+\frac{C_1}{\lambda}, \frac{5C_1}{\lambda}+2\max(0,\log\frac{4}{m})\} \le \frac{6C_1}{\lambda}$. □

We also have $\theta_\lambda(\omega) \in B_\lambda$ by the definition of $\theta_{n,\lambda}(\omega),\ \theta_\lambda(\omega)$, and

$$\begin{aligned} L_0(\theta_{n,\lambda}) &\le L_\lambda(\theta_{n,\lambda}) \le L_{n,\lambda}(\theta_{n,\lambda}) - D_{n,\lambda} \\ &\le L_{n,\lambda}(\theta_\lambda) - D_{n,\lambda} \le L_\lambda(\theta_\lambda) - 2D_{n,\lambda} \le L_\lambda(\theta_0) - 2D_{n,\lambda}. \end{aligned} \tag{3.11}$$

Let $\eta_n(t) = \int_0^t [I_n(v) - E(I_n(v))]d(v)$. Integrating by parts gives

$$\begin{aligned} & L_\lambda(\theta) - L_{n,\lambda}(\theta) \\ = \ & 2\int_0^\pi [I_n(\omega) - E(I_n(\omega))]e^{(-\theta(\omega)}d\omega + \int_{-\pi}^\pi [E(I_n(\omega)) - f(\omega)]e^{(-\theta(\omega)}d\omega \\ = \ & 2\eta_n(\pi)e^{(-\theta(\pi))} + 2\int_0^\pi \eta_n(\omega)e^{(-\theta(\omega))}\theta'(\omega)d(\omega) + \int_{-\pi}^\pi [E(I_n(\omega)) - f(\omega)]e^{(-\theta(\omega)}d\omega. \end{aligned}$$

Denote by $Q_1,\ Q_2,\ Q_3$, respectively, the last terms in the above equation. By computing the moments of $\eta_n(t)$, it can be shown that

Lemma 3.3.5. *For $s = 1, 2, \ldots, E(D_{n,\lambda}^{2s}) \le C_s/(n\lambda^2/\exp(12C_1/\lambda))^s$, where C_s is a constant depending only on s and M. Let c, δ be two constant with $0 < \delta < 1$, and*

$$\lambda(n) = \frac{12C_1}{(1-\delta)\log n}.$$

If s is large enough, then we have the property that$E(\sum_n (\frac{D_{n,\lambda(n)}}{\lambda(n)})^{2s}) \le \infty$ implying $P(D_{n,\lambda(n)} \to_n 0) = 1$.

Proof. The proof is based on [38], [18]. It is known from Lemma 8.4 of [38] and the proof of Lemma 2.6 in [18] that for $s = 1,\ 2,\ \ldots$

$$E|\eta_n(t)| \le C_s M^{2s}/n^s,$$

where C_s is a constant depending only on s. Hence for each $\theta \in B_\lambda$,

$$E|Q_1|^{2s} \le C_1 M^{2s} e^{(12sC_1/\lambda)}/n^s, \tag{3.12}$$

and by the generalized Hölder's inequality

$$\begin{aligned} E|Q_2|^{2s} &\le 2^{2s}\pi^{s-1}\Big(\int_0^\pi e^{(-2\theta(\omega))}(\theta'(\omega))^2 d\omega\Big)^s \cdot E\Big(\int_0^\pi \eta_n^{2s}(\omega)d\omega\Big) \\ &\le C_2\Big(e^{(12C_1/\lambda)}/(n\lambda^2)\Big)^s. \end{aligned} \tag{3.13}$$

As to the nonrandom term Q_3, it follows from the equation (1.7) of [38] and the equation (5.3) of [18] that there exists a constant C_3 such that

$$\begin{aligned} |Q_3| &\le (C_3 M \log n)\int_{-\pi}^{\pi} |\theta'(\omega)| \exp(-\theta(\omega))d\omega/n \\ &\le C_3 M \log n \exp(C/\lambda)\int_{-\pi}^{\pi} |\theta'(\omega)|d\omega/n \\ &\le C_4 M \log n e^{(6C_1/\lambda)}/(n\lambda) < C_4 M e^{(6C_1/\lambda)}/(\sqrt{n}\lambda). \end{aligned} \tag{3.14}$$

Combine (3.12), (3.13) and (3.14) together and the first of this lemma follows with a different constant C_s.

If we take $\lambda(n) = \frac{12C_1}{(1-\delta)\log n}$, then we have

$$E[(\frac{D_{n,\lambda(n)}}{\lambda(n)})^{2s}] = O((\frac{(\log n)^2}{n^\delta})^s). \tag{3.15}$$

Take s large enough, such as $s > \frac{1}{\delta}$. The sequence $\{D_{n,\lambda(n)}\}$ has the desired property that $E(\sum_n (\frac{D_{n,\lambda(n)}}{\lambda(n)})^{2s}) \le \infty$ implying $P(\frac{D_{n,\lambda(n)}}{\lambda(n)} \to_n 0) = 1$. □

Theorem 3.3.6. *Let $\lambda(n)$ be given in Lemma (3.3.5). The estimator $e^{\theta(\omega)}$ is a strongly L_1 consistent estimator to the spectral density f.*

Proof. For each θ_λ, $L_\lambda(\theta_\lambda) \leq L_\lambda(\theta_0) \to_\lambda L_0(\theta_0)$; therefore

$$\limsup_\lambda L_0(\theta_\lambda) \leq \limsup_\lambda L_\lambda(\theta_\lambda) \leq \min_{\theta \in A} L_0(\theta),$$

and then by Lemma 3.3.3

$$\lim \int_{-\pi}^{\pi} |e^{-\theta_\lambda(\omega)} f(\omega) - 1| d\omega = 0. \tag{3.16}$$

From the definition of $\theta_{n,\lambda}$ θ_λ, we can obtain

$$L_0(\theta_{n,\lambda}) \leq L_\lambda(\theta_{n,\lambda}) \leq L_{n,\lambda}(\theta_{n,\lambda}) + D_{n,\lambda} \leq L_{n,\lambda}(\theta_\lambda) + D_{n,\lambda} \leq L_\lambda(\theta_\lambda) + 2D_{n,\lambda}.$$

By using Lemma 3.3.5, there is a sequence $\lambda(n)$ such that

$$P(\lim_n L_0(\theta_{n,\lambda(n)}) = \min L_0(\theta)) = 1;$$

it follows from Lemma (3.3.3) that

$$P(\lim_m \int_{-\pi}^{\pi} |e^{-\theta_{n,\lambda(n)}(\omega)} f(\omega) - 1| d\omega \to_n 0) = 1.$$

Note that, if the $\lambda(n)$ satisfy the condition of lemma 3.3.5, then we can easy show that there is a large constant $N > 0$, such that for any n we have $|\theta_{n,\lambda(n)}(\omega)| < N$ with probability 1, hence we have

$$\begin{aligned} P(\lim_m \int_{-\pi}^{\pi} |e^{\theta_{n,\lambda(n)}(\omega)} - f(\omega)| d\omega \to_n 0) &= 1, \\ \lim_n E(\int_{-\pi}^{\pi} |e^{\theta_{n,\lambda(n)}(\omega)} - f(\omega)| d\omega) \to_n 0 &= 1, \end{aligned}$$

where the second one is from f is also a bounded function. That is, $e^{\theta_{n,\lambda(n)}(\omega)}$ is a strongly L^1 consistent estimator of f. Therefore only the condition

$$P(|\theta_{n,\lambda(n)}(\omega)| < N) = 1,$$

needs to be verified. By the inequality (3.11) and using $L_\lambda(\theta_{n,\lambda}) \leq L_\lambda(\theta_0) - 2D_{n,\lambda}$, we have

$$\lambda(n) \int_{-\pi}^{\pi} |\theta^{(m)}_{n,\lambda(n)}(\omega)| d(\omega) \leq \lambda(n) \int_{-\pi}^{\pi} |\theta^{(m)}_{0}(\omega)| d(\omega) - 2D_{n,\lambda(n)}. \tag{3.17}$$

Both side divided by $\lambda(n)$, we have the $\int_{-\pi}^{\pi} |\theta^{(m)}_{n,\lambda(n)}(\omega)| d(\omega) \leq 2\pi ||\theta_0||^2_m + 1$ with the large n. Note that the $\int_{-\pi}^{\pi} |\theta^{(m)}_{n,\lambda(n)}(\omega)| d(\omega)$ has been bounded by a constant, not depended on the $\lambda(n)$. Applying the Fourier series expansion of $\theta_{n,\lambda(n)}(\omega)$ again as the proof of Lemma 3.3.4, we have $||\theta_{n,\lambda(n)}||_\infty \leq \max\{M + 1 + l,\ 2\sqrt{M+1} + l, 5l + 2\max(0, \log \frac{4}{m})\}$ by replacing C_1/λ as l, where $l = 2\pi||\theta_0||^2_m + 1$. □

3.4 Numerical Illustration

3.4.1 Simulation Study

In this section we investigate the finite sample behavior of our nonparametric total variation penalty estimator (PTVE) by comparing with following three existing estimator of the spectral density.

Penalized Whittle Likelihood Estimator (PLE). This is the Whittle likelihood-based estimator proposed by Pawitan and O'Sullivan [62]. The estimate is obtained with the automatic selection of the smoothing parameter by using the iterative least squares method to find the solution.

Smoothed Periodogram Estimate (SPE). The periodogram at the discrete Fourier frequencies are smoothed using the Daniell kernel.

Autoregression Spectral Estimator (ARE). An autoregression(AR) model is fitted to the data, with the order selected by the Akaike information criterion (AIC), and the spectral density of the fitted model is used as estimated.

Four different time series models are considered for the simulation study:

(M1): $X_t = 1.372X_{t-1} - 0.677X_{t-2} + \varepsilon_t$, $\varepsilon_t \sim N(0, 0.4982)$,

(M2): $X_t = -.9X_{t-4} + 0.7X_{t-8} - 0.63X_{t-12} + \varepsilon_t$, $\varepsilon_t \sim N(0, 1)$,

(M3): $X_t = 2.7607X_{t-1} - 3.8106X_{t-2} + 2.6535X_{t-3} - 0.9238X_{t-4} + \varepsilon_t$, $\varepsilon_t \sim N(0, 1)$,

(M4): $X_t = Y_t + Z_t$, where $Y_t + 0.2Y_{t-1} + 0.9Y_{t-2} = \varepsilon_t + \varepsilon_{t-2}$, and $\varepsilon_t \sim N(0, 1)$, $Z_t \sim N(0, 0.25)$.

The four processes have been used for simulation studies by various authors. Model M1 has been considered in simulation studies in Künsch [46] and Bühlmann [9]. Model M2 was first given by Wahba [80] and studied by Choudhuri, Ghosal, and Roy [17] later. Percival and Walden [63] and Bühlmann [9] considered the model (M3). Neumann [59] smoothed the spectral density of Model (M4), and Davies and Kovac [20] investigated the numbers of peaks in the spectral density function. From the Figure (3.1), the AR(2) process in M1 has relatively smooth spectral density and exhibits a 'pseudo-periodic behavior', whereas other processes have few sharp spikes. The AR(4) process in M3 has two closed peaks and there is a big jump in the spectral density of the model M4.

Dataset are generated for sample size $n = 128, 256, 512$ and 1024 with 1000 replicates in each case. In each replicate, the first 20,000 values of the generated time series data are discarded to reach stationarity. Unfortunately, we do not have a

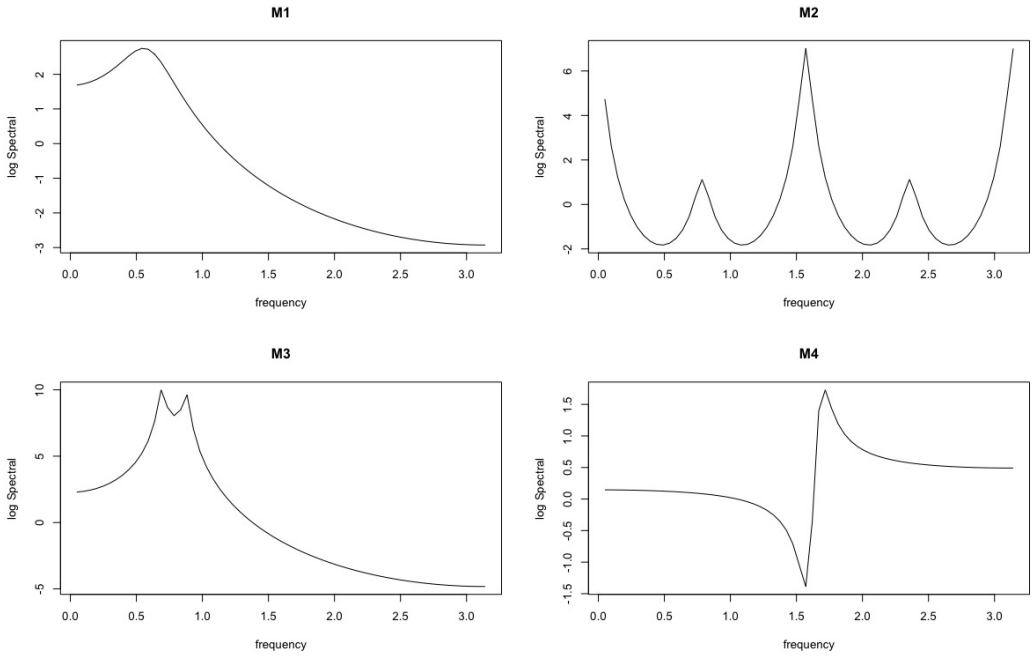


Figure 3.1: Spectral density functions of Model (M1)–(M4).

theoretically optimal value for the smoothing parameter λ. To make comparisons, we add a constraint condition that the Whittle likelihood of our estimator equals to the likelihood of penalized Whittle likelihood estimator in each replication. All of the estimates are computed for each sampled and compared in terms of L_1 error or integrated absolute error (IAE),

$$\text{IAE} = \|f_0 - \hat{f}_0\|_1 = \int_{-\pi}^{\pi} |f_0(\omega) - \hat{f}_0(\omega)| d(\omega).$$

Table (3.1) shows the mean L_1 errors form these 1000 and their boxplots are presented in Fig(3.2).

As expected, the nonparametric estimators, PTVE, PLE, and SPE, outperform the ARE in Model M4 but underperform in the Model M1, which is not that surprising, given it is an autoregressive model. A bit surprising is the performance

that our method PTVE also outperform the ARE in the AR(12) process with the small sample size. Perhaps one reason is that the ARE often underestimates the order of the model, resulting in large errors. Although the PTVE mostly detects the peaks correctly, it underestimates the magnitude of a sharp peaks, thus leading a higher L_1-error. In contrast, the PLE detects false spikes in small samples. Visually, the PTVE provides a better fit to the true density than SPE for all sample size and outperforms the PLE for all but very large samples.

