Topics on Nonparametric Methods for Longitudinal Data Analysis and Jumps Detection

By

Shengji Jia

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

Doctor of Philosophy
(Statistics)

at the
UNIVERSITY OF WISCONSINCMADISON
2019

Date of final oral examination: 11/14/2019

The dissertation is approved by the following members of the Final Oral Committee:

Chunming Zhang, Professor, Statistics

Zhengjun Zhang, Professor, Statistics

Xiaoxia Shi, Professor, Economics

Yazhen Wang, Professor, Statistics

Nicolas Garcia Trillos, Assistant Professor, Statistics

To my family

Acknowledgements

First and foremost, I would like to express my sincere appreciation and gratitude to my advisor Professor Chunming Zhang, for the continuous support of my Ph.D. study and related research, for her patience, insightful guidance, and immense knowledge. It has been an honor to be her Ph.D. student. Her advice on both research as well as on life have been priceless, whether in academia or another setting. I immensely learned from every meeting I had with her. Her encouragement, motivation, and support are pivotal throughout my Ph.D. life.

Besides, I would like to thank my committee members professor Yazhen Wang, Professor Zhengjun Zhang, Professor Xiaoxia Shi, Professor Nicolas Garcia Trillos for their time, stimulating discussions, and brilliant comments on the thesis. I would also like to thank Professor Jun Shao for his invaluable feedback.

My sincerest thanks and appreciation also go to the Department of Statistics at the University of Wisconsin-Madison for providing the high-quality education and making my study and research experience great and fruitful. My thanks are also extended to all my academic siblings, Yi Chai, Cheng Chen, Xiao Guo, Lilun Du, Muhong Gao, Yongsu Li, Yanbo Shen, Bowen Zhang, Yongfeng Wu, Taiyu Ye, for their valuable opinions and insightful questions

at our weekly group meetings. Additionally, I am truly thankful to my fellow students for their kind support, generous help, and constant encouragement, and all the great memories we shared.

Furthermore, I want to give special thanks to Yongsu Li, for all his computational techniques support in R programming. I would also like to thank all of my friends for always being there to listen and advise.

Last but not least, I want to thank my loving parents, Difang Jia and Zhaofang Chen. Their love and faith were what sustained me thus far. Words cannot express how grateful I am to my parents for their never-ending encouragement and love, and for always believing in me throughout my life.

Contents

List of Figures

List of Tables

Abstract

In this thesis we investigate the nonparametric methods applied for longitudinal data analysis and jumps detection. The first part is about efficient semi-parametric regression for longitudinal data with nonparametric covariance estimation. Improving estimation efficiency for regression coefficients is an important issue in the analysis of longitudinal data, which involves estimating the covariance matrix of errors. But challenges arise in estimating the covariance matrix of longitudinal data collected at irregular or unbalanced time points. We develop a regularization method for estimating the covariance function and a stepwise procedure for estimating the parametric components efficiently in the varying-coefficient partially linear model. This procedure is also applicable to the varying-coefficient temporal mixed effects model. Our method utilizes the structure of the covariance function and thus has faster rates of convergence in estimating the covariance functions and outperforms the existing approaches. The second part is about adaptive jumps detection via nonparametric screening and multiple testing procedure. In many applications, it may appear that a regression function is smooth except at several points where jump discontinuities occur. But challenges arise when the number of jumps is quite large and unknown. We develop a jumps detection procedure

via nonparametric screening and multiple testing. The candidates of jumps are first detected through screening and then a multiple testing procedure is applied to rule out the noises. Our proposed method is quite robust in jumps detection and doesn't depend on the choice of tuning parameter and threshold in the screening procedure. All the two procedures are easy to implement and their numerical performance are investigated using both simulated and real data.

Key words and phrases: Bootstrap; Change points; FDR; Local linear regression; Multiple testing; Copy number variations; Covariance function; Method of regularization; Profile weighted least squares; Semiparametric varying-coefficient partially linear model; Sobolov space; tensor product space.

Chapter 1

Introduction

This thesis contains two parts about different topics of nonparametric techniques: efficient semiparametric regression for longitudinal data with nonparametric covariance estimation and adaptive jumps detection via screening and multiple testing procedure.

For the first part about longitudinal data analysis, there has been substantial recent interest in nonparametric and semiparametric methods for longitudinal or clustered data with dependence within subjects (or clusters) (see [14];[3];[20]). Improving estimation efficiency is an important issue in the analysis of longitudinal data. In the nonparametric setting, Lin and Carroll [14] recommended an approach which ignores the within-subject correlation completely and treats the data as if they were independent. However, Wang *et al.* [18] showed that, in the semiparametric setting, the estimator for parametric component in the model will achieve the semiparametric efficiency bound if the within-subject correlation structure is specified correctly. Thus estimating the covariance function is an important issue in the semiparametric model for

longitudinal data.

Many authors have investigated the problem of within-subject correlation in longitudinal data. For example, Wu and Pourahmadi [22] proposed nonparametric estimation of large covariance matrices using a two step estimation procedure. But their method can only deal with balanced or nearly balanced longitudinal data. Challenges arise in estimating the covariance function if the data are collected at irregular or subject specific time points. Wu and Zhang (see [19];[21]) proposed another method called local polynomial linear mixed-effects model (LLME) to analyze longitudinal data, which estimated the within-subject error directly instead of estimating the covariance function of the errors. Other methods for modeling the covariance function include the functional principal component analysis (FPCA) proposed by Yao *et al.* [23].

In this part, we consider a semiparametric varying-coefficient partially linear model

$$Y(t) = \boldsymbol{X}(t)^T\boldsymbol{\alpha}(t) + \boldsymbol{Z}(t)^T\boldsymbol{\beta} + \eta(t) + \zeta(t), \tag{0.1}$$

where $\boldsymbol{\alpha}(t)$ comprises p unknown smooth functions, $\boldsymbol{\beta}$ is a q-dimensional unknown parameter vector, and $\eta(t)$ captures the within-subject dependence with smooth covariance function $R_\eta(t_1, t_2)$, $\zeta(t)$ is just the measurement error with covariance function $\sigma_\zeta^2(t_1)\mathrm{I}(t_1 = t_2)$. All the temporal correlations are relegated to $\eta(t)$, so this decomposition is unique. Nonparametric models for longitudinal data can be viewed as special cases of model (0.1). Moreover, model (0.1) is an extension of the partially linear model and the time-varying coefficient model.

We focus on estimating the covariance function $R_\eta(t_1, t_2)$ (defined in (1.4)) when observations are collected at irregular and possibly subject-specific time

points. In this paper,

(i) A *varying-coefficient temporal mixed effects model* is introduced as a good approximation of model (0.1). The within-subject correlated error $\eta(t)$ can be considered as a combination of some common random factors (not related to t) and some temporal functions.

(ii) A general framework of the regularization method is applied to estimate the covariance function, which can be viewed as an extension of one-dimensional smoothing splines. This method is introduced through a careful characterization of the function space (tensor product of Hilbert space) in which the covariance function $R_\eta(t_1, t_2)$ resides, and thus has faster rates of convergence compared to other methods which only assume $R_\eta(t_1, t_2)$ is a bivariate continuous function.

(iii) An explicit spectral decomposition of the estimated covariance function is established, and we can easily guarantee the estimated covariance function to be positive definite after truncating the negative eigenvalues. To improve the efficiency of estimating the regression coefficients, the weight matrix is chosen by the inverse of the adjusted covariance matrix in the weighted least squares method.

(iv) Our proposed method can be applied to both the sparse longitudinal data and the densely sampled functional data. Besides, our method also works quite well for the missing longitudinal/functional data, which can be considered as a special case of longitudinal/functional data collected at irregular time points.

There also has been a vast volume of work on modelling covariance functions in longitudinal data in literature. Fan *et al.* ([6];[7]) proposed a quasi-maximum likelihood method to model covariance function of $\eta(t)$. In their method, the variance function $\text{var}\{\eta(t)\}$ is modeled nonparametrically, but the correlation function $\text{corr}\{\eta(t_1), \eta(t_2)\}$ is assumed to be a member of a known family of parametric correlation functions (e.g., an AR or ARMA correlation structure). The quasi-maximum likelihood method relies on correctly assuming the form of the correlation functions and can be misspecified easily. Li [13] applied bivariate kernel smoothing techniques to estimate covariance functions. But the techniques of bivariate kernel smoothing do not utilize the structures of the covariance functions, and thus have slower rates of convergence compared with our method. Besides, the kernel covariance estimator is not guaranteed to be positive semidefinite, and an adjustment procedure is required by discretizing the kernel estimator on a dense grid, followed by taking the eigenvalue decomposition of the resulting covariance matrix. This discretizing procedure is quite subjective, since it depends heavily on the choice of dense grids.

For the second part about jumps detection, very long and noisy sequence data arise in many areas such as genetics (Fearnhead and Liu 2007), environmental statistics (Lu *et al.* 2010) and finance (Fan and Wang 2007) including high throughput data in genomics and stock prices in econometrics. Usually, such data are studied in order to identify and understand shifts in trends, for example, from a bull market to a bear market in finance or from a normal number of chromosome copies to an excessive number of chromosome copies in genetics. Thus, identifying multiple jump points in a regression function for

a long sequence is an important issue.

Many authors have investigated the issue of single jump detection and estimation in literature. Estimation of the single jump point in nonparametric regression has been studied by Müller (1992) and Loader (1996) based on kernel estimates. A different approach based on semi-parametric model was developed by Eubank and Speckman (1994). Müller and Song (1997) and Gijbels *et al.* (1999) proposed two-step estimation procedures. These methods can be generalized and applied easily to the case when there are finite and known number of jump points.

In contrast, when the number of jump points is large and unknown, the problem is much more challenging. Wu and Chu (1993) considered the estimation of the number of jumps by a sequence of hypothesis tests. Wavelet based method (Wang 1995) is one of the most commonly used methods when there are multiple jumps.

In this part, we propose a robust jumps detection method via screening and multiple testing. We consider that the nonparametric regression function $m(x)$ is smooth except at points $\tau_1, ..., \tau_q$, where the number q of jump points is finite but unknown. In the first step, we use screening method to get a set of initial estimates of the jump points $\{\widehat{\tau}_1, ..., \widehat{\tau}_{\widehat{q}}\}$. After we guarantee that all the true jump points are contained in this set, we conduct multiple testing to rule out the noises and select a subset of $\{\widehat{\tau}_1, ..., \widehat{\tau}_{\widehat{q}}\}$ as the refined estimates of jump points. The advantage of this method is that we do not need to estimate the number $\widehat{q}$ in the first step exactly, thus there is a wide choice for the tuning parameter and threshold in the screening procedure.

Our proposed method is also adaptive to a special case of the multiple

jumps detection, which is called *multiple change-point model* (MCM). The MCM can be presented as

$$Y_i = \theta_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad 1 \leq i \leq n,$$

with

$$\theta_1 = \theta_2 = \ldots = \theta_{\tau_1} \neq \theta_{\tau_1+1} = \ldots = \theta_{\tau_2} \neq \theta_{\tau_2+1} = \ldots = \theta_{\tau_q} \neq \theta_{\tau_q+1} = \ldots = \theta_n.$$

In other words, we assume that the signal $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^T$ is piecewise constant with jumps or drops at $\boldsymbol{\tau} = (\tau_1, ..., \tau_q)^T$. This model is also widely used in the signal processing of computer science. Moreover, this distinct feature results in some efficient algorithms which are quite different from that in multiple jumps detection problems, such as $\boldsymbol{l}_1$-Penalization (Harchaoui and Lévy 2010) and SaRa algorithm (Niu and Zhang 2012).

The rest of the thesis is organized as follows. In Chapter 2, Section 2.1 describes the semiparametric varying-coefficient model for longitudinal data and decomposition of covariance function. A nonparametric estimation of the covariance function and an efficient estimation procedure for parameters based on profile least squares techniques is described in Section 2.2. Sampling properties of the proposed procedure are presented in Section 2.3. In Section 2.4 and Section 2.5 the proposed method is illustrated via simulation studies and real data examples, respectively. In Chapter 3, Section 3.1 describes the model and set up for jumps detection. A two-step method including screening and multiple testing is proposed in Section 3.2. Sampling properties of the proposed procedure are presented in Section 3.3. In Section 3.4 and Section 3.5 the proposed method is illustrated via simulation studies and real data examples, respectively. All technical proofs are relegated to Appendix.

Chapter 2

Efficient Semiparametric Regression for Longitudinal Data with Nonparametric Covariance Estimation

2.1 Model and covariance structure

Suppose all longitudinal observations from different subjects (or clusters) are made on a fixed time interval $\mathcal{T} \subset \mathbb{R}$, e.g., $\mathcal{T} = [0, 1]$. The data consist of n independent subjects. For the ith subject, $i = 1, \ldots, n$, the response variable $Y_i(t)$ and the covariates $\{\boldsymbol{X}_i(t), \boldsymbol{Z}_i(t)\}$ are collected at time points $t = t_{i,j}$, $j = 1, \ldots, J_i$, where J_i is the total number of observations for the ith subject. In this article, we consider a semiparametric varying-coefficient partially linear

model

$$Y(t) = \boldsymbol{X}(t)^T \boldsymbol{\alpha}(t) + \boldsymbol{Z}(t)^T \boldsymbol{\beta} + \varepsilon(t)$$

i.e.

$$Y_i(t_{i,j}) = \boldsymbol{X}_i(t_{i,j})^T \boldsymbol{\alpha}(t_{i,j}) + \boldsymbol{Z}_i(t_{i,j})^T \boldsymbol{\beta} + \varepsilon_i(t_{i,j}), \quad i = 1, \ldots, n, \ j = 1, \ldots, J_i, \tag{1.1}$$

where $\boldsymbol{\alpha}(t)$ comprises p unknown smooth functions, $\boldsymbol{\beta}$ is a q-dimensional unknown parameter vector, and $\{\varepsilon_i(t) : i = 1, \ldots, n\}$ are i.i.d. error processes with $\mathrm{E}\{\varepsilon_i(t) \mid \boldsymbol{X}_i(t), \boldsymbol{Z}_i(t)\} = 0$. To consider the within-subject dependence, we assume that $\varepsilon_i(t)$ can be decomposed into two independent error processes:

$$\varepsilon_i(t) = \eta_i(t) + \zeta_i(t),$$

where $\{\eta_i(t) : i = 1, \ldots, n\}$ are i.i.d. mean zero error processes capturing the within-subject dependence or temporal correlation, and $\{\zeta_i(t) : i = 1, \ldots, n\}$ are the i.i.d. measurement error (see [23];[10]). For $t_1 \in \mathcal{T}$ and $t_2 \in \mathcal{T}$, suppose

$$\begin{aligned} \mathrm{cov}\{\eta_i(t_1), \eta_i(t_2)\} &= R_\eta(t_1, t_2), \\ \mathrm{cov}\{\zeta_i(t_1), \zeta_i(t_2)\} &= \sigma_\zeta^2(t_1)\mathrm{I}(t_1 = t_2), \quad i = 1, \ldots, n, \end{aligned} \tag{1.2}$$

where $\mathrm{I}(\cdot)$ is an indicator function, $R_\eta(\cdot, \cdot)$ and $\sigma_\zeta^2(\cdot)$ are smooth functions. Then the covariance function $R(t_1, t_2)$ of $\varepsilon_i(t)$ is given by

$$R(t_1, t_2) \equiv \mathrm{cov}\{\varepsilon_i(t_1), \varepsilon_i(t_2)\} = R_\eta(t_1, t_2) + \sigma_\zeta^2(t_1)\mathrm{I}(t_1 = t_2), \quad i = 1, \ldots, n, \tag{1.3}$$

which is a smooth surface except on the diagonal points where $t_1 = t_2$. In Section 2.2, we will estimate $R_\eta(\cdot, \cdot)$ and $\sigma_\zeta^2(\cdot)$ separately.