Method	n=128	n=256	n=512	n=1024
M1 Model				
PTVE	0.952	0.712	0.509	0.390
PLE	0.948	0.699	0.497	0.383
ARE	0.793	0.581	0.399	0.289
SPE	0.880	0.699	0.566	0.505
M2 Model				
PTVE	36.706	25.270	19.221	14.247
PLE	39.067	26.101	20.899	15.937
ARE	44.689	27.442	18.351	13.049
SPE	51.603	42.559	34.786	25.155
M3 Model				
PTVE	679.567	506.266	473.628	432.377
PLE	669.560	531.387	516.272	469.315
ARE	647.494	434.230	297.374	206.267
SPE	784.127	631.770	546.992	397.393
M4 Model				
PTVE	0.504	0.400	0.317	0.241
PLE	0.520	0.429	0.350	0.271
ARE	0.623	0.515	0.415	0.322
SPE	0.540	0.440	0.356	0.293

Table 3.1: Mean L_1-error from the 1,000 Monte Carlo Replicates.

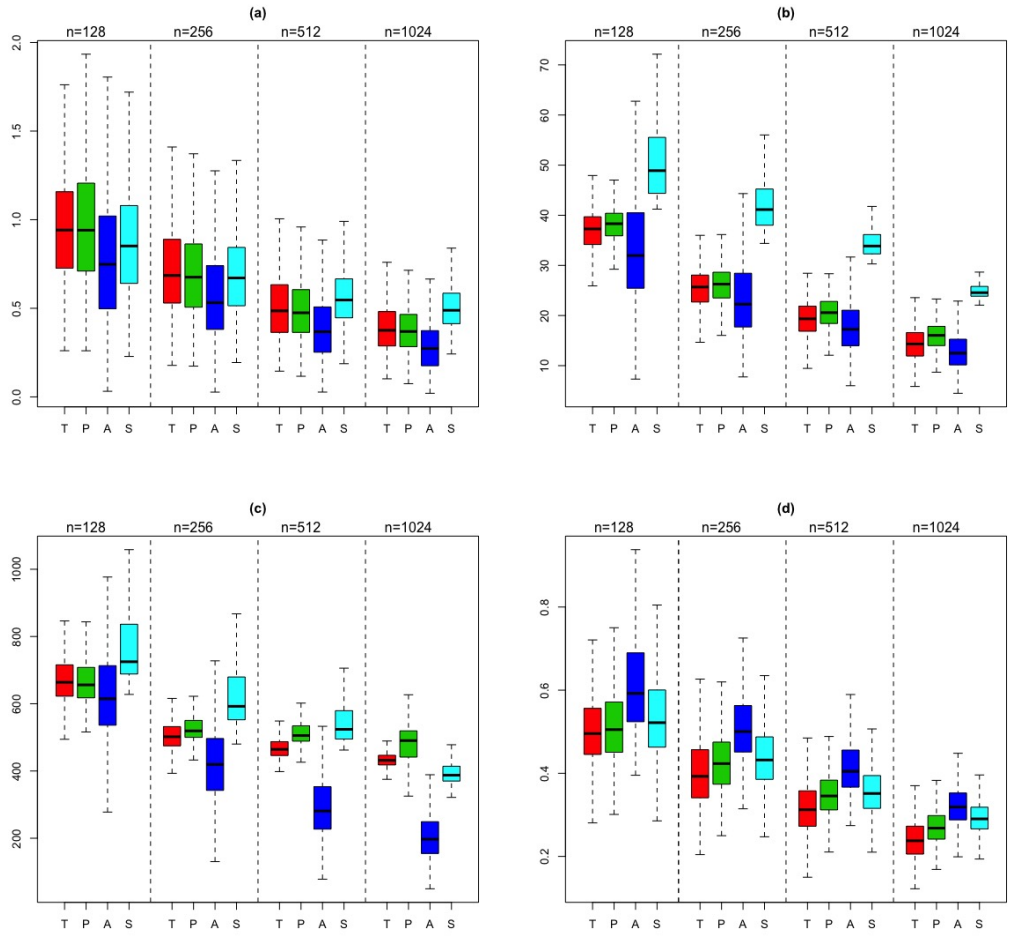


Figure 3.2: The panels show boxplots of the the L_1-errors from the 1,000 Monte-Carlo Replicates. The Panel (a) shows the result for AR(2) process, (b) for AR(12), (c) for AR(4) and (d) for ARMA(2,2). The letter beneath the boxplots indicate the method; T stands for PTVE, introduced in this chapter, P, S, A, indicate respectively PLE, SPE and ARE methods.

3.4.2 Analysis of the Sunspot Dataset

Here, we analyze the well-known yearly sunspot dataset, which consists of the annual average value of the daily index of the number of sunspots for the year 1700-1987. Following the data processing in [17], we take the square root transformation and subtract the mean to make the data look more symmetric and stationary. We set $\lambda = 0.039594$ for PTVE such that this estimator has the same Whittle likelihood value of the PLE. Apply the four procedures to produce the corresponding spectral estimates. The estimates are plotted in Figure (3.3). PTVE and PLE reveal a large peak of the spectral density at about $\omega_0 = 0.0903 \times 2\pi$ which consist with the peak in periodogram $I_N(\omega)$. However, the other two estimates display a peak at about $\omega = 0.09375 \times 2\pi$ which is different with the peak in periodogram. Also, the peak indicates a strong periodic cycle of period length $2\pi/\omega_0 \approx 11$.

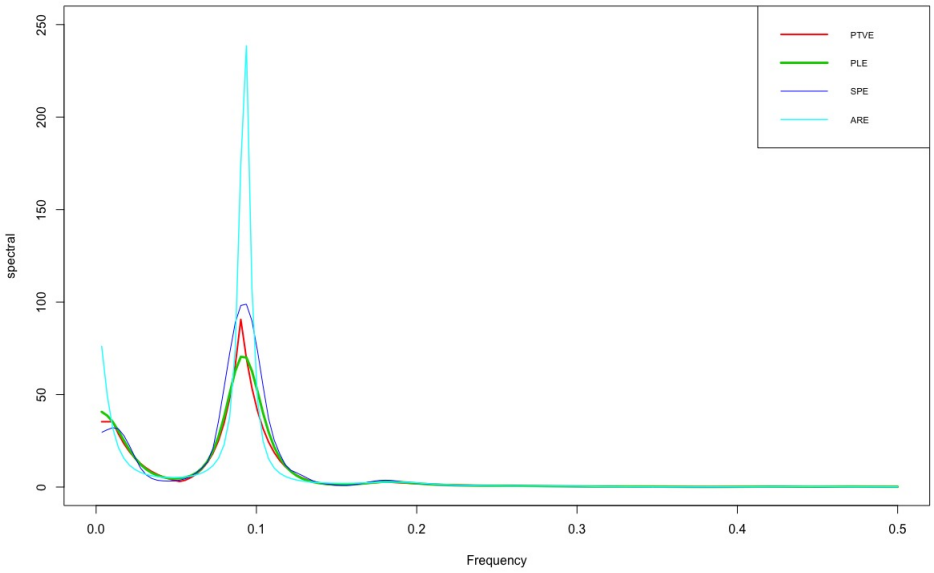


Figure 3.3: All four spectral estimates of the square root transformed Sunspot Data.

3.4.3 Analysis of the Váh River Dataset

In this study, we analyze the average yearly discharge time series of the Váh River, see 3.4, Liptovský Mikuláš water gauge (1931-2002, 72 measurements). The Váh River is the longest river in Slovakia. It is created by the confluence of tributaries Biely and Čierny Váh. Biely Váh flows from the hillside of Kriváň in High Tatras, Čierny Váh originates under Král'ová Hol'a hill in Low Tatras. The Figure 3.5 shows the four estimators, comparing with the periodogram. Previous result showed that there is a strong period of 3.5 years in the Váh river dataset. Clearly, the periodogram reveals a large peak of the spectral density at $\omega_1 = 0.27778 \times 2\pi$, which corresponds to the period length $2\pi/\omega_1 = 3.6$ closed to the period 3.5 years. We obtain the ARE based on the MLE method as well as the AIC criterion, and apply the automatic selecting smoothing parameter for PLE. Then, we select the tuning parameter $\lambda = 0.059914$. PTVE reveals a peak of the spectral density at ω_1 consistent with the peak of the periodogram and previous result. However, the SPE and ARE show a little move of the peak, see the Figure 3.5. From the result, total variation penalty method appears to have some distinct advantages when estimating spectral density functions with sharply defined features, such as locally peaks.

We are grateful to Dr. Pavla Pekárová for kindly sending us the Váh river data.

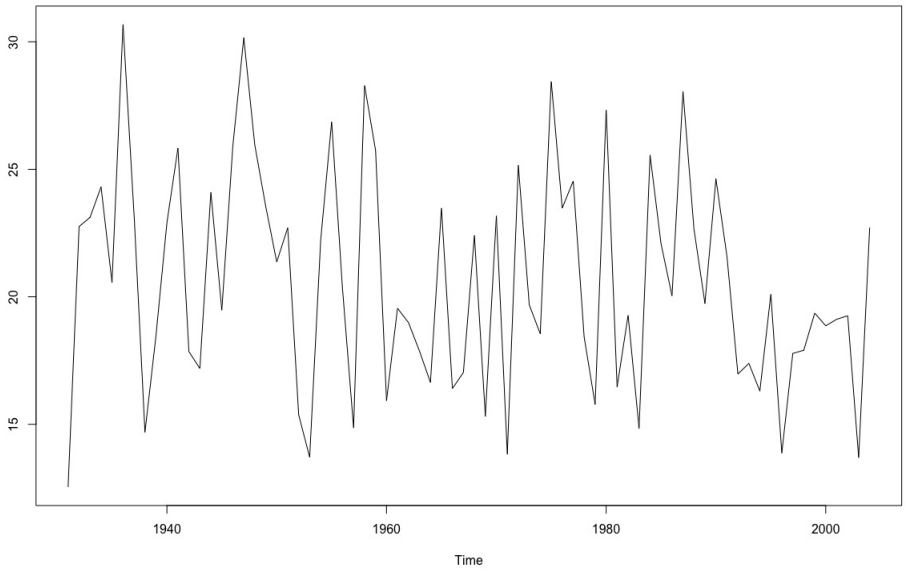


Figure 3.4: The yearly discharge time series of the Váh, period 1931-2002.

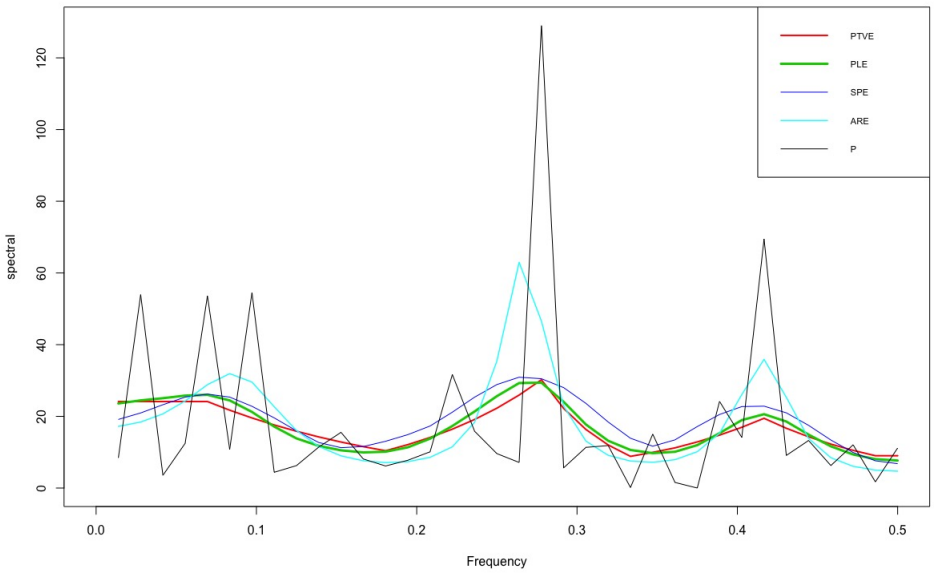


Figure 3.5: All four spectral estimates of the Váh Data. The letter P stands for periodogram. The PTVE and periodogram reveal a peak at $0.27778 \times 2\pi$. The PLE, ARE and SPE indicate a peak at $\omega = 0.27778 \times 2\pi$, $0.26389 \times 2\pi$ and $0.26389 \times 2\pi$, respectively.

Chapter 4

Sparse Wavelet Quantile Regression with Multiple Predictive Curves

We propose a penalized quantile regression method for fitting functional linear models with scalar outcomes and multiple functional predictors. The functional data are approximated by the wavelet basis, and the sparse Group Lasso penalty is imposed to control the smoothness of coefficient functions and the sparseness of the model. By utilizing wavelet bases, the approach can be extended to the setting of two-dimensional predictors. We transfer our problem to an equivalent standard second-order cone program and solve it. We also discuss asymptotic properties of the proposed estimation, investigate finite sample performance with simulation studies, and illustrate its application using a real data set from ADHD studies.

4.1 Introduction

Quantile regression, as introduced by Koenker and Bassett [41], plays an important role in contemporary statistical learning and scientific discoveries. It has been

widely used in various areas of economics [42], genetics [8], ecology [11] and other disciplines. As an extension of ordinary least squares regression which focus on the conditional means, quantile regression aims at estimating either the conditional median or other quantiles of the response variable. There are at least three advantages to consider conditional quantiles instead of conditional means. First, quantile regression, in particular median regression, offers a more robust objective compared with mean regression, in a sense that it is more resistant against outliers in responses. It is more efficient than mean regression when the errors follow a distribution with heavy tails. Second, quantile regression is capable of dealing with hereroscedasticity, the situation where the error variances depend on some covariates. Finally, quantile regression can give a more complete picture on how the responses are affected by the covariates especially when the tail behavior is conditional on covariates. The monograph by Koenker [40] provides an excellent summary of the history and recent development in quantile regression.