There are many different methods to analyze the within-subject error $\eta_i(t)$. For example, Wu and Zhang [19] used the local polynomial method to decompose $\eta_i(t)$. Here let us consider functional principal component analysis model for $\eta_i(t)$. Let $\lambda_{(1)} \geq \lambda_{(2)} \geq \cdots \geq 0$ be ordered values of the eigenvalues of the linear operator determined by $R_\eta(\cdot,\cdot)$ with $\sum_{k=1}^{\infty} \lambda_{(k)} < \infty$, and the $\psi_k(\cdot)$'s be the corresponding orthonormal eigenfunctions or principal components, see [10]. Then, $R_\eta(\cdot,\cdot)$ admits the spectral decomposition:

$$R_\eta(t_1, t_2) = \sum_{k=1}^{\infty} \lambda_{(k)} \psi_k(t_1) \psi_k(t_2), \tag{1.4}$$

and $\eta_i(t)$ admits the Karhunen-Loeve expansion as follows,

$$\eta_i(t) = \sum_{k=1}^{\infty} \xi_{i,k} \psi_k(t),$$

where $\xi_{i,k} = \int_{t \in \mathcal{T}} \eta_i(t) \psi_k(t) dt$ are uncorrelated random variables with $\mathrm{E}(\xi_{i,k}) = 0$ and $\mathrm{E}(\xi_{i,k}\xi_{i,j}) = \lambda_{(k)}\mathrm{I}(j = k)$. If $\lambda_{(k)} \approx 0$ for $k \geq L + 1$, then model (1.1) can be approximated by

$$Y_i(t_{i,j}) \approx \boldsymbol{X}_i(t_{i,j})^T \boldsymbol{\alpha}(t_{i,j}) + \boldsymbol{Z}_i(t_{i,j})^T \boldsymbol{\beta} + \sum_{k=1}^{L} \xi_{i,k} \psi_k(t_{i,j}) + \zeta_i(t_{i,j}). \tag{1.5}$$

Model (1.5) can be regarded as a *varying-coefficient temporal mixed effects model*, since $\xi_{i,k}$ are random variables and $\psi_k(t)$ are "unknown" but fixed basis functions.

As for the estimation of $\sigma_\zeta^2(\cdot)$ and $R_\eta(\cdot,\cdot)$, we assume that $\sigma_\zeta^2(\cdot)$ is a smooth function so that smoothing techniques such as the local linear regression can be applied to estimate the variance function $\sigma_\zeta^2(\cdot)$. By the covariance function decomposition (1.4), we assume that $R_\eta(\cdot,\cdot)$ resides in a tensor product of

Hilbert space $\mathcal{W}_2^2(\mathcal{T}) \otimes \mathcal{W}_2^2(\mathcal{T})$, which is the closure of the following linear space

$$\text{span}\{f(s)g(t) : \ f(\cdot), g(\cdot) \in \mathcal{W}_2^2(\mathcal{T})\}, \tag{1.6}$$

where $\mathcal{W}_2^2(\mathcal{T}) = \{f : f', f'' \text{ are absolutely continuous, } f'' \in L_2(\mathcal{T})\}$ is a Sobolev space endowed with the squared norm $\int_{\mathcal{T}}(f'')^2$. Because $\mathcal{W}_2^2(\mathcal{T}) \otimes \mathcal{W}_2^2(\mathcal{T})$ is dense in the continuous bivariate function space, we can find an element in $\mathcal{W}_2^2(\mathcal{T}) \otimes \mathcal{W}_2^2(\mathcal{T})$ that approximates any continuous bivariate function very well.

2.2 Proposed methodology for estimation

In practice, estimation of $\{\boldsymbol{\alpha}(t), \boldsymbol{\beta}\}$ must be done in multiple steps. Their initial estimates are constructed by ignoring within-subject correlation. With the initial estimates of $\{\boldsymbol{\alpha}(t), \boldsymbol{\beta}\}$, we can estimate $R(\cdot, \cdot)$ based on residuals. Finally, we can estimate $\{\boldsymbol{\alpha}(t), \boldsymbol{\beta}\}$ more efficiently by using the estimate of $R(\cdot, \cdot)$. In this section, we propose the efficient estimates for $\{\boldsymbol{\alpha}(t), \boldsymbol{\beta}\}$ using profile weighted least squares techniques.

2.2.1 Step 1: Initial estimator

For a given $\boldsymbol{\beta}$, model (1.1) reduced to a varying-coefficient model:

$$Y_i(t_{i,j}) - \boldsymbol{Z}_i(t_{i,j})^T\boldsymbol{\beta} = \boldsymbol{X}_i(t_{i,j})^T\boldsymbol{\alpha}(t_{i,j}) + \eta_i(t_{i,j}) + \zeta_i(t_{i,j}). \tag{2.1}$$

Ignoring the within-subject correlation or $\eta_i(t)$, we use the profile local linear regression to get initial estimates of $\{\boldsymbol{\alpha}(t), \boldsymbol{\beta}\}$, see [5]. For any t in the neighborhood of t_0, the lth component $\alpha_l(t)$ of $\boldsymbol{\alpha}(t)$, admits Taylor's expansion as

follows:

$$\begin{aligned}\alpha_l(t) &\approx \alpha_l(t_0) + \alpha_l'(t_0)(t - t_0) \\ &\equiv a_l + b_l(t - t_0), \quad \text{for } l = 1, \ldots, p.\end{aligned}$$

Let $K(\cdot)$ be a kernel function and h_1 be a bandwidth. Thus we can find the local linear estimator $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(t_0)$ of $\boldsymbol{\alpha}(t_0)$, where $\boldsymbol{\alpha}(t)$ is the true varying function in model (2.1). Let $(\widehat{a}_1, \ldots, \widehat{a}_p, \widehat{b}_1, \ldots, \widehat{b}_p)$ be the minimizer of

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i}\Big[Y_i(t_{i,j}) - \boldsymbol{Z}_i(t_{i,j})^T\boldsymbol{\beta} - \sum_{l=1}^{p}\{a_l + b_l(t_{i,j} - t_0)\}X_{il}(t_{i,j})\Big]^2 K_{h_1}(t_{i,j} - t_0), \quad (2.2)$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, and $\boldsymbol{X}_i(t) = (X_{i1}(t), \ldots, X_{ip}(t))^T$. Then $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(t_0) = (\widehat{a}_1, \ldots, \widehat{a}_p)^T$. Note that the profile least squares estimator of $(\boldsymbol{\alpha}(t), \boldsymbol{\beta})$ has a closed form using the following matrix notation. Let

$$\begin{aligned}\boldsymbol{Y} &= (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_n^T)^T, & \boldsymbol{Y}_i &= (Y_i(t_{i,1}), \ldots, Y_i(t_{i,J_i}))^T, \\ \mathbf{Z} &= (\mathbf{Z}_1^T, \ldots, \mathbf{Z}_n^T)^T, & \mathbf{Z}_i &= (\boldsymbol{Z}_i(t_{i,1}), \ldots, \boldsymbol{Z}_i(t_{i,J_i}))^T, \\ \boldsymbol{m} &= (\boldsymbol{m}_1^T, \ldots, \boldsymbol{m}_n^T)^T, & \boldsymbol{m}_i &= (\boldsymbol{X}_i(t_{i,1})^T\boldsymbol{\alpha}(t_{i,1}), \ldots, \boldsymbol{X}_i(t_{i,J_i})^T\boldsymbol{\alpha}(t_{i,J_i}))^T.\end{aligned}$$

Then model (2.1) can be written as

$$\boldsymbol{Y} - \mathbf{Z}\boldsymbol{\beta} = \boldsymbol{m} + \boldsymbol{\eta} + \boldsymbol{\zeta}, \quad (2.3)$$

where $\boldsymbol{\eta} = (\eta_1(t_{1,1}), \ldots, \eta_n(t_{n,J_n}))^T$ and $\boldsymbol{\zeta} = (\zeta_1(t_{1,1}), \ldots, \zeta_n(t_{n,J_n}))^T$. Since the estimator $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(\cdot)$ is linear in $\boldsymbol{Y} - \mathbf{Z}\boldsymbol{\beta}$, given $\boldsymbol{\beta}$, the estimator of $\boldsymbol{m}$ is of the form $\widehat{\boldsymbol{m}} = \mathbf{S}(\boldsymbol{Y} - \mathbf{Z}\boldsymbol{\beta})$, where $\mathbf{S}$ is a smoothing matrix of the local linear smoother, see [29]. Substituting $\widehat{\boldsymbol{m}}$ into model (2.3) results in the linear model,

$$(\mathbf{I} - \mathbf{S})\boldsymbol{Y} \approx (\mathbf{I} - \mathbf{S})\mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\zeta},$$

where $\mathbf{I}$ is an identity matrix. So an initial estimator for $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}^{\text{ini}} = \{\mathbf{Z}^T(\mathbf{I}-\mathbf{S})^T(\mathbf{I}-\mathbf{S})\mathbf{Z}\}^{-1}\mathbf{Z}^T(\mathbf{I}-\mathbf{S})^T(\mathbf{I}-\mathbf{S})\boldsymbol{Y}. \tag{2.4}$$

Then the profile least squares estimator for the nonparametric component $\boldsymbol{\alpha}(\cdot)$ is just $\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\beta}}^{\text{ini}}}(\cdot)$.

2.2.2 Step 2: Covariance estimator

After we get the initial estimators $\widehat{\boldsymbol{\beta}}^{\text{ini}}$ and $\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\beta}}^{\text{ini}}}(t)$ in Step 1, the residuals are

$$\widehat{\varepsilon}_i(t_{i,j}) = Y_i(t_{i,j}) - \boldsymbol{X}_i(t_{i,j})^T\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\beta}}^{\text{ini}}}(t_{i,j}) - \boldsymbol{Z}_i(t_{i,j})^T\widehat{\boldsymbol{\beta}}^{\text{ini}}, \quad i = 1,\ldots,n,\ j = 1,\ldots,J_i. \tag{2.5}$$

Then we will derive the nonparametric estimator of $R(\cdot,\cdot)$ based on $\widehat{\varepsilon}_i(t_{i,j})$. Since there are too many parameters in $R_\eta(\cdot,\cdot) \in \mathcal{W}_2^2(\mathcal{T}) \otimes \mathcal{W}_2^2(\mathcal{T})$, similar to the idea of smoothing spline, a penalty for over-parametrization is imposed to regularize the covariance function. Let $\widehat{R}_\eta(s,t) \in \mathcal{W}_2^2(\mathcal{T}) \otimes \mathcal{W}_2^2(\mathcal{T})$ be the minimizer of

$$\frac{1}{\sum_{i=1}^n J_i(J_i-1)} \sum_{i=1}^n \sum_{1\le j_1\ne j_2\le J_i} \left\{\widehat{\varepsilon}_i(t_{i,j_1})\widehat{\varepsilon}_i(t_{i,j_2}) - R_\eta(t_{i,j_1}, t_{i,j_2})\right\}^2 + \lambda_n P(R_\eta) \tag{2.6}$$

where $\lambda_n \ge 0$ is a tuning parameter, and $P(R_\eta)$ is a penalty function for $R_\eta(s,t) = \sum_{j\ge 1} a_j f_j(s) g_j(t)$ defined in (1.6).

The diagonal element of $R(\cdot,\cdot)$ requires a special treatment since it involves both $R_\eta(t,t)$ and $\sigma_\zeta^2(t)$. Denote $\sigma^2(t) \equiv R(t,t)$, which can be estimated by an one-dimensional local linear smoother. Let $(\widehat{\gamma}_0, \widehat{\gamma}_1)$ be the minimizer of

$$\sum_{i=1}^n \sum_{j=1}^{J_i} \{\widehat{\varepsilon}_i^2(t_{i,j}) - \gamma_0 - \gamma_1(t_{i,j}-t)\}^2 K_{h_2}(t_{i,j}-t), \tag{2.7}$$

where h_2 is a new bandwidth which can be different from h_1 in Step 1. Define $\widehat{\sigma}^2(t) = \max\{\widehat{\gamma}_0, 0\}$, where we take the maximum to avoid a negative $\widehat{\gamma}_0$. According to the definition of $R(s,t)$ in (1.3), $R(s,t)$ can also be written as

$$R(s,t) = R_\eta(s,t)\mathrm{I}(s \neq t) + \sigma^2(t)\mathrm{I}(s = t),$$

that is, $R(s,t)$ is $R_\eta(s,t)$ when $s \neq t$ and $\sigma^2(t)$ when $s = t$. So the estimator $\widehat{R}(\cdot,\cdot)$ of the covariance function is a combination of $\widehat{R}_\eta(s,t)$ and $\widehat{\sigma}^2(t)$:

$$\widehat{R}(s,t) = \widehat{R}_\eta(s,t)\mathrm{I}(s \neq t) + \widehat{\sigma}^2(t)\mathrm{I}(s = t). \tag{2.8}$$

By the definition of $\sigma^2(t)$, $\sigma^2(t) = R_\eta(t,t) + \sigma_\zeta^2(t) > R_\eta(t,t)$. Since $\widehat{\sigma}^2(t)$ is a consistent estimator of $\sigma^2(t)$, it can be shown that with probability tending to 1, $\widehat{\sigma}^2(t) > \widehat{R}_\eta(t,t)$.

2.2.3 Step 3: Refined estimator

First, let Σ_i and $\widehat{\Sigma}_i$ be the true and estimated covariance matrix within the ith subject, i.e.,

$$\Sigma_i = \big[R(t_{i,j}, t_{i,k})\big]_{j,k=1}^{J_i}, \quad \widehat{\Sigma}_i = \big[\widehat{R}(t_{i,j}, t_{i,k})\big]_{j,k=1}^{J_i}. \tag{2.9}$$

Since we ignore the within-subject correlation or $\eta_i(t)$ in Step 1, the initial least squares estimator $\widehat{\boldsymbol{\beta}}^{\mathrm{ini}}$ in (2.4) is not efficient. To improve the efficiency for estimating $\boldsymbol{\beta}$, we use the profile weighted least squares estimator as follows:

$$\{\mathbf{Z}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\mathbf{Z}\}^{-1}\mathbf{Z}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\boldsymbol{Y}, \tag{2.10}$$

where $\mathbf{W}$ is a weight matrix, called a working covariance matrix. Then the initial estimate $\widehat{\boldsymbol{\beta}}^{\mathrm{ini}}$ in (2.4) is just a special case of (2.10) with $\mathbf{W}$ being an

identity matrix. As usual, misspecification of the working covariance matrix does not affect the consistency of the resulting estimate, but does affect the efficiency. Fan *et al.* [6] has shown that the most efficient estimator for $\boldsymbol{\beta}$ among the profile weighted least squares estimates given in (2.10) is the one that uses the inverse of the true variance-covariance matrix of errors as the weight matrix, that is, $\mathbf{W} = \text{diag}(\Sigma_1^{-1}, \ldots, \Sigma_n^{-1})$. Because Σ_i's are unknown, we can use the estimators $\widehat{\Sigma}_i$ in (2.9), so the final refined estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \{\mathbf{Z}^T(\mathbf{I}-\mathbf{S})^T\widehat{\mathbf{W}}(\mathbf{I}-\mathbf{S})\mathbf{Z}\}^{-1}\mathbf{Z}^T(\mathbf{I}-\mathbf{S})^T\widehat{\mathbf{W}}(\mathbf{I}-\mathbf{S})\boldsymbol{Y}, \tag{2.11}$$

where $\widehat{\mathbf{W}} = \text{diag}(\widehat{\Sigma}_1^{-1}, \ldots, \widehat{\Sigma}_n^{-1})$. The profile least squares estimator for the nonparametric component is simply $\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\beta}}}(\cdot)$.

2.2.4 Adjusted covariance function estimator

It can be shown that $\widehat{R}_\eta$ in Step 2 is a consistent estimator of R_η, but is not guaranteed to be positive semidefinite, and therefore some adjustment is needed to enforce the positive semidefinite condition, which is particularly important when the sample size is relatively small. The idea is to take a spectral decomposition of $\widehat{R}_\eta$ and truncate the negative components.

Computing the eigenvalues and eigenfunctions of a symmetric bivariate function is generally non-trivial. Typically this is done by discretizing the covariance function estimation and approximating its eigenfunctions by the respective eigenvectors (see [11]). However, discretizing the covariance function is quite subjective since it depends heavily on the choice of dense grids. Fortunately, if the choice of the penalty function $P(R_\eta)$ for $R_\eta(s,t) =$

$\sum_{j\geq 1} a_j f_j(s) g_j(t)$ follows [2],

$$P(R_\eta) \;=\; \|R_\eta\|^2_{\mathcal{W}_2^2(\mathcal{T})\otimes\mathcal{W}_2^2(\mathcal{T})}, \tag{2.12}$$

then

$$\begin{aligned} P(R_\eta) &= \int\int_{t\in\mathcal{T},s\in\mathcal{T}} \big\{\frac{\partial^4 R_\eta(s,t)}{\partial^2 s \partial^2 t}\big\}^2 ds dt \\ &= \int\int_{t\in\mathcal{T},s\in\mathcal{T}} \big\{\sum_{j\geq 1} a_j f_j''(s) g_j''(t)\big\}^2 ds dt \\ &= \sum_{i,j\geq 1}\sum a_i a_j \int_{\mathcal{T}} \{f_i''(s) f_j''(s)\} ds \int_{\mathcal{T}} \{g_i''(t) g_j''(t)\} dt, \end{aligned}$$

and the minimizer $\widehat{R}_\eta$ of (2.6) must have the following form

$$\widehat{R}_\eta(s,t) = \sum_{i=1}^n \boldsymbol{H}_i(s)^T \widehat{\mathbf{A}}_i \boldsymbol{H}_i(t) = \sum_{i=1}^n \sum_{j,k=1}^{J_i} \widehat{\mathbf{A}}_i(j,k) H(s,t_{i,j}) H(t,t_{i,k}), \tag{2.13}$$

where $\widehat{\mathbf{A}}_i$ is a $J_i \times J_i$ symmetric matrix, $\widehat{\mathbf{A}}_i(j,k)$ is the element in the j-th row and k-th column of $\widehat{\mathbf{A}}_i$, and

$$\boldsymbol{H}_i(s) = \big(H(s,t_{i,1}), \ldots, H(s,t_{i,J_i})\big)^T \in \mathbb{R}^{J_i},$$

where $H(s,t) = \frac{1}{4} B_2(s) B_2(t) - \frac{1}{24} B_4(|s-t|)$ and B_r is the rth Bernoulli polynomial, see [16]. Thanks to the representation (2.13), the eigenvalues and eigenfunctions of $\widehat{R}_\eta(s,t) = \sum_{i=1}^n \boldsymbol{H}_i(s)^T \widehat{\mathbf{A}}_i \boldsymbol{H}_i(t)$ can actually be computed explicitly. Let $\boldsymbol{h}(\cdot) = (\boldsymbol{H}_1(\cdot)^T, \ldots, \boldsymbol{H}_n(\cdot)^T)^T$, and

$$\widehat{\mathbf{A}} = \begin{pmatrix} \widehat{\mathbf{A}}_1 & 0 & \ldots & 0 \\ 0 & \widehat{\mathbf{A}}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \widehat{\mathbf{A}}_n \end{pmatrix}_{N\times N}$$

where $\widehat{\mathbf{A}}_i$'s are defined in (2.13) and $N = J_1 + \ldots + J_n$ is the total number of observations.

Assume that $\widehat{\mathbf{A}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{U}}^T$ is the eigenvalue decomposition of $\widehat{\mathbf{A}}$, where $\widehat{\mathbf{\Lambda}} = \text{diag}\{\widehat{\lambda}_{(1)}, \widehat{\lambda}_{(2)}, \ldots, \widehat{\lambda}_{(N)}\}$ is the diagonal matrix of the decreasing eigenvalues $\widehat{\lambda}_{(1)} \geq \widehat{\lambda}_{(2)} \geq \ldots \geq \widehat{\lambda}_{(N)}$, and $\widehat{\mathbf{U}} \equiv (\widehat{\boldsymbol{u}}_1, \widehat{\boldsymbol{u}}_2, \ldots, \widehat{\boldsymbol{u}}_N)$ is the matrix of the corresponding eigenvectors. Then $\widehat{R}_\eta(s,t)$ in (2.13) admits the following spectral decomposition:

$$\begin{aligned}
\widehat{R}_\eta(s,t) &= \boldsymbol{h}(s)^T\widehat{\mathbf{A}}\boldsymbol{h}(t) \\
&= \boldsymbol{h}(s)^T\big(\widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{U}}^T\big)\boldsymbol{h}(t) \\
&= \boldsymbol{h}(s)^T\Big\{\sum_{k=1}^{N}\widehat{\lambda}_{(k)}\widehat{\boldsymbol{u}}_k\widehat{\boldsymbol{u}}_k^T\Big\}\boldsymbol{h}(t) \\
&\equiv \sum_{k=1}^{N}\widehat{\lambda}_{(k)}\widehat{\psi}_k(s)\widehat{\psi}_k(t)
\end{aligned}$$

where $\widehat{\psi}_k(\cdot) = \widehat{\boldsymbol{u}}_k^T\boldsymbol{h}(\cdot)$ is the estimator of $\psi_k(\cdot)$ in (1.5). Since $\widehat{\lambda}_{(k)}$'s are not necessarily positive, we can first truncate the negative eigenvalues. Then the adjusted estimators for R_η and R are defined as

$$\widetilde{R}_\eta(s,t) = \sum_{k=1}^{N}\widehat{\lambda}_{(k)}\text{I}(\widehat{\lambda}_{(k)} > \tau)\widehat{\psi}_k(s)\widehat{\psi}_k(t), \quad s,t \in \mathcal{T}, \tag{2.14}$$

$$\widetilde{R}(s,t) = \widetilde{R}_\eta(s,t)\text{I}(s \neq t) + \max\{\widetilde{R}_\eta(t,t), \widehat{\sigma}^2(t)\}\text{I}(s = t), \tag{2.15}$$

where $\tau \geq 0$ is a predetermined threshold for the eigenvalues (e.g., 0.01) or a percentage (e.g., 1 percent) of the sum of all the positive eigenvalues. It will be shown in the next section that $\widetilde{R}(s,t)$ is positive definite. Thus when we want to take the inverse of $\widehat{\Sigma}_i$ in (2.9) to estimate $\boldsymbol{\beta}$, it is better to replace $\widehat{R}(t_{i,j}, t_{i,k})$ by $\widetilde{R}(t_{i,j}, t_{i,k})$ in the expression of $\widehat{\Sigma}_i$ in (2.9) and $\widehat{\mathbf{W}}$ in (2.11).

2.3 Theoretical results

In this section we investigate sampling properties of the covariance function estimator as $n \to \infty$. The proposed estimation procedures are applicable for various formulations for collecting longitudinal data. To facilitate the presentation, we assume that $\{J_i : i = 1, \ldots, n\}$ are independent and identically distributed random variables with $0 < \mathrm{E}(J_i) < \infty$, and $\{t_{i,1}, \ldots, t_{i,J_i} \mid J_i\}_{i=1}^n$ are independent and identically distributed on $\mathcal{T}$ with a density $f_T(t)$, see [6]. In this section and Sections 2.4–2.5, the penalty function $P(R_\eta)$ takes the special form (2.12).

First, we show that the residuals $\widehat{\varepsilon}_i(t_{i,j})$ in (2.5) are uniformly consistent estimators for the true errors $\varepsilon_i(t_{i,j})$.

Proposition 1 *Assume regularity conditions* C1–C5 *in the Appendix. If* $\mathrm{E}\{\boldsymbol{X}(t)\boldsymbol{X}(t)^T\}$ *is positive definite for each* $t \in \mathcal{T}$*, then*

$$\sup_{i,j} |\widehat{\varepsilon}_i(t_{i,j}) - \varepsilon_i(t_{i,j})| = O_{\mathrm{P}}(h_1^2 + \{-\log(h_1)/(nh_1)\}^{1/2}).$$

From the proof of Proposition 1 and the definition of $\widehat{\boldsymbol{\beta}}$ in (2.11), it is easy to derive the consistency of the proposed estimators of $\boldsymbol{\alpha}(t)$ and $\boldsymbol{\beta}$ in Corollary 1.

Corollary 1 *Under the conditions of Proposition* 1*, we have*

$$\begin{aligned} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= O_{\mathrm{P}}(n^{-1/2}); \\ \sup_{t \in \mathcal{T}} \left| \widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\beta}}}(t) - \boldsymbol{\alpha}(t) \right| &= O_{\mathrm{P}}(h_1^2 + \{-\log(h_1)/(nh_1)\}^{1/2}). \end{aligned}$$

Next, define $\overline{\sigma}^2(t) = \overline{\gamma}_0$, where $(\overline{\gamma}_0, \overline{\gamma}_1)$ minimize

$$\sum_{i=1}^{n} \sum_{j=1}^{J_i} \{\varepsilon_i^2(t_{i,j}) - \gamma_0 - \gamma_1(t_{i,j} - t)\}^2 K_{h_2}(t_{i,j} - t) \tag{3.1}$$

and $\overline{R}_\eta \in \mathcal{W}_2^2(\mathcal{T}) \otimes \mathcal{W}_2^2(\mathcal{T})$ be the minimizer of

$$\frac{1}{\sum_{i=1}^n J_i(J_i-1)} \sum_{i=1}^n \sum\sum_{1\le j_1 \ne j_2 \le J_i} \left\{\varepsilon_i(t_{i,j_1})\varepsilon_i(t_{i,j_2}) - R_\eta(t_{i,j_1}, t_{i,j_2})\right\}^2 + \lambda_n P(R_\eta). \tag{3.2}$$

So (2.7) and (2.6) are data versions of (3.1) and (3.2) after we replace the unobserved $\varepsilon_i(t_{i,j})$ by residuals $\widehat{\varepsilon}_i(t_{i,j})$. $\overline{\sigma}^2$ and $\overline{R}_\eta$ are called pseudo-estimators for σ^2 and R_η. The next proposition showed the consistency of $\overline{\sigma}^2$ and $\overline{R}_\eta$.

Proposition 2 (i) *Under the regularity conditions* C1–C7 *in the Appendix,*

$$\sup_{t\in\mathcal{T}} \left|\overline{\sigma}^2(t) - \sigma^2(t)\right| = O_{\mathrm{P}}(h_2^2 + \{-\log(h_2)/(nh_2)\}^{1/2}). \tag{3.3}$$

(ii) *Under the regularity conditions* C1–C7 *in the Appendix,*

$$\|\overline{R}_\eta - R_\eta\|_{L_2} = O_{\mathrm{P}}\left(\{\log(n)/n\}^{2/5}\right), \tag{3.4}$$

if the tuning parameter $\lambda_n \asymp \{\log(n)/n\}^{2/5}$. *Here* $\|R_\eta(\cdot,\cdot)\|_{L_2} = \left\{\int\int_{s,t\in\mathcal{T}} R_\eta^2(s,t)dsdt\right\}^{1/2}$ *denotes the integrated squared norm of a bivariate function, see [1].*

Now let's consider the rates of convergence for the estimators $\overline{\sigma}^2$ and $\overline{R}_\eta$. From (3.3), when $h_2 \asymp \{\log(n)/n\}^{1/5}$, we have $\sup_{t\in\mathcal{T}} \left|\overline{\sigma}^2 - \sigma^2\right| = O_{\mathrm{P}}\left(\{\log(n)/n\}^{2/5}\right)$. Thus $\overline{\sigma}^2$ has the same optimal rates of convergence as $\overline{R}_\eta$ in (3.4). On the other hand, if we use two-dimensional smoothing techniques to estimate $R_\eta(\cdot,\cdot)$, the optimal L_2-convergence rate is $n^{-1/3}$, which is much larger than the convergence rate $\{\log(n)/n\}^{2/5}$ in Proposition 2. This is because the bivariate continuous functions space with continuous second derivatives $C^{(2)}(\mathcal{T})$ is much larger than the tensor product Hilbert space $\mathcal{W}_2^2(\mathcal{T})\otimes\mathcal{W}_2^2(\mathcal{T})$.

Finally, we show in the next proposition that the adjusted covariance function estimator defined in Section 2.2.4 is positive definite.

Proposition 3 *The adjusted covariance function estimator* $\widetilde{R}(s,t)$ *in* (2.15) *is positive definite almost surely.*

2.4 Simulation study

2.4.1 Simulation 1

In this section, we investigate the finite sample properties of the estimators proposed in Sections 2.2 through Monte Carlo simulations. Suppose the data are generated from the following model:

$$Y_i(t_{i,j}) = \boldsymbol{X}_i(t_{i,j})^T\boldsymbol{\alpha}(t_{i,j})+\boldsymbol{Z}_i(t_{i,j})^T\boldsymbol{\beta}+\eta_i(t_{i,j})+\zeta_i(t_{i,j}),\ \ i=1,\ldots,n,\ j=1,\ldots,J_i. \tag{4.1}$$

We set the sample size n be 200, and $\{J_i : 1 \le i \le n\}$ be independent discrete uniform random variables on $\{5,6,7\}$. Let $\mathcal{T} = [0,1]$ and given J_i, the observation times $\{t_{i,j} : 1 \le i \le n,\ 1 \le j \le J_i\}$ be independent variables with uniform distribution on $\mathcal{T}$. We let the coefficients of $\boldsymbol{\alpha}(t)$ and $\boldsymbol{\beta}$ be two dimensional in our simulation, and further set $X_1(t) \equiv 1$ to include an intercept term. We generate the covariates in the following way: For a given t, $(X_2(t), Z_1(t))^T$ follows a bivariate normal distribution with mean 0, variance 1, and correlation 0.5, and $Z_2(t)$ is a Bernoulli distributed random variable with success probability 0.5 and independent of $X_2(t)$ and $Z_1(t)$. In this simulation we set $\boldsymbol{\beta} = (1,2)^T$, $\alpha_1(t) = \sqrt{t}$, and $\alpha_2(t) = \sin(2\pi t)$. For $i = 1,\ldots,n$, the within-subject errors $\eta_i(t)$ are generated from a temporal mixed effects model:

$$\eta_i(t) = \sum_{k=1}^{L} \xi_{i,k}\psi_k(t) \tag{4.2}$$

where $\xi_{i,k}$'s are independent standard normal random variables. Thus the covariance function is $R_\eta(s,t) = \sum_{k=1}^{L} \psi_k(s)\psi_k(t)$. We set $L = 1$, and $\psi_1(t) = \cos(\pi t)$. Finally we assume the measurement errors $\zeta_i(t_{i,j})$ follow $N(0, \left(\sqrt{0.1}\right)^2)$ and are independent of $\eta_i(t_{i,j})$.

For comparison, in each simulated dataset, we fit the model using four different estimators of the covariance function:

Method I. Working independence (WI) estimator (see [14]).

Method II. Our proposed method.

Method III. True covariance function.

Method IV. Misspecified ARMA$(1,1)$ method. We assume the covariance function $R(s,t)$ for $\varepsilon_i(t) = \eta_i(t) + \zeta_i(t)$ is of the form

$$R(s,t) = R_\eta(s,t)\mathrm{I}(s \neq t) + \sigma^2(t)\mathrm{I}(s = t), \tag{4.3}$$

where $R_\eta(s,t) = \sigma(s)\sigma(t)\gamma\rho^{|s-t|}$, has the misspecified ARMA$(1,1)$ structure. The parameters $\gamma \in [0,1]$ and $\rho \in [0,1]$ in the Method IV are estimated using the quasi maximum likelihood method (QL), see [6]. For a fair comparison, we use the same bandwidth $h_1 = 0.1$ when estimating $\boldsymbol{\alpha}(\cdot)$ for all three estimators. The bandwidth h_2 in (2.7) is selected by the plug-in method (Fan and Gijbels [29]). Throughout the simulations and the real data examples in the next section, we use the Epanechnikov kernel, and the tuning parameter λ_n in (2.6) is selected automatically by the package "ssfcov", which is based on the tuning parameter selection method in smoothing splines.

Table 2.1 summarizes of the results over 200 simulations. We assess the performance of different approaches by calculating the bias and standard errors of 200 estimates. In the table, "Bias" represents the median of the 200

Table 2.1: **(Simulation 1)** Compare the performance of $\widehat{\boldsymbol{\beta}}$ using different methods (with $n = 200$)

	$\widehat{\beta}_1$			$\widehat{\beta}_2$		
Method	Bias	SD	RE	Bias	SD	RE
Method I	-.0021	.0255	.33	-.0002	.0378	.22
our Method II	.0013	.0184	.63	-.0033	.0294	.37
Method III	.0005	.0146	1.0	-.0008	.0179	1.0
Method IV	-.0031	.0198	.54	-.0034	.0298	.36

estimates subtracting the true value, "SD" represents the median absolute deviation of the 200 estimates divided by a factor of 0.6745, and "RE" of a current estimator represents the relative efficiency between the oracle estimator (Method III) and the current estimator, which is defined as SD^2(oracle estimator)/SD^2(current estimator). Intuitively, if the relative efficiency of a method is larger, the SD of the coefficient using this method is smaller, so this method will be better (i.e., more efficient).

From Table 2.1, all parameter estimators considered are asymptotically unbiased, which is confirmed from the numerical results: the biases are much smaller than the standard errors in all cases. In terms of the efficiency, theoretically, the oracle estimator (Method III) using the true covariance function should be the best, and our proposed approach (Method II) should be much better than that using both working independence structure which ignored $\eta_i(t)$ (Method I) and the misspecified ARMA$(1, 1)$ correlation structure for $\eta_i(t)$ (Method IV). The results in Table 2.1 agreed with this conjecture. For example, our proposed method has 30% efficiency gain over the estimator as-

suming working independence (Method I) for $\widehat{\beta}_1$, while the Method IV has only 21% efficiency gain over the estimator assuming working independence (Method I) for $\widehat{\beta}_1$.