Consider the functional linear quantile regression model in which the conditional quantile of the response is modeled as a linear function of a set of scalar and functional covariates. In particular, for given $\tau \in (0,\ 1)$, the functional linear quantile regression model is of the form

$$Q_\tau(y_i|\boldsymbol{u}_i, \boldsymbol{x}_i(t)) = \alpha_\tau + \boldsymbol{u}_i^T\boldsymbol{\gamma}_\tau + \int_0^1 \boldsymbol{x}_i^T(t)\boldsymbol{\beta}_\tau(t)dt \quad \text{for i} = 1, \ldots, \text{n}, \tag{4.1}$$

where $Q_\tau(y_i|\boldsymbol{u}_i, \boldsymbol{x}_i(t))$ is the τ-th conditional quantile of y_i given covariates $\boldsymbol{u}_i$ and $\boldsymbol{x}_i(t)$, α_τ is the intercept, $\boldsymbol{\gamma}_\tau = (\gamma_{1\tau}, \ldots, \gamma_{p\tau})^T$ is a $p \times 1$ vector of scalar coefficients, $\boldsymbol{u}_i(t) = (u_{i1}, \ldots, u_{ip})^T$ is a $p\times1$ vector of scalar covariates, $\boldsymbol{\beta}_\tau(t) = (\beta_{1\tau}(t), \ldots, \beta_{m\tau}(t))^T$ is a $m \times 1$ vector of functional coefficients, and $\boldsymbol{x}_i(t) = (x_{i1}(t), \ldots, x_{im}(t))^T$ is a $m \times 1$ vector of functional covariates. This model is an extension of functional lin-

ear regression to the quantile regression case. To facilitate the estimation of $\beta_\tau(t)$, we usually require it to satisfy certain smoothness conditions and restrict it onto a functional space. For example, we may require that its second derivative exist and it belong to the Sobolev space $\mathbb{W}^{2,2}$; see Yuan and Cai [88]. Even in such a case, the estimation is still an infinite dimensional problem.

To estimate functional coefficients $\beta_\tau(t)$, the common practice is to project each functional element of $\beta_\tau(t)$ into a functional space with a finite functional basis. Most of the existing methods are based on the functional principal component basis ([14]; [58]; [23]; [12]; [32]), and partial least square basis [22]. The success of these approaches hinges on the availability of a good estimate of the functional principal components for functional covariates or the qualitative responses. We can also generate the basis independently of the data to avoid these constraints, such as the B-spline basis of Cardot, Ferraty, and Sarda [13], wavelet basis of [92] and others.

Our choice here is to represent the functional covariates in terms of the wavelet basis. Although using a wavelet basis initially provides no dimension reduction, the method we use to achieve dimension reduction relies on this sparsity property. For a large variety of functions, we know that the wavelet decomposition allows good representation of the function by using only a relatively small number of coefficients. Moreover, wavelets are particularly good at handling sharp, highly localized features, including feature changes in space or time. Also, the wavelet transform is computationally efficient and it is easy to extend our approach to multidimensional functional variables by using the tensor product techniques.

Since wavelet bases are well suited for sparse representation of functions, recent work considered combining them with sparsity-inducing penalties for regression with functional predictors by Zhao, Ogden, and Reiss [92], Wang, Nan, Zhu,

et al. [83] and Zhao, Chen, and Ogden [91]. All of these have focused on L_1 penalization, also known as the Lasso, in the case of one functional predictor under the wavelet domain. However, in many real-world problems, it is common to generate hundreds or thousands of functional explanatory variables for one subject. Many of the functional predictors may be unrelated to the responses. In these cases, we can impose the shrinkage penalties on the effects of functional predictors to achieve model selection and enhance interpretability and predictive capability. Gertheiss, Maity, and Staicu [30] proposed an approach using a penalized likelihood method that simultaneously controls the sparsity of the model and the smoothness of the corresponding coefficient functions. Oliva, Póczos, Verstynen, *et al.* [60] presented the functional shrinkage and selection operator (FussO), a functional analogue to the Lasso, which efficiently finds a sparse set of functional response covariates to regress a real-valued response against. The general idea behind these methods is Group Lasso technique, regularization of each functional coefficient as a whole. We extend these ideas to the setting of functional predictors by using sparse Group Lasso. The L_1 penalty can help us to identify the sparse representation of the relevant functional coefficients and smooth these coefficients. Note that without the smoothing property, interpretation of the influence of functional predictors on the response is meaningless.

There are two major contributions of this chapter. First, we develop a variable selection method for functional linear quantile model (4.1) based on penalized quantile regression with wavelet basis. Our variable selection method combines selection of the functional covariates and estimation of the smooth effects for the chosen subset of covariates. By using wavelet basis, it is possible to extend our method to deal with multidimensional functional covariates, such as 2D image predictors. Under homoscedasticity assumption, we extend our techniques to penalized com-

posite quantile regression. An efficient algorithm has been developed. Second, we investigate asymptotic properties of our estimator when the functional covariates are increasingly densely observed as the sample size increases.

The rest of the chapter is organized as follows. In Section 4.2, we review some necessary background on wavelets and convert the functional linear quantile regression problem to a multiple linear quantile regression problem. In Section 4.3, we develop one algorithm to solve our optimization problems and discuss the selection of tuning parameters. Consistency properties of our estimation are provided in Section 4.4. The proposed method is illustrated numerically in simulation results and an application to a real data set from an ADHD study in section 4.5.

4.2 Wavelet-based Sparse Group LASSO

In this section, we first provide some necessary background on wavelets and then project our data into the space generated by the wavelet basis. After that, we take advantage of the sparse representation of the functions in the wavelet domain to derive our penalized objective function.

4.2.1 Some Background on Wavelets

Wavelets are basis functions that can be used to efficiently approximate particular classes of functions with few nonzero wavelet coefficients. The construction of a wavelet basis for $L^2[0,1]$ starts with two orthonormal basic functions: a scaling function, $\varphi(t)$, and a wavelet function, $\psi(t)$, satisfying $\int_0^1 \varphi(t) = 1$ and $\int_0^1 \psi(t) = 0$. The dilated and translated versions of the scaling and wavelet function are given by

$$\varphi_{jk}(t) = \sqrt{2^j}\varphi(2^j t - k), \quad \psi_{jk}(t) = \sqrt{2^j}\psi(2^j t - k),$$

where the integer j refers to the dilation and k is an integer that serves as a translation index. Given a primary resolution level j_0, the wavelet bases define an orthonormal bases of $L^2[0,1]$ via dilation and translation of φ and ψ as the collection

$$\left\{\{\varphi_{j_0,k}\}_{0\le k\le 2^{j_0}-1},\ \{\psi_{j,k}\}_{j_0\le j,\ 0\le k\le 2^j-1}\right\}. \tag{4.2}$$

Moreover, the coefficient functions $\beta_{l\tau}(t)$ in (4.1) can be expanded in the above wavelet series by

$$\beta_{l\tau}(t) = \sum_{k=0}^{2^{j_0}-1} a^l_{j_0k}\varphi_{j_0k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} d^l_{jk}\psi_{jk}(t) \quad \text{for } 1 = 1,\ldots,\mathrm{m}, \tag{4.3}$$

where $a^l_{j_0k} = \int_0^1 \beta_{l\tau}(t)\varphi_{j_0,k}(t)dt$ and $d^l_{jk} = \int_0^1 \beta_{l\tau}(t)\psi_{jk}(t)dt$, which are the approximation coefficients at the coarsest resolution j_0 and the detail coefficients that characterize the fine structures of $\beta_{l\tau}(t)$, respectively.

Suppose that the functional predictors are discretely observed in the same set of $N = 2^J$ equally spaced points. In such a case, we can not extract the local information of the curve finer than the resolution level $J-1$. We obtain the wavelet coefficients via the discrete wavelet transform (DWT); this is not a transformation of the curves in $L^2[0,1]$, but instead it is the transformation of the vector of discrete observations in the curve. Denote $\boldsymbol{W}$ be an $N\times N$ matrix associated with the orthonormal wavelet basis. By using DWT, the functional predictors can be represented by a set of N wavelet coefficients:

$$\boldsymbol{x}_i = (\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{im})^T = (\boldsymbol{W}^T\boldsymbol{C}_{i1},\ldots,\boldsymbol{W}^T\boldsymbol{C}_{im})^T, \tag{4.4}$$

where $\boldsymbol{C}_{il}$ is an $N\times 1$ vector of wavelet coefficients from DWT of $x_{il}(t)$. Similarly,

the wavelet series of coefficient functions $\boldsymbol{\beta}_\tau$ can be written as

$$\boldsymbol{\beta}_\tau = (\boldsymbol{\beta}_{1\tau}, \ldots, \boldsymbol{\beta}_{m\tau}) = (\boldsymbol{W}^T\boldsymbol{\theta}_{1\tau}, \ldots, \boldsymbol{W}^T\boldsymbol{\theta}_{m\tau})^T, \tag{4.5}$$

where $\boldsymbol{\theta}_{l\tau}$ is an $N \times 1$ vector of wavelet coefficients from DWT of $\beta_{l\tau}(t)$. Hereafter, we denote $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_m^T)^T$ and $\boldsymbol{C}_i = (\boldsymbol{C}_{i1}^T, \ldots, \boldsymbol{C}_{im}^T)^T$ be $mN \times 1$ vectors, and $p = mN + q$.

In this chapter, we require an orthonormal wavelet basis on [0, 1], such as those in Daubechies' family. Penalized regression methods introduced to functional linear model are based on the fact that a large class of functions can be well represented by relatively few nonzero wavelet coefficients.

4.2.2 Model Estimation

We exploit the sparseness of wavelet decomposition and tackle our problem by applying L_1 and L_2 penalty to the wavelet coefficients of $\boldsymbol{\beta}(t)$ when fitting model (4.1). Plugging the wavelet expansion (4.4) and (4.5) into (4.1) and using the orthonormality of $\boldsymbol{W}$, we obtain a discrete version of model (4.1); expressed as

$$Q_\tau(y_i|\boldsymbol{u}_i, \boldsymbol{x}_i(t)) = \alpha_\tau + \boldsymbol{u}_i^T\boldsymbol{\gamma}_\tau + \boldsymbol{C}_i^T\boldsymbol{\Theta}_\tau + \varepsilon_i^* \quad \text{for i} = 1, \ldots, \text{n}, \tag{4.6}$$

where ε_i^* is the error term due to the replacement of integrals by averages.

We now aim at estimating the coefficients in (4.6). Meanwhile, we want to detect a few number of functional predictors that are effective in predicting responses and estimate the smooth effects of the chosen subset of functional predictors; sparse Group Lasso penalty is useful for this purpose. The Group Lasso part achieves selection of the functional predictors based on the preset grouping structure of the

parameters $\boldsymbol{\theta}_{j\tau}$s. The Lasso part imposes smoothness of the coefficients $\boldsymbol{\theta}_{j\tau}$s by controlling the number of basis functions. In fact, this sparsity is also implied by using the wavelet basis. Thus, the parameters α_τ, $\boldsymbol{\gamma}_\tau$, and $\boldsymbol{\Theta}_\tau$ can be estimated by minimizing the loss function of quantile regression with some shrinkage constraints. That is

$$(\hat{\alpha}_\tau, \hat{\boldsymbol{\gamma}}_\tau, \widehat{\boldsymbol{\Theta}}_\tau) = \underset{\alpha,\boldsymbol{\gamma},\boldsymbol{\Theta}}{\arg\min} \sum_{i=1}^{n} \rho_\tau(y_i - \alpha - \boldsymbol{u}_i^T\boldsymbol{\gamma} - \boldsymbol{C}_i^T\Theta) + P_{\lambda_1,\lambda_2}(\boldsymbol{\Theta}), \tag{4.7}$$

where $P_{\lambda_1,\lambda_2}(\Theta) = \lambda_1 \sum_{l=1}^{m} \|\boldsymbol{\theta}_l\|_1 + \lambda_2 \sum_{l=1}^{m} \|\boldsymbol{\theta}_l\|_2$ is the penalty function indexed by the two tuning parameters $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$.

To combine the information from multiple quantile levels, in particular, under the condition that the effects of covariates are piecewise constant or continuous across different quantile levels, Zou and Yuan [94] proposed composite quantile regression, to simultaneously consider multiple quantile regression models at different levels. Such a regression is more efficient by combining the strength across multiple quantile regression models. In particular, under the homoscedasticity assumption (the model errors do not depend on covariates), all conditional regression quantiles are parallel and we have the same coefficients but different intercepts. Similarly, we can add the sparse Group Lasso penalty in composite quantile regression models. Denote $0 < \tau_1 < \cdots < \tau_k < 1$. The composite quantile regression estimates of $(\boldsymbol{\alpha}, \boldsymbol{\Theta}, \boldsymbol{\gamma})$ with a sparse Group Lasso penalty are then

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\Theta}}) = \underset{\boldsymbol{\alpha},\boldsymbol{\Theta},\boldsymbol{\gamma}}{\arg\min} \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \alpha_k - \boldsymbol{u}_i^T\boldsymbol{\gamma} - \boldsymbol{C}_i^T\Theta) + P_{\lambda_1,\lambda_2}(\boldsymbol{\Theta}), \tag{4.8}$$

where $\boldsymbol{\alpha} = (\alpha_{\tau_1}, \ldots, \alpha_{\tau_{K-1}})$ is the intercept vector. Typically, we use the equally spaced quantiles: $\tau_k = \frac{k}{K}$, $k = 1, 2, \ldots, K-1$. For notation simplicity, we denote

$L_n(\boldsymbol{\alpha}, \Theta, \boldsymbol{\gamma}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \alpha_k - \boldsymbol{u}_i^T \boldsymbol{\gamma} - C_i^T \Theta)$ to be the objective function of composite quantile regression, which is a mixture of the objective functions from different quantile regression models. Since (4.7) is a special case of (4.8), we focus hereafter on the composite quantile regression with the sparse Group Lasso.

4.3 Implementations

Two essential points need to be addressed in solving the optimization problem (4.8). First, an efficient optimization algorithm must be derived to cope with the large scale case at low computational cost. Moreover, a practical rule is needed for choosing tuning parameters λ_1 and λ_2.

4.3.1 Algorithm

In this section, we transform the original problem to the standard form of second-order cone programing (SOCP). We can then use some R packages to solve it.

Denote $\Theta^+ = (\boldsymbol{\theta}_1^+, \ldots, \boldsymbol{\theta}_m^+)$ and $\Theta^- = (\boldsymbol{\theta}_1^-, \ldots, \boldsymbol{\theta}_m^-)$ to be the positive and negative part of Θ in the element-wise sense. Then we have $\Theta = \Theta^+ - \Theta^-$ and $\|\Theta\|_1 = \|\Theta^+\|_1 + \|\Theta^-\|_1$. Each term in the objective function of composite quantile regression can be also written as a linear constraint; for example,

$$-u_{ki}^- \leq y_i - \alpha_k - \boldsymbol{u}_i^T \boldsymbol{\gamma} - \boldsymbol{C}_i^T (\Theta^+ - \Theta^-) \leq u_{ki}^+, \quad u_{ki}^- \geq 0,\ u_{ki}^+ \geq 0,$$

where u_{ki}^+ and u_{ki}^- are two nonnegative slack variables. Finally, we deal with the L_2 type penalty in $P_{\lambda_1,\lambda_2}(\Theta)$. We rewrite $\|\boldsymbol{\theta}_l\|_2$ in the constrain condition by introducing

m slack variables to construct m standard quadratic cones; for instance

$$\mathbb{Q}_l^{2N+1} = \left\{(\nu_l, \boldsymbol{\theta}_l^+, \boldsymbol{\theta}_l^-) \in \mathbb{R}^{2N+1} | \nu_l \geq \sqrt{\|\boldsymbol{\theta}_l^+\|_2^2 + \|\boldsymbol{\theta}_l^-\|_2^2}\right\},$$

where ν_l is a slack variable. Consequently, our convex optimization problem (4.8) can be expressed as the following standard SOCP,

$$\begin{aligned} \min \quad & \sum_{k=1}^{K}\sum_{i=1}^{n}(\tau_k u_{ki}^+ + (1-\tau_k)u_{ki}^-) + \lambda_1 \sum_{l=1}^{m}(\|\boldsymbol{\theta}_l^+\|_1 + \|\boldsymbol{\theta}_l^-\|_1) + \lambda_2 \sum_{l=1}^{m}\nu_l \\ \text{subject to} \quad & -u_{ki}^- \leq y_i - \alpha_k - \boldsymbol{u}_i^T\boldsymbol{\gamma} - \boldsymbol{C}_i^T(\Theta^+ - \Theta^-) \leq u_{ki}^+ \\ & \sqrt{\|\boldsymbol{\theta}_l^+\|_2^2 + \|\boldsymbol{\theta}_l^-\|_2^2} \leq \nu_l \\ & \Theta^+ \geq 0,\ \Theta^- \geq 0,\ \nu_l \geq 0,\ u_{ki}^- \geq 0,\ u_{ki}^+ \geq 0. \end{aligned} \tag{4.9}$$

Once we get the solutions of Θ^+ and Θ^-, then we can return back to Θ by the equation $\Theta = \Theta^+ - \Theta^-$.

The above optimization problem is the exact equivalent of the optimization problem (4.8) by the fact that either $\theta_{l_i}^+ = 0$ or $\theta_{l_i}^- = 0$ would be held in the above optimization problem. Otherwise, suppose there exist l and s_0 such that $\theta_{l_{s_0}}^+ > 0$ and $\theta_{l_{s_0}}^- > 0$, then we can reset

$$\theta_{l_{s_0}}^+ = \begin{cases} 0 & if\ \theta_{l_{s_0}}^+ < \theta_{l_{s_0}}^-, \\ \theta_{l_{s_0}}^+ - \theta_{l_{s_0}}^- & otherwise, \end{cases} \quad \theta_{l_{s_0}}^- = \begin{cases} 0 & if\ \theta_{l_{s_0}}^+ > \theta_{l_{s_0}}^-, \\ \theta_{l_{s_0}}^- - \theta_{l_{s_0}}^+, & otherwise. \end{cases}$$

Under taking the new $\theta_{l_{s_0}}^+$ and $\theta_{l_{s_0}}^-$, we can see the value of objective function (4.9) in the first part has no changes, however the value of rest part in the objective function decrease, which is a contradiction. Therefore, these two expressions, $(\|\boldsymbol{\theta}_l^+\|_1 + \|\boldsymbol{\theta}_l^-\|_1)$ and $\sqrt{\|\boldsymbol{\theta}_l^+\|_2^2 + \|\boldsymbol{\theta}_l^-\|_2^2}$, are in fact $\|\boldsymbol{\theta}_l^+ - \boldsymbol{\theta}_l^-\|_1$ and $\|\boldsymbol{\theta}_l^+ - \boldsymbol{\theta}_l^-\|_2$ which are the L_1 and L_2

norms of the θ_l, respectively. This implies (4.8) and (4.9) are identical.

Since there are various packages in R to solve the SOCP problem, such as the package **Rmosek**, we can obtain the estimations by solving the equivalent form (4.9).

4.3.2 Selection of Tuning Parameters

The proposed method involves tuning parameters that control the model complexity. Specifically, λ_1 and λ_2 control the sparsity within group and group-wise sparsity, respectively. In practice, for problems where we expect strong overall sparsity and would like to encourage grouping we can set relative large λ_1 and small λ_2. In contrast, if we expect strong group-wise sparsity, but only mild sparsity within group, we use small λ_1 and large λ_2. Until now, we do not have a certain model selection for λ_1 and λ_2. In general, people usually pre-specify a finite set of values for the regularization parameters, then use either a validation dataset or a certain model selection criterion to pick the regularization parameters. Although many existing criteria including AIC and $K-$fold cross validation could be potentially employed to select the tuning parameters, Wang, Li, and Tsai [81] and Zhang, Li, and Tsai [89] showed that the tuning parameters selected by AIC and cross validation may fail to consistently identify the true model. Zhang, Li, and Tsai [89] introduced employing the generalized information criterion (GIC), encompassing the commonly used AIC and BIC, for selecting the regularization parameter. Recently, Zheng, Peng, and He [93] proposed a GIC-type uniform selector of the tuning parameters for a set of quantile levels to avoid some of the potential problems with model selection at individual quantile levels under the high dimension setting. Motivated by these results, we select the practically optimal tuning parameters by minimizing a

GIC given as

$$(\hat{\lambda}_1, \hat{\lambda}_2) = \arg\min_{\lambda_1,\lambda_2} \frac{1}{K} \sum_{k=1}^{K} \ln\left(\frac{1}{n} \sum_{i=1}^{n} \rho_{\tau_k} (y_i - \hat{y}_{ki}) \right) + \phi_n \|\widehat{\boldsymbol{\Theta}}_{\lambda_1,\lambda_2}\|_0,$$

where $\widehat{\boldsymbol{\Theta}}_{\lambda_1,\lambda_2}$ is the part of solution in the problem (4.8) by setting tuning parameter as λ_1 and λ_2. Here $\|\cdot\|_0$ denotes L_0 norm (total number of non-zero elements in a vector), and $\hat{y}_{ki}$ is calculated from (4.6) with $\tau = \tau_k$.

In addition, we can also select the tuning parameters by the validation set, see [86], [50]. Simulation studies in Section 5 demonstrate satisfactory behavior of the proposed parameters selection method, compared it with the validation dataset method to select tuning parameters.

4.4 Consistency of the Wavelet-based Group Lasso Estimator

In this section, we investigate the behavior of our wavelet-based spare Group Lasso estimator when both $n \to \infty$ and $N \to \infty$, meaning that the sample size n increases and the curves, $\boldsymbol{x}_i(t)$s, are also more densely observed, respectively. Let N_n be the number of discrete points at which the functional predictors $\boldsymbol{x}_{ij}(t)$ are observed with the sample size n. In order to derive the convergence rate of $\hat{\beta}_j(t)$ to $\beta_j(t)$, we need the following assumptions as in [92].

A1 $\varepsilon_1, \ldots, \varepsilon_n$ are independent with the same distribution F, with a density function $f(\cdot)$ that is bounded away from zero and has a continuous and uniformly bounded derivative.

A2 There are two constants c_1 and c_2 such that the regression matrix $\boldsymbol{A}$

satisfies the eigenvalue condition

$$0 < c_1 < \lambda_{\min}(\frac{1}{n}\boldsymbol{A}^T\boldsymbol{A}) \leq \lambda_{\max}(\frac{1}{n}\boldsymbol{A}^T\boldsymbol{A}) < c_2 < \infty,$$

where the ith row of $\boldsymbol{A}$ is $\boldsymbol{A}_i = (1, \boldsymbol{C}_i^T, \boldsymbol{u}_i^T)$.

A3 There exists a constant M such that $\|\boldsymbol{A}_i\|_2 < M$ for all i.

A4 $\beta_j(t)$ is a d times differentiable function in the Sobolev sense and the wavelet basis has v vanishing moments, where $v > d$.

A5 $\lambda_1 = O(\sqrt{n})$, $\lambda_2 = O(\sqrt{n})$ and $n = O_p(N^{4d})$.

A6 $\frac{N_n}{n} \to 0$.

The above regularity conditions are reasonable for quantile regression with possibly growing number of parameters. Condition A1 is standard for quantile regression, see [41], [40], [90]. Condition A2 is a classical condition that has been assumed in the linear model literature. Condition A3 can be found in [92]. Condition A4 guarantees that the space spanned by the wavelet basis is good to estimate the smooth function, for example the approximation error goes to 0. The wavelet has v vanishing moments if and only if its scaling function φ can generate polynomials of degree smaller than or equal to v. Then we have the following theorem.

Theorem 4.4.1. *Let $\hat{\beta}_j$ be the estimator resulting from (4.8) and β_j is the truth coefficient function. If the assumptions A1-A6 hold, then*

$$\|\hat{\beta}_j - \beta_j\|_2^2 = O_p(\frac{N_n}{n}) + o_p(\frac{1}{N_n^{2d}}).$$

Proof. First, we introduce some notation. The orthonormal wavelet basis set of $L^2[0,1]$ is defined as $\{\varphi_{j_0k}, k = 1, \ldots, 2^{j_0}\} \cup \{\psi_{jk},\ j \geq j_0,\ k = 1, \ldots, 2^j\}$. Without

loss of generality, the wavelet bases are ordered according to the scales from the coarsest level J_0 to the finest one. Let $\mathbb{V}_{N_n} := \text{Span}\{\varphi_1, \ldots, \varphi_{N_n}\}$ be the space spanned by the first N_n basis function, for example, if $N_n = 2^{j_0+t}$, then the collection $\{\varphi_{j_0k}, k = 1\ldots, 2^{j_0}\} \cup \{\psi_{jk}, j_0 \leq j \leq j_0 + t - 1, k = 1, \ldots, 2^j\}$ is the basis of $\mathbb{V}_{N_n}$. Let $\boldsymbol{\theta}^j_{N_n}$ be an $N_n \times 1$ parameter vector with elements $\theta^j_k = \langle \beta_j(t), \varphi_k \rangle$. In addition, let $\beta^j_{N_n}$ be the functions reconstructed from the vector $\boldsymbol{\theta}^j_{N_n}$. Here $\beta^j_{N_n}$ is a linear approximation to β_j by the first N_n wavelet coefficients, while $\hat{\beta}_j$ denotes the functions reconstructed from the wavelet coefficients $\hat{\boldsymbol{\theta}}_j$ from (4.8).

By the Parseval theorem, we have $\|\hat{\beta}_j - \beta_j\|^2_{L_2} = \|\hat{\theta}^j_{N_n} - \theta^j_{N_n}\|^2_2 + \sum_{k=N_n+1}^{\infty} \theta^{j2}_k$. To derive the convergence rate of $\hat{\beta}_j$ to β_j, we bound the error in estimating $\beta^j_{N_n}$ by $\hat{\beta}_j$ and the error in approximating β_j by β_{N_n}. By the Theorem 9.5 of Mallat [53], the linear approximation error goes to zero as

$$\sum_{k=N_n+1}^{\infty} \theta^{j2}_t = o(N_n^{-2d}). \tag{4.10}$$

Let $\Upsilon^0 = (\boldsymbol{\alpha}^0, \boldsymbol{\gamma}^0, \Theta^0)$ be the true coefficients with $\Theta^0 = (\boldsymbol{\theta}^1_{N_n}, \ldots, \boldsymbol{\theta}^m_{N_n})$. To obtain the result, we show that for any given $\varepsilon > 0$, there exists a constant C such that

$$\Pr\left\{ \inf_{\|\boldsymbol{v}\|=C} L_n\left(\Upsilon^0 + r_n \boldsymbol{v}\right) + P_{\lambda_1,\lambda_2}\left(\Theta^0 + r_n \boldsymbol{v}_\theta\right) > L_n\left(\Upsilon^0\right) + P_{\lambda_1,\lambda_2}\left(\Theta^0\right) \right\} \geq 1 - \varepsilon, \tag{4.11}$$

where $r_n = \sqrt{N_n/n}$ and $\boldsymbol{v} = (v_1, \ldots, v_k, \boldsymbol{v}_\gamma, \boldsymbol{v}_\theta)$ is a vector with the same length of vector Υ^0. This implies that there exists a local minimizer in the ball $\{\Upsilon^0 + r_n v : \|v\| \leq C\}$ with probability at least $1 - \varepsilon$. Hence, there is a local minimizer $\widehat{\Upsilon}$ such that $\|\widehat{\Upsilon} - \Upsilon^0\| = O_p(r_n)$.

To show (4.11), we compare $L_n(\boldsymbol{\Upsilon}^0) + P_n(\boldsymbol{\Theta}^0)$ with $L_n(\boldsymbol{\Upsilon}^0 + r_n \boldsymbol{v}) + P_n(\boldsymbol{\Theta}^0 + r_n \boldsymbol{v}_\theta)$.

By using the Knight identity,

$$\rho_\tau(u-v)-\rho_\tau(u) = -v\varrho_\tau(u) + \int_0^v (I(u \le t) - I(u \le 0))dt,$$

where $\varrho_\tau(u) = \tau - I(u < 0)$, we have

$$\begin{aligned} I &:= L_n(\boldsymbol{\Upsilon}^0 + r_n\boldsymbol{v}) - L_n(\boldsymbol{\Upsilon}^0) \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n}[\rho_{\tau_k}(e_{ki}-b_{ki}) - \rho_{\tau_k}(e_{ki})] \\ &= -\sum_{k=1}^{K}\sum_{i=1}^{n}[-b_{ki}\varrho_{\tau_k}(e_{ki})] + \sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^{b_{ki}}(I(e_{ki}\le t) - I(e_{ki}\le 0))dt \\ &= I_1 + I_2, \end{aligned}$$

where $e_{ki} = y_i - \alpha^0_{\tau_k} - \boldsymbol{u}_i^T\boldsymbol{\gamma}^0 - \boldsymbol{C}_i^T\boldsymbol{\Theta}^0$ and $b_{ki} = r_n v_k + r_n\boldsymbol{u}_i^T\boldsymbol{v}_u + r_n\boldsymbol{C}_i^T\boldsymbol{v}_\theta$. Note that $e_{ki} = \varepsilon_i - F^{-1}(\tau_k) + o(N_n^{-2d})$, hence we have $E(\varrho_{\tau_k}(e_{ki}) = o(N_n^{-2d})$. By the definition of b_{ki}, we obtain $I_1 \le r_n\|\boldsymbol{v}\|(\sum_{k=1}^{s}\|\sum_{i=1}^{n}\varrho_{\tau_k}(e_{ki})\boldsymbol{A}_i^T\|)$ and

$$\begin{aligned} E\|\sum_{i=1}^{n}\varrho_{\tau_k}(e_{ki})\boldsymbol{A}_i\|^2 &= E\|\sum_{j=1}^{mN_n+1}\sum_{i=1}^{n}\sum_{l=1}^{n}a_{ij}a_{lj}\psi_{\tau_k}(e_{ki})\psi_{\tau_k}(e_{kl})\| \\ &= O_p(nN_n), \end{aligned}$$

which leads to $E(I_1) \le O_p(r_n\sqrt{nN_n})\|\boldsymbol{v}\| = O_p(nr_n^2)\|\boldsymbol{v}\|$.