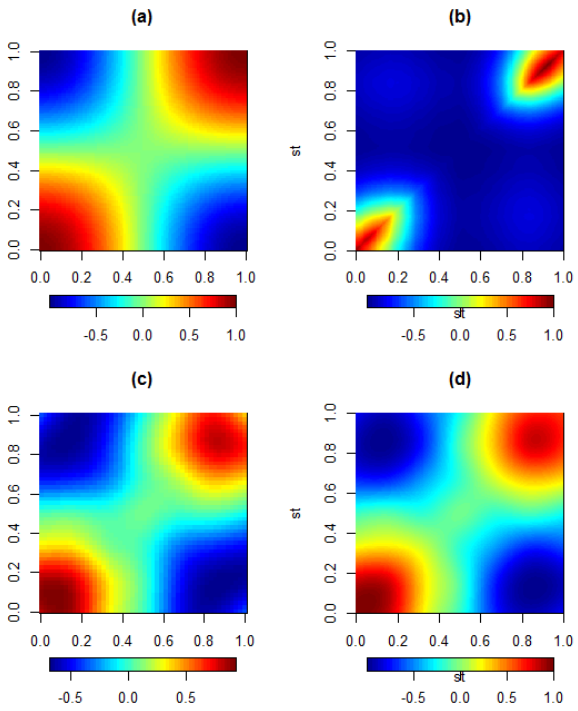

Figure 1: **(Simulation 1)** Panel (a): true covariance function $R_\eta(\cdot,\cdot)$; panel (b): estimated covariance function $\widehat{R}_\eta(\cdot,\cdot)$ using ARMA$(1,1)$ model for $\eta_i(t)$ (Method IV); panel (c): estimated covariance function $\widehat{R}_\eta(\cdot,\cdot)$ using our proposed method (Method II); panel (d): adjusted covariance function estimator $\widetilde{R}_\eta(\cdot,\cdot)$ based on (2.14) for one simulated data (with $n = 200$).

Figure 1 shows the advantage of our proposed method. From the plot, the true covariance function $R_\eta(s,t)$ for $\eta_i(t)$ in Figure 1(a) is very complicated and cannot be estimated by any parametric model such as AR(1) or ARMA$(1,1)$

model. Figure 1(b) shows the parametric ARMA$(1,1)$ estimator of $R_\eta(s,t)$ for $\eta_i(t)$ and Figure 1(c) shows the nonparametric estimator $\widehat{R}_\eta(s,t)$ using our proposed method for one simulated data. Figure 1(d) shows the adjusted estimator $\widetilde{R}_\eta(s,t)$ in (2.14), which is very close to $\widehat{R}_\eta(s,t)$ in Figure 1(c) but is positive definite. From the plot, our proposed method captures the structure of covariance function very well, and the estimators $\widehat{R}_\eta(s,t)$ and $\widetilde{R}_\eta(s,t)$ are obviously consistent.

2.4.2 Simulation 2

In simulation 2, we study the robustness of our proposed method. The data are generated with the same setup as in the previous simulation, except that $\eta_i(t)$ is a Gaussian process with the covariance function $R_\eta(s,t)$ in (4.3). We set the sample size n be 100 and the marginal variance $\sigma^2(t) = 1$, and set $\rho = 0.35$ and $\gamma = 0.75$ in (4.3). We apply the semiparametric regression methods assuming working independence (Method I), nonparametric covariance (Method II), true covariance (Method III) and ARMA$(1,1)$ covariance structure for $\eta_i(t)$ (Method IV) to the simulated data and repeat the simulation 200 times. The selection methods of h_1, h_2 and λ_n are the same as in the previous simulation. The results for estimating $\boldsymbol{\beta}$ are summarized in Table 2.2.

From Table 2.2, again all parameter estimators are asymptotically unbiased since the biases are much smaller than the standard errors. When we compare the efficiencies of the estimates, theoretically, since the true correlation structure for $\eta_i(t)$ is ARMA$(1,1)$ model, the efficiency of the estimator using ARMA$(1,1)$ correlation structure (Method IV) should be close to the oracle estimator (Method III) using the true covariance function, and both of

Table 2.2: **(Simulation 2)** Compare the performance of $\widehat{\boldsymbol{\beta}}$ using different methods (with $n = 100$)

	$\widehat{\beta}_1$			$\widehat{\beta}_2$		
Method	Bias	SD	RE	Bias	SD	RE
Method I	.0055	.0517	0.45	.0043	.0819	0.60
our Method II	.0052	.0415	0.70	.0005	.0742	0.74
Method III	-.0013	.0348	1.0	.0050	.0637	1.0
Method IV	-.0008	.0341	1.04	.0057	.0625	1.04

them should be more efficient than our proposed approach (Method II). The estimator using working independence structure (Method I) should be the least efficient. Again the results in Table 2.2 agreed with this conjecture. Our proposed estimator (Method II), on the other hand, performs reasonably well: the standard errors of our estimators are much smaller than those of the working independence estimator (Method I), for example, our proposed method has 25% efficiency gain over the estimator assuming working independence (WI) for $\widehat{\beta}_1$.

2.5 Real data example

2.5.1 Case 1: Multi-Center AIDS Cohort data

We now present an application of our proposed method to the Multi-Center AIDS Cohort study. The dataset comprises the information of 283 subjects who were infected with human immunodeficiency virus (HIV) during the s-

tudy in year 1984-1991. A total of $N = 1817$ observations were made in this study, with between 1 and 14 observations per subject. This dataset was also analyzed by Fan, Huang and Li [6] and Huang, Wu, and Zhou [12]. Our target is to describe the trend in mean CD4 (cluster of differentiation 4) percentage depletion over time and to evaluate the effects of smoking, pre-HIV infection CD4 percentage, and age at infection on the mean CD4 percentage.

We take the response Y to be CD4 cell percentage, X_1 to be the standardized variable for PreCD4, Z_1 to be the smoking status (1 for a smoker and 0 for a nonsmoker) and Z_2 to be the standardized variable for age. The observation time is divided by 6 so that the rescaled observation time t is in between 0 and 1. Now consider a semiparametric varying-coefficient partially linear model

$$Y(t) = \alpha_1(t) + \alpha_2(t)X_1(t) + \beta_1 Z_1(t) + \beta_2 Z_2(t) + \eta(t) + \zeta(t). \tag{5.1}$$

We apply a multifold cross-validation method to select a bandwidth h_1 for $\boldsymbol{\alpha}(t)$. After partitioning the data into 14 groups, we fit model (5.1) for the data excluding the kth-group for each $k = 1, \ldots, 14$. For the computational issue, we minimize the cross-validation (CV) score on a rough grid $h_1 \in \{0.5\kappa^b : b = 0, \ldots, 12\}$, with $\kappa = 0.8$. The resulting optimal bandwidth is $h_1^{\text{opt}} = 0.054$. We can estimate $\alpha_1(t)$ and $\alpha_2(t)$ more precisely by choosing different bandwidths 0.054 and 0.081 for $\alpha_1(t)$ and $\alpha_2(t)$ respectively to avoid the under-smoothness of $\alpha_2(t)$, see Fan and Zhang [8]. As for the estimation of $\sigma^2(t)$, this is a one-dimensional kernel regression of the squared residuals. In this application, we directly use the plug-in bandwidth selector (Fan and Gijbels [29]) and choose the bandwidth $h_2^{\text{opt}} = 0.080$.

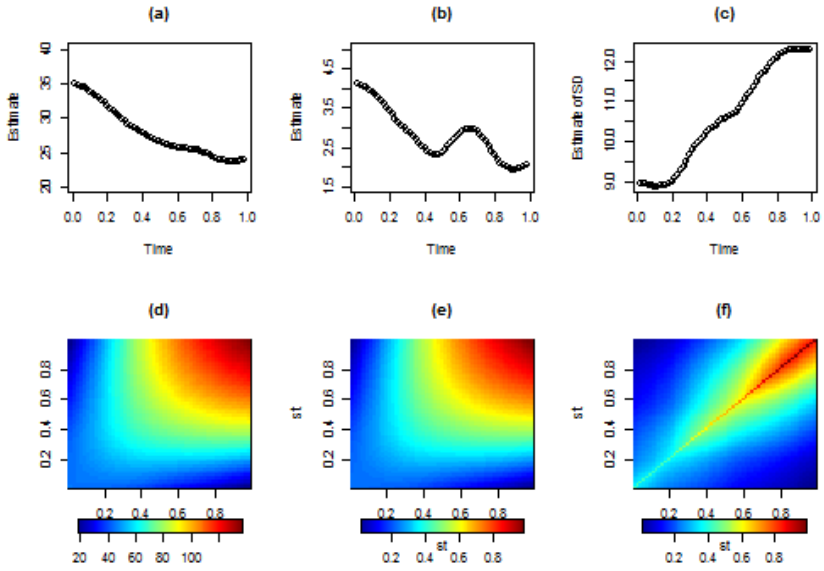

Figure 2: (**Real data case 1**) Panel (a): estimate of $\alpha_1(t)$; panel (b): estimate of $\alpha_2(t)$; panel (c): estimate of $\sigma(t)$. In panels (a)–(c), circles $(\circ)$ represent the estimates of the functions at observation times, with lines $(-)$ connecting them. Panel (d): estimated covariance function $\widehat{R}_\eta(\cdot,\cdot)$ using our proposed method for $\eta_i(t)$; panel (e): adjusted covariance function estimator $\widetilde{R}_\eta(\cdot,\cdot)$ based on (2.14); panel (f): estimated covariance function $\widehat{R}_\eta(\cdot,\cdot)$ using ARMA$(1,1)$ model for $\eta_i(t)$.

The resulting estimate of $\boldsymbol{\alpha}(t)$ is depicted in Figures 2(a) and 2(b). The intercept function decreases with time, implying an overall trend of CD4 cell percentage is decreasing over time. The trend for $\alpha_2(t)$ implies that the impact of PreCD4 on CD4 cell percentage decreases gradually during the first 3 years after infection and then increases a bit. The resulting estimate of $\sigma(t)$ in Figure 2(c), indicates that $\sigma(t)$ seems to be increasing as time increases.

This means predicting the CD4 percentage becomes harder and harder over time. Figure 2(d), 2(e) and 2(f) show the estimates $\widehat{R}_\eta(\cdot,\cdot)$, $\widetilde{R}_\eta(\cdot,\cdot)$ using our proposed method and the estimate $\widehat{R}_\eta(\cdot,\cdot)$ based on ARMA$(1,1)$ structure for $\eta_i(t)$, which characterizes the within subject dependence. From the plot, our estimate of the covariance function $R_\eta(\cdot,\cdot)$ is quite different from that using the ARMA$(1,1)$ structure for $\eta_i(t)$, so the within subject correlation is misspecified if we just use quasi maximum likelihood method (QL).

Table 2.3: **(Real data case 1)** Compare estimates of $\boldsymbol{\beta}$ using different methods

	Method I	Method II	Method IV
$\widehat{\beta}_1$	0.686	0.563	0.734
$\widehat{\beta}_2$	-0.529	0.096	0.045

Table 2.3 shows the estimates of β_1 and β_2 with three different covariance structures: working independence (Method I), our proposed method (Method II) and ARMA$(1,1)$ covariance structure for $\eta_i(t)$ by Fan, Huang and Li [6] (Method IV). The estimates $(\widehat{\beta}_1, \widehat{\beta}_2)$ using Method II and Method IV are quite different, so the ARMA$(1,1)$ covariance structure for $\eta_i(t)$ is misspecified. Finally, the effects of age on the mean CD4 percentage is negative if we assume working independence (Method I) but is positive if we use more efficient estimation method.

2.5.2 Case 2: Progesterone data

We now apply the proposed methods to the longitudinal progesterone data. Progesterone, which is a reproductive hormone, is responsible for normal fer-

tility and menstrual cycling. A longitudinal hormone study on progesterone collected urine samples from 34 healthy women (control group) in a menstrual cycle on alternative days, see [17]. A total of 492 observations were made in this study, with between 11 and 28 observations per subject. The observation time is divided by 30 so that the rescaled observation time t is in between 0 and 1.

Similar to the procedure by Zhang *et al.* [24], a logarithmic transformation is applied on the progesterone level to make the data homoscedastic. We take the response to be the difference between the jth log-transformed progesterone level measured at rescaled time $t_{i,j}$ and the individual's average log-transformed progesterone level. For the ith subject, let X_i and Z_i denote age and body mass index, both of which are standardized to have mean 0 and standard deviation 1. We consider the semiparametric model

$$Y_i(t_{i,j}) = \alpha_1(t_{i,j}) + \beta_1 X_i(t_{i,j}) + \beta_2 Z_i(t_{i,j}) + \eta_i(t_{i,j}) + \zeta_i(t_{i,j}). \tag{5.2}$$

We apply a multifold cross-validation method to select a bandwidth h_1 for $\alpha_1(t)$. We partition the data into 17 groups, and each group contains 2 subjects. We fit model (5.2) for the data excluding the kth-group for each $k = 1, \ldots, 17$. For the computational issue, we minimize the cross-validation (CV) score on a rough grid $h_1 \in \{0.1\kappa^b : b = 0, \ldots, 12\}$, here we choose $\kappa = 0.9$. From Figure 3(a), the resulting optimal bandwidth is $h_1^{\text{opt}} = 0.043$. As for the estimation of $\sigma^2(t)$, this is a one-dimensional kernel regression of the squared residuals. In this application we directly use the plug-in bandwidth selector (Fan and Gijbels [29]) and choose the bandwidth $h_2^{\text{opt}} = 0.096$.

The resulting estimate of $\alpha_1(t)$ is depicted in Figures 3(b). The shape

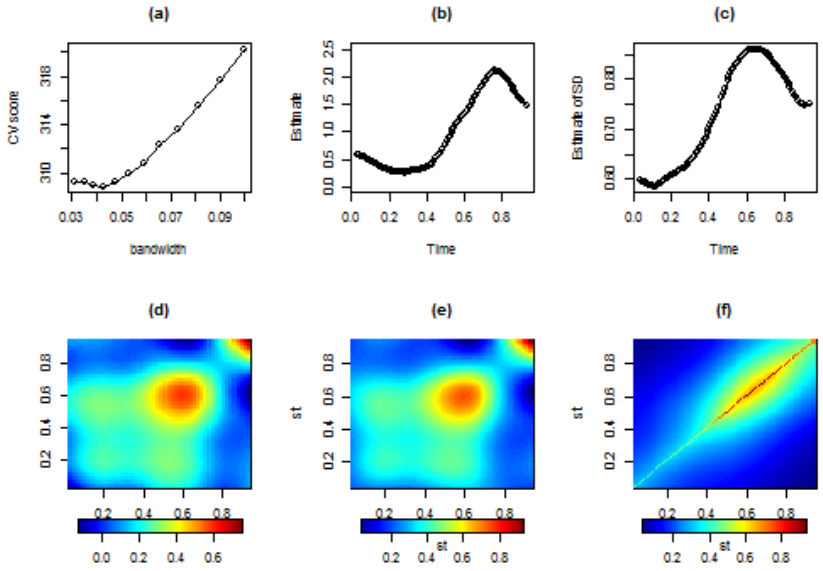

Figure 3: **(Real data case 2)** Panel (a): cross validation score; panel (b): estimate of $\alpha_1(t)$; panel (c): estimate of $\sigma(t)$; panel (d): estimated covariance function $\widehat{R}_\eta(\cdot,\cdot)$ using our proposed method for $\eta_i(t)$; panel (e): adjusted covariance function estimator $\widetilde{R}_\eta(\cdot,\cdot)$ based on (2.14); panel (f): estimated covariance function $\widehat{R}_\eta(\cdot,\cdot)$ using ARMA$(1,1)$ model for $\eta_i(t)$.

of intercept function implies an overall trend of progesterone level over time. The resulting estimate of $\sigma(t)$ in Figure 3(c), indicates that $\sigma(t)$ seems to be increasing as time increases from 0 to 0.6. This means predicting the progesterone level becomes harder and harder over time. Figure 3(d), 3(e) and 3(f) show the estimates $\widehat{R}_\eta(\cdot,\cdot)$, $\widetilde{R}_\eta(\cdot,\cdot)$ using our proposed method and the estimate $\widehat{R}_\eta(\cdot,\cdot)$ based on ARMA$(1,1)$ structure for $\eta_i(t)$, which characterizes the within subject dependence. From the plot, our estimate of the covariance function is quite complicated and different from the ARMA$(1,1)$ or AR(1)

covariance structure for $\eta_i(t)$, so the within subject correlation is misspecified if we use quasi maximum likelihood method (QL).

Table 2.4: **(Real data case 2)** Compare estimates of $\boldsymbol{\beta}$ using different methods

	Method I	Method II	Method IV	Mixed Model [24]
$\widehat{\beta}_1$	0.082	-0.009	0.068	0.925
$\widehat{\beta}_2$	-0.117	-0.187	-0.099	-2.913

Table 2.4 shows the estimates of β_1 and β_2 with four covariance structures: working independence (Method I), our proposed method (Method II), ARMA$(1,1)$ covariance structure for $\eta_i(t)$ (Method IV), and Mixed model by Zhang, et. al. [24], which is a special case of (1.5). The values of $(\widehat{\beta}_1, \widehat{\beta}_2)$ between Method II and Method IV are quite different, showing that the ARMA$(1,1)$ structure for $\eta_i(t)$ is misspecified. Finally, the effects of age on the mean progesterone level is negative if we use our proposed method (Method II), but is positive if we assume other three methods.