Now, we consider the expectation of I_2. Using the expression of e_{ki}, we get

$$\begin{aligned} E(I_2) &= \sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^{b_{ki}}(\Pr(e_{ki}\le t) - \Pr(e_{ki}\le 0))dt \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^{b_{ki}}(F(F^{-1}(\tau_k) + o(N_n^{-2d}) + t) - F(F^{-1}(\tau_k) + o(N_n^{-2d})))dt \\ &= \sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^{b_{ki}}(f(F^{-1}(\tau_k) + o(N_n^{-2d}))t + \frac{f'(\xi)}{2}t^2)dt, \end{aligned}$$

where ξ lies between $F^{-1}(\tau_k) + o(N_n^{-2d})$ and $F^{-1}(\tau_k) + o(N_n^{-2d}) + b_{ki}$. Since there exists M such that $\|A_i\|_2^2 < M$, we have

$$\max_{1\le i\le n} |r_n v_k + r_n \boldsymbol{C}_i^T \boldsymbol{v}_\theta| \to 0.$$

Then, the lower bound of $E(I_2)$ is of the form

$$\begin{aligned} E(I_2) &= \frac{1}{2} r_n^2 \sum_{k=1}^{K} \{[f(F^{-1}(\tau_k) + o(N_n^{-2d})) + o_p(1)](\boldsymbol{g}_k^T A^T A \boldsymbol{g}_k)\} \\ &\ge \frac{c_1 n r_n^2}{2} \|\boldsymbol{v}\|_2^2 \min_k \{f(F^{-1}(\tau_k) + o(N_n^{-2d})) + o_p(1)\}, \end{aligned}$$

where $\boldsymbol{g}_k$ is a vector, such as $\boldsymbol{g}_k = (v_k, \boldsymbol{v}_\theta^T, \boldsymbol{v}_u^T)^T$. Finally, since $r_n \to 0$ and $\|\boldsymbol{v}\|_2 \le C$, we have

$$\begin{aligned} II := P_n(\Theta^0 + r_n \boldsymbol{v}_\theta) - P_n(\Theta^0) &\le \lambda_1 r_n \|\boldsymbol{v}_\theta\|_1 + \lambda_2 r_n \sum_{j=1}^{m} \|\boldsymbol{v}_{\theta_j}\|_2 \\ &\le \lambda_1 r_n \sqrt{mN} \|\boldsymbol{v}_\theta\|_2 + \lambda_2 r_n m \|\boldsymbol{v}_\theta\|_2 \\ &= O_p(n r_n^2 \|\boldsymbol{v}_\theta\|_2). \end{aligned}$$

Since II is bounded by $r_n^2 \|\boldsymbol{v}_\theta\|_2$,we can choose a C such that the II is dominated by the term I_2 on $\|u\| = C$ uniformly. So $Q_n(\Sigma^0 + r_n u) - Q_n(\Sigma^0) > 0$ holds uniformly on $\|u\| = C$. This completes the proof. □

Under some further conditions, the prediction error bound are given in the following theorem.

Theorem 4.4.2. *Suppose $x_j(t)$ are square integrable functions on* $[0, 1]$ *and* $\boldsymbol{F}^{-1}(\tau_l) =$

0. *If the assumptions A1-A6 hold, then*

$$|\hat{y} - y|^2 = O_p(\frac{N_n}{n}) + o_p(\frac{1}{N_n^{2d}}),$$

where y is the true response from (4.1), $\hat{y} = \hat{\alpha}_l + \boldsymbol{u}^T\hat{\boldsymbol{\gamma}} + \int_0^1 \boldsymbol{x}(t)\hat{\boldsymbol{\beta}}(t)dt$.

Proof. The proof follows from 4.4.1 and the Cauchy-Schwarz inequality. We omit the details. □

4.5 Numerical Studies

4.5.1 Simulations

In this section, we compare the finite sample performance of a number of different methods, quantile Lasso and quantile Group Lasso, with regard to the functional estimation error and the prediction error. We also examine the effectiveness of the proposed modeling strategy by investigating whether our method selects functional predictors appropriately. We use least-asymmetric wavelet of Daubechies with 6 vanishing moments for both the simulation study and the real data analysis, and fix the ratio $\lambda_1/\lambda_2 = 0.5$.

The simulation in this section is based on 200 and 400 generated observations of 12 functional covariates and 2 scalar covariates with a scalar response, extending the simulation setup of [19] by including more functional predictors. In particular, the model is of the form

$$y_i \quad = \quad \alpha + \boldsymbol{u}_i^T\boldsymbol{\gamma} + \sum_{l=1}^{12}\int_0^1 x_{il}(t)\beta_l(t)dt + \sigma\varepsilon_i \quad \text{for } i = 1\ldots, n,$$

where $\boldsymbol{u}_i = (u_{i1}, u_{i2})^T$ with $u_{i1} \sim N(0,1)$ and $u_{i2} \sim$ Bernoulli(0.5), and the coefficient

$\gamma = (0.32/256,\ 0.32/256)^T$.

We adapt the setting of [19] to generate $Z_{il}(t)$ on an equally spaced grid of 256 points in $\mathbb{T}_l$ in the following way:

$$Z_{il}(t) = G_{il}(t) + \varepsilon_{il},\ \varepsilon_{il} \sim N\left(0, (.05 r_{x_{il}})^2\right),$$

where $r_{x_{il}} = \max_i(G_{il}(t)) - \min_i(G_{il}(t))$ and

$$\begin{aligned}
G_{i1}(t) &= \cos(2\pi(t - a_1)) + a_2, \mathbb{T}_1 = [0,\ 1],\ a_1 \sim N(-4,\ 3^2),\ a_2 \sim N(7,\ 1.5^2),\\
G_{i2}(t) &= b_1 t^3 + b_2 t^2 + b_3 t, \mathbb{T}_2 = [-1,\ 1],\ b_1 \sim N(-3,\ 1.2^2),\ b_2 \sim N(2, .5^2),\ b_3 \sim N(-2\ , 1),\\
G_{i3}(t) &= \sin(2(t - c_1)) + c_2 t, \mathbb{T}_3 = [0,\ \pi/3],\ c_1 \sim N(-2,\ 1),\ c_2 \sim N(3, 1.5^2),\\
G_{i4}(t) &= d_1 \cos(2t) + d_2 t, \mathbb{T}_4 = [-2,\ 1],\ d_1 \sim U(2,\ 7),\ d_2 \sim N(2, .4^2),\\
G_{i5}(t) &= e_1 \sin(\pi t) + e_2, \mathbb{T}_5 = [0,\ \pi/3],\ e_1 \sim U(3,\ 7),\ e_2 \sim N(0, 1),\\
G_{i6}(t) &= f_1 e^{-t/3} + f_2 t + f_3, \mathbb{T}_6 = [-1,\ 1],\ f_1 \sim N(4,\ 2^2),\ f_2 \sim N(-3, .5^2),\ f_3 \sim N(1, 1),\\
G_{il}(t) &= 5\sqrt{2} \sum_{j=1}^{49} \cos(j\pi t) g_j + 5h, \mathbb{T}_l = [0,\ 1],\ g_j \sim N\left(0,\ \left(\frac{1}{j+1}\right)^2\right),\ h \sim N(0, 1),\ \text{for } l = 7, \ldots, 12.
\end{aligned}$$

Collazos, Dias, and Zambom [19] considered only the first six functions as the predictors. For simplicity, we convert each interval of definition, $\mathbb{T}_l$, into $[0,\ 1]$. To introduce the correlation between each functional variables, the random data G_{il} were converted in to the function data x_{il} through the linear transformations as follows:

$$\begin{aligned}
x_{i1}(t) &= \sqrt{0.84} Z_{i1}(t) + 0.4 Z_{i6}(t), \quad x_{i2}(t) = \sqrt{0.98} Z_{i2}(t) + 0.1 Z_{i1}(t) + 0.1 Z_{i5}(t),\\
x_{i3}(t) &= \sqrt{0.84} Z_{i3}(t) + 0.4 Z_{i4}(t), \quad x_{i5}(t) = \sqrt{0.99} Z_{i5}(t) + 0.1 Z_{i2}(t),\\
x_{il}(t) &= Z_{il}(t) \quad \text{for } l = 4, 6, 7, \ldots, 12.
\end{aligned}$$

We generate the functional coefficients based on the following 4 functions:

$$
\begin{aligned}
f_1(t) &= 0.03 f(t, 20, 60) - 0.05 f(t, 50, 20), \\
f_2(t) &= 4\sin(4\pi x) - \text{sign}(x - 0.3) - \text{sign}(0.72 - x), \\
f_3(t) &= -3\cos(2\pi t) + 3\frac{e^{t^2}}{t^3 + 1}, \\
f_4(t) &= 0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin^2(2\pi t) + 0.4\cos^3(2\pi t) + 0.5\sin^3(2\pi t),
\end{aligned}
$$

where $f(t, \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1}(1-t)^{\beta-1}$. The function $f_1(t)$ has been considered in the article of [92]; HeaviSine function, f_2, is one of Donoho and Johnstone test functions of [24] and is alway studied in the wavelet papers, such as [3]; f_4 is proposed by Lin, Bondell, Zhang, *et al.* [51].

Applying the DWT for $f_1, \ldots, f_4$, we select the wavelet coefficients whose absolute value are greater than .1, then $\beta_1(t), \ldots \beta_4(t)$ are generated based on the inverse DWT of the selected coefficients, respectively. We normalized these slope functions by setting $\|\beta_l(t)\|_2 = 1$ for $l = 1, 2, 3, 4$. The rest of slope functions are zero, for example, $\beta_l(t) = 0$ for $l = 5, \ldots, 12$. Under this setting, there are only first 4 functional variables to be relevant to the response. Sparse structure also exists in the 4 slope functions. Each of functional variables are calculated at $N = 256$ equally spaced time points on $[0, 1]$, and we apply the DWT with periodic boundary correction on each of the functional predictors. The slope functions are plotted in the Figure 4.1. The error term ε is drawn from the 4 type distributions:

1. Standard normal distribution: $N(0, 1)$,

2. Mixed-variance normal distribution: $.95N(0, 1) + .05N(0, 10)$,

3. T distribution with 3 degrees of freedom: t_3,

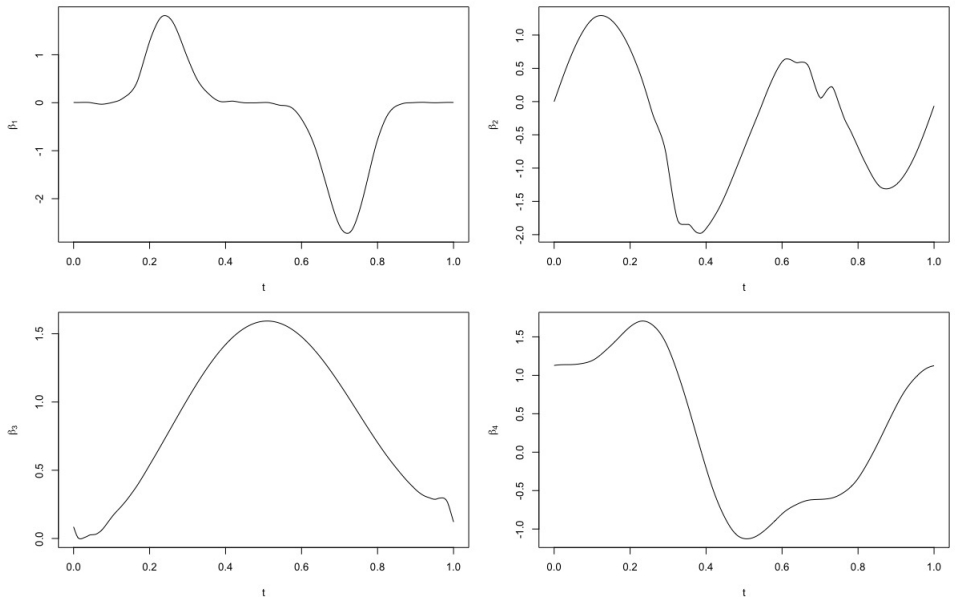


Figure 4.1: Slope fucntions β_1–β_4.

4. Standard Cauchy distribution: $C(0, 1)$.

Signal-to-noise (SNR) is an important measure that compares the level of a desired signal to the level of background noise. In this section, we choose $SNR = \frac{\mu}{\sigma}$, where μ is the mean of signal, and σ is the standard deviation of the noise. In the above four error distribution settings, we take SNR=1,5,10.

Denote the size of training data sets by n. Throughout this section, an independent tuning data set and testing data set of size n and $10n$, respectively, are generated exactly in the same way as the training data set. The tuning parameters are selected via a grid search based on two criteria. First, the tuning parameters for the three methods are selected by minimizing a GIC with $\phi_n = 5p_n,\ 5p_n$ and p_n for the quantile sparse Group Lasso, the quantile Lasso and the quantile Group Lasso, respectively, where $p_n = \log(\log(n))\log(\log(p))/(10n)$. In the real data example, we use exactly the same ϕ_n for these three methods. Second, we use the validation

set to select the gold standard (GS) tuning parameters, see [50], [94], [86] that minimizes the tuning error in terms of the check loss function evaluated on the tuning data. Similarly defined testing errors on the testing data set are reported. More explicitly, a test error refers to the average check loss on the independent testing data set. Using these $\lambda_1,\ \lambda_2$'s, we calculated the prediction errors and the mean absolute deviation with test data for each criterion.

We set $\tau = 0.5$ and use 100 Monte Carlo runs for model assessment. Since inferences on both the functional components β_j and the prediction y are of interest, we report the mean integrated square errors (MISE) of the functional coefficients, which is given by

$$\text{MISE} = \frac{1}{12}\sum_{l=1}^{12}\sum_{j=1}^{256}(\hat{\beta}_l(t_j) - \beta_l(t_j))^2,$$

as well as the individual integrated square error (ISE) for each slope function. We also evaluate the Monte Carlo averages for the proportion of correctly picked up and dropped functional components, called group accuracy (GA) that is defined as $\text{GA} = E\left(\left(|\widehat{M} \cap M_0| + |\widehat{M}^c \cap M_0^c|\right)/12\right)$ where M_0 and $\widehat{M}$ denote the set of indices of the true functional variables and selected functional variables, respectively, such as $M_0 = \{j : \beta_j(t) \neq 0\}$ and $\widehat{M} = \{j : \hat{\beta}_j(t) \neq 0\}$. Analogously, we report the variable accuracy (VA) that is similar to group accuracy through replacing the M_0 and $\widehat{M}$ as the sets of non-zero wavelet coefficients' indices. The mean absolute prediction error (PE) is assessed using an the testing data set of size $10n$ for each Monte Carlo repetition, and is defined as $\text{PE}= E(|\hat{y} - y|)$.