2.6 Discussion

In this article we proposed a class of nonparametric models for the covariance function of longitudinal data at irregular or subject specific time points. We further developed an estimation procedure for $\sigma^2(t) = R(t,t)$ using local linear regression, estimation procedure for $R_\eta(t_1, t_2)$ using regularization approach, and estimation procedure for regression coefficients $\boldsymbol{\alpha}(t)$ and $\boldsymbol{\beta}$ using profile weighted least squares. We also showed that the varying-coefficient temporal

mixed effects model is a good approximation of the semiparametric varying-coefficient partially linear model.

Although we just focused on analyzing the longitudinal data at irregular or subject specific time points, our proposed method can also be applied to equally-spaced or regular time points. In this balanced case, directly estimating the "covariance matrix" of the errors may be better than estimating the "covariance function" of the errors, but our method continues to work reasonably well and will not lose much efficiency.

Several issues are desirable for future research. First, in the presence of outliers, one should consider a robust method to estimate $\boldsymbol{\alpha}(t)$ and $\boldsymbol{\beta}$ instead of using profile weighted least squares. Second, in the simulations and real data examples we just checked the plots of the estimated covariance functions and argued that the covariance function for $\eta_i(t)$ cannot be estimated by any parametric model such as AR(1) or ARMA$(1,1)$ model. It is better to develop a new procedure to test whether the covariance structure for $\eta_i(t)$ has a parametric form such as ARMA$(1,1)$ model. This research topic is beyond the scope of this article and further research is needed.

Chapter 3

Adaptive Jumps Detection via Screening and Multiple Testing Procedure

3.1 Model structure and notations

We assume that a sample of n data pairs $\{(X_1, Y_1), ..., (X_n, Y_n)\}_{i=1}^n$ is observed. After rescaling, we assume that design points X_i's are either regularly spaced on $I = [0, 1]$ or are the order statistics of a random sample from a distribution having a density f supported on $I = [0, 1]$. The data are generated from the model

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i, \qquad i = 1, ..., n, \tag{1.1}$$

where ε_i's are i.i.d. errors, with $\mathbf{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = 1$, and are independent of X_i's. The regression function $m(\cdot)$ is smooth except at finite but unknown

number of points, that means

$$m(x) = m_0(x) + \sum_{j=1}^{q} \gamma_j I_{[\tau_j,1]}(x) \tag{1.2}$$

where $m_0(\cdot)$ is smooth function, q is the number of the jump points, $\tau_1, ..., \tau_q$ are the locations of jump points, and γ_j denotes the jump size at point τ_j, $I_{[a,b]}$ is the indicator function on an interval $[a, b]$. We are interested in the quantities q, τ_j and γ_j. The following conditions are required:

(a). The second order derivative of function $m_0(x)$ is continuous .

(b). The conditional variance $\sigma^2(x) = \text{Var}(Y|X = x)$ is continuous.

(c). $q < \infty$ and the jump points satisfy $\tau_0 \equiv 0 < \tau_1 < ... < \tau_q < 1 \equiv \tau_{q+1}$, and $\xi \equiv \min_{1\le j\le q+1}[\tau_j - \tau_{j-1}] > 0$.

Conditions (a) and (b) are standard assumptions in nonparametric regression and are relatively mild. Here we assume no jumps on the conditional variance $\sigma^2(\cdot)$. Condition (c) means that the jump points are not too close to each other. Of course, condition (c) can be relaxed, as long as the number q of jump points diverges at a slower rate than the sample size n.

Let us introduce some notations. Denote the left and right limit of $m(\cdot)$ at point x by

$$m_-(x) = \lim_{t \nearrow x} m(t), \quad m_+(x) = \lim_{t \searrow x} m(t),$$

then $m_+(\tau_j) - m_-(\tau_j) = \gamma_j$ is the jump size at the jump point τ_j of the regression function, and $m_+(x) - m_-(x) = 0$ when x is a continuous point in $(0, 1)$. If we can estimate $m_+(x)$ and $m_-(x)$ respectively, say $\widehat{m}_+(x)$ and

$\widehat{m}_-(x)$, then we have

$$\begin{aligned} |\widehat{m}_+(x) - \widehat{m}_-(x)| &\approx |\gamma_j|, \quad \text{if } x = \tau_j; \\ |\widehat{m}_+(x) - \widehat{m}_-(x)| &\approx 0, \qquad \text{if } x \text{ is continuous point.} \end{aligned}$$

So the local maximizers of $|\widehat{m}_+(x) - \widehat{m}_-(x)|$ are the candidates the jumps points, see Müller (1992) and Loader (1996).

When estimating the right limit $m_+(x)$ and left limit $m_-(x)$, we only need to use the observations located at the right or left side of point x. The local linear regression (Fan and Gijbels 1996) is proposed to apply since it is known to avoid boundary effects and perform better than the kernel estimators such as Nadaraya-Watson or Gasser-Müller estimator. Morover, the local linear regression is also adaptive to the random design, rather than the fixed design (equally spaced grids) for the wavelet (Wang 1995) in the literatures.

Let $K_-(\cdot)$ be a kernel function which is continuous within its support $[0, 1]$, and let $K_+(x) = K_-(-x)$ for $x \in [-1, 0]$. Then we can replace the symmetric kernel function in the standard local linear smoothing by the one-sided kernels $K_-(\cdot)$ and $K_+(\cdot)$ in the expression of local linear regression, thus we can derive the estimators $\widehat{m}_+(x)$, $\widehat{m}_-(x)$ as follows:

$$\widehat{m}_\pm(x) = \frac{\sum_{i=1}^n w_i^\pm(x) Y_i}{\sum_{i=1}^n w_i^\pm(x)}, \tag{1.3}$$

where

$$w_i^\pm(x) = K_\pm(\frac{x - X_i}{h_n})\big\{S_2^\pm(x) - (x - X_i)S_1^\pm(x)\big\}, \quad i = 1, ..., n \tag{1.4}$$

$$S_l^\pm(x) = \sum_{i=1}^n (x - X_i)^l K_\pm(\frac{x - X_i}{h_n}), \quad l = 0, 1, 2 \tag{1.5}$$

and $h_n > 0$ is the bandwidth.

3.2 Proposed methodology for estimation

In this section, we propose the robust jumps detection procedure in multiple steps. The initial set of the jump points candidates $\{\widehat{\tau}_1, ..., \widehat{\tau}_{\widehat{q}}\}$ is first constructed by screening. With these initial estimates, we then conduct multiple testing to select the real jump points and rule out noises. Finally, we can estimate the jump sizes γ_j based on the refined jump points estimates $\{\widetilde{\tau}_1, ..., \widetilde{\tau}_{\widetilde{q}}\}$.

3.2.1 Step 1: Screening

The estimation of the jump point τ_j and jump size γ_j is based on the difference process:

$$L_n(x) \equiv \widehat{m}_+(x) - \widehat{m}_-(x). \tag{2.1}$$

If there is a unique jump point in the regression function, then the estimator of the unique jump point is just the global maximizer of the process $|L_n(x)|$. Grégoire and Hamrouni (2002) derived the optimal convergence rate $O_p(n^{-1})$ and the asymptotic distribution of the jump point estimator.

When there are multiple jumps, the process $|L_n(x)|$ should have several local maximizers. Then we need a threshold D_n such that any local maximum of $|L_n(x)|$ that is greater than D_n can be viewed as a jump point. Grégoire and Hamrouni (2002) briefly described a method, but it is impossible to apply since the choice of the threshold in their method depends on the minimum of absolute jump sizes: $\min_{1\le j\le q} |\gamma_j|$, which is unknown in advance and cannot be estimated before we know the number of jump points. But we can consider the method of choosing the threshold D_n from another point of view: rather than using the lower bound of the absolute jump sizes $|\gamma_j| = |L_n(\tau_j)|$ at jump

point τ_j, we try to control the upper bound of the difference process $|L_n(x)|$ at continuous points $x \in [0,1]$. As long as we can find the upper bound, any local maximum of $|L_n(x)|$ that is greater than this upper bound can be viewed as a jump point.

In our proposed method, the choice of threshold D_n is not crucial since we only need to make D_n small enough such that all the true jump points are included in the screening. Let $|L_n(x)| = |\widehat{m}_+(x) - \widehat{m}_-(x)|$ be the difference process and D_n be a given threshold (to be specified later on). The initial estimates $\widehat{q}$ and $\widehat{\tau}_j$ can be derived by the following steps. Let

$$\widehat{\omega}_1 = argmax_{x\in[h_n,1-h_n]}|L_n(x)|.$$

If $|L_n(\widehat{\omega}_1)| > D_n$, then denote the neighborhood of $\widehat{\omega}_1$ by $I(\widehat{\omega}_1) = (\widehat{\omega}_1 - 2h_n, \widehat{\omega}_1 + 2h_n)$, let

$$\widehat{\omega}_2 = argmax_{x\in[h_n,1-h_n]\cap I(\widehat{\omega}_1)^c}|L_n(x)|.$$

If $|L_n(\widehat{\omega}_2)| > D_n$, then again denote the neighborhood of $\widehat{\omega}_2$ by $I(\widehat{\omega}_2) = (\widehat{\omega}_2 - 2h_n, \widehat{\omega}_2 + 2h_n)$, and let

$$\widehat{\omega}_3 = argmax_{x\in[h_n,1-h_n]\cap(I(\widehat{\omega}_1)\cup I(\widehat{\omega}_2))^c}|L_n(x)|.$$

$$\vdots$$

We will stop after $\widehat{q}$ steps if $|L_n(\widehat{\omega}_{\widehat{q}})| > D_n$ but $|L_n(\widehat{\omega}_{\widehat{q}+1})| \leq D_n$. Here we search for the local maximizers of $|L_n(x)|$ in the interval $[h_n, 1-h_n]$ other than $[0,1]$ because we want to avoid the boundary points 0 and 1. Let $\widehat{\omega}_{(1)} < \widehat{\omega}_{(2)} < \ldots < \widehat{\omega}_{(\widehat{q})}$ be the order statistics of $\{\widehat{\omega}_1, ..., \widehat{\omega}_{\widehat{q}}\}$, and define the jump points candidates as

$$\widehat{\tau}_1 = \widehat{\omega}_{(1)},\ \ldots\ ,\ \widehat{\tau}_{\widehat{q}} = \widehat{\omega}_{(\widehat{q})}. \tag{2.2}$$

The set of the jump points candidates is defined as $S(D_n) \equiv \{\widehat{\tau}_1, ..., \widehat{\tau}_{\widehat{q}}\}$. The following proposition shows the relationship between different sets $S(D_n)$ of the jump points candidates with different choices of threshold D_n.

Proposition 4 *If d_n is another threshold satisfying $D_n > d_n > 0$, then we have $S(D_n) \subseteq S(d_n)$.*

Instead of the difference process $L_n(x)$ in (2.1), other criterions can also be used in the screening procedure to select the candidates of the jump points. For example, an alternative choice is the local linear estimator of the first derivative, which is studied by Gijbels, Hall and Kneip (1999). Intuitively, if we use a smooth curve to approximate the true regression function $m(\cdot)$, the first derivative or slope of the estimated curve should be large around a jump point and be small elsewhere. Another choice to select the jump points candidates in the screening is to apply the wavelet based method (Wang 1995).

3.2.2 Step 2: Multiple testing

From Proposition 4, as the threshold D_n is decreasing, the set $S(D_n)$ will include more and more candidates of jump points. Of course, some continuous points will also be added into the candidates set $S(D_n)$, especially when the conditional variance function $\sigma^2(x)$ is large at some continuous point $x \in [0, 1]$, see the results in the simulation study in Section 3.4. So we need to conduct multiple testing procedure to check whether each candidate is the true jump point or not.

Consider the multiple testing problem:

$$H_{0j} \quad : \quad m_-(\widehat{\tau}_j) = m_+(\widehat{\tau}_j)$$

$$H_{1j} \quad : \quad m_-(\widehat{\tau}_j) \neq m_+(\widehat{\tau}_j), \qquad j = 1, 2, ..., \widehat{q}. \tag{2.3}$$

We will apply the B-H procedure (Benjamini and Hochberg 1995) to control the false discovery rate (FDR) of the multiple testing. To calculate the p-value of each test, we will apply wild Bootstrap test, which is a modified version of the method proposed by Gijbels and Goderniaux (2004). The p-value p_j for the hypothesis H_{0j} is calculated by the following steps:

Step 1. Compute the residuals $\widehat{\varepsilon}_i = Y_i - \widehat{m}_{US}(X_i)$ for $X_i \in N_j \equiv [\frac{1}{2}\{\widehat{\tau}_{j-1} + \widehat{\tau}_j\}, \frac{1}{2}\{\widehat{\tau}_j + \widehat{\tau}_{j+1}\}]$, where

$$\widehat{m}_{US}(x) = \begin{cases} \widehat{m}_1(x), & \text{if } x \in (\frac{1}{2}\{\widehat{\tau}_{j-1} + \widehat{\tau}_j\}, \widehat{\tau}_j], \\ \widehat{m}_2(x), & \text{if } x \in (\widehat{\tau}_j, \frac{1}{2}\{\widehat{\tau}_j + \widehat{\tau}_{j+1}\}). \end{cases}$$

Here $\widehat{m}_1$ and $\widehat{m}_2$ are the local linear estimators of m on the interval $(\frac{1}{2}\{\widehat{\tau}_{j-1} + \widehat{\tau}_j\}, \widehat{\tau}_j]$ and $(\widehat{\tau}_j, \frac{1}{2}\{\widehat{\tau}_j + \widehat{\tau}_{j+1}\})$, respectively. The subscript US in $\widehat{m}_{US}(x)$ refers to the fact that the resulting estimator is possibly unsmooth at the point $\widehat{\tau}_j$.

Step 2. Obtain samples $\{\nu_i^* : X_i \in N_j\}$ by the following distribution suggested by Mammen(1993):

$$\nu_i^* = \begin{cases} -(\sqrt{5}-1)/2, & \text{with probability}(\sqrt{5}+1)/(2\sqrt{5}), \\ (\sqrt{5}+1)/2), & \text{with probability}(\sqrt{5}-1)/(2\sqrt{5}), \end{cases}$$

and then we can get the bootstrap sample $\chi^* = \{(X_i, Y_i^*) : X_i \in N_j\}$, where

$$Y_i^* = \widehat{m}_S(X_i) + \widehat{\varepsilon}_i \nu_i^*, \quad for\ X_i \in N_j.$$

Here the estimator $\widehat{m}_S$ of m is obtained under the assumption that m is a smooth function having no discontinuities (i.e. $\widehat{\tau}_j = 0$) based on the data $\{(X_i, Y_i) : X_i \in N_j\}$. Construct the bootstrap statistic

$$T^* = \widehat{m}_2^*(\widehat{\tau}_j) - \widehat{m}_1^*(\widehat{\tau}_j) \tag{2.4}$$

where $\widehat{m}_1^*$ and $\widehat{m}_2^*$ are local linear estimator of m on the interval $(\frac{1}{2}\{\widehat{\tau}_{j-1}+\widehat{\tau}_j\}, \widehat{\tau}_j]$ and $(\widehat{\tau}_j, \frac{1}{2}\{\widehat{\tau}_j + \widehat{\tau}_{j+1}\})$ using the bootstrap sample $\chi^* = \{(X_i, Y_i^*) : X_i \in N_j\}$.

Step 3. Repeat Step 2 a large number of times, say B times, and define the p-value p_j via

$$p_j = \frac{\#\{b = 1, ..., B : |T_b^*| \geq |L_n(\widehat{\tau}_j)|\}}{B}, \quad j = 1, ..., \widehat{q}. \tag{2.5}$$

Finally, let us consider the Benjamini-Hochberg (BH) multiple testing procedure (Benjamini and Hochberg 1995) based on the p-values $\{p_1, ..., p_{\widehat{q}}\}$ for a nominal level α, say $\alpha = 0.05$ or 0.1. We reject null hypotheses H_{0j} with $p_j \leq p_{(\widetilde{q})}$, where $\widetilde{q} = \max\{j : p_{(j)} \leq \alpha j/\widehat{q}\}$, and $p_{(1)} \leq \cdots \leq p_{(\widehat{q})}$ denote the ordered p-values $\{p_j\}$. Denote the jump point candidate $\widetilde{\omega}_j \in S(D_n)$ which is corresponding to $p_{(j)}$. Then the refined estimates of the jump points are the order statistics of $\{\widetilde{\omega}_1, ..., \widetilde{\omega}_{\widetilde{q}}\}$:

$$\widetilde{\tau}_1 = \widetilde{\omega}_{(1)}, \ ..., \ \widetilde{\tau}_{\widetilde{q}} = \widetilde{\omega}_{(\widetilde{q})}. \tag{2.6}$$

3.2.3 Step 3: Estimation of jump sizes

Since $m_+(\tau_j) - m_-(\tau_j) = \gamma_j$, we can estimate γ_j based on $\widehat{m}_+(x)$ and $\widehat{m}_-(x)$. After we get the estimators $\widetilde{q}$ and $\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{\widetilde{q}}$, the estimate of the jump size can

be derived as follows:

$$\widehat{\gamma}_j = \widehat{m}_+(\widetilde{\tau}_j) - \widehat{m}_-(\widetilde{\tau}_j), \quad j = 1, \ldots, \widetilde{q}. \tag{2.7}$$

In the volatility analysis for financial data, people are also interested in the jump variation Ψ, which is defined as

$$\Psi = \sum_{j=1}^{q} \gamma_j^2. \tag{2.8}$$

The jump variation Ψ is then estimated by the sum of squares of all of the estimated jump sizes,

$$\widehat{\Psi} = \sum_{j=1}^{\widehat{q}} \widehat{\gamma}_j^2. \tag{2.9}$$

3.3 Theoretical results

In this section, we investigate sampling properties of the estimators as $n \to \infty$. Theorem 1 below justifies the sure screening property of our proposed method: when the threshold D_n is small enough, given any true jump point τ_j, $j = 1, \ldots, q$, we can find a estimated jump point in the candidate set $S(D_n)$ that is very close to τ_j.

Theorem 1 *Assume conditions* (D1)–(D4) *in the Appendix B.*

(a) (*Sure screening property*). *If the threshold* $D_n \to 0$, *then*

$$\lim_{n\to\infty} \mathrm{P}(\widehat{q} \geq q) = 1.$$

Besides, for each jump point τ_j, $j = 1, \ldots, q$, *there exists* $\zeta \in S(D_n)$ *such that*

$$|\zeta - \tau_j| = O_{\mathrm{P}}(n^{-1}), \quad j = 1, \ldots, q.$$

(b) *Under the conditions in part (a), and assuming* $h_n^2 = o(D_n)$ *and* $\log(h_n^{-1})/nh_n = o(D_n^2)$*, we have*

$$\lim_{n\to\infty} \mathrm{P}(\widehat{q} = q) = 1, \tag{3.1}$$

$$|\widehat{\tau}_j - \tau_j| = O_{\mathrm{P}}(n^{-1}), \quad j = 1, \ldots, q. \tag{3.2}$$

Remark 1 *Actually, we can check in the proof of Theorem* 1 *that there exists* $N > 0$*, such that* $\widehat{q} = q$ *when* $n > N$*, if we choose the threshold* D_n *appropriately. So when the sample size is large enough, with probability one, we can estimate the number of jump points without error.*

Remark 2 *Grégoire and Hamrouni* (2002) *have shown that the convergence rate* (n^{-1}) *is the optimal rate for the estimation of jump point* $\widehat{\tau}_j$*. If we use the wavelet based method* (*Fan and Wang* 2007) *to detect jumps, the rate of convergence* $(n^{-1}\log^2 n)$ *is a little bit slower than our proposed method.*

Next, we will derive the asymptotic normality of the jump size estimator $\widehat{\gamma}_j$ in (2.7). Denote $\mu_j = \int_{-\infty}^{0} u^j K_+(u)du$ and $\nu_j = \int_{-\infty}^{0} u^j K_+^2(u)du$.

Theorem 2 *Assume conditions of Theorem 1 and* $nh_n^5 = O(1)$*. Then*

$$\sqrt{nh_n}(\widehat{\gamma}_j - \gamma_j) \xrightarrow{\mathcal{D}} N(0, 2V\sigma^2(\tau_j)/f_X(\tau_j)), \tag{3.3}$$

where

$$V = \frac{\mu_2^2\nu_0 - 2\mu_1\mu_2\nu_1 + \mu_1^2\nu_2}{(\mu_0\mu_2 - \mu_1^2)^2}, \tag{3.4}$$

and the estimate of jump variation $\widehat{\Psi}$ *in* (2.9) *satisfies*

$$\widehat{\Psi} - \Psi = O_{\mathrm{P}}((nh_n)^{-1/2}). \tag{3.5}$$

Remark 3 *One striking result in Theorem* 2 *is that the asymptotic bias term disappeared in* (3.3)*, which is quite different from the standard result in local linear regression in which a stronger condition* $nh_n^5 = o(1)$ *is required. This estimator not only improves the asymptotic bias to the order of typical conditional expectation estimators at an interior point, but actually exhibits further unexpected bias reductions.*

Remark 4 *Under the conditions of Theorem* 2*, the optimal rate of convergence for* $\widehat{\Psi}$ *is* $n^{-2/5}$ *when* $h \asymp n^{-1/5}$*.*

Finally, in the multiple testing procedure, we tested $\widehat{q}$ hypotheses simultaneously using the bootstrap p-values. The following Proposition 5 shows that the false discovery rate FDR can be asymptotically controlled.

Proposition 5 *Assume conditions* (D1)–(D4) *in the Appendix B. The threshold* D_n *is chosen such that* $D_n \to 0$ *and*

$$\max_{1 \le j \le \widehat{q}} |p_j - p_j^0| = o_{\mathrm{P}}(1), \tag{3.6}$$

where p_j^0 *is the true p-value for the hypothesis* H_{0j} *in* (2.3)*. Then the false discovery rate (*FDR*) in the multiple testing procedure can be controlled by level* α *asymptotically.*

3.4 Simulation study

In this section, we investigate the finite sample properties of the estimators proposed in Section 3.2 through Monte Carlo simulations. Suppose the true

data generating process is as follows:

$$Y_i = m_0(X_i) + \sum_{j=1}^{q} \gamma_j \mathrm{I}_{[\tau_j,1]}(X_i) + \sigma(X_i)\varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{4.1}$$

where $m_0(x) = 4x^2 + e^{-x}$ for $x \in [0,1]$, and ε_i's are i.i.d. errors following $N(0, \sigma^2)$ with $\sigma(x) = 0.2\cos(x)$ or $0.5\cos(x)$. We assume the design to be random: $\{X_i : i = 1, \ldots, n\}$ are i.i.d. random variables from the uniform $U[0,1]$ distribution, so that other existing methods can be applied to compare with our proposed method in the simulation study. We set the sample size $n = 5000$ or 10000, the number of jump points $q = 50$, and the jump sizes $\{\gamma_j : 1 \le j \le q\}$ be independent discrete uniform random variables on $\{\pm 1, \pm 0.5\}$. Suppose the location of jump point τ_j can be decomposed into two parts:

$$\tau_j = \Big(-0.005 + \frac{j}{q} \Big) + \frac{\zeta_j}{n}, \qquad j = 1, \ldots, q, \tag{4.2}$$

where ζ_j's are i.i.d. discrete uniform random variables on $\{0,\ \pm 1,\ \ldots,\ \pm 5\}$. That means the true jump point τ_j is a random perturbation of "scheduled" time point $(-0.005 + j/q)$.

Throughout the paper, we will use the one-sided Epanechnikov kernel $K_-(u) = \frac{3}{2}(1-u^2)_+$ for $u \in [0,1]$. Besides, we take the bandwidth $h_n = 0.005$, and the threshold $D_n\{0.5,\ 0.33,\ 0.25,\ 0.2\}$ in the simulation. In the multiple testing procedure, let $B = 200$ in the bootstrap and $\alpha = 0.05$.

Table 3.1 summarizes the results over 100 simulations. For each simulated data with a given sample size n, standard deviation σ and threshold D_n, we record the number of the jump points detected by different methods, and then take the average over 100 simulations. Here the methods "Screening", "*SaMT*" and "Wavelet" refer to applying the screening procedure in Section

Table 3.1: Average number of jump points detected by different methods (with true $q = 50$).

$(n, \sigma(x))$	Method	$D_n = 0.5$	$D_n = 0.33$	$D_n = 0.25$	$D_n = 0.2$
$(5000,\ 0.2\cos(x))$	Screening	40.92 (2.18)	55.88 (2.36)	82.25 (3.37)	103.74 (4.11)
	SaMT	36.62 (1.58)	47.24 (2.32)	53.27 (2.98)	56.74 (3.10)
	Sara	48.25 (3.14)	58.12 (3.17)	90.79 (3.23)	100.07 (4.01)
$(10000,\ 0.2\cos(x))$	Screening	43.26 (2.01)	55.75 (2.12)	62.67 (2.88)	84.47 (3.32)
	SaMT	39.57 (2.26)	45.29 (2.35)	50.68 (2.78)	55.22 (2.66)
	Sara	50.89 (3.15)	54.28 (2.99)	93.20 (3.25)	105.74 (3.89)
$(5000,\ 0.5\cos(x))$	Screening	48.19 (3.91)	58.24 (4.31)	89.97 (5.45)	104.12 (6.01)
	SaMT	33.88 (3.12)	45.68 (3.34)	54.17 (5.44)	59.16 (5.10)
	Sara	59.89 (4.15)	62.12 (4.56)	90.01 (4.89)	110.02 (5.22)
$(10000,\ 0.5\cos(x))$	Screening	47.52 (3.12)	55.66 (3.73)	70.12. (4.13)	85.26 (5.02)
	Sara	40.12 (3.18)	47.98 (4.12)	52.38 (4.35)	56.88 (5.13)
	Wavelet	55.79 (4.25)	61.79 (3.58)	80.41 (3.56)	98.12 (4.92)

3.2.1 (without conducting multiple testing), our proposed method, and the Sara method (Niu 2012) to detect jump points.

From Table 3.1, in the screening procedure, as the threshold D_n is decreasing, the set $S(D_n)$ will include more and more candidates of jump points. On the other hand, the number of the jump points detected by our proposed method is much less than that in the screening procedure, which means some of the candidates in $S(D_n)$ are noises and ruled out. What's more, with appro-

priate choice of threshold D_n (0.2 or 0.25), our proposed method works quite well when $\sigma = 0.2$, but more jump points are detected in the noisier data.

Next, we check the sure screening property mentioned in Theorem 1, that is, given any true jump point τ_j, $j = 1, \ldots, q$, we can find an estimated jump point in the candidate set $S(D_n)$ that is very close to τ_j with high probability. Here we take the jump point τ_5 for example and define the "detection rate" of τ_5 to be the proportion that how many times the event $\{\min_{1\le j\le \widehat{q}} |\widehat{\tau}_j - \tau_5| \le \kappa\}$ will happen out of 100 simulations for a small κ. Here we take $\kappa = 6 \times 10^{-4}$.

From Table 3.2, as the threshold D_n is decreasing, we are more likely to detect the jump points although more noises will be added into the candidate set $S(D_n)$. Besides, we may miss the true jump point when the data are noisier ($\sigma = 0.5$): when the standard deviation σ is large enough to be close to the jump size $\min_{1\le j\le q} |\gamma_j| = 0.5$, it is possible that the absolute difference process $|L_n(x)|$ will be large evaluated at some continuous point and this point may be regarded as the jump candidate, then any true jump point in the neighbor of this continuous point will be missed.

Table 3.2: Detection rate for the jump point τ_5 with $\kappa = 6 \times 10^{-4}$.

$(n, \sigma(x))$	$D_n = 0.5$	$D_n = 0.33$	$D_n = 0.25$	$D_n = 0.2$
$(5000,\ 0.2\cos(x))$	0.65	0.93	0.95	0.98
$(10000,\ 0.2\cos(x))$	1.00	1.00	1.00	1.00
$(5000,\ 0.5\cos(x))$	0.52	0.70	0.77	0.83
$(10000,\ 0.5\cos(x))$	0.58	0.77	0.81	0.88

Finally, we study the impact of the threshold D_n in the screening and

multiple testing procedure in Figure 1. Figure 1(a) shows one simulated data with $n = 10000$ and $\sigma = 0.2\cos(x)$. We plot the absolute difference process $|L_n(x)|$ before and after the screening procedure in Figure 1(b) and (c). The number of the jump points detected applying screening and multiple testing procedure with different choices of threshold $D_n \in \{0.5,\ 0.45,\ 0.4,\ \ldots,\ 0.15\}$ are shown in Figure 1(d).

From Figure 1(d), when the threshold D_n is smaller than some level, say 0.35, the number of the hypotheses we rejected will become stable. That means when the threshold D_n keeps decreasing, only noises will be added into $S(D_n)$. This will result in an adaptive choice of the threshold D_n. The best strategy in the screening procedure is to make the threshold D_n small enough such that all the true jump points are contained in $S(D_n)$. But usually we don't know whether a given threshold is small enough or not. In practice, we can take a decreasing grid search for D_n and then check when the number of the hypotheses we rejected becomes stable.

3.5 Real data example

3.5.1 particulate matter(PM) 2.5 data

We now present an application of our proposed method to particulate matter(PM) 2.5 data. Particulate matter (PM) are microscopic solid or liquid matter suspended in the atmosphere of Earth. They have impacts on climate and precipitation that adversely affect human health. In this example particulate matter (PM) 2.5 data from Shanghai have been investigated to find any

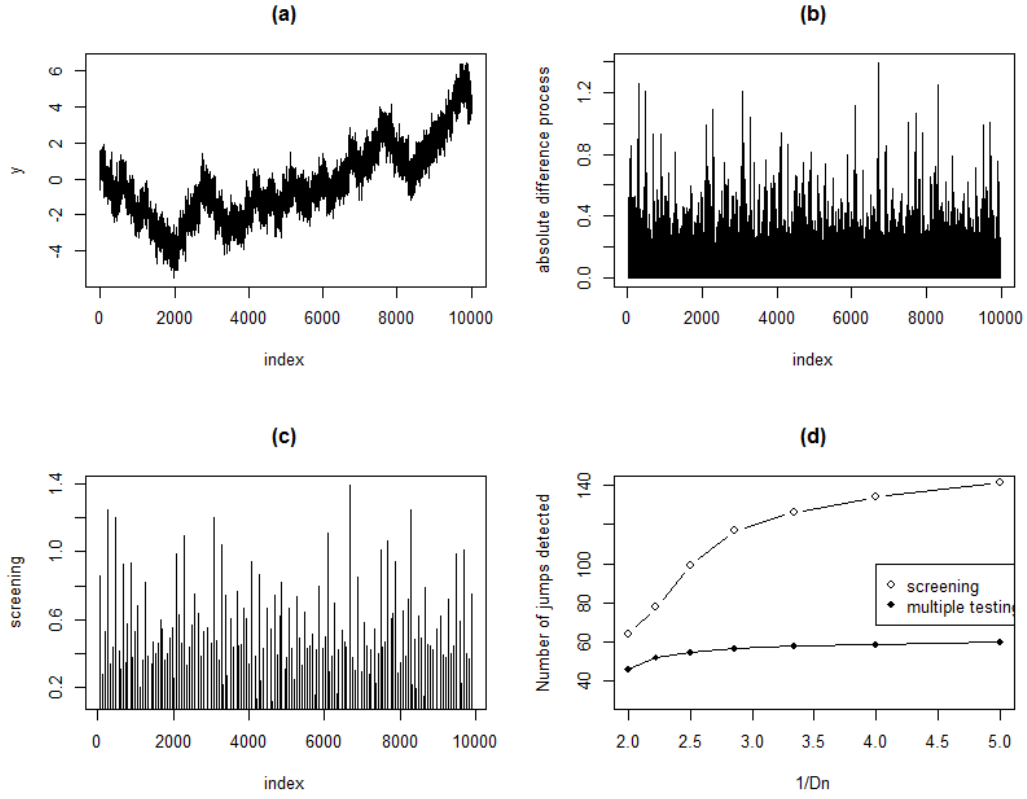

Figure 1: Panel (a): one simulated data with $n = 10000$ and $\sigma(x) = 0.2\cos(x)$; panel (b): absolute value of the difference process $|L_n(x)|$ in (2.1); panel (c): $|L_n(x)|$ evaluated at the initial estimated jump points in $S(D_n)$ with $D_n = 0$; panel (d): number of the jump points detected after screening (circle) and multiple testing procedure (solid) with different choices of threshold D_n.

jumps or changes in structure in this time series. There are totally 1015 data points collected during years 2005–2010.

Several different methods are employed, including SaRa algorithm and the multi-bandwidth SaRa algorithm proposed by Niu and Zhang (2012), Circular

Binary Segmentation (CBS) algorithm proposed by Olshen *et al.* (2004) and our proposed method, to detect possible jumps in this data set. Here, the bandwidths h_n in the three sequences are selected differently such that 50 observations are collected in each neighborhood $(x - h_n, x + h_n)$, and several thresholds $D_n = 0.01, 0.05$ and 0.1 are selected and finally we used $D_n = 0.1$ in the screening step. Similar to the simulation, we choose $B = 500$ and $\alpha = 0.05$ in the multiple testing procedure.

The result for the data analysis is listed in Table 3.3. Among 27 jump points detected by our method, most of them are found to be the missing data during the weekends, which is reasonable. The CBS algorithm detected too many jump points, most of which are most likely to be false positives. Although we detect many jump points candidates in the screening step, most of them are ruled out in the multiple testing procedure. We notice that SaRa and multi-SaRa algorithm detected more number of the jump points than our proposed method. The reason is that the thresholds in those two algorithms are selected too small, although many noises may be also deleted, some jump points with small jump sizes may also be ignored.