To save the space, we only report the results of SNR=5. For the other two SNR situations, the results are both in favor our method and displayed in the Appendix. The simulation results for SNR=5 are summarized in Table 4.1 and 4.2. As shown in Table 4.1, the quantile sparse group lasso in functional linear regression provides

better estimation and prediction than the quantile lasso and the quantile group lasso. In each error case, the gold standard criterion to select the tuning parameter always performs slightly better than the GIC. Table 4.1 also shows that as the sample size increases, the mean integrated errors and prediction error tend to be small, which is consistent with the theoretical results. For the group accuracy, the Group Lasso works better than the sparse Group Lasso and the Lasso in the majority of cases. As expected, one finds that the Lasso performs quite good in terms of variable accuracy. However, when we use the GIC to select the tuning parameters, the sparse Group Lasso outperforms other two methods regarding group accuracy and variable accuracy, especially for the larger sample size. Interestingly, in Table 4.2, the integrated squared error of $\beta_1(t)$ is always less than the error of other slope functions. Perhaps a simple explanation is that the $\beta_1(t)$ is more smooth than other functions, see Figure 4.1. This is also supported by the theoretical analysis in the previous section. Finally, similar comments essentially apply to the $SNR = 1, 10$.

4.5.2 Application to Real Data

We now apply the quantile sparse Group Lasso method to the dataset on attention deficit hyperactivity disorder from the ADHD-200 Sample Initiative Project. ADHD is the most commonly diagnosed mental disorder of childhood and can persist into adulthood. It is characterized by problems related to paying attention, hyperactivity, or impulsive behavior. The dataset is the filtered preprocessed resting state data from New York University Child Study Center using the Anatomical Automatic Labeling atlas. There are 172 equally spaced time courses in the filtering and AAL contains 116 Regions of Interests (ROIs) fractionated into functional space using nearest-neighbor interpolation. After cleaning the raw data that fails in

			GS				GIC			
n	Noise	M	MISE	GA	VA	PE	MISE	GA	VA	PE
200	1	Q	1.449	0.930	0.934	2.600	1.522	0.594	0.840	2.851
		L	3.230	0.919	0.961	2.871	3.159	0.482	0.904	3.205
		G	1.835	1.000	0.082	2.862	2.121	0.970	0.343	4.763
	2	Q	1.372	0.960	0.934	2.466	1.516	0.623	0.835	2.796
		L	3.023	0.932	0.960	2.749	3.086	0.496	0.905	3.142
		G	1.802	1.000	0.082	2.781	2.068	0.973	0.326	4.476
	3	Q	0.598	1.000	0.911	1.436	0.932	0.871	0.857	1.953
		L	1.420	0.985	0.945	1.671	2.487	0.686	0.909	2.654
		G	1.630	1.000	0.065	2.386	1.735	0.993	0.140	2.836
	4	Q	1.284	0.972	0.934	2.326	1.497	0.617	0.829	2.755
		L	2.826	0.927	0.958	2.625	3.135	0.490	0.907	3.145
		G	1.775	1.000	0.075	2.656	2.043	0.976	0.295	4.225
400	1	Q	0.925	0.989	0.915	2.095	1.224	0.911	0.920	2.220
		L	1.774	0.944	0.946	2.187	2.125	0.617	0.898	2.371
		G	1.581	1.000	0.054	2.393	2.246	0.958	0.569	5.240
	2	Q	0.842	0.995	0.911	1.954	1.105	0.967	0.937	2.058
		L	1.640	0.965	0.947	2.040	1.853	0.729	0.912	2.190
		G	1.549	1.000	0.056	2.306	2.263	0.957	0.582	5.294
	3	Q	0.157	1.000	0.875	1.001	0.272	1.000	0.930	1.108
		L	0.285	1.000	0.908	1.026	0.481	0.991	0.943	1.108
		G	1.255	1.000	0.050	1.996	1.438	0.992	0.155	2.472
	4	Q	0.738	0.996	0.909	1.785	0.995	0.983	0.939	1.910
		L	1.469	0.978	0.947	1.860	1.737	0.735	0.906	2.052
		G	1.505	0.999	0.054	2.194	2.102	0.969	0.499	4.490

Table 4.1: Simulation summary of SNR=5. The first column n is the size of training data. The second column is the type of noise. The third column is the method we used, Q for the quantile sparse Group Lasso, L for the quantile Lasso, and G for the quantile Group Lasso. GS means λ was selected by the validation method (gold standard). GIC means λ selected via the GIC criterion. MISE stands for mean integrated errors. PE, GA and VA indicate prediction error, group accuracy and variable accuracy, respectively.

			GS				GIC			
n	Noise	M	ISE1	ISE2	ISE3	ISE4	ISE1	ISE2	ISE3	ISE4
200	1	Q	0.116	0.585	0.331	0.385	0.133	0.550	0.322	0.387
		L	0.289	0.758	1.386	0.734	0.318	0.618	1.136	0.732
		G	0.351	0.675	0.359	0.447	0.372	0.728	0.370	0.648
	2	Q	0.116	0.540	0.322	0.368	0.137	0.560	0.318	0.377
		L	0.283	0.674	1.302	0.703	0.336	0.631	1.049	0.740
		G	0.348	0.665	0.349	0.438	0.367	0.714	0.362	0.621
	3	Q	0.051	0.162	0.163	0.214	0.077	0.311	0.221	0.267
		L	0.105	0.204	0.614	0.468	0.238	0.460	0.939	0.610
		G	0.332	0.605	0.297	0.395	0.342	0.632	0.313	0.446
	4	Q	0.104	0.498	0.304	0.354	0.129	0.551	0.328	0.367
		L	0.248	0.628	1.211	0.679	0.318	0.613	1.157	0.707
		G	0.345	0.657	0.343	0.427	0.367	0.709	0.366	0.597
400	1	Q	0.074	0.321	0.217	0.293	0.091	0.470	0.265	0.353
		L	0.141	0.318	0.729	0.532	0.155	0.377	0.719	0.575
		G	0.325	0.590	0.285	0.381	0.363	0.731	0.399	0.752
	2	Q	0.071	0.274	0.207	0.273	0.088	0.421	0.248	0.331
		L	0.117	0.279	0.695	0.508	0.139	0.324	0.675	0.519
		G	0.321	0.577	0.278	0.372	0.364	0.736	0.401	0.761
	3	Q	0.010	0.018	0.063	0.065	0.016	0.045	0.094	0.115
		L	0.012	0.017	0.139	0.110	0.018	0.034	0.234	0.187
		G	0.295	0.446	0.205	0.308	0.311	0.504	0.244	0.375
	4	Q	0.057	0.220	0.195	0.253	0.071	0.366	0.233	0.312
		L	0.096	0.218	0.643	0.478	0.116	0.273	0.631	0.515
		G	0.319	0.555	0.266	0.363	0.354	0.700	0.383	0.664

Table 4.2: Individual funtional L_2 error of SNR=5. The first column n is the size of training data. The second column is the type of noise. The third column is the method we used. ISE1: $\|\hat{\beta}_1 - \beta_1\|_2^2$; ISE2: $\|\hat{\beta}_2 - \beta_2\|_2^2$; ISE3: $\|\hat{\beta}_3 - \beta_3\|_2^2$; ISE4: $\|\hat{\beta}_4 - \beta_4\|_2^2$.

quality control or has missing data, we include 120 individuals in the analysis. Each 172 time courses is smoothed to 64 equally spaced time points for conveniently applying DWT. We also consider the other 8 scalar covariates, including gender, age, handedness, diagnosis status, medication status, Verbal IQ, Performance IQ and Full4 IQ. Finally, we have 59 ROIs as well as 8 scalar variables; each region has 64 equally spaced time points data. The response of interest is the ADHD index, Conners' parent rating scale-revised.

The objective of this application is to select the ROIs that significantly relate to the ADHD index. We apply the proposed functional variable selection method and compare the results with those of the quantile Lasso and the quantile Group Lasso selection procedures by using the same wavelet basis functions. First, we use the GIC criterion to select tuning parameters for each method. In a simulation of 96 bootstrap samples from the dataset, we perform variable selection using the proposed method, the quantile Lasso and the quantile Group Lasso with the fixed tuning parameter from the previous step. Finally, the empirical distribution of each slope function's L_2 norms is estimated from the 96 bootstrapping. The boxplot of the L_2 norm for each method has been shown in the Appendix, see Figure 4.2, 4.3, 4.4.

We sort the median of each slope function's L_2 norm, then do the selection by setting the thresholding level as 10^{-5}. The functional covariates we considered are the ROIs of cerebellum, temporal, vermis, parietal, occipital, cingulum and frontal, from the suggestion of data description. Table 4.3 shows the seven ROIs are selected or not by the three methods. The right part of cerebellum, temporal and vermis are selected by all methods. Our method also identify the left part. The selected important regions for each method are shown in Table 4.4. They are ordered regarding L_2 norm. Table 4.3 and Table 4.4 reveal that most of our selected ROIs come from

the seven suggested regions. Besides the seven regions, the three methods also give other three common regions; right olfactory, right supraMarginal, and right caudate. There is some evidence to suggest that the affected nodes include these other three regions. Schrimsher, Billingsley, Jackson, *et al.* [68] drew a relationship between caudate asymmetry and symptoms related to ADHD. This correlation is congruent with previous associations of the caudate with attentional functioning. The conclusion in [70] indicate the ROIs of supramarginal gyri is associated with the ADHD symptom scores. Note that the right occipital region is identified by our method, but not other two methods. These results confirm that the proposed variable selection procedure outperforms the other two methods in this dataset.

Significant ROIs	Q	L	G
Cerebellum	R L	R	R
Temporal	R L	R L	R
Vermis	R L	R L	R L
Parietal	R	R L	×
Occipital	R	×	×
Cingulum	×	L	×
Frontal	R L	L	R L

Table 4.3: Selected ROIs for the suggestion 7 regions, R and L indicate the region has been selected from the right brain and left brain, respectively. The symbol × means the brain region has not been chosen.

4.6 Appendix

M	Significant ROIs
Q	”Temporal R” ”Cerebelum R” ”Frontal R” ”Occipital R” ”Olfactory R” ”SupraMarginal R” ”Caudate R” ”Vermis” ”Cuneus L” ”Parietal R” ”Frontal L” ”Precuneus R” ”Temporal L” ”Cerebelum L” ”Precentral R”
L	”Frontal R” ”Caudate R” ”Temporal R” ”Cuneus L” ”SupraMarginal R” ”Parietal R” ”Lingual L” ”Frontal L” ”Precuneus R” ”Vermis” ”Fusiform R” ”Pallidum L” ”Olfactory R” ”Precentral R” ”Cingulum L” ”Cuneus R” ”Parietal L” ”Temporal L” ”Angular L” ”Cerebelum R”
G	”Caudate R” ”Frontal R” ”Cerebelum R” ”Vermis” ”Olfactory R ”Temporal R” ”Precentral R” ”SupraMarginal R” ”Frontal L”

Table 4.4: Selected ROIs for the ADHD-200 fMRI Dataset.

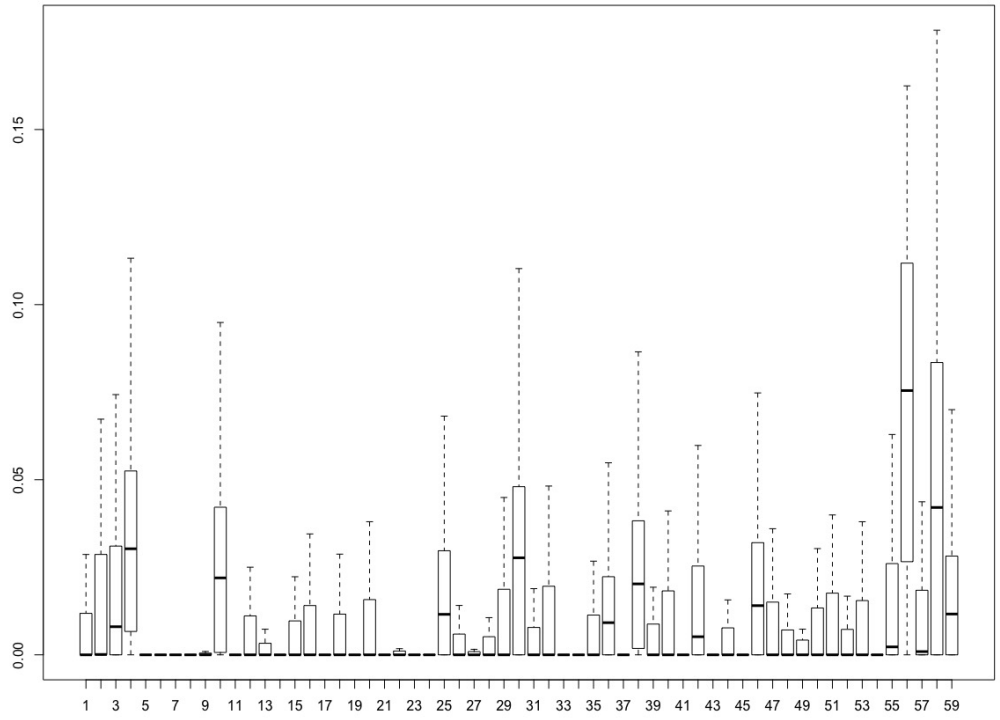


Figure 4.2: Boxplot of L_2 norm for each slope function, by using the quantile spare Group Lasso method.

			GS				GIC			
n	Noise	M	MISE	GA	VA	PE	MISE	GA	VA	PE
200	1	Q	2.426	0.860	0.959	9.557	6.361	0.480	0.854	11.720
		L	5.885	0.965	0.972	9.553	17.062	0.358	0.891	13.134
		G	2.601	0.852	0.118	9.637	4.091	0.708	0.406	12.728
	2	Q	2.322	0.876	0.958	8.833	6.013	0.509	0.857	10.968
		L	5.592	0.968	0.971	8.844	16.564	0.363	0.891	12.760
		G	2.619	0.870	0.123	8.973	4.473	0.704	0.374	11.844
	3	Q	1.063	0.994	0.930	4.200	1.594	0.891	0.908	4.774
		L	2.462	0.978	0.958	4.491	7.252	0.547	0.911	7.466
		G	1.741	1.000	0.073	4.776	3.699	0.857	0.330	7.875
	4	Q	2.252	0.925	0.958	7.967	5.795	0.510	0.856	10.353
		L	5.332	0.983	0.971	8.012	15.874	0.365	0.891	12.401
		G	2.402	0.920	0.113	8.099	4.152	0.751	0.404	11.165
400	1	Q	2.186	0.935	0.954	8.699	2.427	0.959	0.974	9.529
		L	5.246	0.981	0.971	8.756	5.916	0.966	0.970	8.906
		G	2.336	0.944	0.106	8.788	3.450	0.877	0.667	11.703
	2	Q	2.126	0.954	0.954	8.083	2.414	0.963	0.976	9.030
		L	4.962	0.983	0.970	8.153	5.175	1.000	0.974	8.206
		G	2.234	0.973	0.102	8.182	2.742	0.898	0.718	11.403
	3	Q	0.492	1.000	0.883	3.630	1.004	0.999	0.951	3.985
		L	1.035	0.995	0.934	3.698	1.855	0.994	0.965	4.018
		G	1.415	1.000	0.052	4.305	2.394	0.932	0.551	7.679
	4	Q	2.008	0.962	0.950	7.301	2.338	0.965	0.975	8.258
		L	4.602	0.983	0.970	7.394	5.991	0.967	0.968	7.634
		G	2.133	0.983	0.102	7.376	3.250	0.888	0.692	10.880

Table 4.5: Simulation summary of SNR=1, as for Table 4.1.