Table 3.3: Number of jump points detected for the offspring by different methods.

data	SaRa	multi-SaRa	CBS	Screening	SaMT
PM 2.5 before 2010	65	59	19	41	27

3.5.2 Single Nucleotide Polymorphism (SNP) genotyping data

We now present an application of our proposed method to single nucleotide polymorphism (SNP) genotyping data. DNA copy number variation (CNV) refers to deletion or duplication of a region of DNA sequences compared to a reference genome assembly. Identification of CNV is an essential issue in the systematic studies of understanding the role of CNV in human diseases.

In this example SNP genotying data for a father-mother-offspring trio produced by the Illumina 550K platform are collected. The data set can be downloaded along with the PennCNV R-package. The data consist of the measurements of a normalized total signal intensity ratio called the Log R ratio, that is, calculated by $\log_2(R_{obs}/R_{exp})$, where R_{obs} is the observed total intensity of the two alleles for a given SNP, and R_{exp} is the expected intensity computed from linear interpolation of the observed allelic ratio with respect to the canonical genotype clusters, see Niu and Zhang (2012). For each subject, the Log R ratios along Chromosomes 3, 11 and 20 are included in the data set. There are 37,768, 27,272 and 14,296 SNPs in Chromosomes 3, 11 and 20, respectively. Similar to the aCGH data, the segments with concentrated high or low Log R ratios correspond to gains or losses of copy numbers.

Several different methods are employed, including SaRa algorithm and the multi-bandwidth SaRa algorithm proposed by Niu and Zhang (2012), Circular Binary Segmentation (CBS) algorithm proposed by Olshen *et al.* (2004) and our proposed method, to detect CNVs in this data set. Here, the bandwidths h_n in the three sequences are selected differently such that 50 observations

are collected in each neighborhood $(x - h_n, x + h_n)$, and several thresholds $D_n = 0.1, 0.2$ and 0.3 are selected and finally we used $D_n = 0.1$ in the screening step. Similar to the simulation, we choose $B = 200$ and $\alpha = 0.05$ in the multiple testing procedure.

Table 3.4: Number of jump points detected for the offspring by different methods

.

data	Sara	multi-Sara	CBS	Screening	Multiple testing
Chromosome 3	2	19	46	51	17
Chromosome 11	4	2	29	37	9
Chromosome 20	4	7	16	23	11

The result for the offspring is listed in Table 3.4. The CBS algorithm detected too many jump points, most of which are most likely to be false positives. Although we detect many jump points candidates in the screening step, most of them are ruled out in the multiple testing procedure. We notice that SaRa and mtilti-Sara algorithm detected less number of the jump points than our proposed method. The reason is that the thresholds in those two algorithms are selected too large, although many noises may be also deleted, some jump points with small jump sizes may also be ignored. Therefore, our proposed method is more reasonable and works very well.

Bibliography

[1] Bosq, D. (2000), Linear Processes in Function Spaces: Theory and Applications, New York: Springer.

[2] Cai, T. T. and Yuan, M. (2010). Nonparametric covariance function estimation for functional and longitudinal data. *Technical Report.*

[3] Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002). Analysis of Longitudinal Data, second edition. New York: Oxford University Press.

[4] Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and Its Application, London: Chapman & Hall.

[5] Fan, J., and Huang, T. (2005). Profile likelihood inferences on semiparametric varying coefficient partially linear models. *Bernoulli*, 11, 1031–1059.

[6] Fan, J., Huang, T. and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *J. Amer. Statist. Assoc.*, 102, 632–641.

[7] Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrixes for longitudinal data. *J. Amer. Statist. Assoc.*, 103, 1520–1533.

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

[9] Feller, W. (1971). An introduction to probability theory and its applications, *Vol. II*, second edition. John Wiley and Sons, New York.

[10] Hall, P., M*ü*ller, H. G. and Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.*, 34, 1493-1517.

[11] Hall, P., M*ü*ller, H. G. and Yao, F. (2008). Modelling sparse generalized longitudinal observations with latent Gaussian processes. *J. R. Statist. Soc. B.*, 70, 703–723.

[12] Huang, J. Z., Wu, C. O., and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89, 111–128.

[13] Li, Y. (2011). Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika*, 98, 355–370.

[14] Lin, X. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *J. Am. Statist. Assoc.*, 96, 1045–1056.

[15] Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahr. verw. Geb.*, 61, 405–415.

[16] Milton, A. and Irene A. S. (1972). Handbook of mathematical functions with formulas, graphs, and mathematical tables, New York: Dover.

[17] Sowers, M., Randolph, J. F., Crutchfield, M., Jannausch, M. L., Shapiro, B., Zhang, B. and La Pietra, M. (1998). Urinary Ovarian and Gonadotropin Hormone Levels in Premenopausal Women With Low Bone Mass. *Journal of Bone and Mineral Research*, 13, 1191–1202.

[18] Wang, N., Carroll, R. J. and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *J. Am. Statist. Assoc.*, 100, 147–157.

[19] Wu, H. L., and Zhang, J. T. (2002). Local polynomial mixed-effects models for longitudinal data. *J. Am. Statist. Assoc.*, 97, 883–897.

[20] Wu, H. L., and Zhang, J. T. (2002). The study of long-term HIV dynamics using semiparametric nonlinear mixed-effects models. *Statistics in Medicine*, 21, 3655–3675.

[21] Wu, H. L., and Zhang, J. T. (2006). Nonparametric regression methods for longitudinal data: mixed-effects modeling approaches. Wiley, New York.

[22] Wu, W. B., and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90, 831–844.

[23] Yao, F., M*ü*ller, H. G. and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *J. Am. Statist. Assoc.*, 100, 577–590.

[24] Zhang, D., Lin, X., Raz, J. and Sowers, M. (1998). Semiparametric Stochastic Mixed Models for Longitudinal Data. *J. Am. Statist. Assoc.*, 93, 710–719.

[25] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser.* B, 57, 289–300.

[26] Blondin, D. (2007). Rates of strong uniform consistency for local least squares kernel regression estimators. *Statistics & Probability Letters*, 77, 1526–1534.

[27] Cheng, M. Y., Fan, J. and Marron, J. S. (1993). Minimax efficiency of local polynomial fit estimators at boundaries. Institute of Statistics Mimeo Series #2098, University of North Carolina at Chapel Hill.

[28] Eubank, R. L. and Speckman, P. (1994). Nonparametric estimation of functions with jump discontinuities, change-point problems. *IMS Lecture Notes and Monograph Series*, 23, 130–144.

[29] Fan, J. and Gijbels, I. (1996). Local polynomial modelling and its applications. Chapman and Hall, New York.

[30] Fan, J. (2005). A selective overview of nonparametric methods in financial econometrics. *Statistical Science*, 20, 317–337.

[31] Fan, J., Hall, P. and Yao, Q. (2007). To how many simultaneous hypothesis tests can normal, Student's t or bootstrap calibration be applied? *J. Amer. Statist. Assoc.*, 102, 1282–1288.

[32] Fan, J. and Wang, Y. (2007). Multi-scale jump and volatility analysis for high-Frequency financial data. *Jour. Ameri. Statist. Assoc*, 102, 1349–1362.

[33] Fearnhead, P. and Liu, Z. (2007). On-line inference for multiple change-point problems. *J. R. Statist. Soc. B*, 69, 589–605.

[34] Gijbels, I. and Goderniaux, A.C. (2004). Bootstrap test for change-points in nonparametric regression. *Nonparametric Statistics*, 16, 591–611.

[35] Gijbels, I. ,Hall, P. and Kneip, A. (1999). On the estimation of jump points in smooth curves. *The Annals of the Institute of Statistical Mathematics*, 51, 231–251.

[36] Grégoire, G. and Hamrouni, Z. (2002). Change point estimation by local linear smoothing. *Journal of Multivariate Analysis*, 83, 56–83.

[37] Kosorok, Michael and Ma, Shuangge. (2007). Marginal asymptotics for the "large p, small n" paradigm: with applications to microarray data. *Annals of Statistics*, 35, 1456–1486.

[38] Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple changepoint estimation with a total variation penalty. *J. Amer. Statist. Assoc*, 105, 1480–1493.

[39] Loader, C. R. (1996). Change point estimation using nonparametric regression. *Ann. Statist.*, 24, 1667–1678.

[40] Lu, Q., Lund, R. and Lee, T. C. M. (2010). An MDL approach to the climate segmentation problem. *Ann. Appl. Statist.*, 4, 299–319.

[41] Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrscheinlichkeitstheorie Verw. Geb.* 61, 405–415.

[42] Müller, H. G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist.*, 20, 737–761.

[43] Müller, H. G. and Song, K. S. (1997). Two-stage change-point estimators in smooth regression models.*Statistics & Probability Letters*, 34, 323–335.

[44] Müller, H. G. and Stadtmüller, U. (1999). Discontinuous versus smooth regression. *Ann. Statist.*, 27, 299–337.

[45] Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5, 557–572

[46] Niu, Y. S. and Zhang, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Stat.*, 6, 1306–1326.

[47] Qiu, P. (2005). Image processing and jump regression analysis. John Wiley and Sons, New Jersey.

[48] Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika*, 82, 385–397.

[49] Wu, J. S., and Chu, C. K. (1993). Kernel type estimators of jump points and values of a regression function. *Ann. Statist.*, 21, 1545–1566.

[50] Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach. *Bernoulli*, 12, 1019–1043.

Appendix A

Conditions and Proofs of Main Results in Chapter 2

The following technical conditions are imposed. They are not the weakest possible conditions, but they are imposed to facilitate the proofs. For notational convenience, given a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$, define $|\boldsymbol{\alpha}| = (|\alpha_1|, \ldots, |\alpha_p|)^T$.

C1. The density function $f_T(\cdot)$ is Lipschitz-continuous and bounded away from 0 and infinity. The function $K(\cdot)$ is a symmetric density function with a compact support.

C2. $nh_1^8 \to 0$ and $nh_1^2/\{\log(n)\}^3 \to \infty$ as $n \to \infty$.

C3. $\mathrm{E}\{\boldsymbol{X}(t)\boldsymbol{X}(t)^T\}$ and $\mathrm{E}\{\boldsymbol{X}(t)\boldsymbol{Z}(t)^T\}$ are Lipschitz-continuous.

C4. $\boldsymbol{\alpha}(t)$ has a continuous second derivative.

C5. J_i has a finite moment-generating function in some neighborhood of the origin. In addition, $\mathrm{E}\{\|\boldsymbol{X}(t)\|^4\} + \mathrm{E}\{\|\boldsymbol{Z}(t)\|^2\} < \infty$.

C6. $\sigma_\varsigma^2(\cdot)$ has a continuous second derivative, and $\mathrm{E}\{\varepsilon(t)^{4+\delta_0}\} < \infty$ for some $\delta_0 > 0$.

C7. $h_2\{\log(n)\}^2 \to 0$ and $nh_2/\log(n) \to \infty$ as $n \to \infty$.

Proof of Proposition 1. First, By Fan, Huang and Li [6] (Theorem 1 and result (A.1), (A.2)), we have

$$\widehat{\boldsymbol{\beta}}^{\mathrm{ini}} - \boldsymbol{\beta}^\circ = O_{\mathrm{P}}(n^{-1/2}) \tag{A.1}$$

$$\sup_{t\in\mathcal{T}} \big|\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(t) - \boldsymbol{\alpha}^\circ(t) + \big[\mathrm{E}\{\boldsymbol{X}(t)\boldsymbol{X}(t)^T\}\big]^{-1}\big[\mathrm{E}\{\boldsymbol{X}(t)\boldsymbol{Z}(t)^T\}\big](\boldsymbol{\beta} - \boldsymbol{\beta}^\circ)\big| = O_{\mathrm{P}}(c_n) \tag{A.2}$$

where $c_n = h_1^2 + \{-\log(h_1)/(nh_1)\}^{1/2}$, $\boldsymbol{\beta}^\circ$ and $\boldsymbol{\alpha}^\circ(t)$ are true parametric and nonparametric components. Since $\mathcal{T}$ is compact, together with condition C3 and (A.1),

$$\sup_{t\in\mathcal{T}} \big|\big[\mathrm{E}\{\boldsymbol{X}(t)\boldsymbol{X}(t)^T\}\big]^{-1}\big[\mathrm{E}\{\boldsymbol{X}(t)\boldsymbol{Z}(t)^T\}\big](\widehat{\boldsymbol{\beta}}^{\mathrm{ini}} - \boldsymbol{\beta}^\circ)\big| = O_{\mathrm{P}}(n^{-1/2}). \tag{A.3}$$

Utilizing triangle inequality and (A.2), (A.3) yields

$$\sup_{t\in\mathcal{T}} \big|\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\beta}}^{\mathrm{ini}}}(t) - \boldsymbol{\alpha}^\circ(t)\big| = O_{\mathrm{P}}(c_n).$$

Finally, by the definition of $\widehat{\varepsilon}(t)$ in (2.5) and $\varepsilon(t)$, we have

$$\sup_{t\in\mathcal{T}} \big|\widehat{\varepsilon}(t) - \varepsilon(t)\big| = \sup_{t\in\mathcal{T}} \big|\boldsymbol{X}(t)^T\big\{\widehat{\boldsymbol{\alpha}}_{\widehat{\boldsymbol{\beta}}^{\mathrm{ini}}}(t) - \boldsymbol{\alpha}^\circ(t)\big\} + \boldsymbol{Z}(t)^T\big(\widehat{\boldsymbol{\beta}}^{\mathrm{ini}} - \boldsymbol{\beta}^\circ\big)\big| = O_{\mathrm{P}}(c_n).$$

This completes the proof. ■

Proof of Proposition 2. Without loss of generality, suppose $\mathcal{T} = [0, 1]$ and $f_T(\cdot)$ is uniform density on $[0, 1]$. For case (i), by Condition C5, suppose

$\mathrm{E}(e^{tJ_i}) \leq C$ for some $t > 0$. By Markov's inequality, $\mathrm{P}(J_i \geq a) = \mathrm{P}(e^{tJ_i} \geq e^{ta}) \leq \mathrm{E}(e^{tJ_i})/e^{ta} \leq Ce^{-ta}$, then

$$\begin{aligned} & \mathrm{P}(\max_{1\leq i\leq n} J_i \geq d\log(n)) \\ = \; & 1 - \mathrm{P}(\max_{1\leq i\leq n} J_i < d\log(n)) \\ = \; & 1 - \mathrm{P}(J_i < d\log(n))^n \\ = \; & 1 - \left[1 - \mathrm{P}(J_i \geq d\log(n))\right]^n \\ \leq \; & 1 - \left(1 - Ce^{-td\log(n)}\right)^n \\ = \; & 1 - \left(1 - Cn^{-td}\right)^n \to 0 \end{aligned}$$

when d is large enough such that $td > 1$. So $\max_{1\leq i\leq n} J_i = O_{\mathrm{P}}(\log(n))$. Next let $\{T_i^{(1)}, \ldots, T_i^{(J_i)}\}$ be order statistics of $\{t_{i,1}, \ldots, t_{i,J_i}\}$. According to statement by Feller [9] (p.42), for a given J_i,

$$\mathrm{P}(T_i^{(2)} - T_i^{(1)} > 2h_2, \ldots, T_i^{(J_i)} - T_i^{(J_i-1)} > 2h_2) = \left[1 - 2h_2(J_i - 1)\right]_+^{J_i} \quad \text{(A.4)}$$

where $[g]_+$ is the positive part of g. Since $\max_{1\leq i\leq n} J_i = O_{\mathrm{P}}(\log(n))$, with probability tending to 1,

$$\left[1 - 2h_2(J_i - 1)\right]_+^{J_i} \geq \left[1 - 2dh_2\log(n)\right]_+^{2d\log(n)}.$$

By condition C7, we have $\left[1 - 2dh_2\log(n)\right]_+^{2d\log(n)} \to 1$, so it is unlikely for each individual to have more than two observations in the same neighborhood $[t-h_2, t+h_2]$. Thus in what follows, the $\varepsilon_i^2(t_{i,j})$'s can be treated as independent similar to the proof of Theorem 1 in [7]. By classic uniform convergence rates for the kernel smoother ([41]), we have

$$\sup_{t\in\mathcal{T}} \left|\overline{\sigma}^2(t) - \sigma^2(t)\right| = O_{\mathrm{P}}(h_2^2 + \{-\log(h_2)/(nh_2)\}^{1/2}).$$

For case (ii), after we take $\alpha = 2$ in Theorem 4 according to [2], we have

$$\lim_{D\to\infty} \limsup_{n\to\infty} \mathrm{P}\big(\|\overline{R}_\eta - R_\eta\|_{L_2}^2 > D\big(\{\log(n)/(nm)\}^{4/5} + n^{-1}\big)\big) = 0$$

that means

$$\|\overline{R}_\eta - R_\eta\|_{L_2}^2 = O_{\mathrm{P}}\big(\{\log(n)/(nm)\}^{4/5} + n^{-1}\big) \tag{A.5}$$

where $m = \mathrm{E}(J_i)$ is the expected number of observations within each subject. Since condition C5 implies that the first moment $m = \mathrm{E}(J_i)$ is finite, (A.5) becomes $\|\overline{R}_\eta - R_\eta\|_{L_2}^2 = O_{\mathrm{P}}\big(\{\log(n)/n\}^{4/5}\big)$, which is exactly (3.4). ■

Proof of Proposition 3. First, note that $\tau \geq 0$, it is easy to show that $\widetilde{R}_\eta(s,t) = \sum_{k=1}^{N} \widehat{\lambda}_{(k)} \mathrm{I}(\widehat{\lambda}_{(k)} > \tau)\widehat{\psi}_k(s)\widehat{\psi}_k(t)$ defined in (2.14) is positive definite since all the eigenvalues of $\widetilde{R}_\eta(s,t)$ are positive. Secondly, note that

$$\begin{aligned}
\widetilde{R}(s,t) &= \widetilde{R}_\eta(s,t)\mathrm{I}(s \neq t) + \max\{\widetilde{R}_\eta(t,t), \widehat{\sigma}^2(t)\}\mathrm{I}(s = t) \\
&= \widetilde{R}_\eta(s,t) + \max\{0, \widehat{\sigma}^2(t) - \widetilde{R}_\eta(t,t)\}\mathrm{I}(s = t).
\end{aligned}$$

Obviously the second term, $\max\{0, \widehat{\sigma}^2(t) - \widetilde{R}_\eta(t,t)\}\mathrm{I}(s = t)$, is positive semi-definite since it is diagonal and all the diagonal entries are nonnegative. So $\widetilde{R}(s,t)$ is the summation of a positive definite and a positive semi-definite covariance function, which is of course positive definite. ■

Appendix B

Conditions and Proofs of Main Results in Chapter 3

The following technical conditions are imposed. They are not the weakest possible conditions, but they are imposed to facilitate the proofs.