			GS				GIC			
n	Noise	M	ISE1	ISE2	ISE3	ISE4	ISE1	ISE2	ISE3	ISE4
200	1	Q	0.186	0.822	0.673	0.684	0.270	2.415	0.749	0.654
		L	0.629	1.004	3.197	0.987	1.073	3.883	3.019	2.129
		G	0.407	0.901	0.537	0.693	0.529	1.851	0.555	0.868
	2	Q	0.181	0.810	0.642	0.635	0.264	2.494	0.712	0.640
		L	0.592	0.973	2.974	0.989	1.078	4.370	2.808	1.825
		G	0.411	0.971	0.521	0.660	0.530	2.198	0.583	0.861
	3	Q	0.087	0.394	0.252	0.315	0.112	0.646	0.322	0.383
		L	0.197	0.520	1.067	0.640	0.575	1.677	1.619	1.045
		G	0.342	0.645	0.330	0.422	0.438	1.729	0.497	0.737
	4	Q	0.165	0.816	0.646	0.589	0.243	2.383	0.764	0.641
		L	0.552	0.961	2.781	0.986	0.982	4.010	2.891	1.769
		G	0.396	0.858	0.511	0.605	0.509	1.989	0.544	0.862
400	1	Q	0.163	0.830	0.593	0.565	0.189	0.801	0.616	0.817
		L	0.565	0.973	2.692	0.966	0.549	1.176	2.773	1.060
		G	0.387	0.837	0.492	0.598	0.453	1.339	0.422	1.041
	2	Q	0.165	0.814	0.579	0.540	0.194	0.795	0.619	0.803
		L	0.513	0.966	2.501	0.938	0.523	0.970	2.679	0.982
		G	0.383	0.797	0.476	0.563	0.420	0.900	0.375	1.007
	3	Q	0.038	0.133	0.137	0.177	0.070	0.393	0.241	0.298
		L	0.065	0.147	0.456	0.350	0.123	0.404	0.814	0.500
		G	0.312	0.516	0.242	0.344	0.383	0.778	0.405	0.802
	4	Q	0.146	0.794	0.540	0.502	0.176	0.799	0.604	0.758
		L	0.414	0.952	2.281	0.919	0.461	1.269	2.478	1.140
		G	0.376	0.771	0.448	0.527	0.433	1.177	0.412	1.041

Table 4.6: Individual funtional L_2 error when SNR=1, as for Table 4.2.

			GS				GIC			
n	Noise	M	MISE	GA	VA	PE	MISE	GA	VA	PE
200	1	Q	0.907	0.988	0.906	1.617	0.920	0.935	0.839	1.683
		L	1.962	0.917	0.939	1.835	1.964	0.792	0.910	1.917
		G	1.679	1.000	0.064	2.195	1.743	0.994	0.132	2.578
	2	Q	0.898	0.992	0.912	1.576	0.913	0.943	0.840	1.662
		L	1.866	0.932	0.942	1.784	1.917	0.790	0.912	1.888
		G	1.669	1.000	0.067	2.172	1.779	0.989	0.161	2.857
	3	Q	0.498	1.000	0.903	1.124	0.709	0.943	0.849	1.482
		L	1.203	0.993	0.943	1.325	1.756	0.828	0.914	1.867
		G	1.603	1.000	0.062	2.170	1.659	0.995	0.109	2.465
	4	Q	0.842	0.992	0.915	1.502	0.911	0.943	0.843	1.656
		L	1.774	0.952	0.944	1.709	1.928	0.792	0.913	1.904
		G	1.656	1.000	0.065	2.116	1.722	0.996	0.125	2.420
400	1	Q	0.499	0.999	0.892	1.142	0.610	0.963	0.874	1.222
		L	0.981	0.965	0.932	1.187	1.029	0.838	0.879	1.278
		G	1.371	1.000	0.051	1.684	1.557	0.998	0.208	2.183
	2	Q	0.458	1.000	0.890	1.069	0.565	0.981	0.897	1.145
		L	0.902	0.975	0.933	1.114	0.927	0.867	0.894	1.190
		G	1.361	1.000	0.052	1.665	1.567	0.996	0.216	2.275
	3	Q	0.096	1.000	0.874	0.602	0.167	1.000	0.918	0.671
		L	0.151	1.000	0.903	0.617	0.299	0.999	0.941	0.681
		G	1.220	1.000	0.050	1.679	1.260	1.000	0.081	1.759
	4	Q	0.410	1.000	0.891	0.981	0.515	0.978	0.899	1.067
		L	0.837	0.988	0.934	1.025	0.866	0.898	0.898	1.105
		G	1.336	1.000	0.050	1.627	1.494	0.997	0.175	2.075

Table 4.7: Simulation summary of SNR=10, as for Table 4.1.

n	Noise	M	GS				GIC			
			ISE1	ISE2	ISE3	ISE4	ISE1	ISE2	ISE3	ISE4
200	1	Q	0.080	0.298	0.220	0.286	0.082	0.292	0.222	0.284
		L	0.166	0.340	0.819	0.570	0.165	0.317	0.799	0.569
		G	0.334	0.625	0.312	0.407	0.338	0.637	0.318	0.449
	2	Q	0.081	0.299	0.216	0.282	0.087	0.294	0.218	0.277
		L	0.158	0.315	0.776	0.559	0.177	0.321	0.746	0.565
		G	0.334	0.621	0.310	0.403	0.342	0.641	0.318	0.477
	3	Q	0.040	0.117	0.146	0.188	0.061	0.206	0.182	0.233
		L	0.077	0.148	0.540	0.415	0.141	0.265	0.737	0.512
		G	0.330	0.597	0.289	0.387	0.333	0.607	0.296	0.423
	4	Q	0.072	0.270	0.211	0.271	0.080	0.293	0.227	0.273
		L	0.137	0.293	0.751	0.543	0.171	0.308	0.788	0.549
		G	0.333	0.618	0.306	0.397	0.337	0.630	0.319	0.435
400	1	Q	0.038	0.119	0.145	0.188	0.050	0.164	0.156	0.214
		L	0.052	0.109	0.440	0.349	0.056	0.119	0.415	0.334
		G	0.307	0.501	0.229	0.333	0.316	0.562	0.279	0.400
	2	Q	0.036	0.100	0.141	0.173	0.046	0.146	0.157	0.202
		L	0.044	0.094	0.412	0.327	0.050	0.099	0.385	0.309
		G	0.305	0.498	0.227	0.330	0.316	0.560	0.278	0.413
	3	Q	0.005	0.007	0.043	0.040	0.008	0.017	0.069	0.072
		L	0.007	0.007	0.076	0.059	0.009	0.013	0.154	0.121
		G	0.291	0.430	0.198	0.301	0.294	0.445	0.209	0.312
	4	Q	0.028	0.080	0.135	0.160	0.038	0.122	0.150	0.191
		L	0.039	0.076	0.397	0.306	0.043	0.085	0.380	0.294
		G	0.302	0.485	0.223	0.325	0.311	0.532	0.263	0.388

Table 4.8: Individual funtional L_2 error when SNR=10, as for Table 4.2.

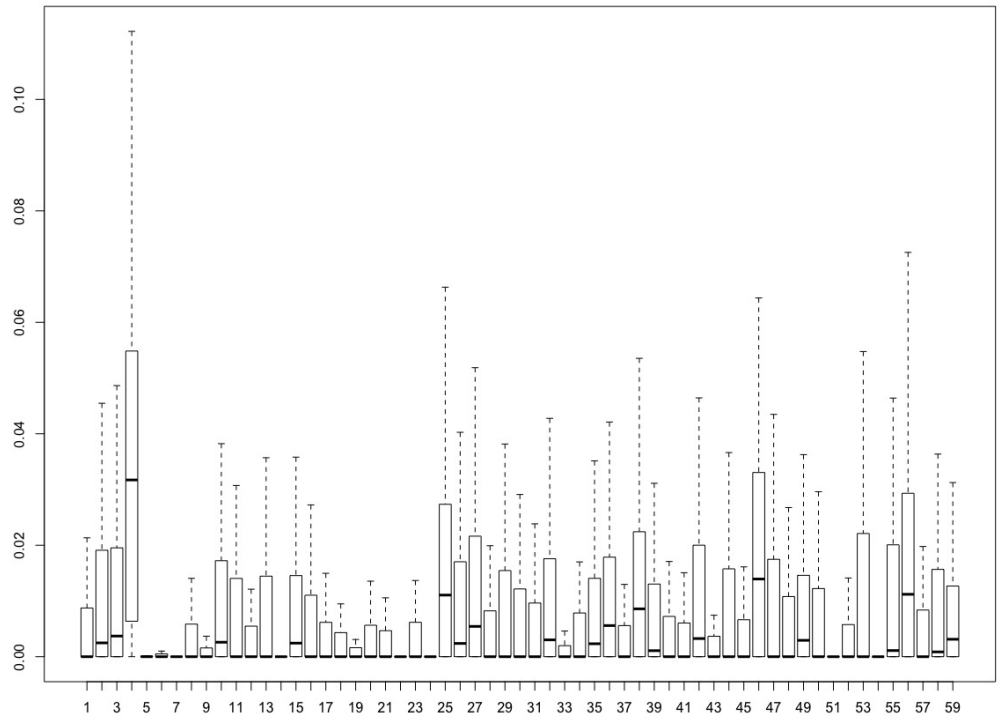


Figure 4.3: Boxplot of L_2 norm for each slope function, by using the quantile Lasso method.

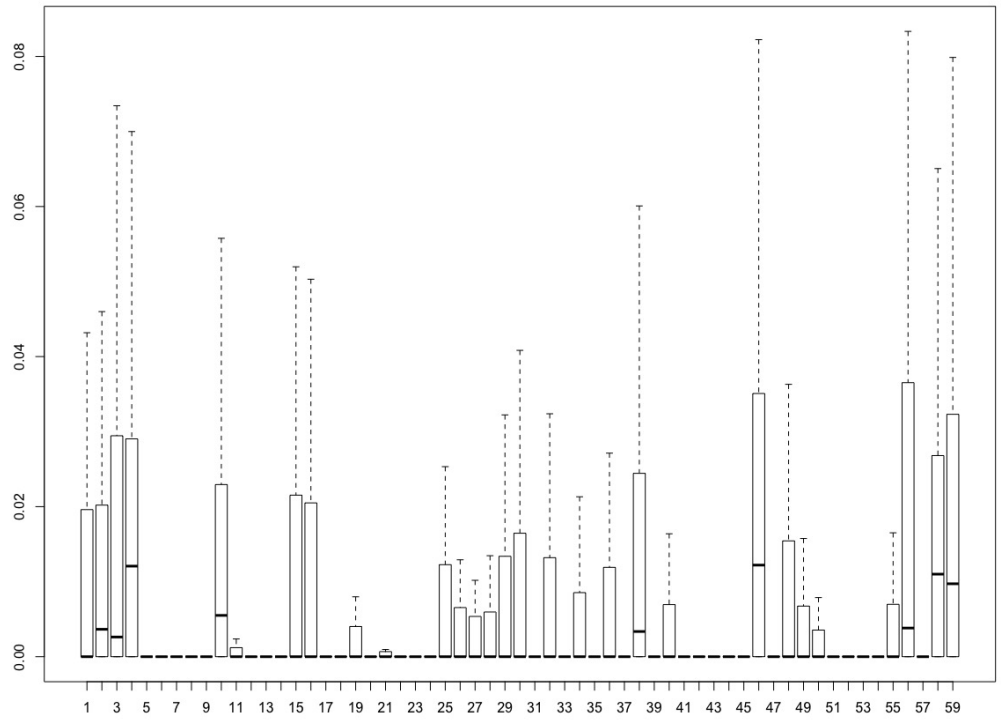


Figure 4.4: Boxplot of L_2 norm for each slope function, by using the quantile Group Lasso method.

Chapter 5

Conclusion

This thesis concentrates on the penalty methods used in the time series and functional data analysis.

In Chapter 2, we present two periodogram-like functions based on Group Lasso regularization in the context of least squares and quantile regression to estimate frequencies from unevenly spaced time series. Theorem 2.3.1 was proved to show that the Group Lasso penalty simultaneously shrinks and selects frequencies, owing to the property that Group Lasso penalty attempts to shrink some frequencies' regularized least squares periodogram toward exactly zero. A data-dependent procedure for selection of tuning parameter is given. These methods are validated on the simulation and real data, and its superiority is verified by comparing with adjust periodogram and robust periodogram methods.

The application of the total variation penalized Whittle likelihood to nonparametric spectral estimation is developed in Chapter 3. The most important feature of the proposed method is capable of capturing sharp changes in the target spectral density function while still maintaining a general smoothing objective. We give the strict mathematical proofs for the convergence of the estimator and establish an L_1

consistency result for the estimator. By simulations and real data, we show that the penalized method has the potential to capture local features in the density more efficiently than do more global approximations methods, such as the AR spectral estimators.

Chapter 4 expresses discrete time series data as a smooth function and then draws information from the collection of functional data. We restrict the coefficient function to the span of a wavelet basis and consider the variable selection problem for quantile regression where variables are given as functional forms. We convert the functional regression problem into a high dimensional variable selection problem by transforming each functional predictors into a set of wavelet coefficients and then selecting good predictors of the response variable from among these via sparse Group Lasso. We investigate asymptotic properties of the estimated regression functions through establishing an L_2 consistency result. Our numerical study and real data application suggest the promising performance of the procedure in variable selection and function estimation.

Bibliography

[1] M. Ahdesmäki, H. Lähdesmäki, A. Gracey, l. Shmulevich, and O. Yli-Harja, "Robust regression for periodicity detection in non-uniformly sampled time-course gene expression data," *BMC Bioinformatics*, vol. 8, no. 1, pp. 1–16, 2007.

[2] D. Anevski and P. Soulier, "Monotone spectral density estimation," *Annals of Statistics*, vol. 39, no. 1, pp. 418–438, 2011.