(D1) $K_{-}(\cdot)$ is a right-continuous function with bounded variation on $\mathbb{R}$;

(D2) $K_{-}(\cdot)$ is compactly supported with $K_{-}(0) > 0$ and $\int K_{-}(u)du = 1$;

(D3) $h_n \to 0$, $nh_n/\log(h_n^{-1}) \to \infty$ and $\log(h_n^{-1})/\log\log(n) \to \infty$;

(D4) The marginal density f_X of covariate X is continuous and bounded away from 0.

To control the upper bound of the difference process $L_n(x)$ at continuous points, strong uniform consistency for local linear regression estimator is important. Blondin (2007) derived the strong uniform consistency based upon modern empirical process theory.

Lemma 1 *Under conditions* (D1)–(D3)*, we have*

$$\Big|\Big\{\frac{nh_n}{2\log(h_n^{-1})}\Big\}^{1/2}\sup_{x\in[0,1]}|\widehat{m}_-(x)-\mathrm{E}\{\widehat{m}_-(x)\}|-\Lambda\Big|=o(1)\ \ \text{a.s.},\qquad\text{(B.1)}$$

where

$$\Lambda=\sup_{x\in[0,1]}\Big[f_X(x)^{-1}\sigma^2(x)\int\{K_-(u)\}^2du\Big]^{1/2}.$$

Proof: This result can be derived directly from [26] (Theorem 3.1). ■

When we evaluate $\widehat{m}_+(x)$ in (1.3) at the jump point τ_j, we only use the data located at the right of point τ_j. This is equivalent to the estimation at the boundary point in standard local linear regression. The following lemma provides the asymptotic normality for $\widehat{m}_+(\tau_j)$.

Lemma 2 *Let* $\widehat{m}_+(x)$ *be defined in* (1.3)*, under the conditions of Theorem* 2*,*

$$\sqrt{nh_n}\Big\{\widehat{m}_+(\tau_j)-m_+(\tau_j)-C\cdot\frac{h_n^2}{2}m_0''(\tau_j)\Big\}\xrightarrow{\mathcal{D}}N(0,V\sigma^2(\tau_j)/f_X(\tau_j)),\quad\text{(B.2)}$$

where V *is defined in* (3.4)*, and*

$$C=\frac{\mu_2^2-\mu_1\mu_3}{\mu_0\mu_2-\mu_1^2}$$

with μ_j *and* ν_j *defined in Theorem* 2*.*

Proof: Taking $c=0$ in Theorem 3.3 of [29] completes the proof. ■

Lemma 3 *If* $q=1$*, or equivalently, there is unique change point* τ*. Suppose* $K_-(0)>0$*, and* $h_n\to0$*,* $nh_n\to\infty$*, then the maximizer of* $|L_n(x)|$*, say* $\widehat{\tau}$*, satisfies*

$$nh_n\{L_n(\widehat{\tau})-L_n(\tau)\}=O_{\mathrm{P}}(1).\qquad\text{(B.3)}$$

Proof: The result holds according to [36] (Theorem 3.1 and Lemma 3.1) after we take $\alpha(n,h_n)=\beta(n,h_n)=nh_n$ therein. ■

Proof of Theorem 1. We only need to prove part (b). Let $\mathcal{I}_j \equiv (\tau_j - h_n, \tau_j + h_n)$ be the neighbor of jump point τ_j, $j = 1, \ldots, q$. Let $\mathcal{I}_0 \equiv [0,1] \cap \{\mathcal{I}_1 \cup \cdots \cup \mathcal{I}_q\}^c$.

For any $x \in \mathcal{I}_0$, the estimator $\widehat{m}_-(x)$ in (1.3) does not involve any jump points, this is the standard local linear smoothing and the asymptotic bias is $|\mathrm{E}\{\widehat{m}_-(x)\} - m_-(x)| = O(h_n^2)$ (see [29]), together with Lemma 1, we get

$$\sup_{x \in \mathcal{I}_0} |\widehat{m}_-(x) - m_-(x)| = O\Big\{ h_n^2 + \sqrt{\frac{2\log(h_n^{-1})}{nh_n}} \Big\} \quad \text{a.s.}, \tag{B.4}$$

similarly,

$$\sup_{x \in \mathcal{I}_0} |\widehat{m}_+(x) - m_+(x)| = O\Big\{ h_n^2 + \sqrt{\frac{2\log(h_n^{-1})}{nh_n}} \Big\} \quad \text{a.s.}. \tag{B.5}$$

Since all the points in $\mathcal{I}_0$ are continuous points, $\forall x \in \mathcal{I}_0$, $m_-(x) = m_+(x)$, by (B.4), (B.5) and triangular inequality, we have

$$\sup_{x \in \mathcal{I}_0} |L_n(x)| = \sup_{x \in \mathcal{I}_0} |\widehat{m}_-(x) - \widehat{m}_+(x)| = O\Big\{ h_n^2 + \sqrt{\frac{2\log(h_n^{-1})}{nh_n}} \Big\} = o(D_n) \quad \text{a.s} \tag{B.6}$$

By the definition of $\widehat{\tau}_j$, we have $|L_n(\widehat{\tau}_j)| > D_n$, $j = 1, \ldots, \widehat{q}$. From (A.4), we have $|L_n(\widehat{\tau}_j)| > \sup_{x \in \mathcal{I}_0} |L_n(x)|$, that means

$$\widehat{\tau}_j \in \mathcal{I}_0^c = \mathcal{I}_1 \cup \cdots \cup \mathcal{I}_q \quad as \qquad (j = 1, \ldots, \widehat{q}). \tag{B.7}$$

Since $\widehat{\omega}_k \in \mathcal{I}(\widehat{\omega}_j)^c$ for $k \neq j$, $|\widehat{\tau}_{j+1} - \widehat{\tau}_j| = |\widehat{\omega}_{(j+1)} - \widehat{\omega}_{(j)}| > 2h_n = |\mathcal{I}_j|$ together with (B.7) means that each $\mathcal{I}_j$ can not contain more than one $\widehat{\tau}_j$, so

$$\widehat{q} \leq q, \quad \text{a.s.}. \tag{B.8}$$

On the other hand, since $h_n \to 0$, there exists $N > 0$ such that $9h_n < \xi = \min_{1 \leq j \leq q+1}(\tau_j - \tau_{j-1})$ for any $n > N$, and

$$|L_n(\tau_j)| = |\widehat{m}_+(\tau_j) - \widehat{m}_-(\tau_j)| \geq \frac{\gamma_j}{2} \gg D_n \quad \text{a.s.}. \tag{B.9}$$

From the definition of $\widehat{\omega}_{\widehat{q}+1}$,

$$|L_n(x)| \le |L_n(\widehat{\omega}_{\widehat{q}+1})| \le D_n, \quad \forall\, x \in [0,1] \cap \big(\mathcal{I}(\widehat{\omega}_1) \cup \cdots \cup \mathcal{I}(\widehat{\omega}_{\widehat{q}})\big)^c, \quad \text{(B.10)}$$

so by (B.9) and (B.10), we have $|L_n(\tau_j)| > \sup_{x \in [0,1] \cap \left(\mathcal{I}(\widehat{\omega}_1) \cup \cdots \cup \mathcal{I}(\widehat{\omega}_{\widehat{q}})\right)^c} |L_n(x)|$ when $n > N$, that means

$$\tau_j \in \mathcal{I}(\widehat{\omega}_1) \cup \cdots \cup \mathcal{I}(\widehat{\omega}_{\widehat{q}}) \ \text{ a.s.} \qquad (j = 1, \ldots, q), \quad \text{(B.11)}$$

when $n > N$. Since $\min_{1 \le j \le q+1}(\tau_j - \tau_{j-1}) = \xi > 4h_n = |\mathcal{I}(\widehat{\omega}_k)|$, again each $\mathcal{I}(\widehat{\omega}_k)$ can not contain more than one τ_j, that means

$$q \le \widehat{q}, \quad \text{a.s.}, \quad \text{when } n > N. \quad \text{(B.12)}$$

So (3.1) follows by (B.8) and (B.12). To show (3.2), by (3.1), it suffices to consider what happens on event $\{\widehat{q} = q\}$, which means each $\mathcal{I}_j$ will contain only one $\widehat{\tau}_j$. Now consider the single jump point detection on the intervals $H_j \equiv [\tau_j - \xi/3, \tau_j + \xi/3]$, it's easy to check that

(i) H_j's are disjoint and there is only one jump point τ_j in each H_j.

(ii) Denote the number of observations over the interval H_j by $m_n = \sum_{i=1}^n I[X_i \in H_j]$, then $\frac{m_n}{n} \to \int_{H_j} f_X(x)dx > |H_j| \min_{x \in H_j} f_X(x) > 0$ a.s., so m_n and n are of the same order, and thus $m_n h_n \to \infty$ since $nh_n \to \infty$.

Now all the conditions in the Theorem 3.2 of [36] are satisfied, and we have $m_n|\widehat{\tau}_j - \tau_j| = O_{\mathrm{P}}(1)$ and thus (3.2) follows. ■

Proof of Theorem 2. We first show (3.3). Since $K_-(x) = K_+(-x)$, replace $K_+(-x)$ by $K_-(x)$ in Lemma 2 shows the asymptotic normality for $\widehat{m}_-(\tau_j)$:

$$\sqrt{nh_n}\Big\{(\widehat{m}_-(\tau_j) - m_-(\tau_j) - C \cdot \frac{h_n^2}{2} m_0''(\tau_j)\Big\} \xrightarrow{\mathcal{D}} N(0, V\sigma^2(\tau_j)/f_X(\tau_j)). \quad \text{(B.13)}$$

Since $\widehat{m}_+(\tau_j)$ and $\widehat{m}_-(\tau_j)$ only use the data located at the right and left of point τ_j, they are independent, the difference of (B.2) and (B.13) becomes

$$\sqrt{nh_n}\{L_n(\tau_j) - \gamma_j\} \xrightarrow{\mathcal{D}} N(0, 2V\sigma^2(\tau_j)/f_X(\tau_j)). \tag{B.14}$$

Again, consider the interval $H_j \equiv [\tau_j - \xi/3, \tau_j + \xi/3]$. In the proof of Theorem 1, we have shown that all the conditions in Lemma 3 hold on H_j. Thus by Lemma 3,

$$\sqrt{nh_n}\{L_n(\widehat{\tau}_j) - L_n(\tau_j)\} = \frac{1}{\sqrt{nh_n}} \cdot nh_n\{L_n(\widehat{\tau}_j) - L_n(\tau_j)\} = o_{\mathrm{P}}(1). \tag{B.15}$$

Now (3.3) follows from (B.14) and (B.15). Next, to show (3.5), for any $d > 0$,

$$\mathrm{P}(|\widehat{\Psi} - \Psi| > d(nh_n)^{-1/2}) \le \mathrm{P}(\widehat{q} = q, |\widehat{\Psi} - \Psi| > d(nh_n)^{-1/2}) + \mathrm{P}(\widehat{q} \neq q).$$

From (3.1), we have

$$\lim_{d\to\infty}\lim_{n\to\infty} \mathrm{P}(|\widehat{\Psi} - \Psi| > d(nh_n)^{-1/2}) \le \lim_{d\to\infty}\lim_{n\to\infty} \mathrm{P}(\widehat{q} = q, |\widehat{\Psi} - \Psi| > d(nh_n)^{-1/2}) \tag{B.16}$$

With $\widehat{q} = q$, we have

$$\widehat{\Psi} - \Psi = \sum_{j=1}^{q}(\widehat{\gamma}_j^2 - \gamma_j^2) = \sum_{j=1}^{q}\{(\widehat{\gamma}_j - \gamma_j)^2 + 2\gamma_j(\widehat{\gamma}_j - \gamma_j)\}.$$

By (3.3), $|\widehat{\gamma}_j - \gamma_j| = O_{\mathrm{P}}((nh_n)^{-1/2})$, thus

$$\sum_{j=1}^{q} 2|\gamma_j(\widehat{\gamma}_j - \gamma_j)| = O_{\mathrm{P}}((nh_n)^{-1/2}), \qquad \sum_{j=1}^{q}(\widehat{\gamma}_j - \gamma_j)^2 = O_{\mathrm{P}}((nh_n)^{-1}).$$

So as n and d sequentially go to infinity,

$$\begin{aligned} & \mathrm{P}(\widehat{q} = q, |\widehat{\Psi} - \Psi| > d(nh_n)^{-1/2}) \\ \le\ & \mathrm{P}\Big(\sum_{j=1}^{q}(\widehat{\gamma}_j - \gamma_j)^2 + \sum_{j=1}^{q} 2|\gamma_j(\widehat{\gamma}_j - \gamma_j)| > d(nh_n)^{-1/2}\Big) \end{aligned}$$

$$\begin{aligned} &\leq \mathrm{P}\Big(\sum_{j=1}^{q}(\widehat{\gamma}_j-\gamma_j)^2 > \frac{d}{2}(nh_n)^{-1/2}\Big) + \mathrm{P}\Big(\sum_{j=1}^{q} 2|\gamma_j(\widehat{\gamma}_j-\gamma_j)| > \frac{d}{2}(nh_n)^{-1/2}\Big) \\ &\to 0. \end{aligned} \tag{B.17}$$

So (3.5) follows by (B.16) and (B.17). ■

Proof of Proposition 4. For any $\widehat{\tau} \in S(D_n) = \{\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{q}}\} = \{\widehat{\omega}_1, \ldots, \widehat{\omega}_{\widehat{q}}\}$, without loss of generality, let $\widehat{\tau} = \widehat{\omega}_j$, $1 \leq j \leq \widehat{q}$, then by the definition of $\widehat{\omega}_j$, we have

$$\widehat{\tau} = \widehat{\omega}_j = \arg \max_{x \in [h_n, 1-h_n] \cap (\mathcal{I}(\widehat{\omega}_1) \cup \cdots \cup \mathcal{I}(\widehat{\omega}_{j-1}))^c} |L_n(x)|, \tag{B.18}$$

and $|L_n(\widehat{\tau})| > D_n$. Since $D_n > d_n$, we get $|L_n(\widehat{\tau})| > d_n$. Together with (B.18), we have $\widehat{\tau} \in S(d_n)$, and thus $S(D_n) \subseteq S(d_n)$. ■

Proof of Proposition 5. According to (B.12), when $n > N$, we have

$$\widehat{q} \geq q, \text{ a.s.}, \tag{B.19}$$

and there is at most one true jump point in the interval $N_j = [\frac{1}{2}\{\widehat{\tau}_{j-1} + \widehat{\tau}_j\}, \frac{1}{2}\{\widehat{\tau}_j + \widehat{\tau}_{j+1}\}]$. Since the intervals N_j's are disjoint, and the bootstrap p-value p_j only depends on the observations over N_j, the p-values $\{p_j : 1 \leq j \leq \widehat{q}\}$ are independent. Finally by [37] and (3.6), the result follows. ■