[3] A. Antoniadis, J. Bigot, and T. Sapatinas, "Wavelet estimators in nonparametric regression: A comparative simulation study," *Journal of Statistical Software*, vol. 6, pp–1, 2001.

[4] S. Baisch and G. H. Bokelmann, "Spectral analysis with incomplete time series: An example from seismology," *Computers Geosciences*, vol. 25, no. 7, pp. 739 –750, 1999.

[5] P. J. Bickel and B. Li, "Regularization in statistics," *Test*, vol. 15, no. 2, pp. 271–344, 2006.

[6] S Bourguignon, H Carfantan, and T Böhm, "Sparspec: A new method for fitting multiple sinusoids with irregularly sampled data," *Astronomy and Astrophysics*, vol. 462, no. 1, pp. 379–387, 2007.

[7] S. Bourguignon and H. Carfantan, "New methods for fitting multiple sinusoids from irregularly sampled data," *Statistical Methodology*, vol. 5, no. 4, pp. 318 –327, 2008.

[8] L. Briollais and G. Durrieu, "Application of quantile regression to recent genetic and-omic studies," *Human Genetics*, vol. 133, no. 8, pp. 951–966, 2014.

[9] P. Bühlmann, "Locally adaptive lag-window spectral estimation," *Journal of Time Series Analysis*, vol. 17, no. 3, pp. 247–270, 1996.

[10] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science and Business Media, 2011.

[11] B. S. Cade and B. R. Noon, "A gentle introduction to quantile regression for ecologists," *Frontiers in Ecology and the Environment*, vol. 1, no. 8, pp. 412–420, 2003.

[12] T. T. Cai and P. Hall, "Prediction in functional linear regression," *Annals of Statistics*, pp. 2159–2179, 2006.

[13] H. Cardot, F. Ferraty, and P. Sarda, "Spline estimators for the functional linear model," *Statistica Sinica*, vol. 13, no. 3, pp. 571–591, 2003.

[14] H. Cardot, F. Ferraty, and P. Sarda, "Functional linear model," *Statistics and Probability Letters*, vol. 45, no. 1, pp. 11 –22, 1999.

[15] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[16] Z.-G. Chen, "An alternative consistent procedure for detecting hidden frequencies," *Journal of Time Series Analysis*, vol. 9, no. 3, pp. 301–317, 1988.

[17] N. Choudhuri, S. Ghosal, and A. Roy, "Bayesian estimation of the spectral density of a time series," *Journal of the American Statistical Association*, vol. 99, no. 468, pp. 1050–1059, 2004.

[18] Y.-S. Chow and U. Grenander, "A sieve method for the spectral density," *Annals of Statistics*, vol. 13, no. 3, pp. 998–1010, 1985.

[19] J. A. Collazos, R. Dias, and A. Z. Zambom, "Consistent variable selection for functional regression models," *Journal of Multivariate Analysis*, vol. 146, pp. 63–71, 2016.

[20] P. L. Davies and A. Kovac, "Densities, spectral densities and modality," *Annals of Statistics*, vol. 32, no. 3, pp. 1093–1136, Jun. 2004.

[21] P. De Cat and C. Aerts, "A study of bright southern slowly pulsating B stars-II. the intrinsic frequencies," *Astronomy and Astrophysics*, vol. 393, no. 3, pp. 965–981, 2002.

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

[23] A. Delaigle, P. Hall, and T. V. Apanasovich, "Weighted least squares methods for prediction in the functional data linear model," *Electronic Journal of Statistics*, vol. 3, pp. 865–885, 2009.

[24] D. L. Donoho and J. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[25] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, *et al.*, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[26] J. Fan and J. Lv, "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, vol. 20, no. 1, pp. 101–148, 2010.

[27] G. B. Folland, *Real analysis: Modern techniques and their applications*. John Wiley & Sons, 2013.

[28] P. Fortin and M. C. Mackey, "Periodic chronic myelogenous leukaemia: Spectral analysis of blood cell counts and aetiological implications," *British Journal of Haematology*, vol. 104, no. 2, pp. 336–345, 1999.

[29] G. Froster, "The cleanest fourier spectrum," *The Astronomical Journal*, vol. 109, no. 4, pp. 1889–1902, 1995.

[30] J. Gertheiss, A. Maity, and A.-M. Staicu, "Variable selection in generalized functional linear models," *Stat*, vol. 2, no. 1, pp. 86–101, 2013.

[31] I. Good, "Non-parametric roughness penalty for probability densities," *Nature*, vol. 229, no. 1, pp. 29–30, 1971.

[32] P. Hall and J. L. Horowitz, "Methodology and convergence rates for functional linear regression," *Annals of Statistics*, vol. 35, no. 1, pp. 70–91, 2007.

[33] P. Hall and M. Li, "Using the periodogram to estimate period in noparametric regression," *Biometrika*, vol. 93, no. 0, pp. 411–424, 2006.

[34] P. Hall and J. Yin, "Nonparametric methods for deconvolving multiperiodic functions," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 65, no. 4, pp. 869–886, 2003.

[35] E. J. Hannan and D. Nicholls, "The estimation of the prediction error variance," *Journal of the American Statistical Association*, vol. 72, pp. 834–840, 1977.

[36] T. Hesterberg, N. H. Choi, L. Meier, C. Fraley, *et al.*, "Least angle and L1 penalized regression: A review," *Statistics Surveys*, vol. 2, pp. 61–93, 2008.

[37] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[38] I. A. Ibragimov, "On estimation of the integrals of the spectral function of a stationary gaussian process," *Theory of Probability and Its Applications*, vol. 8, pp. 366–401, 1963.

[39] R. Koenker and I. Mizera, "Density estimation by total variation regularization," *Advances in Statistical Modeling and Inference, Essays in Honor of Kjell A. Doksum*, 2006.

[40] R. Koenker, *Quantile regression*. Cambridge university press, 2005.

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

[42] R. Koenker and K. F. Hallock, "Quantile regression," *Journal of Economic Perspectives*, vol. 15, no. 4, pp. 143–156, 2001.

[43] R. Koenker and I. Mizera, "Convex optimization in R," *Journal of Statistical Software*, vol. 60, no. 5, pp. 1–23, 2014.

[44] R. Koenker, P. Ng, and S. Portnoy, "Quantile smoothing splines," *Biometrika*, vol. 81, no. 4, pp. 673–680, 1994.

[45] ——, "Quantile smoothing splines," *Biometrika*, vol. 81, no. 4, pp. 673–680, 1994.

[46] H. Künsch, "The jacknife and the boostrap for general stationary observations," *Annals of Statistics*, vol. 17, pp. 1217–1241, 1989.

[47] C. Lévy-Leduc, E. Moulines, and F. Roueff, "Frequency estimation based on the cumulated lamb-scargle periodogram," *Journal of Time Series Analysis*, vol. 29, no. 6, pp. 1104–1131, 2008.

[48] T. Li, "Laplace periodogram for time series analysis," *Journal of the Americans Statistical Association*, vol. 103, no. 482, pp. 757–768, 2008.

[49] ——, "Quantile periodogram," *Journal of the Americans Statistical Association*, vol. 107, no. 498, pp. 765–776, 2012.

[50] Y. Li, Y. Liu, and J. Zhu, "Quantile regression in reproducing kernel Hilbert spaces," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 255–268, 2007.

[51] C.-Y. Lin, H. Bondell, H. H. Zhang, and H. Zou, "Variable selection for non-parametric quantile regression via smoothing spline analysis of variance," *Stat*, vol. 2, no. 1, pp. 255–268, 2013.

[52] N. Lomb, "Least-squares frequency analysis of unequally spaced data," *Astrophysics and Space Science*, vol. 39, pp. 447–462, 1976.

[53] S. Mallat, *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*, 3rd. Academic Press, 2008.

[54] N. Meinshausen and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 4, 2010.

[55] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *Annals of Statistics*, pp. 246–270, 2009.

[56] J. S. Morris, "Functional regression," *Annual Review of Statistics and its Applications*, vol. 2, 2015.

[57] M. Mudelsee, D. Scholz, R. Röthlisberger, D. Fleitmann, A. Mangini, and E. W. Wolff, "Climate spectrum estimation in the presence of timescale errors," *Nonlinear Processes in Geophysics*, vol. 16, no. 1, pp. 43–56, 2009.

[58] H.-G. Müller and F. Yao, "Functional additive models," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1534–1544, 2008.

[59] M. H. Neumann, "Spectral density estimation via nonlinear wavelet methods for stationary non-gaussian time series," *Journal of Time Series Analysis*, vol. 17, no. 6, pp. 601–633, 1996.

[60] J. B. Oliva, B. Póczos, T. Verstynen, A. Singh, J. G. Schneider, F.-C. Yeh, and W.-Y. I. Tseng, "Fusso: Functional shrinkage and selection operator.," in *AISTATS*, 2014, pp. 715–723.

[61] Y. Paloyelis, M. A. Mehta, J. Kuntsi, and P. Asherson, "Functional MRI in ADHD: A systematic literature review," *Expert Review of Neurotherapeutics*, vol. 7, no. 10, pp. 1337–1356, 2007.

[62] Y. Pawitan and F. O'Sullivan, "Nonparametric spectral density estimation using penalized whittle likelihood," *Journal of the American Statistical Association*, vol. 89, pp. 600–610, 1994.

[63] D. B. Percival and A. T. Walden, *Spectral analysis for physical applications*. Cambridge University Press, 1993.

[64] J. O. Ramsay, *Functional data analysis*. Wiley Online Library, 2006.

[65] T. Ruf, "The Lomb-Scargle periodogram in biological rhythm research analysis of incomplete and unequally spaced time-series," *Biological Rhythm Research*, vol. 30, no. 2, pp. 178–201, 1999.

[66] J. Scargle, "Studies in astronomical time-series analysis II: Statistical aspect of spectral ananlysis of unevenly spaced data," *The Astrophysical Journal*, vol. 263, pp. 835–853, 1982.

[67] M. Schimmel, "Emphasizing difficulties in the detection of rhythms with Lomb-Scargle periodograms," *Biological Rhythm Research*, vol. 32, no. 3, pp. 341–346, 2001.

[68] G. W. Schrimsher, R. L. Billingsley, E. F. Jackson, and B. D. Moore, "Caudate nucleus volume asymmetry predicts attention-deficit hyperactivity disorder (ADHD) symptomatology in children," *Journal of Child Neurology*, vol. 17, no. 12, pp. 877–884, 2002.

[69] R. H. Shumway and A. S. Stoffer, *Time Series Analysis and Its applications with R examples*. Springer Science, 2005.

[70] J. Sidlauskaite, K. Caeyenberghs, E. Sonuga-Barke, H. Roeyers, and J. R. Wiersema, "Whole-brain structural topology in adult attention-deficit/hyperactivity disorder: Preserved global disturbed local network organization," *NeuroImage: Clinical*, vol. 9, pp. 506 –512, 2015.

[71] B. W. Silverman, "On the estimation of a probability density function by the maximum penalized likelihood method," *Annals of Statistics*, pp. 795–810, 1982.

[72] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[73] M. Taniguchi, "On estimation of parameters of gaussian stationary processes," *Journal of applied Probability*, vol. 16, pp. 575–591, 1979.

[74] ——, "On estimation of the integrals of certain functions of spectral density," *Journal of applied Probability*, vol. 17, pp. 73–83, 1980.

[75] A. M. Thieler, M. Backes, R. Fried, and W. Rhode, "Periodicity detection in irregularly sampled light curves by robust regression and outlier detection," *Statistical Analysis and Data Mining*, vol. 6, no. 1, pp. 73–89, 2013.

[76] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pp. 267–288, 1996.

[77] A. Tikhonov, "On the stability of inverse problems," in *C. R. (Dokl.) Acad. Sci. URSS (NS)*, vol. 39, 1943, pp. 176–179.

[78] D. Vidaurre, C. Bielza, and P. Larrañaga, "A survey of L1 regression," *International Statistical Review*, vol. 81, no. 3, pp. 361–387, 2013.

[79] C. Waelkens, "Slowly pulsating B-stats," *Astronomy and Astrophysics*, vol. 246, pp. 453–468, 1991.

[80] G. Wahba, "Automatic smoothing of the log periodogram," *Journal of the American Statistical Association*, vol. 75, no. 369, pp. 122–132, 1980.

[81] H. Wang, R. Li, and C.-L. Tsai, "Tuning parameter selectors for the smoothly clipped absolute deviation method," *Biometrika*, vol. 94, no. 3, pp. 553–568, 2007.

[82] J.-L. Wang, J.-M. Chiou, and H.-G. Müller, "Review of functional data analysis," *Annual Review of Statistics and its Applications*, vol. 1, p. 41, 2015.

[83] X. Wang, B. Nan, J. Zhu, and R. Koeppe, "Regularized 3D functional regression for brain image data via haar wavelets," *The Annals of Applied Statistics*, vol. 8, no. 2, p. 1045, 2014.

[84] P. Whittle, "Curve and periodogram smoothing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 19, no. 1, pp. 38–63, 1957.

[85] ——, "Gaussian estimation in stationary time series," *Bulletin of the International Statistical Institute*, vol. 39, no. 1, pp. 105–129, 1962.

[86] Y. Wu and Y. Liu, "Variable selection in quantile regression," *Statistica Sinica*, vol. 19, no. 2, p. 801, 2009.

[87] M. Yuan and Y. Liu, "Model selection and estimation in regression with grouped," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 68, pp. 49–67, 2006.

[88] M. Yuan and T. T. Cai, "A reproducing kernel Hilbert space approach to functional linear regression," *Annals of Statistics*, vol. 38, no. 6, pp. 3412–3444, 2010.

[89] Y. Zhang, R. Li, and C.-L. Tsai, "Regularization parameter selections via generalized information criterion," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 312–323, 2010.

[90] W. Zhao, R. Zhang, and J. Liu, "Sparse group variable selection based on quantile hierarchical lasso," *Journal of Applied Statistics*, vol. 41, no. 8, pp. 1658–1677, 2014.

[91] Y. Zhao, H. Chen, and R. T. Ogden, "Wavelet-based weighted lasso and screening approaches in functional linear regression," *Journal of Computational and Graphical Statistics*, vol. 24, no. 3, pp. 655–675, 2015.

[92] Y. Zhao, R. T. Ogden, and P. T. Reiss, "Wavelet-based lasso in functional linear regression," *Journal of Computational and Graphical Statistics*, vol. 21, no. 3, pp. 600–617, 2012.

[93] Q. Zheng, L. Peng, and X. He, "Globally adaptive quantile regression with ultra-high dimensional data," *Annals of Statistics*, vol. 43, no. 5, p. 2225, 2015.

[94] H. Zou and M. Yuan, "Composite quantile regression and the oracle model selection theory," *Annals of Statistics*, vol. 36, no. 3, pp. 1108–1126, 2008.